MASTER THESIS

Optimizing Classification of Small Cloud Particles Using Neural Networks

MATTIS J. DEISEN

FRANKFURT SCHOOL OF FINANCE &
MANAGEMENT

MASTER THESIS

Optimizing Classification of Small Cloud Particles Using Neural Networks

---

Author:
Mattis J. DEISEN

Supervisors:
Prof. Dr. Jan NAGLER
Dr. Jan HENNEBERGER

Master in Applied Data Science

in the

Deep Dynamics Group
Centre for Human and Machine intelligence

Kölner Straße 58
60327 Frankfurt
8303580

Frankfurt, January 2021

Acknowledgements

I want to thank my supervisor Prof. Dr. Jan Nagler from the deep dynamics group at Frankfurt School of Finance & Management for the opportunity to write this thesis in the Deep Dynamics Group and for sharing his enthusiasm and advice regarding machine learning methodologies.

I am grateful to my co-supervisor Dr. Jan Henneberger and Annika Lauber from the Atmospheric Physics Group at ETH Zürich for providing this interesting topic, the cloud particle data and for welcoming me in the holography group. Further I appreciate their introduction to the topic of cloud particles and helpful feedback throughout the duration of the thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Understanding the microphysics of clouds is important for understanding precipitation formation and radiation feedback of clouds. The most important variables are size and concentration of cloud particles. They can be studied by using holographic imagers, which measure their concentration, shape and size. While holographic imagers offer the needed information, analyzing the holograms is time consuming. Recorded holograms can contain thousands of cloud particles. An algorithm extracts the particles, which can be displayed as pictures. The extracted particles need to be classified into droplets and ice crystals while artifacts need to be sorted out. Artifacts are particles that were mistakenly detected as cloud particles by the software. Since a typical field campaign produces millions of these particle pictures, a robust classification algorithm is needed. Previous research used convolutional neural networks to classify particles of a size between 25 µm and 250 mm with high accuracy (Touloupas et al., 2020). Classification performance for particles smaller than 25 µm was less reliable. Automated and reliable classification of cloud particles accross all particle size ranges is needed to support research of cloud microphysics.

1.2 Problem Statement

The problem at hand is an image classification task. When using holographic cloud data, captured images of cloud particles need to be classified. As there are up to hundreds of thousand images in a single data set, manual classification is time consuming. Previous work (Touloupas et al., 2020) achieved high accuracies by applying deep neural networks to classify particles larger than 25 µm into droplets, ice crystals and artifacts. Currently, a robust method of classifying particles smaller than 25 µm is missing. The aim of this work is to create a

method to robustly classify these small particles especially when confronted with new datasets. Previous work has shown that the existing methods have trouble generalizing to unseen datasets. This thesis focuses on the classification of cloud particles recorded by holographic imagers that are smaller than 25 µm.

1.3 Thesis Outline

In chapter 2, the data used to train and evaluate the classifiers is introduced. Following a summary of the data collection process, the seven datasets from different data collection campaigns are introduced. This introduction includes the review of particle size and class distributions as well as a look at examples of images contained in the data. Chapter 3 introduces neural networks and convolutional neural networks for image recognition as well as the metrics used to assess their performance. Chapter 4 presents the network architecture used for the classification. To find a benchmark for later generalization experiments, networks were optimized on all of the source dataset. Optimal input image size, the use of metadata and class rebalancing methods were evaluated. Further, chapter 4 begins with the assessment of generalization performance of classifiers when predicting new datasets. In chapter 5, the generalization efforts are extended upon by combining training data from various datasets. Datasets are combined and images transformed to understand which factors matter for robust generalization of classifiers onto new datasets. Chapter 6 summarizes the findings of the experiments and lays out ideas for further improvements.

Chapter 2

Data

2.1 Data Source

Image and particle data was selected from seven datasets that were generated using holographic imaging. The technology enables capturing three dimensional position as well as two dimensional shape for each particle in the observation volume (Baumgardner et al., 2011).

Image Extraction The working principle of a holographic imager is visualized in figure 2.1. The observation volume is irradiated by a laser. Particles scatter the laser and the resulting intereference pattern is captured by a camera. Each capture (hologram) contains information about a large number of particles. The hologram is processed by the HOLOSUITE software (Fugal et al., 2009) that detects particles within the observation volume using various thresholds that are set by a human. The thresholds for a particular hologram depend on the level of noise. A tradeoff between excluding more noise and including cloud particles of interest is made. The noise that is not filtered by the set threshold is found in the particle data in the form of artifacts. Artifacts can be caused by only partly captured particles, interference between refractions from multiple particles and particles mistakenly identified as cloud particles by the software (Henneberger et al., 2013). The captured holographic images allow reconstruction of amplitude and phase information at any position in the measurement volume. The in-focus position of a particle in the hologram is determined by an edge detection algorithm. The lower detection limit for particles lies at 6 µm due to camera resolution. Sample amplitude and phase information of a droplet and artifact scaled to 10x10 pixels can be seen in figure 2.2.

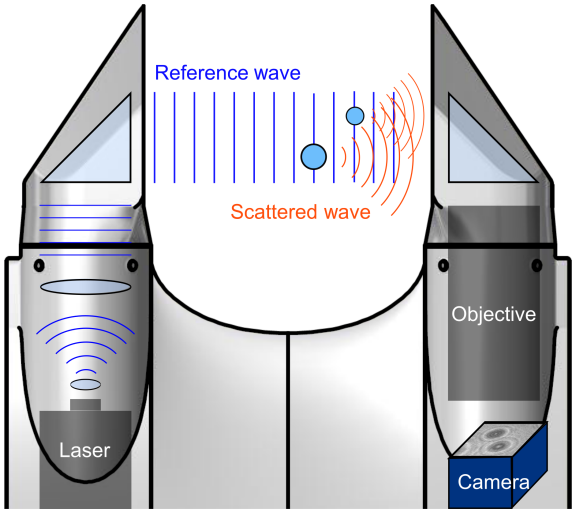
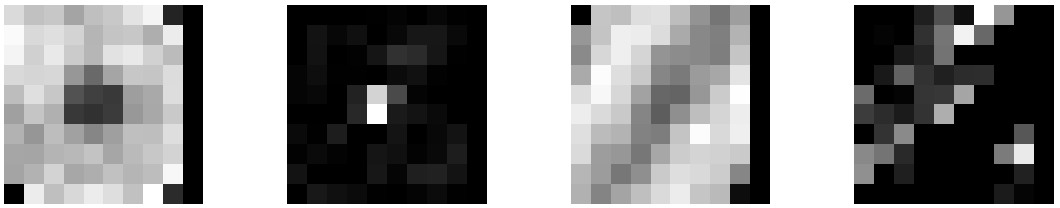


FIGURE 2.1: "Schematic of the working principle of digital in-line holography. A collimated laser beam is scattered by two particles. The scattered waves interfere with the reference wave and form an interference pattern (i.e., a hologram) which is recorded by a digital camera." Reprinted from Ramelli et al. (2020)



(A) Droplet Amplitude (B) Droplet Phase (C) Artifact Amplitude (D) Artifact Phase

FIGURE 2.2: droplet and artifact amplitude and phase images extracted by HOLOSUITE. Scaled to 10x10 pixels.

2.2 Data Description

Seven datasets that were collected between 2016 and 2019 in several field measurements were provided by the Atmospheric Physics Group at ETH Zürich. In this thesis, only particles with a major-axis smaller than 25 µm were considered, as an automated classification is already successfully applied to particles larger than 25 µm in their major-axis (Touloupas et al., 2020). When combined, the seven datasets contain a total of 44,358 labeled entries of particles smaller than 25 µm in their major-axis. Of these, 20,033 are artifacts and 24,325 are droplets. The size of the datasets as well as distribution of artifacts and droplets within each is shown in figure 2.3 and table (2.1). About half of the overall data comes from dataset 1 and 4. The two largest datasets 1 and 4 contain slightly more artifacts than droplets with droplet to artifact ratios of 0.89 and 0.93, respectively. The other five datasets contain more droplets than artifacts with ratios between 1.03 and 8.32. Datasets 5 and 7 have the strongest class imbalance.
Each entry in the dataset consists of a complex image, a label and metadata that was calculated from the image. The complex image can be split and depicted as amplitude and phase images. The label contains the manual classification of the researcher who evaluated the dataset. Due to the resolution and particle shape, particles can only be classified as droplet or artifact (Henneberger et al., 2013). Since the data was classified by humans, incorrectly classified and labeled particles are possible. Mislabeled particles can be caused by errors in handling the classification software as well as misinterpretation of the particle data. In addition, classification was performed by varying researchers depending on the dataset. Touloupas et al. (2020) estimate that for classification of droplets larger than 25 µm, there is a variation in classification decisions of ± 4 % depending on the person labeling the data.

Dataset	Particle Count	Droplets (D)	Artifacts (A)	D:A
1	12469	5870	6599	0.89
2	3135	1589	1546	1.03
3	3709	2147	1562	1.37
4	11118	5353	5765	0.93
5	4575	4084	491	8.32
6	5794	3008	2786	1.08
7	3558	2274	1284	1.77
Total	44358	24325	20033	1.21

TABLE 2.1: Dataset size and class ratio. The table shows the overall particle as well as droplet and artifact counts and their ratio.

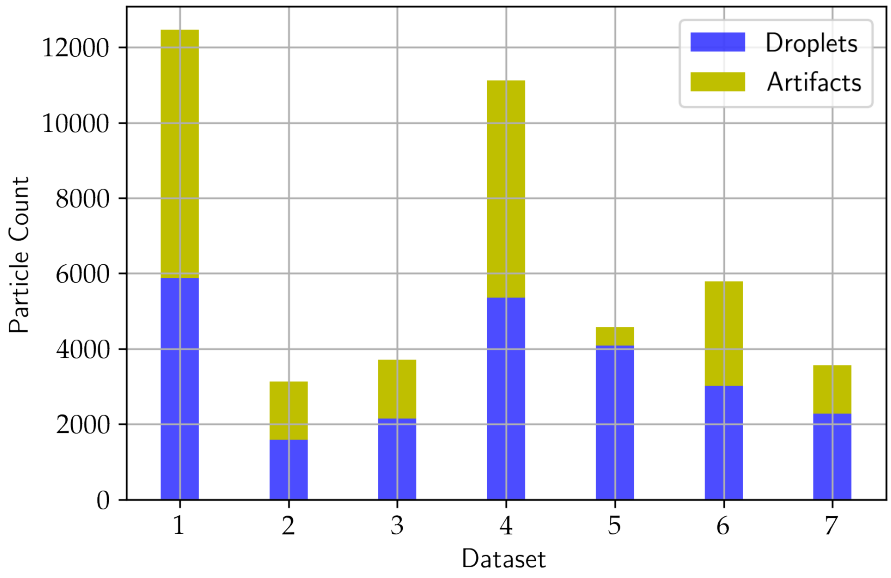


FIGURE 2.3: Artifact and droplet count by dataset. The figure shows the size differences between datasets as well as droplet and artifact proportions within each.

2.2.1 Image Data

Like in Touloupas et al., 2020, the 2D amplitude and phase images were used as input for the neural network. When extracted by HOLOSUITE, the dimensions of the images depend on the shape and size of the particle. The neural network architecture used to classify images in this thesis requires the images to have the same size and dimensions. This is done by selecting a size, zooming the image to the desired size and squaring the image by padding it with zeros. Unknown pixel values are also set to zero. For the larger particles a size of 32x32, pixels was chosen (Touloupas et al., 2020). The image size used for the smaller particles is determined experimentally as part of this thesis in section 4.1.2.

2.2.2 Metadata

In addition to the complex image data, the datasets contain between 48 and 113 additional metadata entries for each particle calculated by the software from the image data. A full list and description can be found in Schlenczek, 2018. Previously, metadata was used with decision trees and support vector machines to predict particle classes (Touloupas et al., 2020). However, the results achieved were surpassed by the use of convolutional neural networks (Touloupas et al., 2020). Whether supplementing image data with metadata improves classification performance for small particles, will be investigated in section 4.2 and section 5.2.

2.3 Data Distribution

Differences or shifts in data distribution between the datasets can hinder generalization performance between them (Storkey, 2009; Corfield, 2009). When assessing the data distributions, differences in dataset size, particle size and the ratio between droplets and artifacts are apparent. Additionally, pixel intensities or metadata between datasets might be scaled inconsistently or shifted. The possible influence of all of these needs to be accounted for in the eventual training data. Zadrozny, 2004 and Fan et al., 2007 describe sample selection bias in machine learning. If the training and evaluation data is not balanced properly, the classifier will be biased or evaluated unreliably. As a remedy for unbalanced data, Zadrozny, 2004 names resampling or cost sensitive methods. Types of both methods will be evaluated later on.

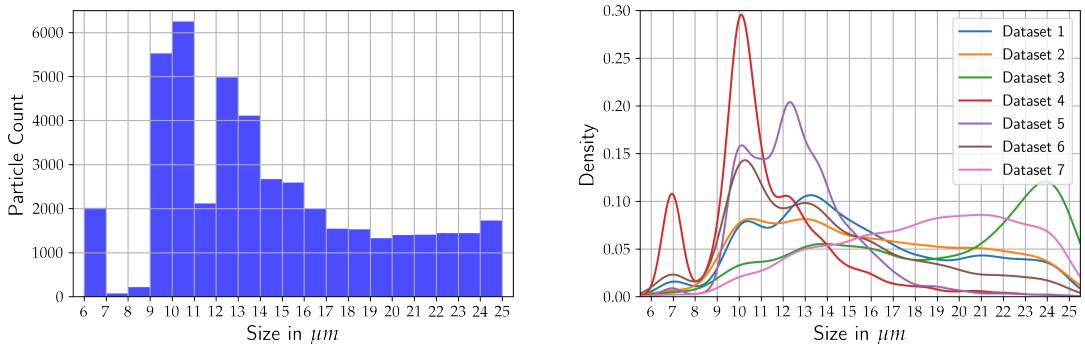
2.3.1 Dataset Size

Datasets 1 and 4 contain about half of the particles of all datasets with 12,469 and 11,118 particles, respectively. The other datasets are smaller and contain between 3,135 (dataset 2) and 5,794 particles (dataset 6). In order to assess generalization ability between datasets, classifiers will be trained on six of the seven datasets and evaluated on the seventh. A classifier trained on multiple datasets will be biased towards predicting the distribution covered by the larger datasets (here: dataset 1 and 4). The size differences between datasets will be addressed by resampling.

2.3.2 Particle Size Distribution

Particle major-axis size-distribution in the range between 6 µm and 25 µm is shown in figure 2.4a. The particles are binned in 1 µm size bins. There are peaks between 9 µm and 11 µm as well as between 12 µm and 14 µm. The peak between 9 µm and 11 µm is mostly caused by dataset 4. A kernel density estimation (KDE) can be used to estimate unknown probability density distributions (Silverman et al., 1989). Computing the KDE of the particle size for all datasets (figure 2.4b) reveals that the underlying distributions between datasets vary strongly. Datasets 1, 2 and 6 contain particles in all size ranges. Dataset 4 and 5 contain very few particles larger than 15 µm and datasets 3 and 7 contain mostly larger particles in the range between 15 µm to 25 µm.

Droplet and Artifact Size Distribution The artifact to droplet ratio for the combined dataset is dropping from small particle major axis-size to larger major-axis



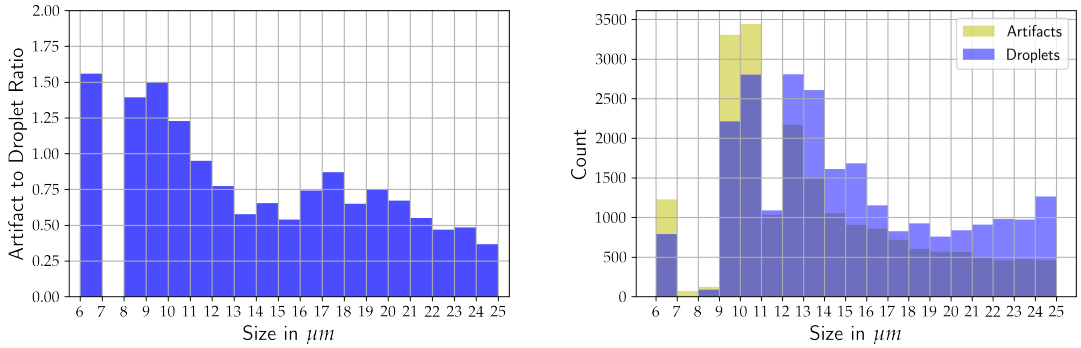
(A) Particle Distribution by Major Axis Size (B) KDEs by Particle Major Axis Size

FIGURE 2.4: Particle size distribution and particle density by size and dataset. Subfigure A shows the particle count accross all datasets in size bins. Subfigure B shows KDEs for the particle size distribution of all datasets. It reveals that particle size distribution between datasets varies significantly and that the peaks between 9 µm and 11 µm seen in subfigure A are caused mostly by dataset 4.

particle size as depicted in figure 2.5a. The artifact to droplet ratio peaks at 1.55 in the size bin from 6 µm to 7 µm and then steadily drops to 0.37 in the size bin between 24 µm to 25 µm. At all size bins larger than 11 µm, droplets outnumber artifacts. The gap of the ratio at the size bin 7 µm to 8 µm can be explained by a total lack of droplets at that size (figure 2.5b). Figure 2.6 shows artifacts and droplets binned by major axis size for all datasets. Artifacts are mostly found at the smaller sizes up to 15 µm as best seen in datasets 1, 4 and 6. Datasets 2 (figure 2.6a) and 4 (figure 2.6d) are balanced between droplets and artifacts across size bins. Dataset 3, (figure 2.6c) dataset 5 (figure 2.6e) and dataset 7 (figure 2.6g) are most unbalanced between classes. In addition to their class imbalance, dataset 3 and dataset 7 contain a lot more droplets in the size bins larger than 21 µm for dataset 3 and larger than 15 µm for dataset 7 compared to their smaller size bins. The particles in dataset 5 are focused between 9 µm and 17 µm and are mostly droplets with an overall droplet to artifact ratio of 8.32.

2.3.3 Pixel Intensity Distribution

To ensure that the datasets pixel intensities are comparable, pixel intensity distributions for each datasets amplitude and phase information were plotted. This was done by flattening the pixel intensities of calculated 20x20 images to a 1D list with 400 values for every particle. Figure 2.7 shows the pixel intensity probability distributions for phase and amplitude images for datasets 1 and 2. The

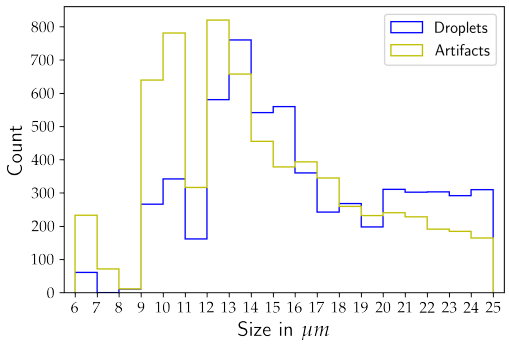


(A) Artifact to Droplet Ratio by Particle Major Axis Size

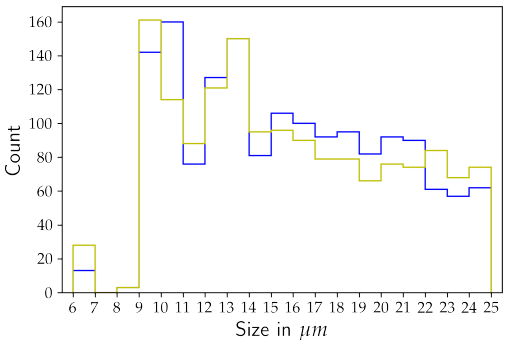
(B) Artifact and Droplet Distribution by Particle Major Axis Size

FIGURE 2.5: Artifact and droplet ratio and distribution. The artifact to droplet ratio in subfigure A shows, that the ratio drops steadily as particle major axis size rises. Subfigure B explains the gap at 7 µm to 8 µm which is due to a total lack of droplets at that size.

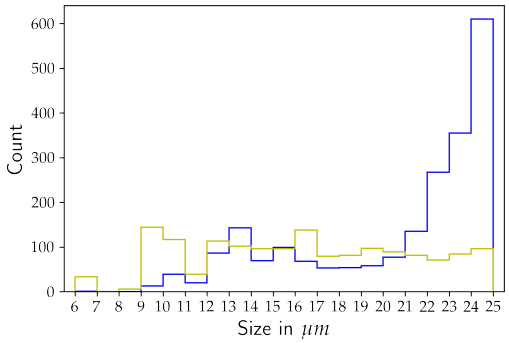
distributions for all other datasets can be found in 1. While the phase intensity distributions are very similar between all datasets, the amplitude intensity distributions vary between datasets. The amplitude intensities of Datasets 1, 3 and 7 reach values up to 0.9. Datasets 4 and 5 go up to 0.6 and dataset 6 only reaches up to 0.5. The expected intensity values for amplitude images are 0 to 1 and -1 to 1 for phase images. Every dataset contains amplitude images with intensities lower than 0 and outliers for the phase image intensities up to 4.74. The vast majority of pixel intensities lie in the expected range. The effect of shifting image intensities to a shared range needs to be explored later on during generalization experiments.



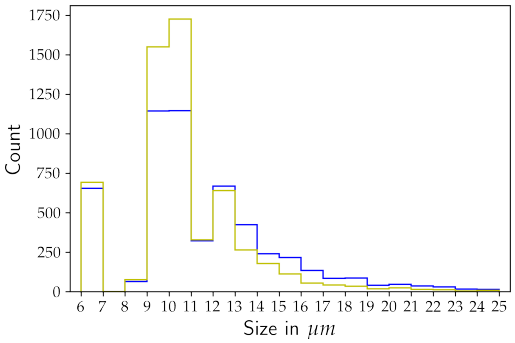
(A) Dataset 1



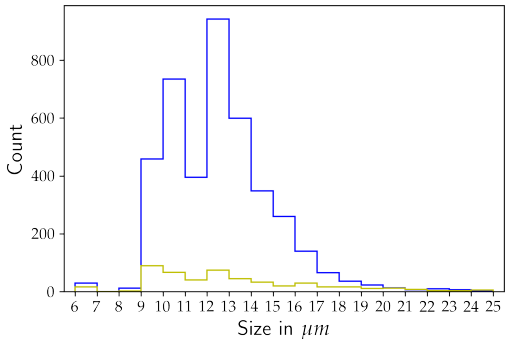
(B) Dataset 2



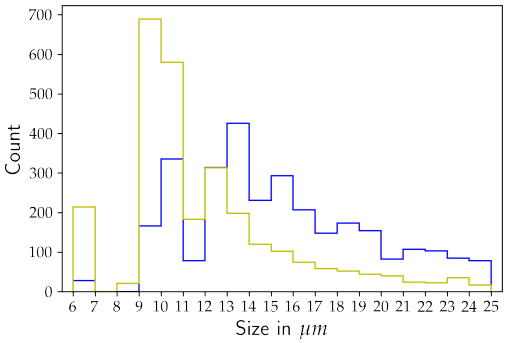
(C) Dataset 3



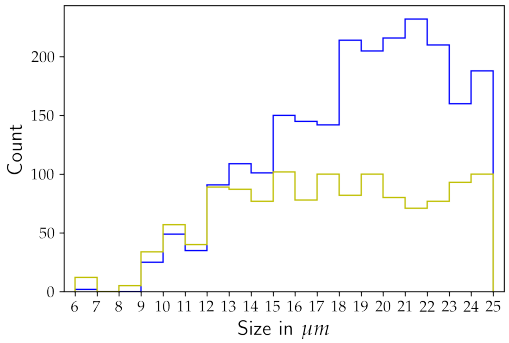
(D) Dataset 4



(E) Dataset 5

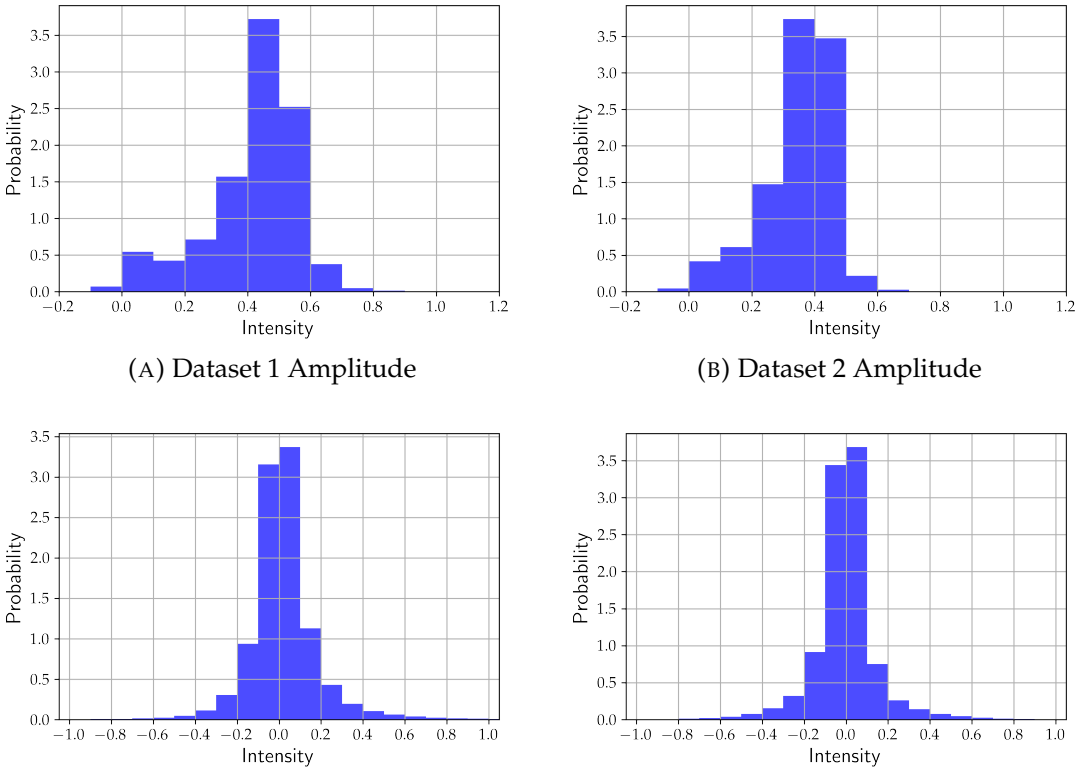


(F) Dataset 6



(G) Dataset 7

FIGURE 2.6: Droplet, artifact distribution by dataset and major axis size. The majority of artefacts are found at sizes below 15 µm. Subfigures C, E and G show the datasets with the highest class imbalance. In case of C and G the imbalance is caused by a larger number of droplets at the larger particle sizes.



(A) Dataset 1 Amplitude

(B) Dataset 2 Amplitude

(C) Dataset 1 Phase

(D) Dataset 2 Phase

FIGURE 2.7: Pixel intensity distributions for datasets 1 and 2. Differences in the amplitude intensity distribution are visible. Pixels of amplitude images in dataset 1 are much more likely to have an intensity between 0.5 and 0.8 than pixels of dataset 2. The amplitude image intensity peak for dataset 1 lies between 0.4 and 0.6 and between 0.3 and 0.5 for dataset 2.

Chapter 3

Neural Networks for Image Recognition

This chapter gives an overview over the techniques used to classify images using neural networks. The basic functionality of a neural network as well as the more task specific convolutional neural network will be introduced.

3.1 Neural Networks

In the following, neurons, dense layers and their training process in a network are layed out.

3.1.1 Neurons

The base component of a neural network is the neuron. Inspired by the biological neurons in the human brain, fundamentally, they take a number of inputs, process them and produce an output (figure 3.1). Their output is mathematically described by:

$$f(b + \sum w_i x_i)$$

The sum in the term describes the combination of inputs x_i. They are weighted under utilization of the weights w_i. Following this step, the bias b is added to the sum. The resulting value is passed to f, the activation function of the neuron which sets the output value of the neuron.

Activation Functions Figure 3.2 shows the commonly used Sigmoid and ReLU (Nair et al., 2010) activation functions.

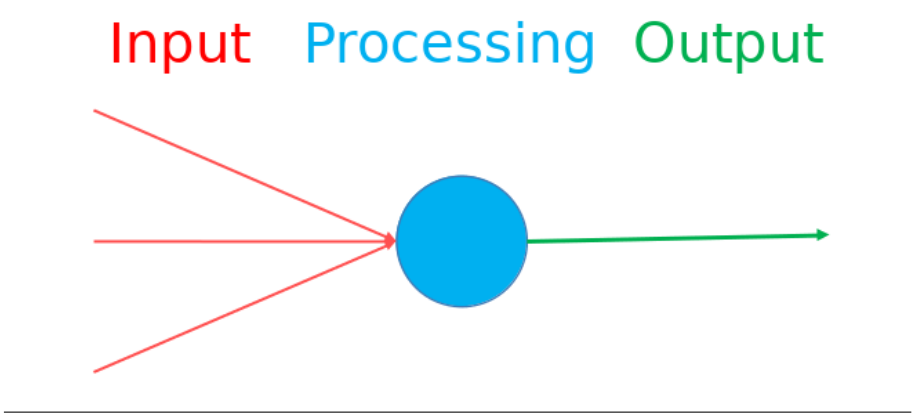


FIGURE 3.1: Neuron

Sigmoid The Sigmoid function (figure 3.2a), often referred to as logistic function, squishes the output of a neuron into the range [0,1]. It is defined as:

$$S(x) = \frac{1}{1 + e^{-x}}$$

The sigmoid function is commonly used in feed forward neural networks (Han et al., 1995) but is less suited for deep neural networks (Glorot et al., 2010a). Because outputs between 0 and 1 are ideal for binary classification tasks, a single neuron with a sigmoid activation function will be used for the classification between droplets and artifacts.

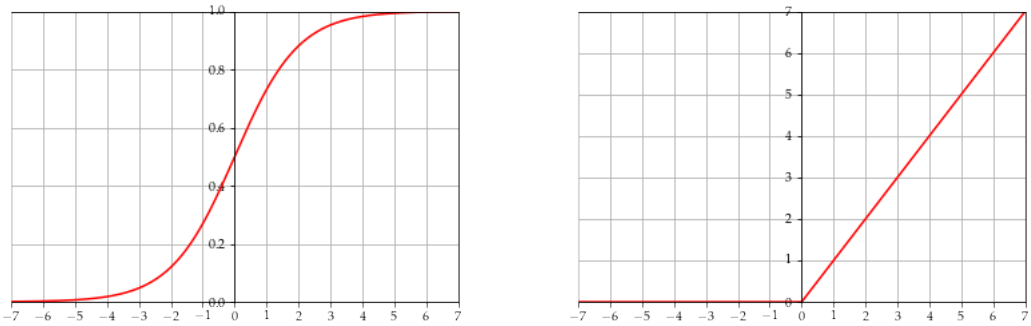
ReLU The relu function 3.2b defined as:

$$rectifier(x) = max(0, x)$$

has been shown to perform well in deep neural networks. Due to the lack of an exponential function, it is computationally cheap (Glorot et al., 2010b) (for example compared to the sigmoid function). The ReLU activation function is currently the most widely used and successful activation function (Ramachandran et al., 2017). It will be used in the neural networks fully connected layers which are introduced in the next section.

3.1.2 Dense Layer

The combination of multiple neurons in parallel is called a layer. Neurons within a layer do not interact with each other. The most basic form of a layer is the "dense" or "fully connected" layer. As implied by the name, every neuron in the layer is connected to every output from the previous layer.



(A) Sigmoid (B) ReLU

FIGURE 3.2: Activation Functions

3.1.3 Training Neural Networks

During the training process of a neural network, the weights and biases of all neurons in the network can be adjusted. Training a neural network is an optimization problem. When trained successfully, a network is able to model the relation between a set of inputs (input layer) and the outputs (output layer). Since the underlying relationship is not necessarily known, optimization happens by minimizing network loss which is calculated using a loss function. The loss function produces a scalar (loss) that summarizes network performance. To decide which changes should be performed to the neural network during training, the gradient of the network loss is calculated and weights and biases of the individual neurons adjusted to minimize loss. In the case of a labeled test data set, loss describes in a single number how close the predictions of the network are to the actual classes of the test data.

Underfitting and Overfitting Since the aforementioned loss is optimized on the training data, there is a risk to overfit it. This happens when the network learns relationships that are only predictive of the desired output in the training data but not in the real world. A possible solution to this is early stopping or the use of dropout layers.

3.1.4 Dropout Layer

In contrast to a fully connected layer, a dropout layer has a chance to randomly drop a previous layers' output and its connections, thereby randomly removing neurons. For every neuron in the previous layer there a chance to be removed, therefore removing its further influence during the current forward pass and back

propagation (Krizhevsky et al., 2012). The use of dropout layers to prevent overfitting and push the network to learn more robust features by diminishing the ability to co-adapt has been shown to be effective (Krizhevsky et al., 2012; Hinton et al., 2012).

3.2 Convolutional Neural Networks

3.2.1 Convolutional Layer

The convolutional layer is a fundamental component of a convolutional neural network that can be used to extract features from images. It uses convolutions which is a specialized type of linear operation (Yamashita et al., 2018). In contrast to a dense layer, the neurons in a convolutional layer are not connected to every input from the previous layer. In the image classification task at hand, the convolutional layers are used to learn regional features in the images. A detailed explanation can be found in LeCun et al., 1995.

3.2.2 Pooling Layer

Pooling layers are used to merge similar features from the convolutional layers and reduce dimensionality of the output (LeCun et al., 2015). In the network architecture used in this thesis (figure 4.1a), a convolutional layers is followed by a pooling layer which is in turn fed into a dropout layer.

3.2.3 Flattening Layer

After the convolutional and pooling layers, a flattening layer can be used to convert the last layers multidimensional output into an one-dimensional array that can be fed into a dense layer.

3.3 Classifier Evaluation

For the purpose of estimating particle concentrations in clouds, overestimating droplets is equally harmful to underestimating them. In order to understand whether the model over or underestimates the number of droplets, precision and recall for the droplet class are calculated. For model evaluation, their harmonic mean, the F1 score is used.

3.3.1 Confusion Matrix

To visualize precision, recall and F1, a confusion matrix for a given prediction can be calculated (see 3.1). For binary classification, it consists of four entries. In the this context where droplets are the positive class, the values can be interpreted as follows:

True positives (TP) - Droplets that were correctly predicted as droplets.
False positives (FP) - Artifacts that were falsely predicted as droplets.
False negatives (FN) - Droplets that were falsely predicted as artifacts.
True negatives (TN) - Artifacts that were correctly predicted as artifacts.

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{3.1}$$

3.3.2 Precision

Precision is the ratio between correctly predicted occurrences and total predicted occurrences. In the application to water droplets, the number of correctly detected droplets (TP) are divided by the number of correctly detected droplets (TP) added to the number of erroneously detected droplets (FP) in the dataset.

$$precision = \frac{TP}{TP + FP} \tag{3.2}$$

The value ranges from zero to one. The closer it is to one, the fewer false positives were predicted. A value of one means, that the prediction contained no false positives. If the precision has the value 0.4, it implies that of the droplets that were predicted as positive, only $40\,\%$were actually droplets.

3.3.3 Recall

The ratio between the correctly predicted occurrences of a class and the total actual occurrences of the class is called recall. Applying this to water droplets gives a measure of how high the chance of detection of droplets is. A value of one means that all droplets were found.

$$recall = \frac{TP}{TP + FN} \tag{3.3}$$

If the recall has the value 0.4 , it implies, that only 40 %of the droplets in the data were found.

3.3.4 F1 Score

Because for the application at hand, over and underestimating the number of water droplets is equally harmful. A combination of precision and recall can be used to take either into account. The F1 score is the harmonic mean of precision and recall.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3.4}$$

Here, precision and recall are equally weighed and lower extremes of either lead to degradation of the F1 score. Therefore it presents a balanced compromise between overestimating and underestimating concentration of droplets.

Chapter 4

Single Dataset Experiments

This chapter describes the experiments conducted to find preprocessing steps and network settings to fit all of the seven datasets well individually without considering generalization between datasets.

4.1 Single Dataset Optimization

To better understand the nature of the seven source datasets and get a benchmark to compare the generalization between datasets, classifiers were trained and optimized on all datasets. Each dataset was randomly split into training and test set (80-20 split). Training and hyperparameter optimization was performed by using only the training set. The test sets of each dataset are used for final evaluation after all model and hyperparameter tuning is finished. If the amounf of training data allows for it, classifiers are trained and evaluated using 5-fold validation (Burman, 1989) and the results for the folds are averaged. In this way, the results are always derived from all particles in the training data. Classification was performed using either the phase images, the amplitude images, metadata or combinations thereof. Figures shown are generally of the results for dataset 1 and only irregularities in dataset performance pointed out. In this chapter, the results for dataset 1 are shown as an example for all datasets. Irregularities in the results of other datasets are pointed out and results for datasets 2 to 7 shown in the appendix.

4.1.1 Network Architecture

The network architecture used is shown in figure 4.1. It contains a total of twelve layers. There are two convolutional layers, two pooling layers, four dropout layers, a flattening layer and three dense layers. The network has an input shape of (image width, image height, channel size). Because the images are squared, image width equals image size. The channel size depends on whether amplitude

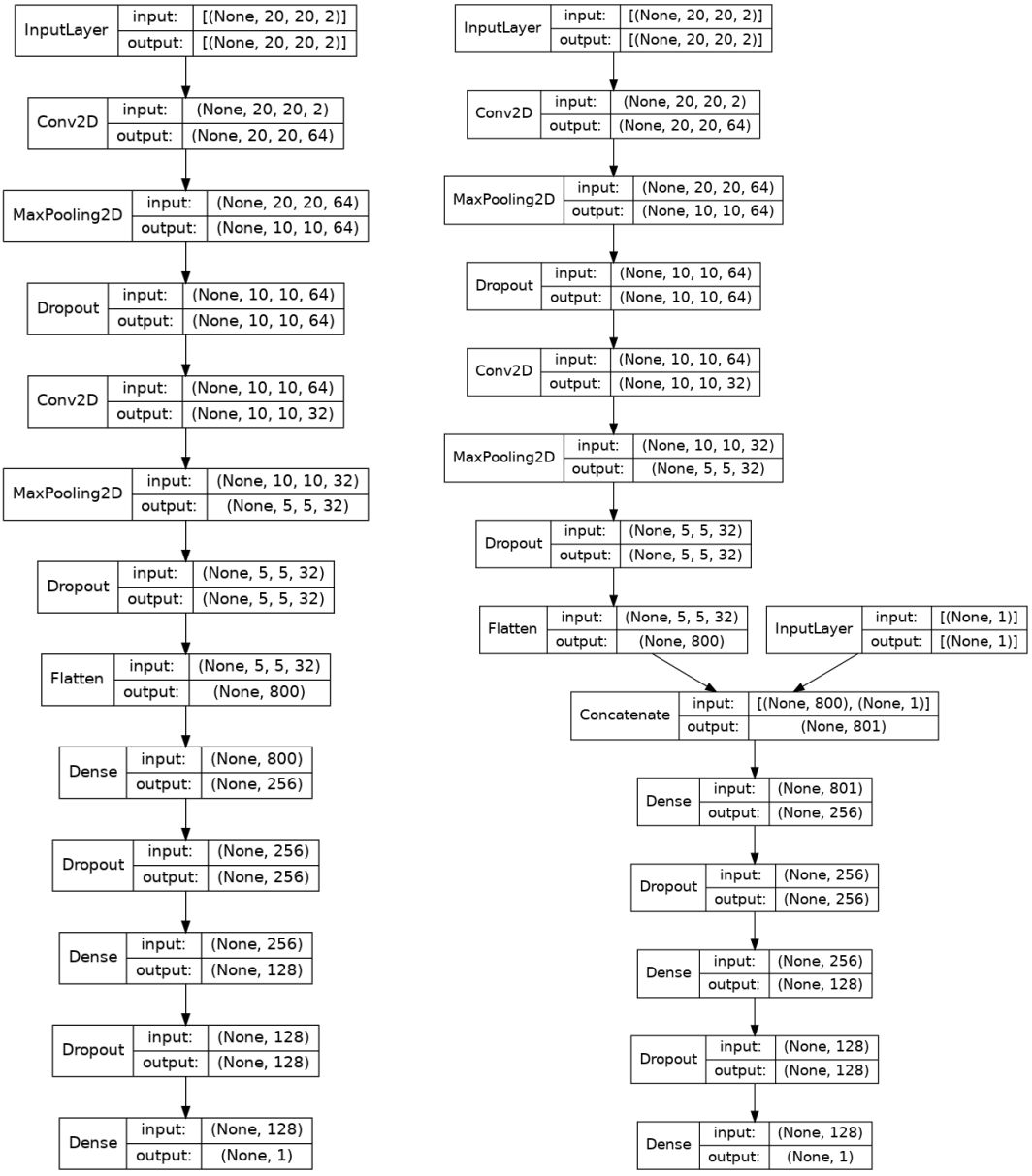
images, phase images or the combination of both is used and is equal to 1 in the first two cases and equal to 2 in the latter case. All dropout layers are set to use a dropout chance of 30%, meaning that 30% of neurons in dense layers during training are inactive. The output layer consists of a dense layer with a single neuron that uses a sigmoid activation function. If its activation is equal to or larger than 0.5, the network assigns the droplet class, if it is smaller, it assigns the artifact class to a given particle.

Network Extension Additional inputs using metadata or calculated features can be added by merging the output of any layer with additional features and feeding them into the next layer. Figure 4.1b shows the original network architecture as well as an example, where a one-dimensional feature was added.

4.1.2 Selecting Image Size and Image Data

The pixel size of the images is about 3 µm per pixel (Touloupas et al., 2020) on the x and y axis. As the particles classified are smaller than 25 µm in their major-axis, the largest particles in the dataset can be no larger than nine pixels across. During the particle extraction process from the hologram, four pixels of the background around the detected particle are added. Therefore, the maximum width of the images used as input is seventeen pixels. The images can be re-sized by zooming in, zooming out, or by cutting away the sides. Cutting the sides of particles to reach a specific image size might remove information that is useful for classification. To empirically find which image size works best and whether there are differences between the datasets, the previously introduced network without additional features (figure 4.1a) was trained and validated with five folds for all image sizes from eight to twenty pixels for all datasets. The process was executed a total of three times for amplitude images, phase images and the combination of both. A total of 1,365 models (7 datasets * 13 image sizes * 3 image configurations * 5 folds) were trained and evaluated.

Amplitude and Phase Image The five folds average recall, precision and F1 for the amplitude (figure 4.2a) and phase images (figure 4.2b) for dataset 1 shows a small improvement in F1 score as the image size gets closer to the maximum of twenty pixels. The results for datasets 2 to 7 can be seen in the appendix 2. Whether using the amplitude or phase images yields better results depends on the dataset. For all datasets, except dataset 2, using the amplitude images produces better results than using phase images. Except for dataset 6, droplet recall



(A) Base Model (B) Model With Added Input

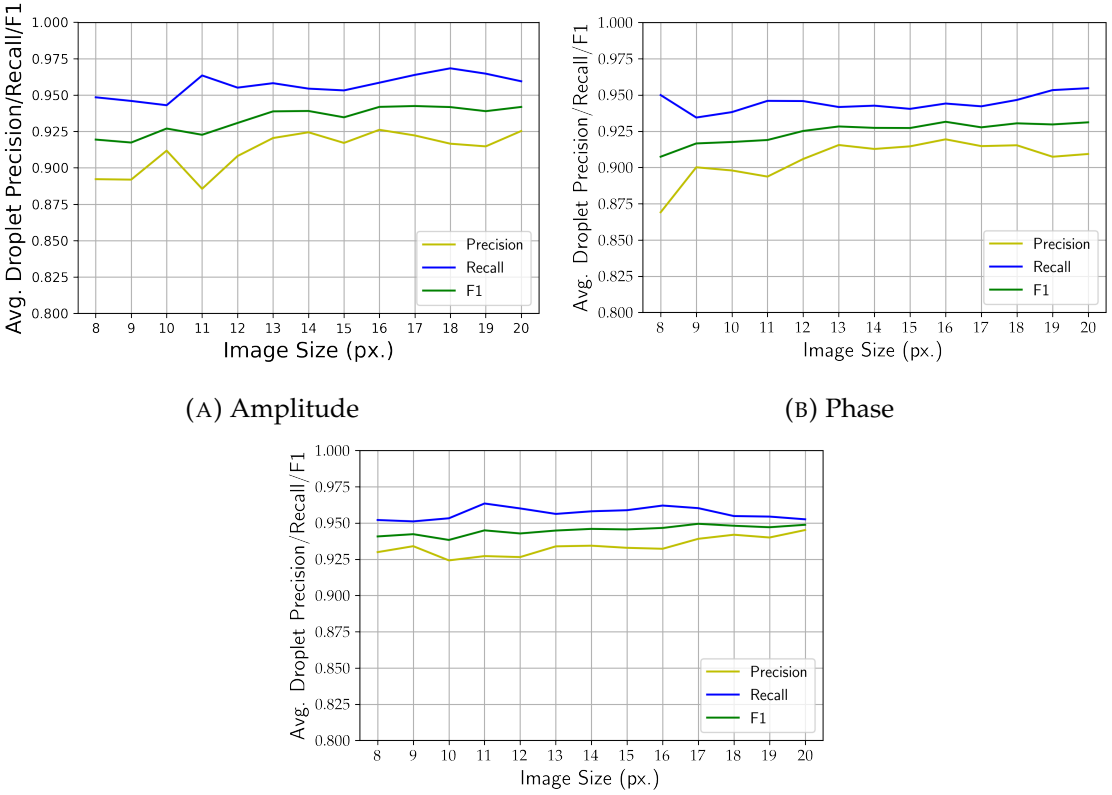
FIGURE 4.1: Neural network architectures used. Subfigure A shows the base model without added metadata. Subfigure B shows the same model as A but extended with a second input layer that is concatenated with the flattening layer that follows the convolutional section. The added input layers shape depends on the number of features that is added.

is higher than droplet precision for both amplitude and phase images. Their classifiers are very likely to find the droplets (high recall) but are prone to overestimating the number of droplets (lower precision) by misidentifying artifacts as droplets.

Combined Amplitude and phase images can be used simultaneously by adding another channel to the network's input layer. The input shape changes from (image width, image height, 1) to (image width, image height, 2). Figure 4.2c shows that the network trained in such a way on dataset 1 performs better throughout all image sizes and metrics than the network trained on either amplitude images or phase images alone. The F1 score is improved and recall and precision fall closer together on an overall higher level. Most of the improvement is caused by gains in precision. Enabling use of both channels helps to eliminate false positives. Again, using larger images slightly improves performance. Similar results are observed for datasets 2-7 (appendix 2). Datasets 2 and 7 are an exception, where using the combination of amplitude and phase images slightly reduced performance when compared to the better performance of either amplitude or phase images alone. The results when using both amplitude and phase images were more consistent across datasets when compared with the use of only one of them. Since consistency across datasets is important, both channels will be used. Image size has little impact on performance, in some cases, using larger images yields slightly better but never worse performance. **All further experiments will use an image size of *20x20* pixels and both phase and amplitude images.**

4.1.3 Addressing Class Imbalance in Training Data

As visualized in chapter 2, the source datasets vary strongly in their size and distribution. In the following, the more common class will be referred to as majority class and the less common class as minority class. In datasets 1 and 4, artifacts are the majority class. In all other datasets, droplets are. Class imbalance can lead to worsened prediction performance for the minority class (Nitesh V. Chawla et al., 2004) which often is the one of interest (López et al., 2012). In addition to the class imbalance, the distribution of droplets and artifacts by size is strongly imbalanced for all datasets. While evaluating different input image sizes, no mitigation for class imbalance was performed. Dataset 5 will be used in the following to show the effects of addressing class imbalance as it is the most imbalanced dataset with a ratio of droplets to artifacts of 8.32. During the image size evaluation on the amplitude image, droplet recall for dataset 5 was consistently equal to



(A) Amplitude

(B) Phase

(C) Combined

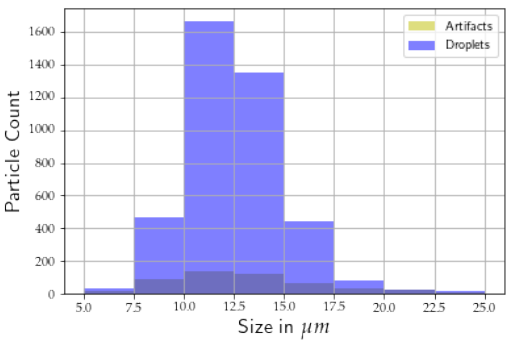
FIGURE 4.2: Image size selection. The figure shows the results for dataset 1 for networks trained on amplitude images, phase images and the combination of both. For all of them, performance increases slightly with image size. The combined amplitude and phase images produce the best results with precision and recall lying closer together and an overall higher F1 score.

1 (figure 2j). The network has a strong bias towards the majority class (droplets). Additionally, figures 4.3b and 4.3c show how the network performs when the particles are binned in eight equally spaced 2.5 µm size bins. The large class imbalance in the training data leads to the expected preference for droplets which is reflected in the lower droplet precision as well as the low artifact recall. Next to the particle distributions, figure 4.3a shows that droplet precision is closely linked to the droplet to artifact ratio. In this case of the presented major imbalance, only assessing droplet metrics can be misleading as the lack of artifacts can lead to high scores even if the classifier recognizes no artifacts at all. In further experiments, the training data imbalance will be adressed. Methods for balancing training data can be categorized into preprocessing and cost-sensitive learning (Haixiang et al., 2017). While preprocessing is concerned with resampling and feature selection, cost-sensitive learning assigns higher weights to the minority class and lower weights to the majority class to balance their impact during training.

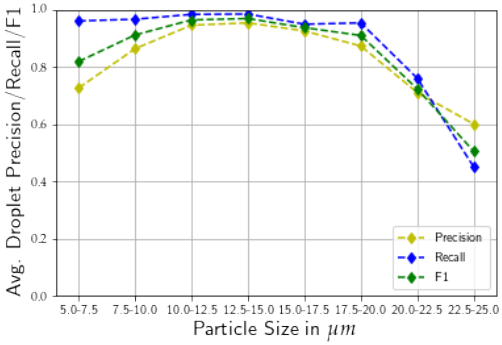
4.1.3.1 Resampling

Resampling aims to negate the imbalance by either reducing the number of samples of the majority class (undersampling), increasing the number of samples of the minority class (oversampling) or a combination of both (López et al., 2012). Resampling has the advantage, that it benefits a wide range of classifiers (López et al., 2013). Whether undersampling or oversampling should be chosen depends on the amount of available data per class (Loyola-González et al., 2016). If training data is abundant, applying undersampling methods can speed up classifier training without significantly reducing performance. If training data is scarce, further shrinking it by undersampling should be avoided and oversampling methods might be more appropriate. The following presents the results of undersampling and oversampling the training data when compared to performing no mitigation for the class imbalance. All tests were done using 5-fold cross-validation. The presented performance metrics are averaged over the five folds. As done previously, a 20% test set is kept aside from all training and optimizing activity for each dataset.

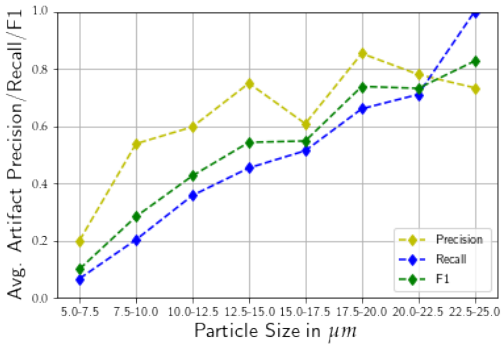
No Resampling As a benchmark for performance on non class-balanced training data, the averaged results for droplet and artifact precision, recall and F1 can be seen in figure 4.5. Throughout all datasets, the validation performance is very high with droplet F1 scores close to or above 0.9. Most of the datasets are very balanced (table 2.1) and similar precision and recall scores across both classes are



(A) Artifact/Droplet Distribution Dataset Five

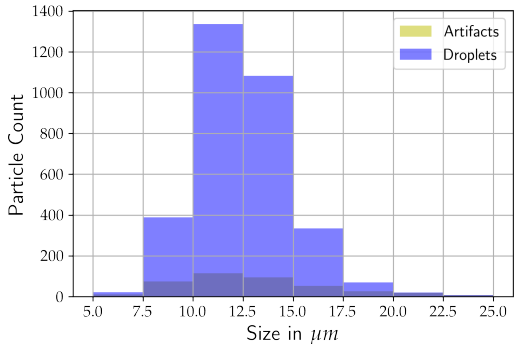


(B) Droplet Performance Metrics by Particle Size

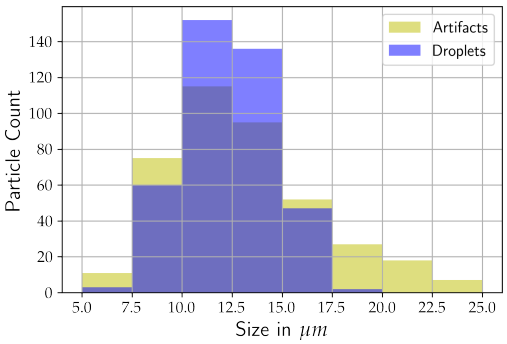


(C) Artifact Performance Metrics by Particle Size

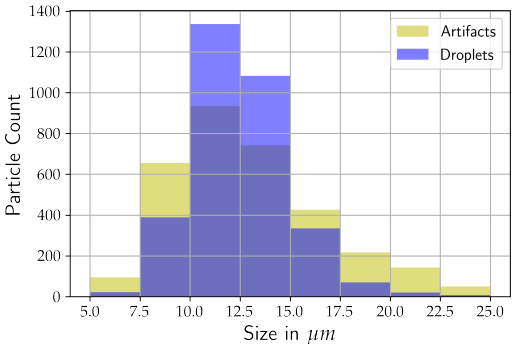
FIGURE 4.3: Neural network performance on the class imbalanced dataset 5. Subfigure A shows the distribution of droplets and artifacts in the dataset by major axis particle size. Especially artifact and droplet data is sparse in the size bins from 5 µm to 7.5 µm and in the size bins larger than 17.5 µm. Subfigure B shows, that droplet recall is close to one for most size bins and droplet precision depends on the ratio of droplets to artifacts in the size bin.



(A) No Resampling

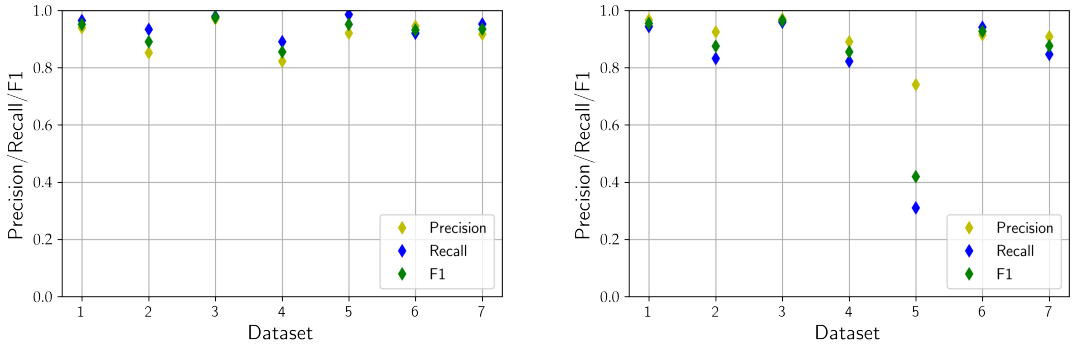


(B) Random Undersampling



(C) Random Oversampling

FIGURE 4.4: Dataset 5 particle distribution for different resampling methods. Subfigure A shows the class imbalanced unchanged dataset 5 artifact and droplet count in 2.5 µm major-axis particle size bins. For all size bins, the share of artifacts is small. Subfigure B shows the same dataset with applied random undersampling. The overall particle count is strongly reduced due to the randomly removed droplets. Subfigure C shows that random oversampling increases the number of artifacts without changing the droplets. While the distribution of droplets and artifacts accross size bins between random undersampling and random oversampling appears similar, the number of particles in each bin is substantially higher when random oversampling is applied.

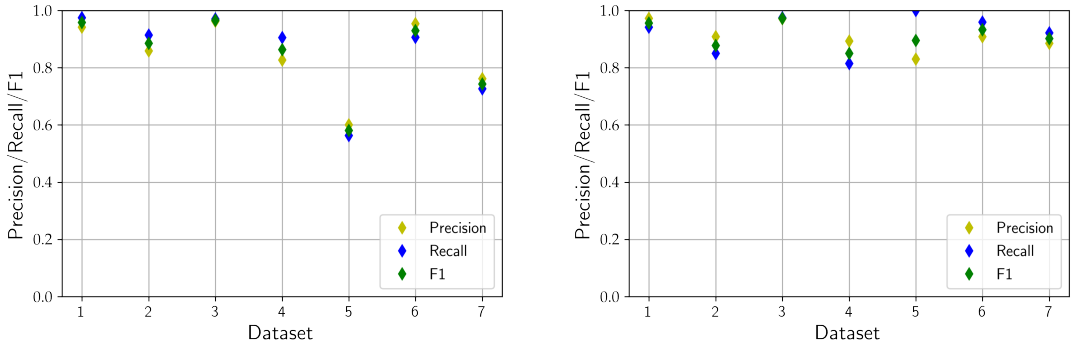


(A) Droplet Performance by Dataset (B) Artifact Performance by Dataset

FIGURE 4.5: Droplet and artifact performance on class imbalanced datasets. The classifiers for datasets 1, 3, 6 and 7 are very consistent for both droplets and artifacts. The classifiers for datasets 2 and 4 show a small bias towards predicting droplets. The classifier for dataset 5, which is the most imbalanced towards droplets, has a strong droplet recall and precision but very low artifact recall.

not surprising. Generally, droplet precision is a little higher than droplet recall while artifact precision is lower than its recall. This suggests that the models are classifying some artifacts as droplets and therefore the results slightly overestimate the droplet class count. As expected, the largest deviation in performance can be observed in dataset 5, which has the largest class imbalance. All artifact metrics for this dataset are low compared to the other datasets but due to the small number of artifacts, droplet metrics are barely affected.

Undersampling Random undersampling was tested to downsample the training datasets. To achieve class balance, random entries of the majority class are dropped until there is an equal number of samples for both classes (Batista et al., 2004). Doing so for the individual datasets leads to balanced classes as shown in figure 4.4b at the cost of drastically reducing training data especially for the imbalanced datasets. Since some datasets contain only few particles, effective 5-fold cross validation was not possible for many of the datasets when random undersampling was applied. Since a main advantage of undersampling lies in the lower computational cost (Liu et al., 2009) and currently training time is not an issue but lack of training data is, undersampling results are not shown. Another drawback of undersampling is the removal of possibly useful data (Liu et al., 2009; López et al., 2013). Mitigations by smarter selection of dropped samples such as condensed-nearest-neighbor have been proposed (Tomek, 1976).



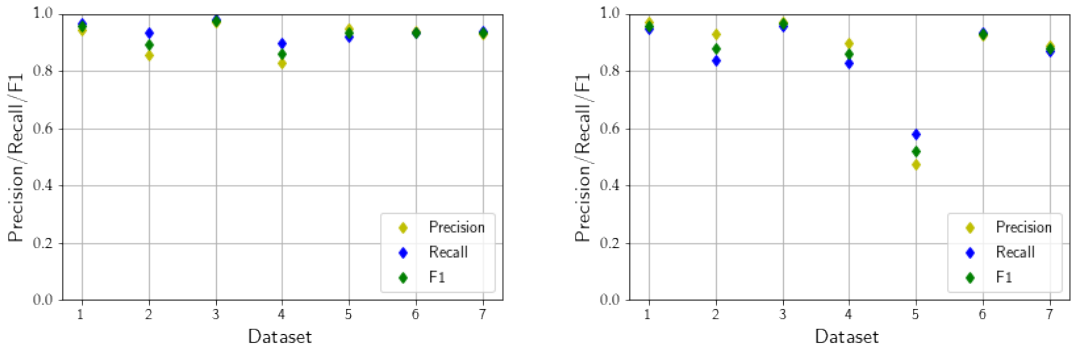
(A) Droplet Performance by Dataset (B) Artifact Performance by Dataset

FIGURE 4.6: Droplet and artifact performance on class balanced datasets (oversampled). Subfigure A shows, that the droplet performance of the most imbalanced datasets 5 and 7 decreased significantly compared to the results without oversampling. The biggest change is visible in the artifact metrics for dataset 5. Artifact recall for dataset 5 sits at 1 at the cost of overestimating artifacts and underestimating droplets.

Oversampling Random oversampling was used, which follows the same principle as random undersampling. Instead of dropping entries of the majority class, entries of the minority class are duplicated until class balance is achieved (Batista et al., 2004). According to Batista et al. (2004), random oversampling yields results that are competitive to more complicated oversampling methods (N. V. Chawla et al., 2002). Applying random oversampling to the datasets had little effect on most performance metrics. The most notable change can be observed in dataset 5 where the artifact performance improved substantially. The oversampled distribution for dataset 5 is shown in 4.4c. The improved artifact metric incur the cost of drastically lowered performance for the droplet class. When no resampling was applied, the network was biased towards droplets while now it is biased towards artifacts. Due to the initially low number of training samples for the artifact in this dataset, random oversampling might exaggerate anomalies or human misclassifications for this class as the individual artifacts have a much higher impact on the training process. Overall, while the results are more balanced than those seen without resampling, the loss in droplet performance is limiting the usefulness of the classifier for estimating cloud particle concentrations.

4.1.3.2 Cost-Sensitive Learning

As alternative to resampling, training and evaluation were repeated using adjusted class weight (Zadrozny et al., 2003). Before each training run, a class weight score was calculated for the training data. The weight of a given class



(A) Droplet Performance by Dataset (B) Artifact Performance by Dataset

FIGURE 4.7: Droplet and artifact metrics with cost sensitive learning. Subfigure A shows that the classifiers trained with class weights have high droplet F1 performances (above 0.85) and is very consistent between droplet precision and recall for all datasets except 2 and 4. Artifact performance is generally above 0.85 but less consistent between datasets than with random oversampling. Artifact metrics for dataset 5 are more balanced between precision and recall than without class weights but lower than when oversampling was applied

was calculated as:

$$classweight = \frac{totalcount}{2 * classcount}$$

The results shown in figure 4.7 are the most consistently high and balanced between recall and precision so far. The spread between precision and recall is minimized to a degree that was not achieved by either resampling method. Notably, using cost-sensitive learning did not improve the dataset 5 artifact F1 score but equalized precision and recall without hurting droplet performance which was the case when oversampling was used. For the intent, which focuses on recognizing droplets, the observed behaviour of minimizing spread between recall and precision, while focusing on droplet performance is ideal. Unless other reasons make the use of resampling beneficial, **further experiments use class weights during training**.

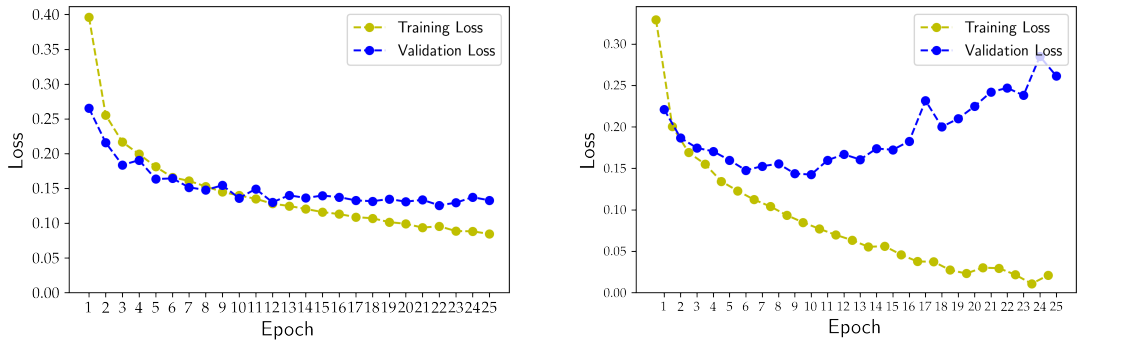
4.1.4 Early Stopping and Dropout to Prevent Overfitting

Training and validation loss averaged across the five folds were plotted with epochs on the x axis (figure 4.8a for dataset 1 and appendix 3 for datasets 2-7). Depending on the dataset, the models' validation loss (loss was introduced in section 3.1.3) stops to decrease after five to ten epochs. Training loss drops further, the longer the model is trained. Notably, for all datasets, validation loss

appears to start lower than training loss in epochs 1-4. This effect might be explained by three factors. The use of dropout layers, the time at which validation loss is calculated and the validation sets being easier to predict than the training sets. The dropout layers are active during training but not during the validation which might in part explain the lower loss during validation. While training loss is calculated continuously during the training epoch, validation loss is calculated at the end of the epoch. Since the network is continuously optimized, loss might be higher at the beginning of an epoch. This explanation can be confirmed by setting the dropout rate to 0% and shifting training loss 0.5 epochs to the left on the x axis. Setting the dropout rate to 0% prevents neurons from being dropped by the dropout layers during training. Shifting the training loss on the x axis by 0.5 epochs offsets the difference in time at which the network is evaluated. The possibility of easier validation data remains unaccounted for. Figure 4.8b shows that when the dropout is set to 0% and the training loss is shifted, validation loss lags slightly behind training loss, which is the expected behaviour. Since dropout rates and other hyperparameters can be subject to change, all models will be trained for 20 epochs. **Validation loss will be tracked during training and the model with the lowest validation loss will be used.** In addition to explaining most of the unexpected loss behaviour, plotting loss with and without dropout revealed, that the higher dropout rate prevents overfitting to some degree. Throughout all datasets, the spread between training and validation performance is larger in the later epochs when dropout is set to 0%. **The dropout rate of 30% will be kept in further experiments.**

4.1.5 Current Network Performance

To better understand how the networks perform and which particles are misclassified, the performances of the networks were plotted against the size of the evaluated particles. The validation data was binned in eight equally spaced 2.5 µm size bins. For every bin and dataset, the **median** value of the five validation folds is shown. In this case the median instead of the average is shown because the number of particles for each bin between validation folds might differ and taking the unweighted average would distort the results. The distribution of particles and artifacts in dataset 1 is shown in figure 4.9a. Graphs for the other datasets can be viewed in the appendix 4. As previously seen, artifacts are more concentrated in the smallest particle ranges and their numbers drop when getting closer to 25 µm. The metrics vary strongly between datasets and runs. For the single dataset valiation runs, splitting by size is unrealiable due to a lack of validation



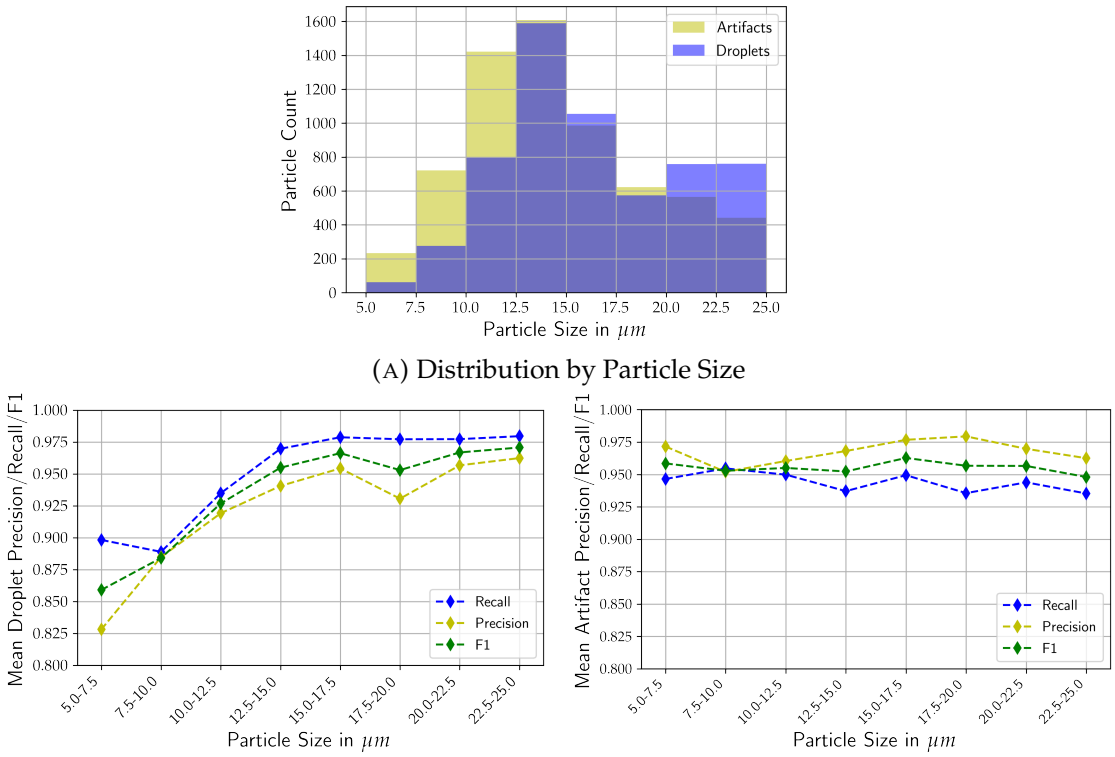
(A) 30% Dropout Chance (B) 0% Dropout Chance, Training Loss Shifted

FIGURE 4.8: Dataset 1 training and validation loss by epoch. The figure shows that the network overfits the training data very early (epoch five to eight) when dropout is set to 0%. Subfigure A shows unexpected behaviour of the validation loss as it is lower than training loss in the early epochs. Subfigure B shows that this effect might be caused by the 30% dropout rate used in subfigure A. Additionally in subfigure B, training loss is shifted by 0.5 to the left on the x-axis to account for the point in time at which loss is calculated during training.

data. Overall, the results show no particularly unexpected or low performance in any size range that can not be explained by data availability.

Droplet Metrics Figure 4.9b shows median precision, recall and F1 for droplets in dataset 1 for each size bin. Droplet performance is worst on droplets smaller than 12.5 µm. At sizes larger than 12.5 µm, performance levels off and slightly improves further moving towards the maximum particle size of 25 µm. It is noticeable that the three worst performing bins for the droplets are also those, where the droplets are outnumbered by artifacts by far. The smallest bin from 5 µm to 7.5 µm contains so few droplets, that the validation performance between folds varies strongly. In this regard, the unbinned performances are more reliable as they are derived from a larger number of predicted particles and more stable. The lower performance in the small particle size bins below 12.5 µm for droplets holds true even when small droplets are available as is the case in dataset 4 (figure in appendix 4h).

Artifact Metrics Performance on artifacts smaller than 12.5 µm in their major-axis is more consistent across particle sizes for dataset 1 (figure 4.9c) and higher than that of droplets in the smaller bins for datasets 2-7 (figures in appendix 4). This might simply be due to the larger data availability as there are more artifacts in the smaller size bins. Precision and recall scores lie very closely together for



(A) Distribution by Particle Size

(B) Droplet Performance Binned

(C) Artifact Performance Binned

FIGURE 4.9: Particle size distribution and prediction performance (dataset 1). The figure shows the major axis size distribution for droplets and artifacts as well as performance metrics for both classes binned by major-axis particle size. While droplet F1 is very high across all particle sizes (above 0.85, subfigure B), it is significantly lower (between 0.85 and 0.95) for the size bins smaller than 12.5 µm compared to the larger size bins (F1 score above 0.95).

droplets and artifacts throughout all datasets and size bins, which is desirable as it implies that even if the network produces some errors in classification, the overall number of droplets and artifacts can be estimated accurately.

4.2 Additional Metadata

The performance measures for single dataset training and validation reach median droplet F1 scores ranging from 0.847 to 0.975 and artifact F1 scores from 0.507 (outlier, dataset 5) to 0.957 (table 4.1). The precision and recall values are close to each other, meaning that cloud particle concentration could be estimated accurately and the F1 score for droplets is high for all datasets. Up to this point no metadata was added to the image information. Since depending on the dataset

	Droplet			Artifact		
Dataset	F1	Precision	Recall	F1	Precision	Recall
1	0.952	0.940	0.965	0.955	0.967	0.944
2	0.900	0.857	0.948	0.885	0.941	0.836
3	0.976	0.967	0.985	0.966	0.979	0.952
4	0.847	0.827	0.870	0.852	0.874	0.831
5	0.924	0.946	0.903	0.484	0.437	0.584
6	0.935	0.944	0.927	0.931	0.922	0.940
7	0.931	0.919	0.944	0.872	0.896	0.851
Avg.	0.924	0.914	0.935	0.849	0.859	0.848

TABLE 4.1: Averaged validation performance without added metadata. For all datasets except dataset 4, droplet F1 scores are above 0.90 with an average of 0.924. Artifact F1 scores are slightly lower at an average of 0.849 and all datasets except dataset 5 above 0.85. Precision and recall are very balanced for both the droplet and artifact class for all datasets.

there are between 48 and 113 entries available, this section evaluates whether adding metadata can improve performance.

4.2.1 Recommended Metadata

The Atmospheric Physics Group at ETH Zürich recommended the use of "underthresh", "asprat" and "dsqoverlz" to aid in classification, described by Schlenczek, 2018 as shown in table 4.2. Training and evaluation for the individual datasets was repeated with an additional input layer containing the three features as shown in figure 4.1b. Dataset 2 did not contain the "dsqoverlz" data and therefore it was not used in training and validation for that dataset. The averaged performance values (table 4.3) are very close to the results without the additional metadata. While the additional metadata might not further improve performance in this experiment, it might still be useful in generalization between datasets later on.

4.2.2 Feature Selection

The addition of the recommended metadata did not improve performance for the single dataset evaluation. To evaluate the other 48 to 113 additional metadata entries for every dataset individually, their predictive value was evaluated using decision trees. A decision tree seperates data into classes based on the data's features and their values. A decision tree consists of one root, branches, nodes and leafs. Each node presents a decision point at which the data is split. To decide

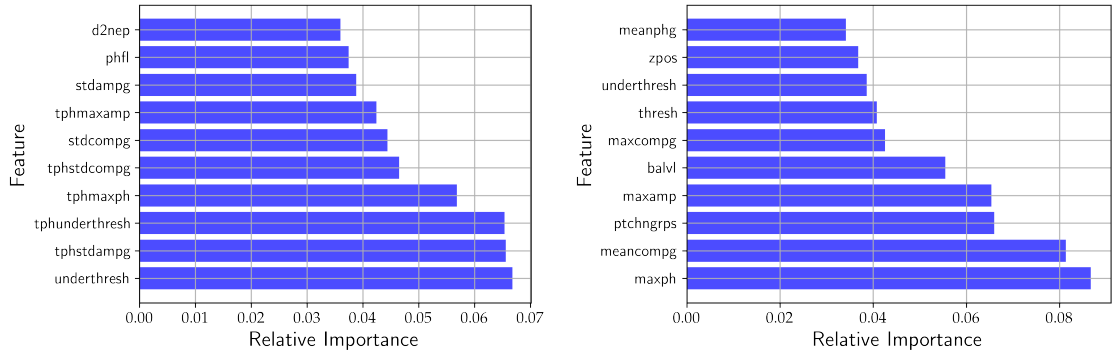
Acronym	Calculated from	Description
asprat	Lateral distribution of threshholded pixels	Particle aspect ratio, defined as the major axis length divided by the minor axis length
dsqoverlz	numzs, eqsiz	Ratio of particle area to the number of slices in the actual linked patch group
underthresh	thresh, minamp	Minimum amplitude value below threshhold

TABLE 4.2: Recommended metadata description (reprinted from Schlenczek, 2018)

	Droplet			Artifact		
Dataset	F1	Precision	Recall	F1	Precision	Recall
1	0.949	0.942	0.956	0.953	0.959	0.947
2	0.900	0.856	0.949	0.884	0.941	0.835
3	0.973	0.969	0.978	0.963	0.969	0.957
4	0.851	0.822	0.882	0.852	0.883	0.824
5	0.906	0.958	0.860	0.489	0.383	0.694
6	0.934	0.949	0.920	0.930	0.915	0.946
7	0.936	0.957	0.916	0.892	0.861	0.926
Avg.	0.921	0.922	0.923	0.852	0.845	0.875

TABLE 4.3: Averaged validation performance with recommended metadata. The table shows very similar results to those without the addition of any metadata.

which features should be used to seperate the data, the tree calculates information gain of the candidate features (Sugumaran et al., 2007). The use of a feature in a tree provides information about a feature's predictive value (Peng et al., 2008). The extracted relative feature importances from a random forest classifier with 20 trees for datasets 1 and 2 are shown in figure 4.10. The relative importance scores cannot be compared between datasets as the number of features varies between them and relative performance adds up to 1 for all features of a dataset. The recommended features *asprat* and *underthresh* can be found in most of the top ten charts 5. The third recommended feature, *dsqoverlz* is only found within the top ten for dataset 4. Models for the datasets were trained using the displayed ten most important features for each dataset.As was the case when adding recom-



(A) Dataset 1 (B) Dataset 2

FIGURE 4.10: Feature importance for datasets 1 and 2. The features are explained in detail by Schlenczek, 2018.

	Droplet			Artifact		
Dataset	F1	Precision	Recall	F1	Precision	Recall
1	0.954	0.943	0.965	0.957	0.967	0.948
2	0.897	0.866	0.931	0.885	0.924	0.850
3	0.972	0.969	0.975	0.961	0.965	0.957
4	0.847	0.846	0.849	0.858	0.861	0.857
5	0.936	0.939	0.933	0.491	0.492	0.509
6	0.931	0.939	0.924	0.926	0.919	0.933
7	0.930	0.919	0.942	0.871	0.894	0.852
Avg.	0.924	0.917	0.931	0.850	0.860	0.844

TABLE 4.4: Averaged validation performance with most predictive metadata. The table shows that with the top ten most predictive metadata added, the performance stays very close to the performance without the addition of any metadata.

mended features, averaged performance does not show any improvement when the top ten features with the highest predictive value according to a random forest model are added (table 4.4). **While the metadata has predictive value, it might not go beyond what is already extracted from the images by the convolutional layers.**

4.2.3 Circular Intensity Average

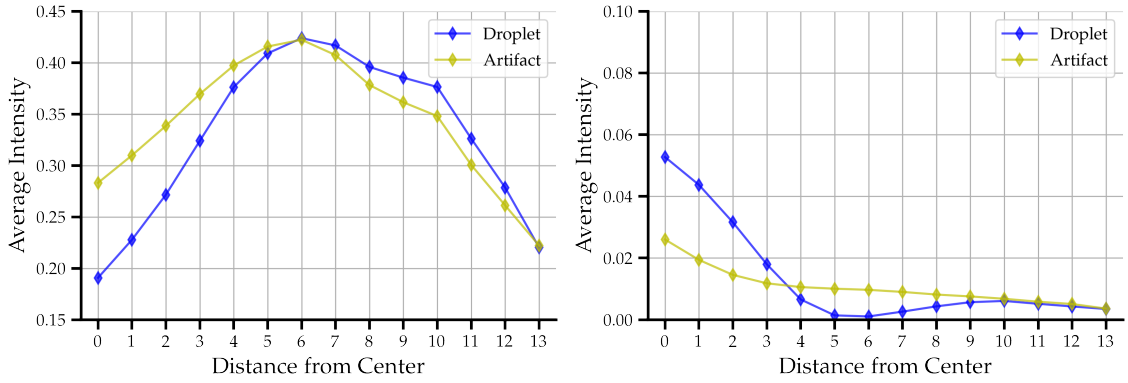
When manually deciding whether a droplet is a particle or artifact, the circularity of the particle or gradient between particle border and background might be considered. In order to numerically express the gradient between particle border and background, all image data for amplitude and phase was converted to a one-dimensional vector containing the average brightness of the circles with radius r

from the center. For a droplet, a low intensity center is expected with a gradual increase in brightness to the sides. For an artifact, a more random distribution of intensities is expected.

For the 20x20 pixel images, this process yields 14 values since the corners are a distance of 14 pixel widths away from the center. The result of averaging the circular intensities for droplets and artifacts for amplitude and phase images across all datasets is shown in figure 4.11. The drop-off in intensity at the larger distances from center is likely due to the zero padding that is applied during the prior re-sizing of the images. Plotting the average radial intensities for all of the seven datasets individually (figures in appendix 6), shows that dataset 2 behaves differently than all other datasets. Except for dataset 2, the average droplet intensities for amplitude images at distances from center up to 4-6 pixels are lower than those of artifacts. For phase images, it is reversed with the intensities for artifacts being lower than those for droplets in the center. Dataset 2 behaves in the opposite way. Additionally, for all of the other datasets, the average intensities for droplets and artifacts approach or cross each other when the distance from center is larger than 6 pixels. This does not happen in dataset 2. Instead droplet intensity is continually higher than artifact intensity for the amplitude images and the other way around for phase images. The differences between datasets and particularly the unusual behaviour of dataset 2 might lead to trouble in generalization. Calculating the circular intensities for every image and using it as an additional feature during training and prediction, had no noticeable effect on prediction performance.

4.3 Misclassifications and Confidence

The networks trained and validated on the single datasets appear to have reached their upper performance limit with droplet F1 scores between 0.847 and 0.976 (table 4.1). Adding metadata or calculating and adding average circular intensities did not substantially improve or alter performance. To understand the performance of the networks better, assessing the prediction confidences might be helpful. The output layer of the networks consists of a single sigmoid neuron which returns values between zero and one. The value can be interpreted as the confidence of the network for a given prediction. Values close to zero imply high confidence in an artifact prediction and values close to one imply a high confidence in the prediction of a droplet. An activation value close to 0.5 implies that the network is uncertain between the two classes for a particle. To express this behaviour on a single scale, the absolute difference between 0.5 and the activation



(A) Amplitude (B) Phase

FIGURE 4.11: Averaged radial intensities for amplitude and phase images accross all datasets. The unit for the distance is equal to the width of a pixel. Subfigure A shows the average for droplets and artifacts in amplitude images. Droplets have a lower intensity closer to the center. Subfigure B shows that in phase images, droplets have higher intensities in the center which quickly drops off to zero.

value was calculated and multiplied by two for every particle prediction:

$$confidence = |0.5 - activation| * 2$$

Regardless of particle class, the minumum possible confidence for a prediction is 0 and the maximum confidence 1. The following sections will examine whether the activation of the output layer's sigmoid neuron contains meaningful information that goes beyond the decision for a class. For example, whether mislabeled particles can be detected by reevaluating particles that were wrongly predicted with a high confidence.

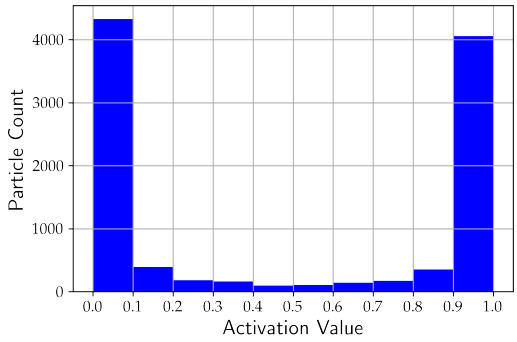
Distribution of Predictions Figure 4.12a shows the activation values for all particles of dataset 1 during validation. Of the activations for the correctly classified particles, 87.86% are either below 0.1 or above 0.9 implying a very high confidence. For the misclassified particles this value drops to 23.08%. The activations for falsely classified particles in dataset 1 are shown in figure 4.12b. **The figures show that the confidences given by the network for predictions of dataset 1 are higly reliable.** The networks predictions with high confidence are mostly correct and the networks predictions with lower confidence (below 0.8) are often incorrect. The confidences for dataset 1 binned by major-axis particle size for correct and incorrect predictions as well as for droplets and artifacts are shown in figure 4.12c. The activation distributions for all particles, incorrectly classified particles as well as confidences for datasets 2 to 7 are shown in the appendix, figure 8. For

all datasets, correctly classified particles have higher confidence values than incorrectly classified particles which confirms that the confidences are meaningful accross all datasets.

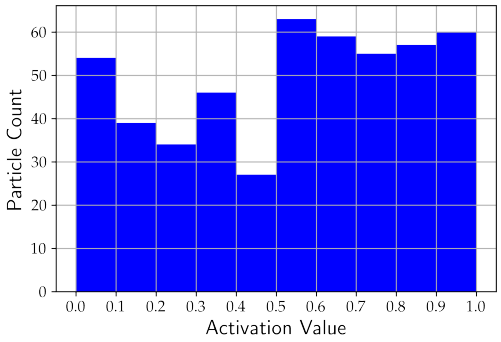
Correctly Predicted Particle Examples Figure 4.13 shows droplets and artifacts that were correctly predicted with a confidence of 1. Figure 4.13a shows the very clear amplitude image of a droplet. Its phase image is equally distinct 4.13b. Similarly, 4.13c and 4.13d show the amplitude and phase image of the most confidently and correctly predicted artifact. In contrast to the droplet, the boundaries of the artifact are not clear and there are multiple bright and dark spots throughout.

Wrongly Predicted Particle Examples Having examined the examples for correctly classified droplets and artifacts that were predicted with the highest confidence, figure 4.14 shows the falsely classified particles from dataset 1 that were classified with the highest confidence. For example, if the particles true class is droplet, and the network predicted artifact with a high confidence. This case is shown in figures 4.14a and 4.14b. The sigmoid had very low activation of 8.74×10^{-12} which is close to perfect confidence for a predicted artifact. However, the true class of the particle as labeled by the human is droplet. In this case, the human classification might need to be questioned. However, it is possible that the human had access to additional metadata which clearly identified the particle as droplet. From the amplitude and phase images, the presumed misclassification by the network seems very understandable. An example of the opposite case can be seen in 4.14c and 4.14d. Here, the network predicted a droplet with an activation of 1 while it was labeled as artifact. Again, the image appears to be a perfect example of a droplet and the human classification needs to be questioned. This might hint, that in some cases, the network is more consistent than its human counterparts which might falsely label particles due to clicking on a wrong button.

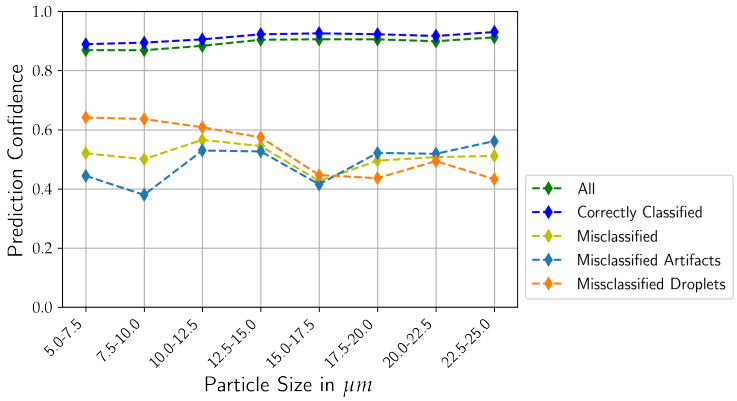
Edge Cases Having looked at correctly and falsely predicted particles (compared to the label assigned by a human which is assumed to be always correct), Figure 4.15 shows the least confident correct predictions of a droplet and an artifact. The droplet was predicted with an activation of 0.5266 and the artifact with an activation of 0.4984. As expected, both appear to be in between of clear artifacts and droplets. They are not as obviously round or messy as the previous examples. In these cases, the decision might be unclear for a human as well.



(A) All Particles



(B) Misclassified Particles



(C) Confidence by Particle Size

FIGURE 4.12: Output neuron activation and confidence (dataset 1). Subfigure A shows that the overwhelming majority of activations are below 0.1 or above 0.9 which means that the network is confident in the predicted class. Subfigure B shows the activations for the misclassified particles. Contrary to subfigure A, the confidences of the misclassifications are much more evenly distributed in the whole space between 0 and 1. This suggests that the network was less confident about the misclassified particles than the correctly classified particles. This is further visualized in subfigure C. The correctly classified particles are classified with confidence values around 0.90 while the misclassified particles are predicted with confidences around 0.5. Subfigure C shows that the confidences values work well for all particle size bins as well as both the artifact and droplet class.

(A) Droplet Amplitude (B) Droplet Phase (C) Artifact Amplitude (D) Artifact Phase

FIGURE 4.13: Sample amplitude and phase images of correctly classified particles with high confidence. These are very clear examples of the classes that network classified correctly.

(A) Mislabeled Droplet Amplitude (B) Mislabeled Droplet Phase (C) Mislabeled Artifact Amplitude (D) Mislabeled Artifact Phase

FIGURE 4.14: Sample amplitude and phase images of misclassified particles with high confidence. The subfigures show amplitude and phase images of wrongly classified particles. Subfigure A and B are of a particle that was labeled as droplet but was predicted by the model as an artifact with high confidence. Subfigures C and D show the same for a predicted droplet that was labeled as an artifact. In both cases, the classifier uncovered wrongly labeled particles in the dataset. **The networks' predictions are correct but were marked as false due to the erroneous label.**

(A) Droplet Amplitude (B) Droplet Phase (C) Artifact Amplitude (D) Artifact Phase

FIGURE 4.15: Sample amplitude and phase images of correctly classified particles with low confidence.

Having at some instances of wrongly and correctly predicted particles, all decisions by the network seem reasonable and possibly more consistent than those of the human who labeled them originally especially in the cases where the network has a high confidence in its prediction.

4.4 Generalization Between Datasets

After having achieved high performance results on all single datasets, and shown that the confidence values provided by the network correlate to the correctness of a prediction, generalization between the classifiers that are trained on single dataset and the other datasets can be tested. The same network architecture as before was trained on the data used for the k-fold validation (80% of the datasets). The resulting networks then were used to predict the validation data of all other datasets and the previously unsused 20% test data of the training dataset. Table 4.5 shows droplet F1 scores of the classifiers trained on datasets 1-7 for all

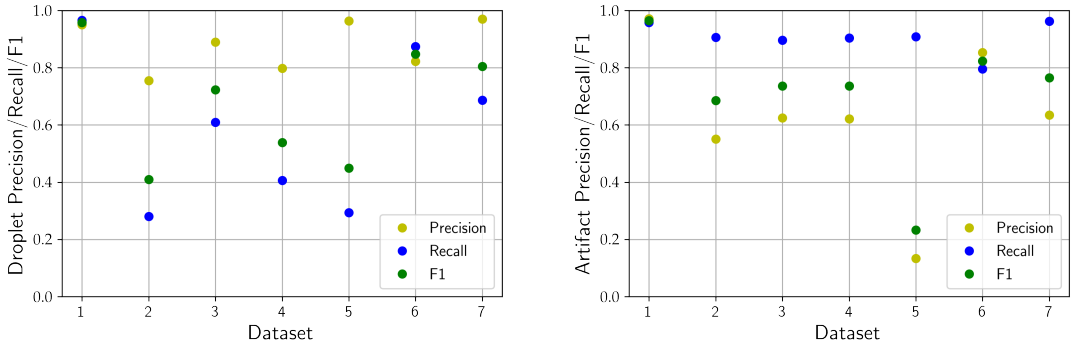
		Predicted Dataset						
		1	2	3	4	5	6	7
Trained Dataset	1	*0.955*	0.251	0.680	0.596	0.512	**0.865**	**0.817**
	2	0.455	*0.877*	0.482	0.229	0.020	0.282	0.555
	3	0.333	0.714	*0.981*	0.394	0.650	0.658	0.619
	4	0.805	0.703	0.839	*0.860*	**0.915**	0.817	0.720
	5	0.759	0.543	**0.840**	**0.784**	*0.928*	0.805	0.674
	6	0.634	0.188	0.707	0.451	0.667	*0.934*	0.549
	7	**0.869**	**0.750**	0.788	0.745	0.876	0.777	*0.946*

TABLE 4.5: Generalization between datasets (droplet F1 score). The diagonal from top left to bottom right (italic) signifies how well the classifiers predict their previously unused test set. Row 1 shows the performance of the classifier trained on dataset 1. It did poorly on dataset 2 (droplet F1 of 0.25) and very well on dataset 6 (droplet F1 of 0.865). The classifer trained on dataset 2 was the worst in terms of generalization with a maximum droplet F1 score of 0.555 on dataset 7. **No single classifier was best at predicting all other datasets.** Dataset 7 had the most consistently high droplet F1 performance for the other datasets.

datasets. Reading the table column wise shows that for every dataset, a classifier trained on another dataset exists that can at least predict it with a droplet F1 score of 0.750 (dataset 2). Except for dataset 2 and 4, all datasets can be predicted with a droplet F1 score of at least 0.817 by a classifier trained on another dataset. Notably, classifiers trained on datasets 2, 3 and 6 are the worst at generalizing to new data. The classifier trained on dataset 7 is the most consistent when predicting other datasets with high performance (lowest droplet F1 score of 0.745 on dataset 4).

4.4.1 Confidences On Generalization

Figure 4.16 shows droplet and artifact metrics for the dataset 1 model in detail. The classifier appears to have trouble finding droplets but on the droplets that it predicts does so with a high precision. This can be seen in the high droplet precision and lower droplet recall scores. While the confidence values for its predictions on its own validation data were very related to the prediction correctness (figure 4.12), the relation of prediction confidence to actual prediction correctness depends on prediction performance when generalizing onto other datasets. If the generalization performance is high, the confidence values are accurate, if it is low, the confidence values are less accurate. This behavior is shown on examples in the following two paragraphs.

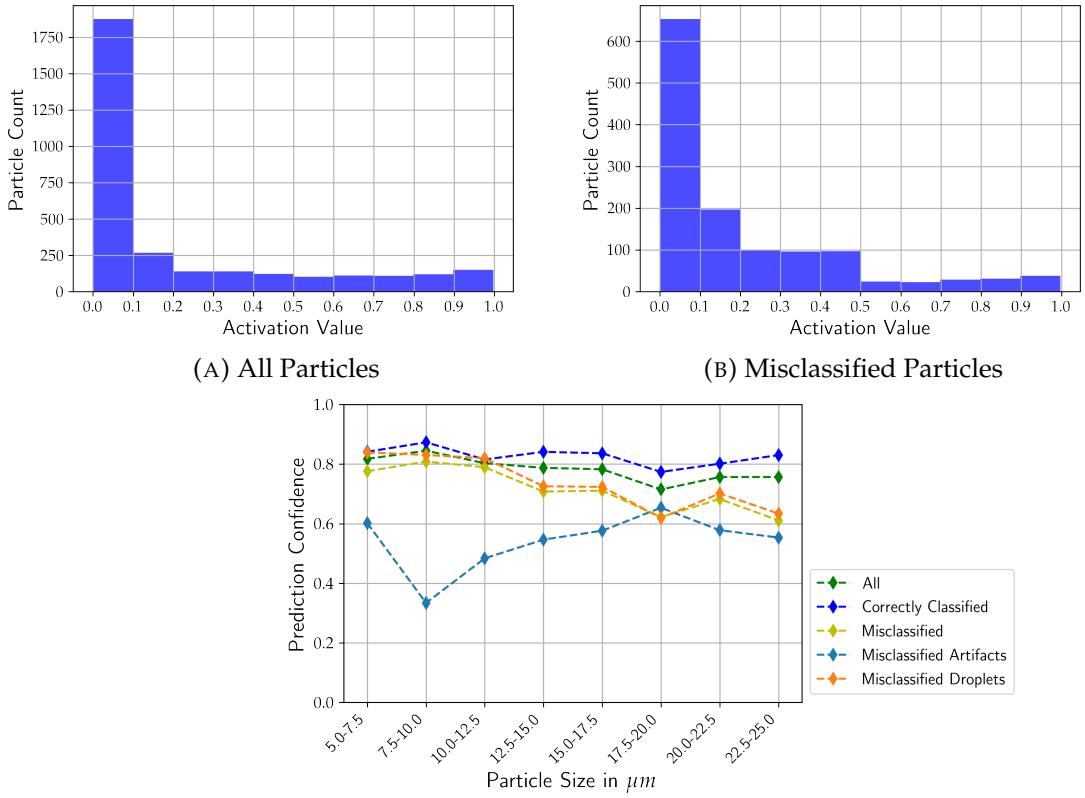


(A) Droplet Metrics (B) Artifact Metrics

FIGURE 4.16: Dataset 1 model droplet and artifact metrics when predicting its test data and datasets 2-7. Subfigure A shows that the model has high droplet precision scores for all other datasets but misses a large number of droplets in most, except dataset 6. Subfigure B shows that the model has high artifact recall but low artifact precision when predicting other datasasets.

Dataset 1 Classifier on Dataset 2 Figure 4.17 shows the activations of the network trained on dataset 1 for the particles of dataset 2. Dataset 2 was chosen because the droplet F1 scores of the classifier trained on dataset 1 were the lowest on dataset 2. The figure shows a strong bias in predictions towards the artifacts class (activations close to zero). The artifact predictions are very confident. Contrary to the previously observed behavior, the misclassified particles were only predicted slightly less confident than the correctly classified particles. The network performs badly and is "unaware" of its low performance. The confidences for misclassified artifacts are an exception to this as they are consistently lower than the confidences for misclassified droplets. **When the classifier predicts a dataset poorly, the prediction confidences are less indicative of the predictions correctness.**

Dataset 1 Classifier on Dataset 6 Repeating the analysis performed on the activations of the classifier trained on dataset 1 when predicting dataset 2 for the predictions of dataset 6 paints a different picture. Dataset 6 was predicted very well by the classifier trained on dataset 1 (droplet F1 of 0.865). The activations and confidences are shown in figure 4.18. They closely resemble the activation distribution and confidences on the validation sets of dataset 1 (figure 4.12). **When the classifier has a high generalization performance onto a new dataset, the confidences are closely related to the predictions correctness.** However, before labeling a dataset by hand, there is no way to tell whether a classifier generalizes well onto a new dataset as there are no labels to compare the predictions to.

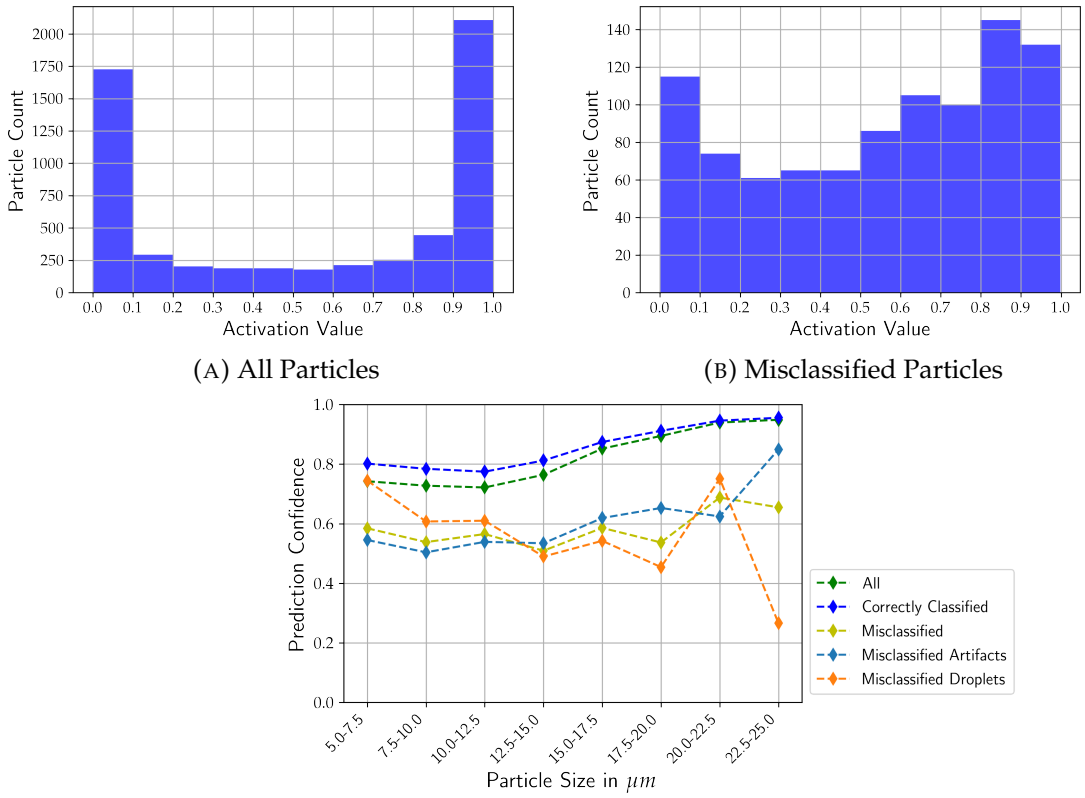


(A) All Particles

(B) Misclassified Particles

(C) Confidence by Particle Size

FIGURE 4.17: Output neuron activation and confidence (dataset 1 classifier on dataset 2). Subfigure A shows, that the classifier predicted mostly artifacts with activations close to zero. The activations for misclassified particles (subfigure B) are only sligtly more spread between 0 and 1, suggesting that the classifier is unaware of the misclassifications. Subfigure C further confirms this with the exception of misclassified artifacts where confidences are lower (misclassified artifacts are labeled artifacts that were predicted as droplets).



(A) All Particles

(B) Misclassified Particles

(C) Confidence by Particle Size

FIGURE 4.18: Output neuron activation and confidence (dataset 1 classifier on dataset 6). Subfigure A shows activations accross the spectrum without a high bias towards one or the other class. Subfigure B shows that the activations for misclassified particles are more evenly distributed between 0 and 1 which shows lower confidences for wrong predictions. Subfigure C confirms this and shows that the confidence values are slightly more accurate in major-axis particle size bins larger than 15 µm.

Checking whether there is a strong bias in activations might provide some information about whether the classifier is able to predict both classes, but needs to be used carefully, as the bias might also be caused by unbalanced classes in the new data.

	Dataset						
	1	2	3	4	5	6	7
Self	0.955	0.877	0.981	0.860	0.928	0.934	0.947
Other	0.869	0.750	0.840	0.784	0.915	0.865	0.817

TABLE 4.6: Benchmark droplet F1 results (single dataset). The table shows the validation droplet F1 performance achieved by a classifier trained on the respective dataset (self) and the best droplet F1 performance achieved when predicted by a classifier trained on another dataset (other). As expected, "other" is always lower than "self".

4.5 Conclusion of the Single Dataset Experiments

The conducted experiments have shown that each of the seven datasets individually can be learned and predicted well. The generalization performance is highly variable and inconsistent. Table 4.6 shows the benchmark droplet F1 scores that will be used to compare the results in the next chapter.

Chapter 5

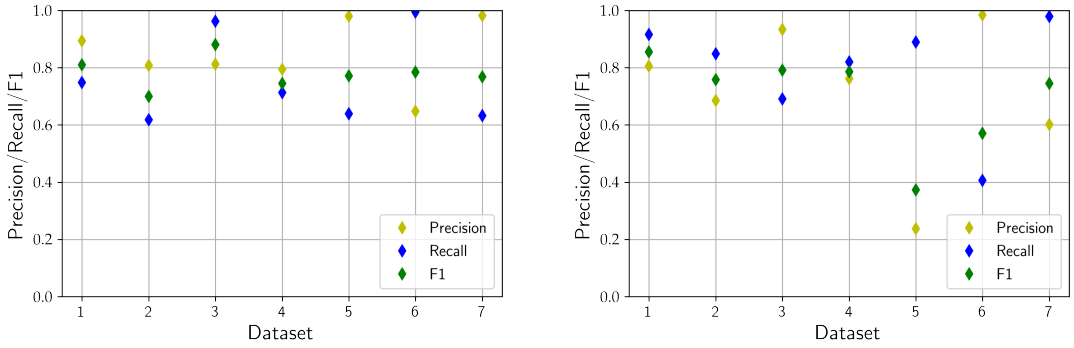
Combined Datasets Experiments

5.1 Combining Training Data

No classifier trained on a single dataset was able to predict all other datasets reliably. Datasets will be combined to achieve better generalization. Previous work (Touloupas et al., 2020) on larger particles has shown, that combining datasets improved prediction performance on new, unseen datasets. Therefore, new classifiers were trained to predict each of the seven datasets. Each classifier is trained using all datasets except the one which it is meant to predict. The classifiers are evaluated on a validation set, which is an unseen subset of the six datasets used to train the classifier (referred to as validation performance) as well as on the unseen seventh dataset (referred to as test performance). Validation performance expresses how well the classifier fits the combined training data and test performance how well the classifier generalizes onto the unseen dataset. All tests are done using 5-fold cross validation using 100% of the six training datasets. Validation and test performance is calculated for each fold and averaged.

5.1.1 Unbalanced Combined Datasets

In the first experiment, the six training datasets for each classifier were combined at their original size. Therefore, a dataset's influence on the trained classifier depends on its size. The larger a dataset, the higher its influence. The generalization performance of the classifier trained in this manner (figure 5.1) is more consistent across datasets than the generalization performance of the classifiers trained on the single datsets, but also does not reach the droplet F1 scores seen in chapter 4 for any of the predicted datasets. Except for datasets 3 and 7, droplet recall is lower than precision, suggesting that the classifier cannot identify some droplets from the unseen datasets. The generalization experiments on the classifiers trained on the single datasets have shown (see table 4.6, row "other"), that all seven datasets can be predicted with higher droplet F1 scores than were achieved



(A) Droplets (B) Artifacts

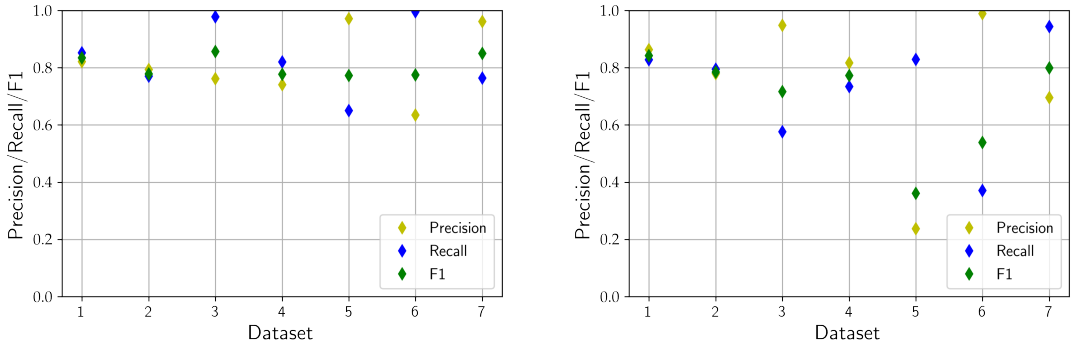
FIGURE 5.1: Droplet and artifact test performance of classifier trained on unbalanced combined datasets. While the droplet F1 scores are more consistent accross datasets than in the single datset experiments, there is still a high variation between dataset results. Except for dataset 2 and dataset 6, droplet precision is higher than droplet recall.

by the combined unbalanced datasets. Ideally, the classifiers trained on the combined training data should be as good or better at predicting any unseen dataset as the best classifier trained on a single dataset. This would show that the method to combine the training data enables the classifier to learn from all datasets and generalize more reliably than the classifiers trained on single datasets. Therefore, to come closer to the generalization results achieved in chapter 4, changing the dataset mix in the training data might lead to improvements.

5.1.2 Balancing Training Dataset Size

To balance the size of the datasets in the combined training data, the previously introduced (section 4.1.3.1) random undersampling and random oversampling were used.

Undersampling The increased influence of larger datasets when combined in the training data can be mitigated by applying random undersampling before training. In chapter 4, resampling was used to address class distribution differences. In this chapter, resampling is used to address dataset size differences. The smallest dataset in the training data contains 3,135 particles (figure 2.3). During each training run, the size of the smallest out of the six used datasets was set as an upper boundary for dataset size. All datasets larger than the smallest were shrunk by randomly removing entries until they reached the set upper bound. In this way, all datasets have an equal influence on the combined training data



(A) Droplets (B) Artifacts

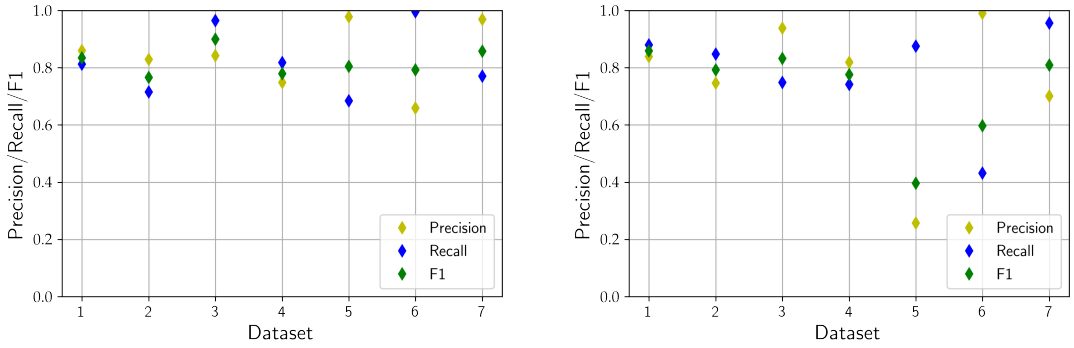
FIGURE 5.2: Droplet and artifact test performance of classifiers trained on balanced (random undersampling) combined datasets. Datasets 1, 2 and 7 show a clear improvement when compared to the results without rebalancing dataset size. Droplet recall and artifact precision for datasets 1, 2, 4 and 7 improve the most.

and the trained classifiers. Especially for datasets 1, 2 and 6 the resulting droplet recall and precision values have moved closer together (see figure 5.2). Droplet recall has improved from an average of 0.758 to 0.833 which is driven by dataset 1, 2, 4 and 7. Datasets 3, 5 and 6 are mostly unaffected by the change. **Overall, the results achieved by resizing the datasets by random undersampling are more consistent between datasets and the average droplet F1 score across datasets rises to 0.806 from 0.780 when no dataset size rebalancing was performed.**

Oversampling The experiment was repeated using random oversampling. Instead of downsampling the larger datasets, the smaller datasets were included multiple times up to the size of the largest dataset in the training run. Compared to the random undersampling approach, using random oversampling did further improve datasets 3, 5, 6 and 7 while reducing dataset 2 droplet F1 performance by 0.012 (figure 5.3). Datasets 1 and 4 are unaffected. Overall, the results are more consistent accross datasets and the average droplet F1 score rises to 0.819 which is the best result so far.

5.1.3 Removing Dataset 2

As shown in chapter 4, the generalization performance of classifiers trained on single datasets depends on the dataset they were trained on. The classifier trained on dataset 2 achieved the worst generalization results of all classifiers trained on single datasets. To test whether including it in the combined datasets hurts generalization performance, the previous experiment was repeated without dataset



(A) Droplets (B) Artifacts

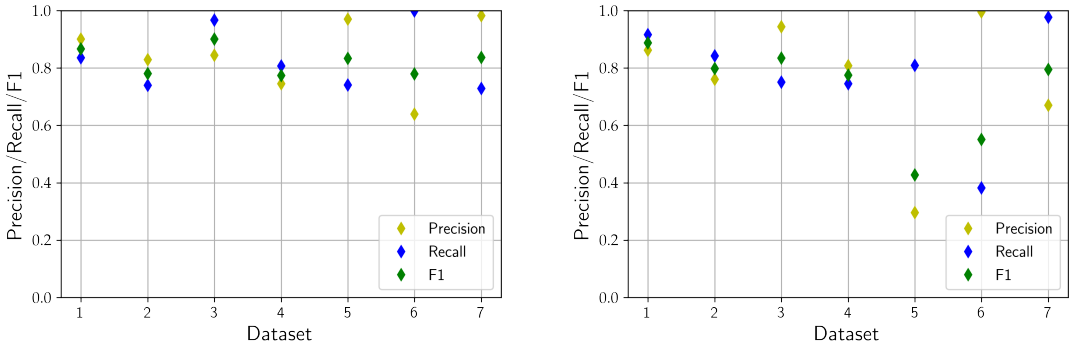
FIGURE 5.3: Droplet and artifact test performance of classifier trained on balanced (random oversampling) combined datasets. Compared to resizing the datasets with random undersampling, random oversampling leads to higher droplet precision scores. Overall, the results are more consistent between datasets on a slightly higher level for both droplet and artifact metrics.

2. Dataset 2 was still predicted but not used during training at all. The resulting droplet and artifact performance metrics of this experiment are shown in figure 5.4. **Removing dataset 2 from the training did not substantially improve or harm performance.**

5.1.4 Shifting Image Intensity Ranges

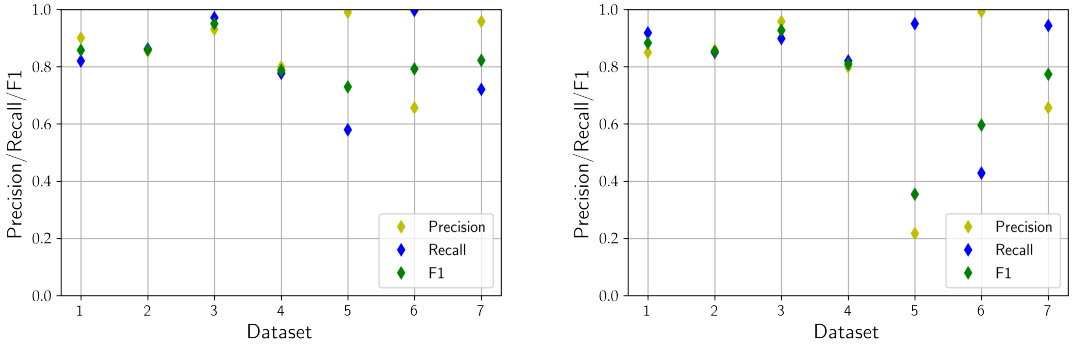
Another factor that might hinder generalization performance between datasets is a shift in image intensity distributions between dataset as observed in chapter 2. To ensure that the image intensities are in a comparable range across datasets, two options to shift intensities were evaluated.

Centering Image Intensities Over Dataset The median intensity values for both phase and amplitude images for each of the seven uncombined datasets was calculated and subtracted from every pixel. The median of all pixel intensities across each dataset is moved to zero. The results for this experiment in figure 5.5 show an improvement in average droplet F1 score which now is at 0.828. Precision and recall for datasets 2 and 3 lies closer together for both artifacts and droplets. Datasets 1-4 are very consistent in recall and precision, they are spread further for datasets 5-7. Especially artifact performance on datasets 5 and 6 is lacking.



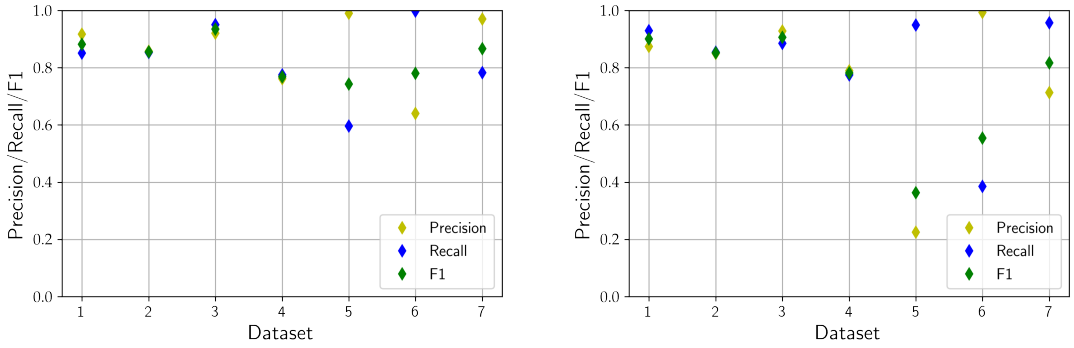
(A) Droplets (B) Artifacts

FIGURE 5.4: Droplet and artifact test performance of classifier trained on balanced (random oversampling) combined datasets without dataset 2. Droplet as well as artifact metrics are comparable for those observed without excluded dataset 2. for all datasets.



(A) Droplets (B) Artifacts

FIGURE 5.5: Droplet and artifact test performance of classifier trained on balanced (random oversampling) combined datasets and image intensities centered over dataset. Droplet and artifact precision and recall for datasets 1-4 are very close together. Especially performance on datasets 2 and 3 is improved by centering the images over the dataset. Datasets 6 and 7 show no large change in performance and dataset 5 is slightly worse.

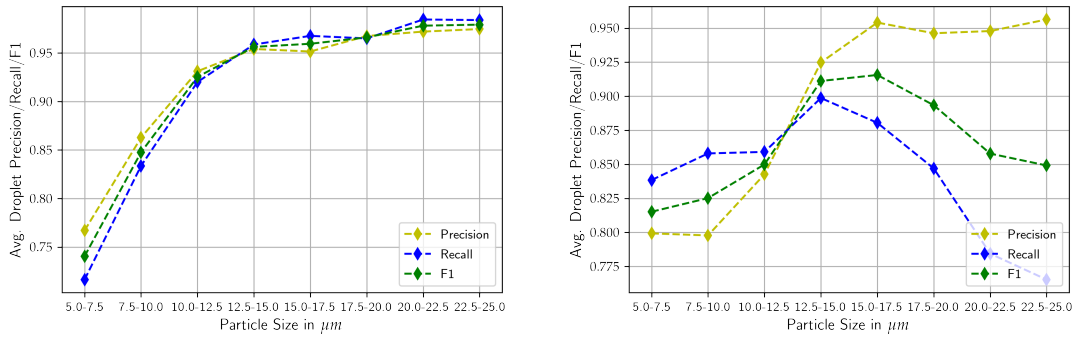


(A) Droplets (B) Artifacts

FIGURE 5.6: Droplet and artifact test performance of classifier trained on balanced (random oversampling) combined datasets and image intensities centered over individual images. The results are almost identical to those achieved when centering the image intensities over the datasets

Centering Image Intensities Individually Another method to center the image intensities around zero is to subtract the median value of the pixel intensities of every individual image from every pixel in that image. While this method might remove useful brightness differences between classes, it ensures that all images are centered around zero. Figure 5.6 shows very similar results to centering the images over the dataset. This suggests that no useful information was lost by centering the images individually. The average droplet F1 score is slightly higher at 0.833.

Performance by Particle Size The validation droplet metrics by particle size (figure 5.7a) next to test performance by particle size (figure 5.7b) for dataset 1, shows that up until a size of 12.5 µm, test performance is similar to validation performance. While droplet precision continues to rise, droplet recall drops when approaching the larger sizes. For the other datasets (figure 7), the validation metrics consistently improve with particle size and low variation between precision, recall and f1 accross training dataset compositions. The test metrics are inconsistent accross datasets. For all datasets except dataset 2, droplet precision rises in the smaller sizes and stays on a very high level at sizes larger than 12.5 µm. Most of the performance variation between datasets is caused by droplet recall. For datasets 2, 4 and 7, droplet recall rises with particle size. Datasets 3 and 6, consistently have a high droplet recall across all sizes. Dataset 5 with its strong class imbalance has a low droplet recall throughout. While no clear pattern is visible, **validation and training performances generally are worst at particle major axis sizes below 12.5 µm.**



(A) Validation (B) Test

FIGURE 5.7: Droplet training vs test performance dataset 1 (grouped balanced training datasets, individually centered images). Subfigure A shows that droplet validation performance is lower at major-axis particle sizes smaller than 12.5 µm. Subfigure B shows that droplet precision but not recall rises with major-axis particle size.

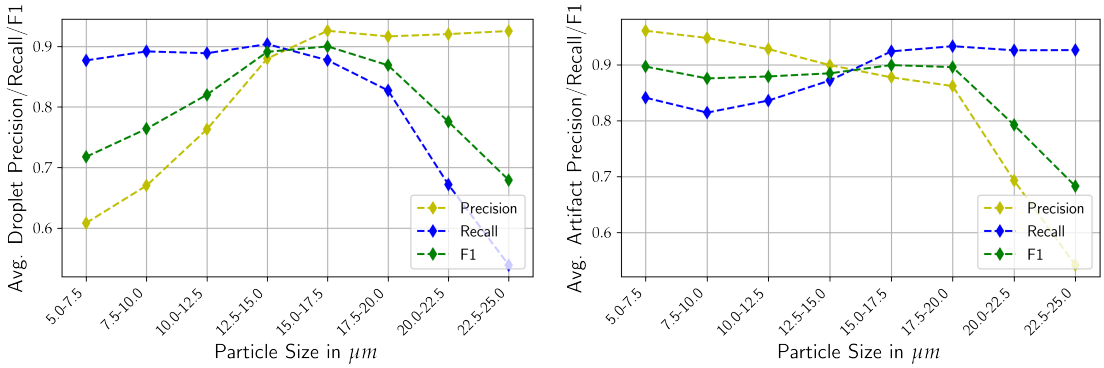
5.2 Adding Metadata

When training individual classifiers on the single datasets, additional metadata or calculated features did not improve performance. To test whether it can help in the generalization between datasets, some of the metadata used in chapter 4 was included again.

5.2.1 Adding Size Information

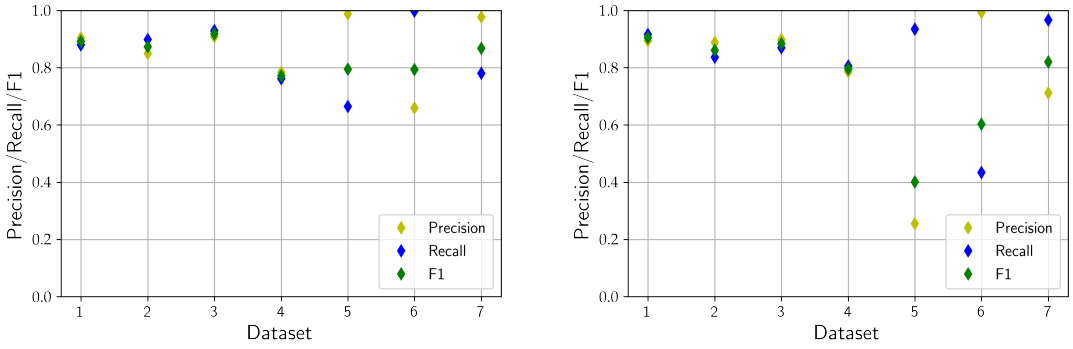
While observing droplet and artifact F1, recall and precision scores througout the previous experiments, it became clear that the performance strongly depends on particle size. In many cases, droplets and artifacts trade places regarding their recall and precision scores as major axis size rises. To test whether adding particle size information aids in the classification, the major-axis particle size was added in the first dense layer of the network. The image intensity centering on the individual images from the previous section was kept as well as training dataset size rebalancing by oversampling.

Figure 5.8a shows the validation and test performance on dataset 1 of the classifier trained on datasets 2-7. Generally, droplet precision is rising with particle size. Artifact precision is not affected by small particle sizes (figure 5.8b). To test whether knowledge about the particle size can remedy this, particle size was added as an input layer and added to the flattening layer output using the keras concatenate layer. No other metadata was added in this experiment.



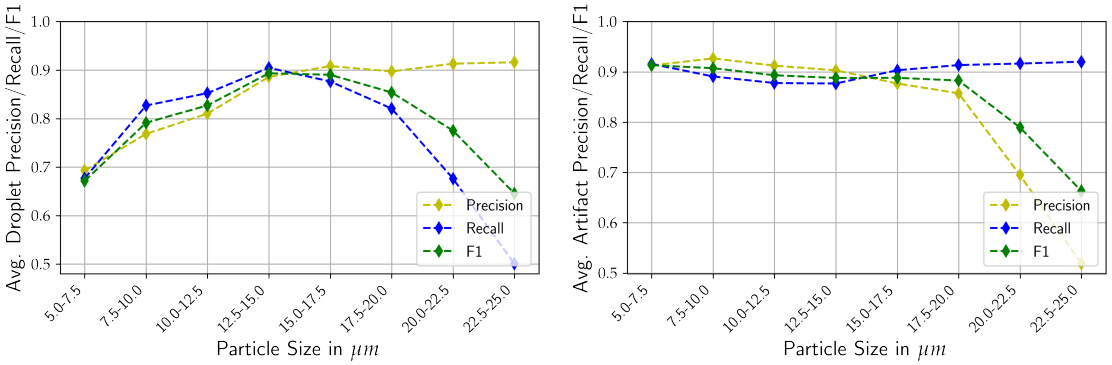
(A) Droplet Metrics Without Size Information (B) Artifact Metrics Without Size Information

FIGURE 5.8: Size binned droplet and artifact performance (Dataset 1) on balanced training datasets. Subfigure A shows a large spread between precision and recall in all particle size bins.



(A) Droplets (B) Artifacts

FIGURE 5.9: Test performance of combined training datasets with added image size. The figure shows similar performance metrics for droplets and artifacts to those observed without added image size.



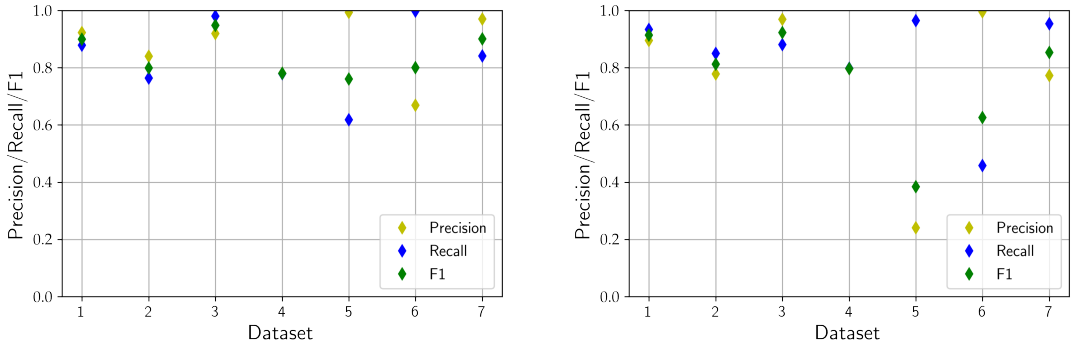
(A) Droplet Metrics With Size Information (B) Artifact Metrics With Size Information

FIGURE 5.10: Size binned droplet and artifact performance (Dataset 1) on balanced training datasets with size information. The subfigures show that the spread between recall and precision in the size bins smaller than 12.5 µm seen in 5.8 has vanished when the particle size information is added. F1 score is mostly unaffected.

Comparing the results without added particle size, to those with added particle size (figure 5.10), addding the particle size does not affect test F1. However, it clearly shows that at least for dataset 1, droplet and artifact precision and recall values are drawn together for the size bins smaller than 12.5 µm. While these results are interesting, there is a concern that a model might consider size and its class distribution in the training data too strongly. As shown in figure 2.5 in chapter 2, artifacts are most common in the smaller major axis size ranges. The classifier should not simply assume that because a particle is small, it is an artifact.

5.2.2 Adding All Available Metadata

To check whether adding all available metadata aids generalization, the metadata shared between all datasets was added to the first dense layer. Of the 112 possible metadata entries, 32 are available for all datasets. The metrics for droplets and artifacts are shown in figure 5.11. The relative performances between datasets do not substantially differ from those seen when no metadata was added (figure 5.6). Since major-axis particle size is also included in the metadata, the same concern about unwanted learned relationship between particle size and class as in the previous section is valid. **Despite reaching the highest average droplet F1 score of 0.841 so far, due to the concern of unwanted learned relations between size and particle class, the inclusion of the metadata is not recommended.**



(A) Droplets (B) Artifacts

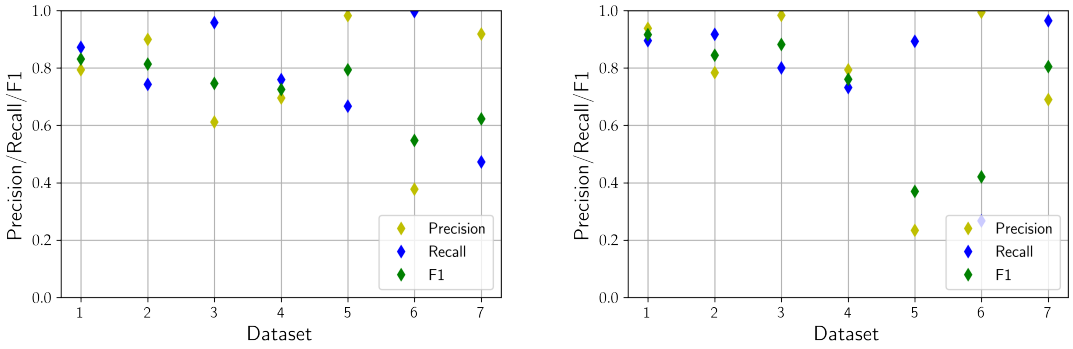
FIGURE 5.11: Droplet and artifact test performance with all available metadata. The figure shows similar performance metrics for droplets and artifacts to those observed without added metadata.

5.3 Splitting at 12.5 µm

A consistent observation throughout all experiments is the lower performance at major axis particle sizes smaller than 12.5 µm. To evaluate this effect, the data was split into particles smaller and larger than 12.5 µm and the training and evaluation repeated. The dataset size was rebalanced by oversampling and the image intensities centered around zero on individual images.

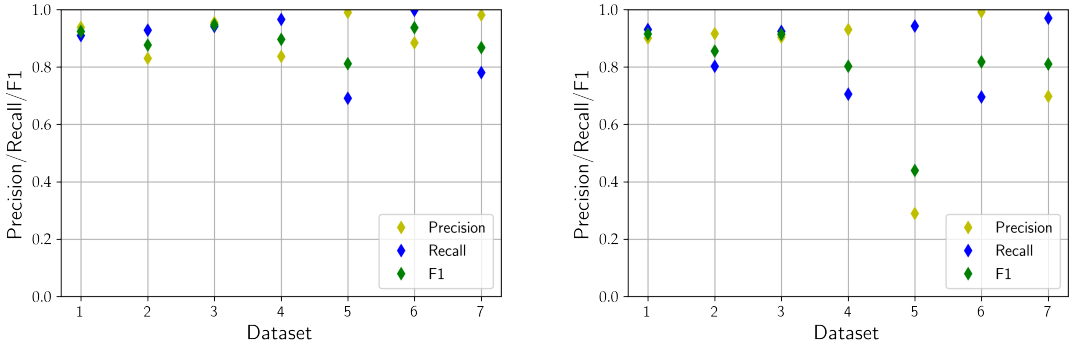
Particles Smaller Than 12.5 µm The test classification performance on the particles smaller than 12.5 µm is lower than that of particles across all sizes (figure 5.12). The averaged droplet test F1 score is 0.725. While the average artifact F1 test score is lower at 0.714, artifact prediction performance is higher than droplet prediction performance for all datasets except datasets 5 and 6, which also had low artifact performance in all previous experiments.

Particles Larger Than 12.5 µm The classification performance on particles larger than 12.5 µm is very high hroughout all datasets (figure 5.13). The minimum droplet F1 score is 0.868 for dataset 7 and the maximum droplet F1 score is 0.948 for dataset 3. **The average droplet test F1 score for particles larger than 12.5 µm is at 0.894 which is 0.061 higher than the highest average score achieved without adding metadata.**



(A) Droplets (B) Artifacts

FIGURE 5.12: Droplet and artifact test performance on particles smaller than 12.5 µm (balanced dataset size, images intensities centered individually). The figure shows the expected drop in performance across all datasets, especially for the droplet metrics compared to the results for all major axis particle sizes.



(A) Droplets (B) Artifacts

FIGURE 5.13: Droplet and artifact test performance on particles larger than 12.5 µm (balanced dataset size, images intensities centered individually). Subfigure A shows very high (above 0.85) droplet F1 scores for all datasets. Subfigure B, shows high (above 0.8) and consistent artifact performance for all datasts except dataset 5.

	Dataset							Avg.	Med.
	1	2	3	4	5	6	7	-	-
Ch. 4 Self	0.96	0.88	0.98	0.86	0.93	0.93	0.95	0.93	0.93
Ch. 4 Other	0.87	0.75	0.84	0.78	0.92	0.87	0.82	0.83	0.84
Unbalanced	0.81	0.70	0.88	0.75	0.77	0.78	0.77	0.78	0.77
Undersampled	0.84	0.78	0.86	0.78	0.77	0.78	0.85	0.81	0.78
Oversampled	0.84	0.77	0.90	0.78	**0.80**	**0.79**	0.86	0.82	0.80
Centered DS	0.86	**0.86**	**0.95**	**0.79**	0.73	**0.79**	0.82	**0.83**	0.82
Centered Ind.	**0.88**	0.86	0.94	0.77	0.74	0.78	**0.87**	**0.83**	**0.86**

TABLE 5.1: Performance comparison all datasets with benchmark. The table shows, the droplet F1 scores for all datasets and most experiments. The rows "Ch.4 Self" and "Ch.4 Other" are the benchmark results from chapter 4.

5.4 Performance Conclusion

The final results of the conducted experiments are summarized in table 5.1. The first two rows "Ch. 4 Self" and "Ch. 4 Other" are the maximum droplet F1 scores achieved by the classifiers from chapter 4. The row "Ch. 4 Self" is the best validation performance of a classifier trained on the dataset that it is validated on. The row "Ch. 4 Other" is the best prediction droplet F1 score for the dataset achieved by any of the six classifiers trained on the other six datasets. Table 5.2 shows the performance delta to the classifier that performed best during generalization for every dataset. **By combining image data from multiple datasets, the predictions by the best classifier trained on a single dataset can be matched for every dataset except datasets 5 and 6.** For datasets 2, 3 and 7, the best results by classifiers trained on single datasets were exceeded by a difference in droplet F1 score of 0.05 to 0.11. While the improvement in average (0.00) and median (0.01) is small, the methodology yields much more robust result than any of the classifiers trained on single datasets achieved. For a new dataset, it is impossible to choose the classifier that works best as was done for the benchmark "Ch. 4 Other" in chapter 4. The classifiers trained in chapter 5 are likely to work for any new dataset with similar performance as can be achieved by choosing the best from six classifiers as was done in chapter 4.

	Dataset							Avg.	Med.
	1	2	3	4	5	6	7	-	-
Ch. 4 Self	0.09	0.13	0.14	0.08	0.01	0.07	0.13	0.09	0.09
Ch. 4 Other	-	-	-	-	-	-	-	-	-
Unbalanced	0.06	0.05	0.04	0.04	0.14	0.08	0.05	0.05	0.05
Undersampled	0.03	0.03	0.02	0.01	0.14	0.09	0.03	0.03	0.01
Oversampled	0.03	0.02	0.06	0.01	**0.11**	**0.07**	0.04	0.02	0.01
Centered DS	0.01	**0.11**	**0.11**	**0.00**	0.19	**0.07**	0.01	**-0.01**	0.00
Centered Ind.	**0.01**	0.11	0.10	0.02	0.17	0.08	**0.05**	**0.00**	**0.01**

TABLE 5.2: Performance comparison all datasets with delta to benchmark. The table shows the droplet F1 scores for all datasets and most experiments compared to the benchmark "Ch.4 Other". The red digits indicates a score lower and the black digits a score higher than "Ch.4 Other". The table shows that the classifiers trained with image centering over datasets and individually reached average and median performances similar or better to the benchmark from chapter 4 except for datasets 5 and 6.

Chapter 6

Conclusion and Outlook

In this work, the performance and generalization performance of convolutional neural networks for classifying small cloud particles at major-axis sizes smaller than 25 µm was evaluated. Seven datasets were studied. Individually, the seven datasets were predicted with droplet F1 scores between 0.85 (dataset 4) and 0.97 (dataset 3). **In generalization experiments with classifiers trained on multiple datasets, four of the seven datasets (datasets 1-3 and 7) were predicted with droplet F1 scores between 0.86 and 0.94. The other three datasets were predicted with droplet F1 scores between 0.74 and 0.78.** The differences in generalization performance might be caused by differences in pixel intensity and particle class distribution between datasets. Prediction performance gains were achieved by adjusting image size, combining amplitude and phase images, altering the dataset size mix and shifting the image intensity ranges around a shared median. The most important findings are summarized below.

Image Size, Input Channels and Use of Metadata It was experimentally determined that scaling the particle images to a size of 20x20 pixels produces the highest droplet F1 scores.
The classifiers trained on amplitude images performed better than those trained on phase images except for dataset 2. Combining amplitude and phase images improved prediction performance for all datasets. With the exception of dataset 7, combining amplitude and phase information reduced the spread between precision and recall.
Adding metadata did not improve the predictions of classifiers trained on single datasets. The generalization results of classifiers trained on multiple datasets were improved (rise in average droplet F1 of 0.011) by adding the metadata shared between all datasets. Due to the concern of fitting unwanted relationships between metadata (such as major-axis size) and class, it was decided not to use metadata.

Use of Output Layer Activation to Detect Falsely Labeled Particles The experiments in chapter 4 showed that the activation of the network's output sigmoid neuron can be used to find falsely labeled particles. **Particles that are predicted with a high confidence by the network but are marked as wrongly classified due to the assigned label were shown to be falsely labeled in some cases.**

Manual labeling of a dataset can be assisted by a network or previously labeled data can be revisited. Training a network and reconsidering all particles where the network prediction with a high confidence did not match the assigned label might help uncover mistakes that were made due to input error by the person labeling the data. Further, the activation values show that the network is able to reliably detect particles that are hard to decide. When the activation of the output neuron is close to 0.5, the decision is difficult for humans as well. A confidence threshold could be introduced and particles where the networks prediction confidence is below the threshold be classified by another classifier or handed over to a person.

Rebalancing Training Dataset Size to Improve Generalization Using combined training datasets to classify unseen datasets produced more reliable and robust results than any classifier trained on a single dataset. The results can be further improved by rebalancing the size of the combined training datasets to have the same impact during training. Without balancing dataset size, an average droplet F1 score of 0.780 was reached. By applying undersampling to balance the dataset size, the average droplet F1 score increased to 0.806. **Balancing the dataset size by oversampling produced the highest average droplet F1 score of 0.819.**

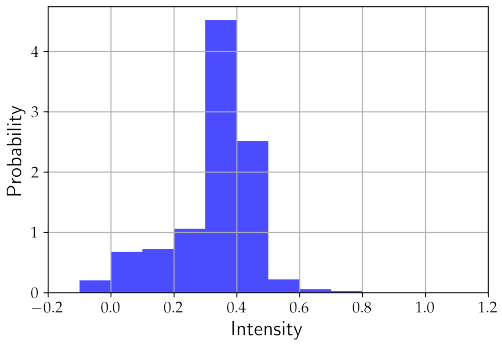
Shifting Image Intensities to Improve Generalization Onto Unseen Data The highest droplet F1 score and most robust generalization performance was achieved by shifting the intensity distributions to a median of zero. Shifting the median to 0 across all pixel intensities of each dataset, lead to an average droplet F1 of 0.828. **Shifting the median to zero on individual images produced an average droplet F1 score of 0.830.** This transformation worked especially well for datasets 1-4 and 7. The datasets that profited the least (5 and 6), were also those where the amplitude intensities covered a smaller range (figure 1). Further generalization improvements might be possible by altering the pixel intensity distributions during preprocessing.

Splitting Classifiers by Particle Size The aim of this work was to improve classification for the smallest cloud particles at major axis sizes smaller than 25 µm. While this succeeded with droplet F1 scores close to or above 0.80 for all datasets, the performance differs strongly between the smallest particles (smaller than 12.5 µm) and the largest of the small particles (larger than 12.5 µm but smaller than 25 µm). For the smaller particles, droplet F1 scores range from 0.547 to 0.831 with an average of 0.725. The droplet F1 scores for the larger particles range from 0.812 to 0.948 with an average of 0.894.

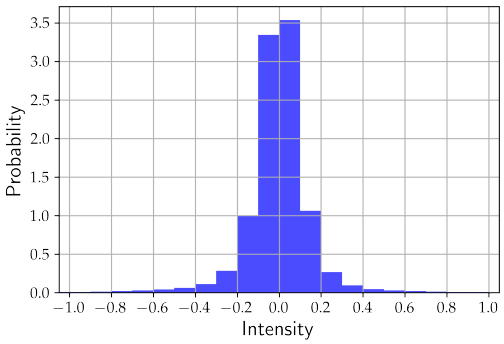
Outlook and Further Potential for Improvement The experiments in this thesis focused on the exploration of differences between the seven datasets and the selection of training data and features to improve generalization performance. The observed differences in pixel intensities (figure 1) and averaged radial intensities (figures 6), especially for the phase images, might explain some of the generalization loss. **Analyzing the source of the differences and preprocessing the images to transform these distributions to be more comparable between datasets might further improve generalization performance.**
The confidences produced by the classifiers trained on single datasets could be used to reevaluate particles that are likely misclassified. The resulting higher correctness of the labels might also enable better training of any classifiers which in turn might further improve generalization performance.
The architecture of the classifier itself can be improved as well. This work only briefly explored the use of dropout layers and early stopping. **Further architecture or hyperparameter tuning might produce better results in the future.**
Previous work (Touloupas et al., 2020) showed that fine-tuning an existing model with a small number of hand-classified particles from a new dataset can substantially improve generalization performance. **Adding the fine-tuning methodology to the optimizations explored in this thesis might further improve prediction performance.**

Appendix A

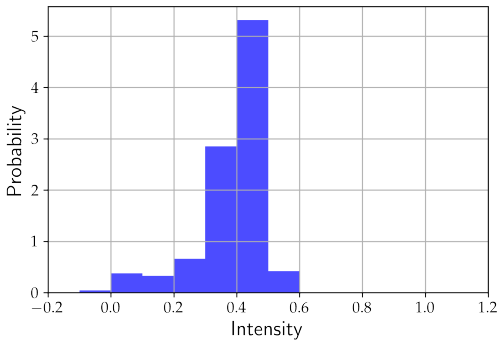
Additional Figures



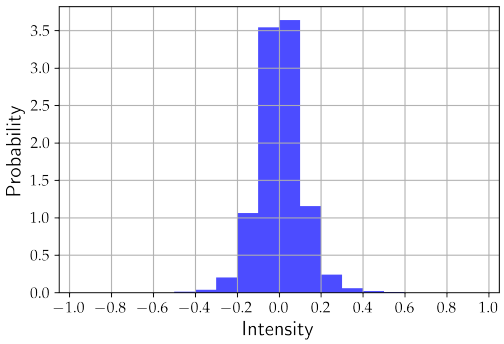
(A) Dataset 3 Amplitude



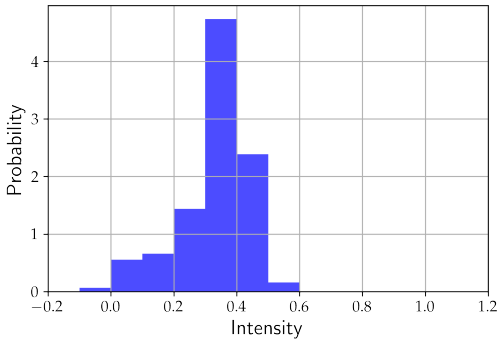
(B) Dataset 3 Phase



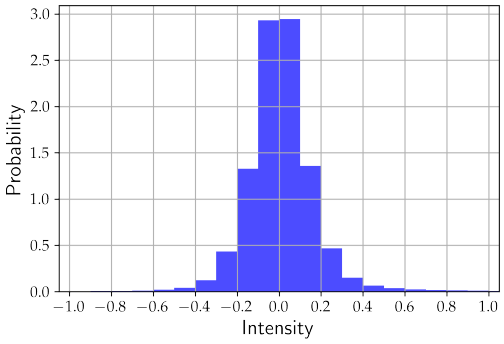
(C) Dataset 4 Amplitude



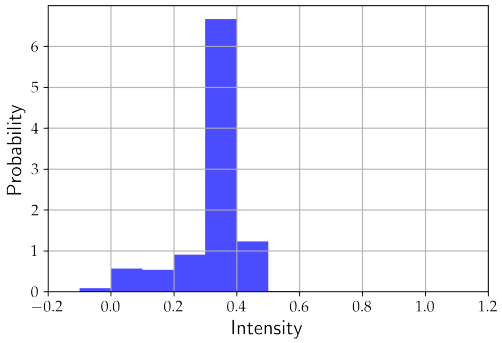
(D) Dataset 4 Phase



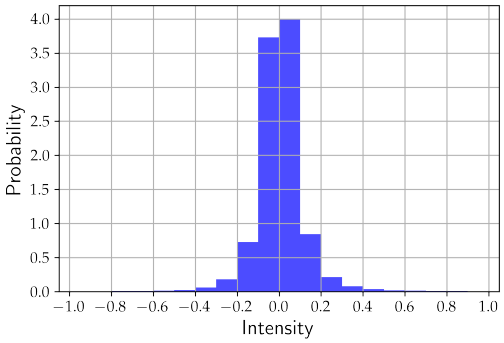
(E) Dataset 5 Amplitude



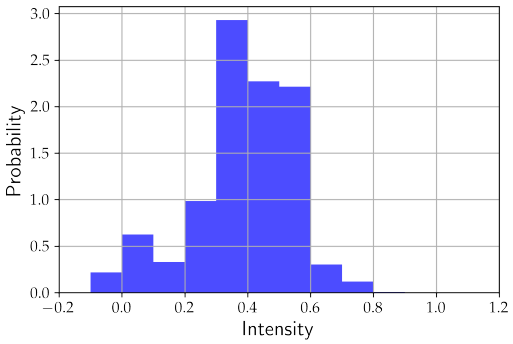
(F) Dataset 5 Phase



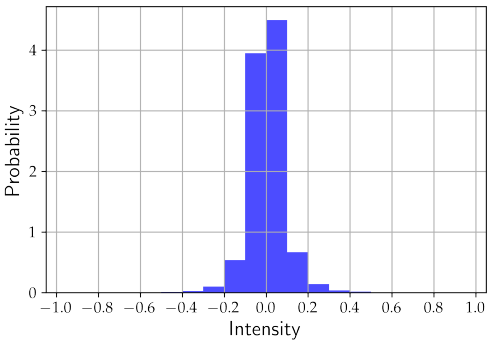
(G) Dataset 6 Amplitude



(H) Dataset 6 Phase

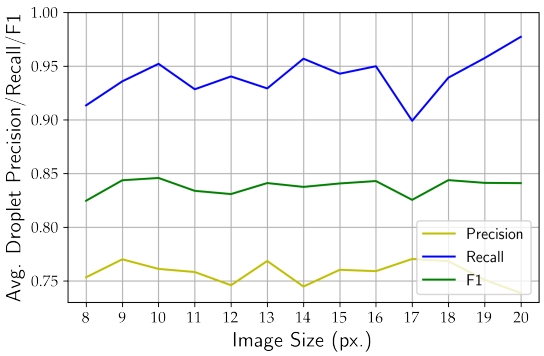


(I) Dataset 7 Amplitude

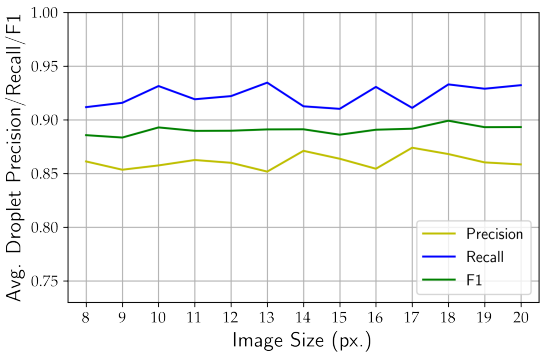


(J) Dataset 7 Phase

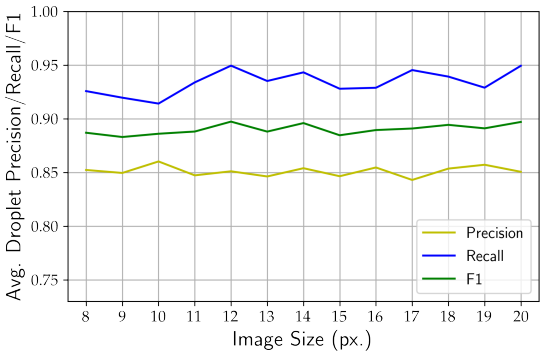
FIGURE 1: Pixel intensity distribution, datasets 3-7. The figures show that the range for intensity values in the amplitude distributions vary between datasets. Datasets 3 and 7 include intensities up to 0.8 while datasets 4 and 5 include intensities up to 0.6 and 0.5 for dataset 6. Further, the intensity distributions of the phase images for datasets 3-5 cover a wider range of values than those of datasets 6 and 7.



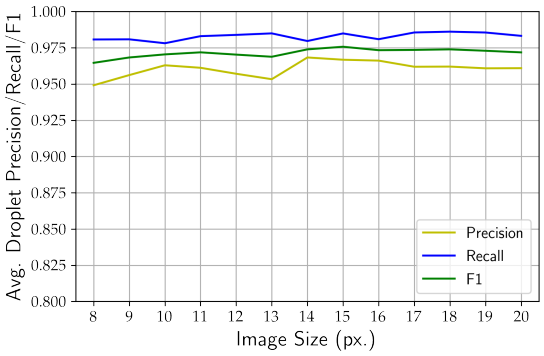
(A) Dataset 2 Amplitude



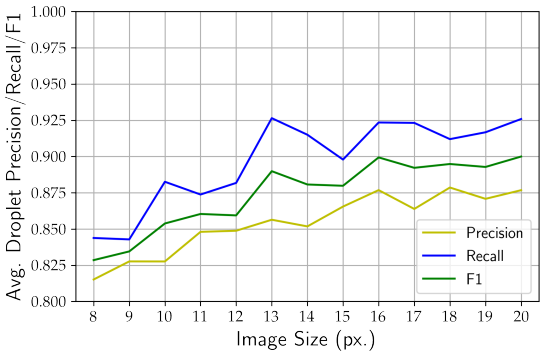
(B) Dataset 2 Phase



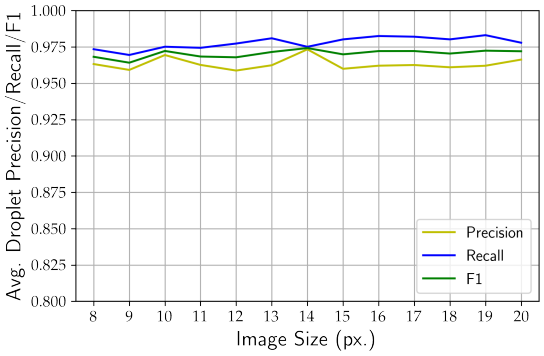
(C) Dataset 2 Combined



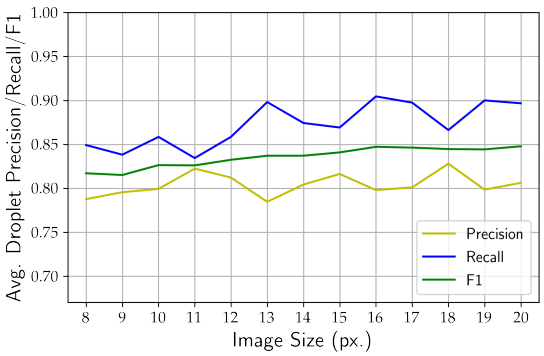
(D) Dataset 3 Amplitude



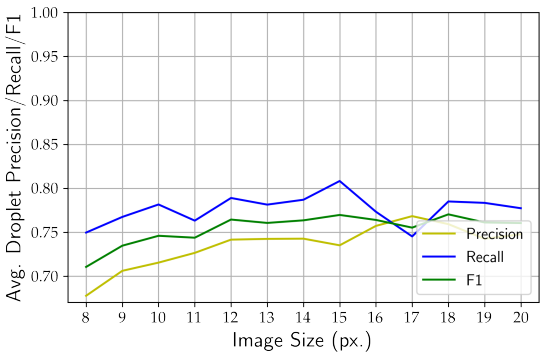
(E) Dataset 3 Phase



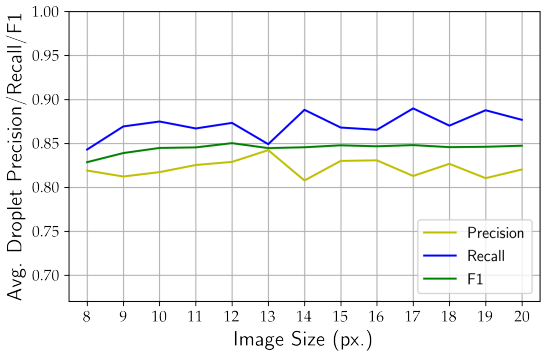
(F) Dataset 3 Combined



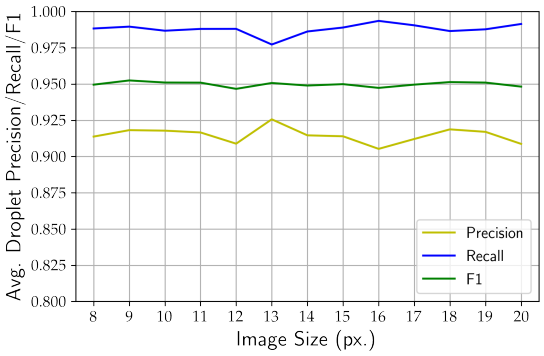
(G) Dataset 4 Amplitude



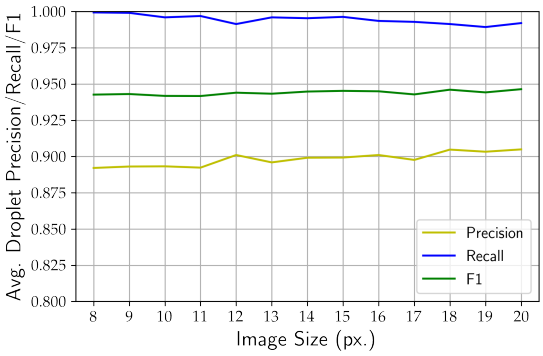
(H) Dataset 4 Phase



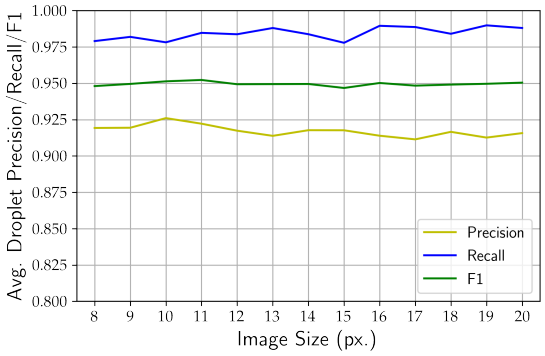
(I) Dataset 4 Combined



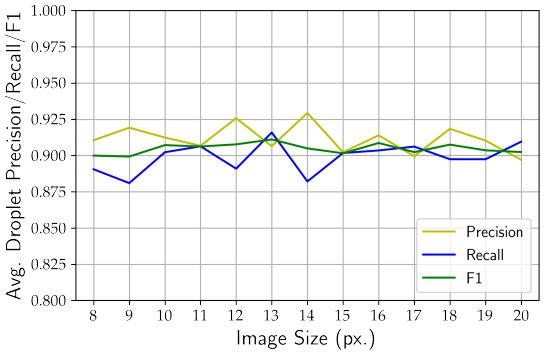
(J) Dataset 5 Amplitude



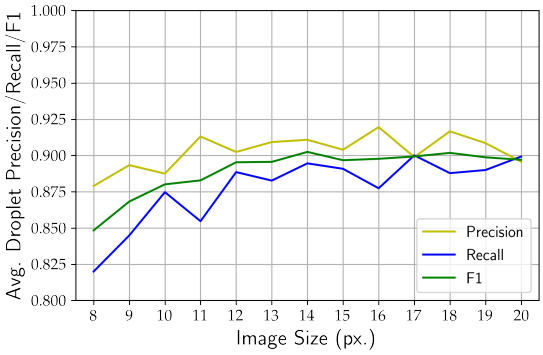
(K) Dataset 5 Phase



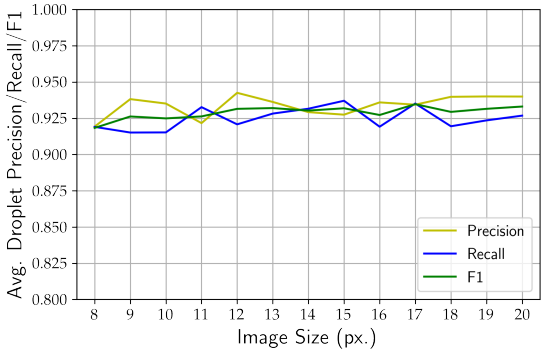
(L) Dataset 5 Combined



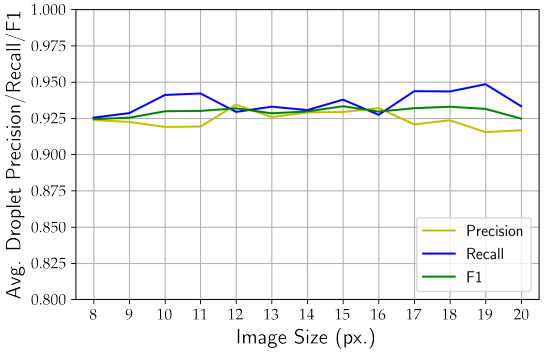
(M) Dataset 6 Amplitude



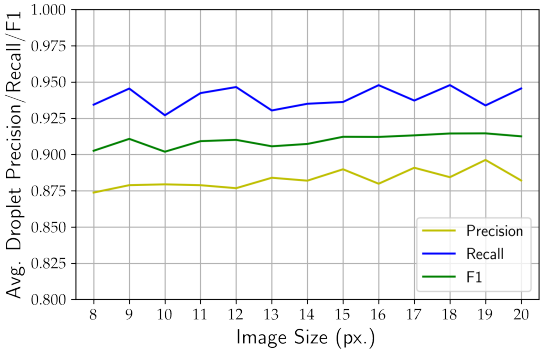
(N) Dataset 6 Phase



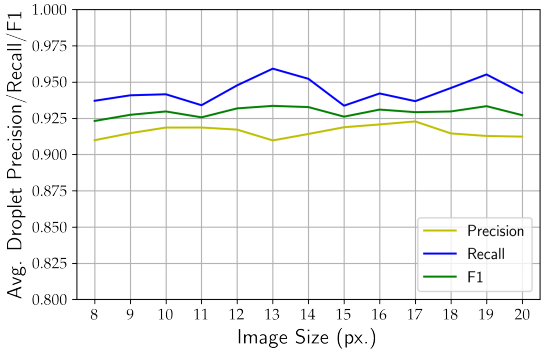
(O) Dataset 6 Combined



(P) Dataset 7 Amplitude

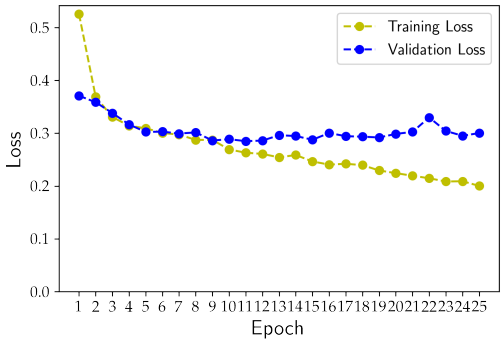


(Q) Dataset 7 Phase

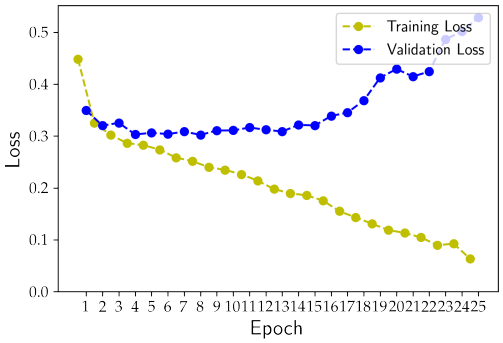


(R) Dataset 7 Combined

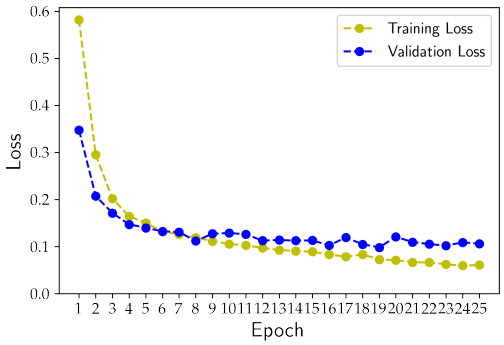
FIGURE 2: Image size selection datasets 2-7. The figures show that, for all datasets, using the combined amplitude and phase images is beneficial or does not negatively impact performance.



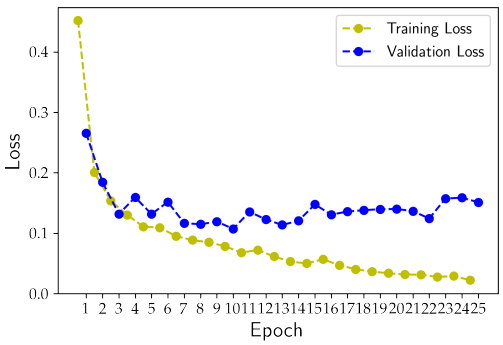
(A) Dataset 2, 30% Dropout Chance



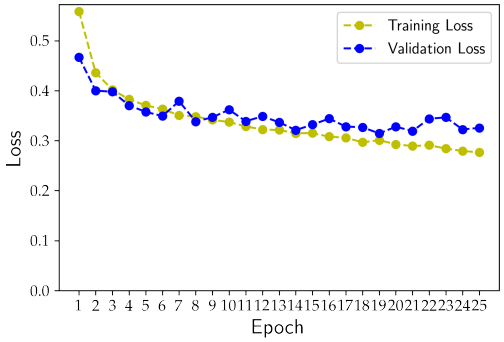
(B) Dataset 2, Training Loss Shifted, 0% Dropout



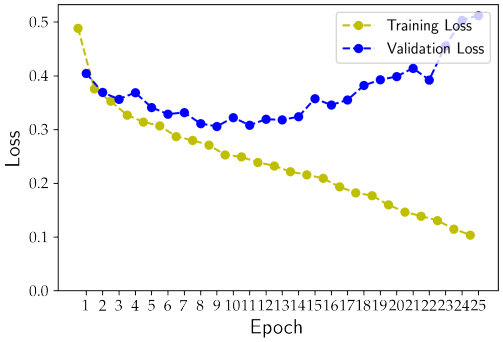
(C) Dataset 3, 30% Dropout Chance



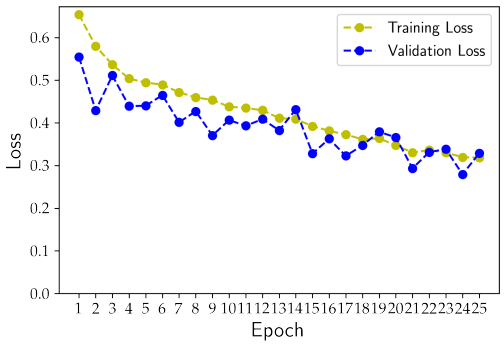
(D) Dataset 3, Training Loss Shifted, 0% Dropout



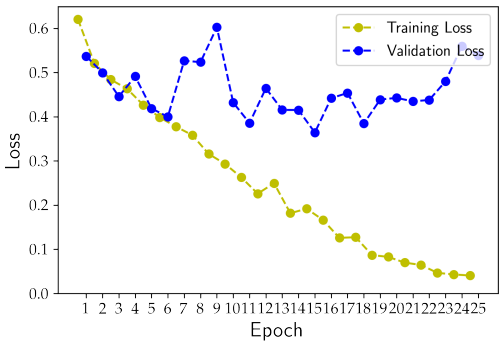
(E) Dataset 4, 30% Dropout Chance



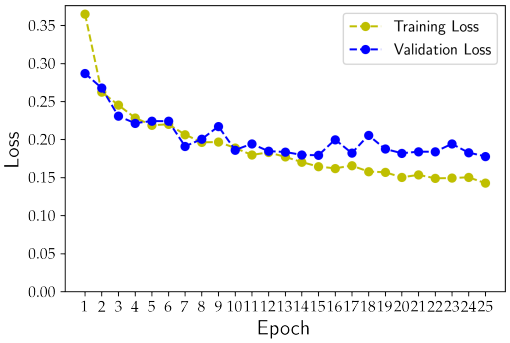
(F) Dataset 4, Training Loss Shifted, 0% Dropout



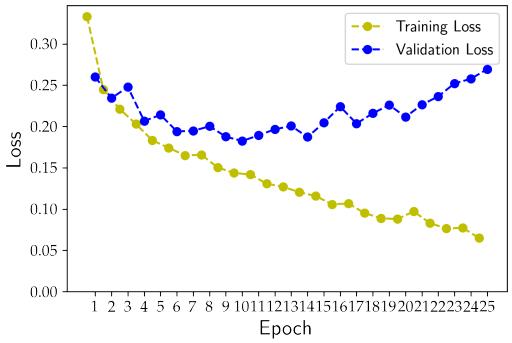
(G) Dataset 5, 30% Dropout Chance



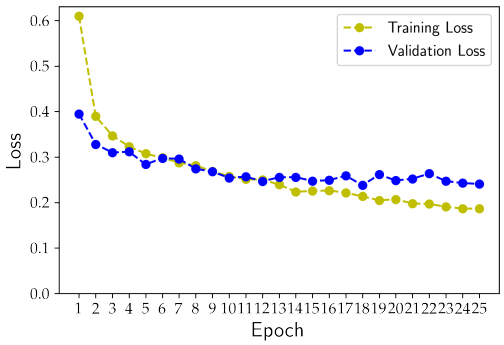
(H) Dataset 5, Training Loss Shifted, 0% Dropout



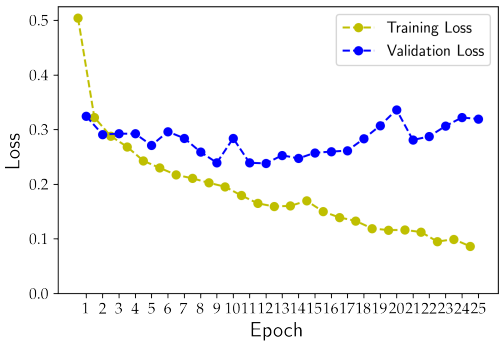
(I) Dataset 6, 30% Dropout Chance



(J) Dataset 6, Training Loss Shifted, 0% Dropout

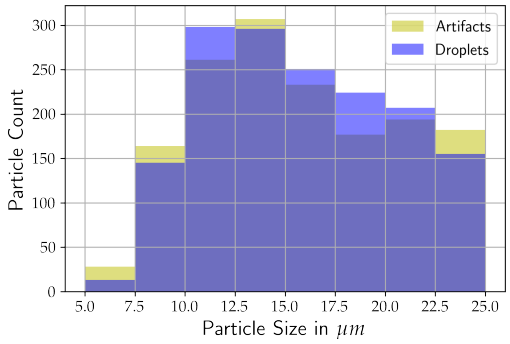


(K) Dataset 7, 30% Dropout Chance

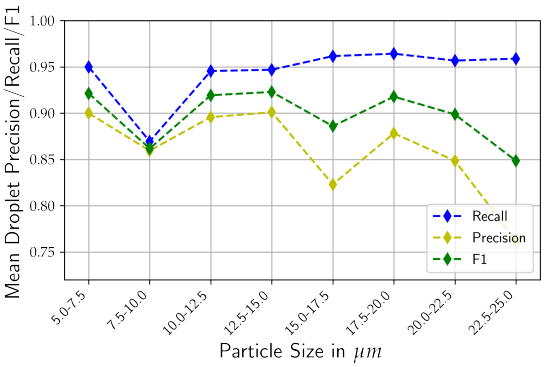


(L) Dataset 7, Training Loss Shifted, 0% Dropout

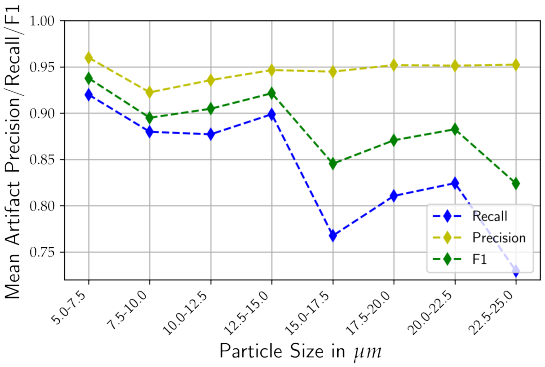
FIGURE 3: Loss by epoch datasets 3-7, with and without dropout. The figures show that, for all datasets, setting the dropout chance to 30% significantly reduces overfitting.



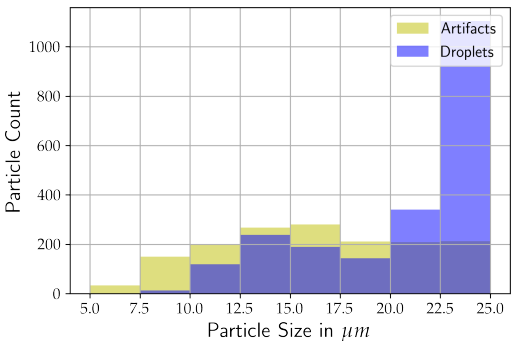
(A) Dataset 2 Distribution



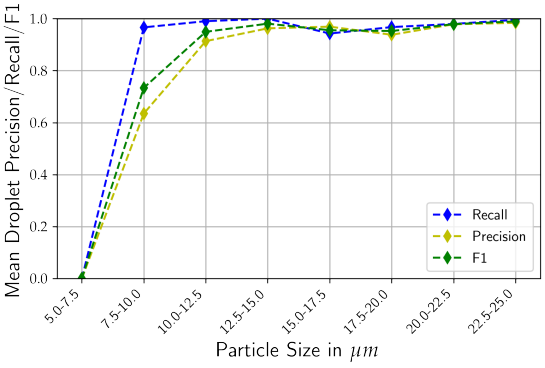
(B) Dataset 2 Droplets



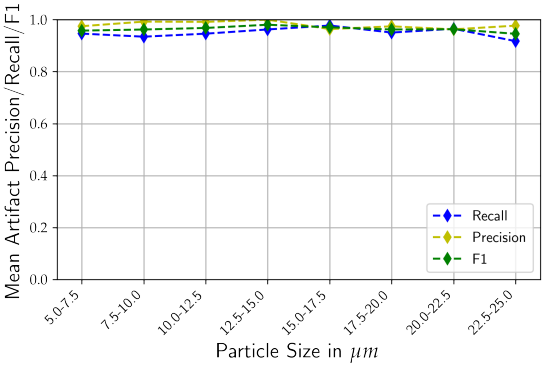
(C) Dataset 2 Artifacts



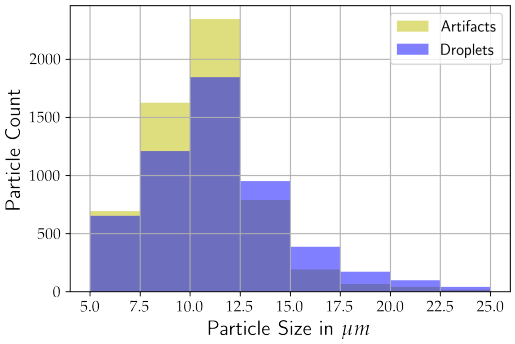
(D) Datast 3 Distribution



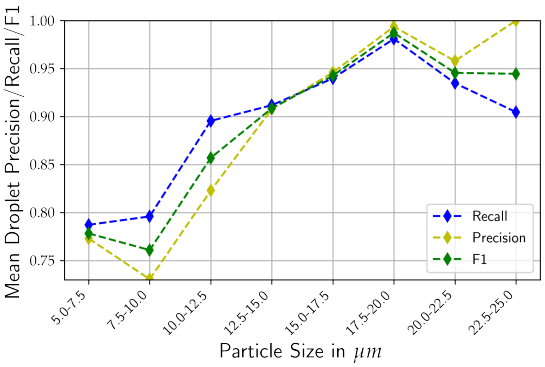
(E) Dataset 3 Droplets



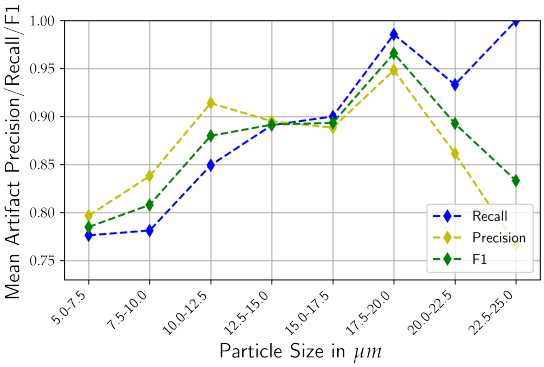
(F) Dataset 3 Artifacts



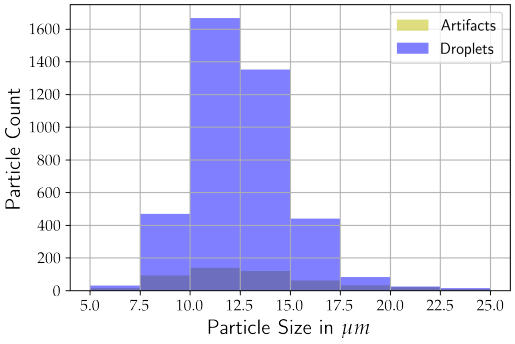
(G) Dataset 4 Distribution



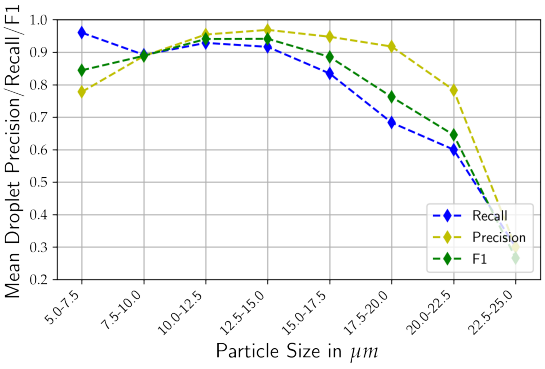
(H) Dataset 4 Droplets



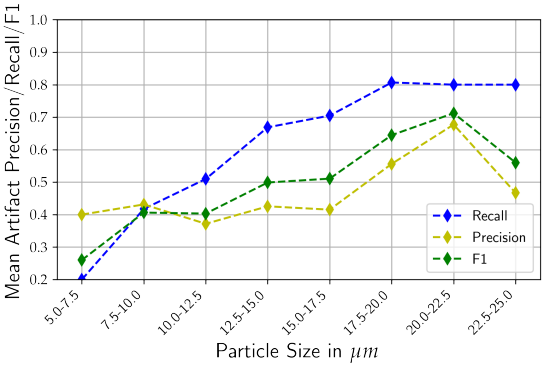
(I) Dataset 4 Artifacts



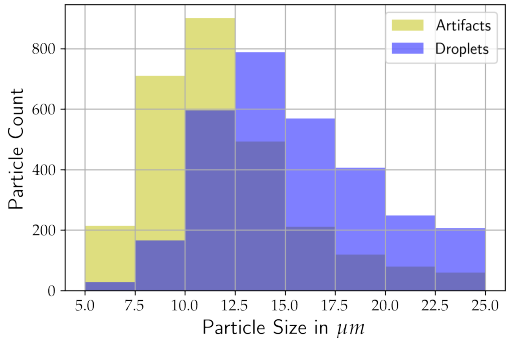
(J) Dataset no. 5 Distribution



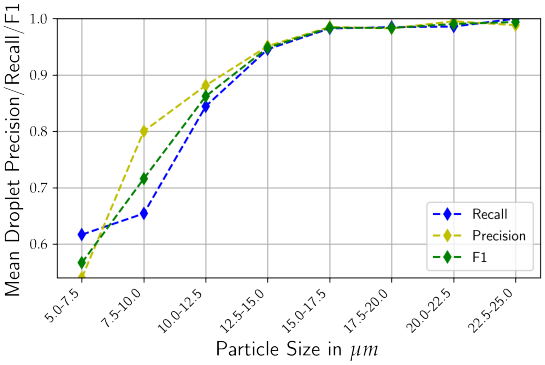
(K) Dataset 5 Droplets



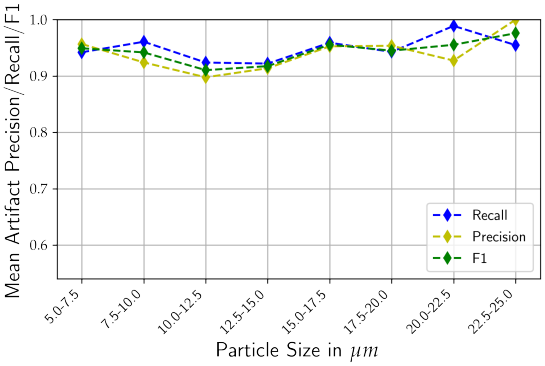
(L) Dataset 5 Artifacts



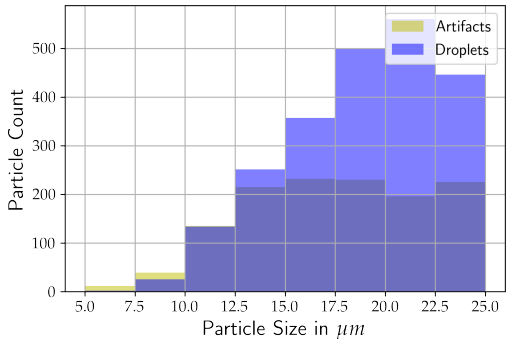
(M) Dataset 6 Distribution



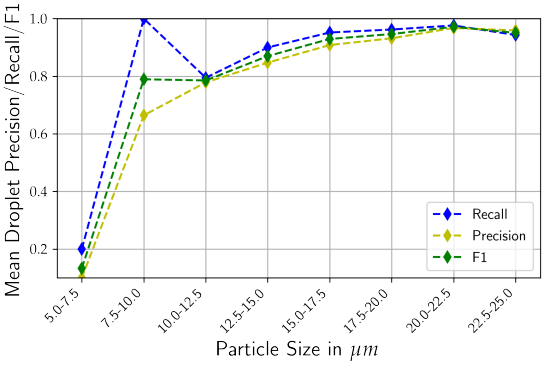
(N) Dataset 6 Droplets



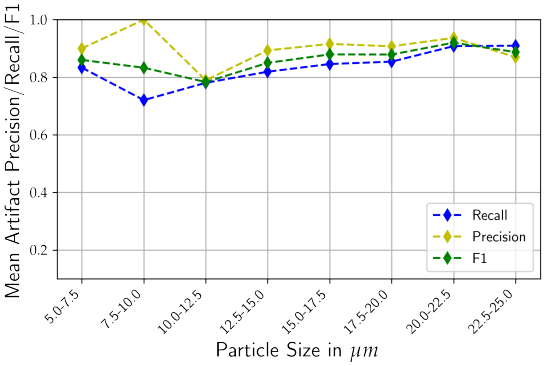
(O) Dataset 6 Artifacts



(P) Dataset 7 Distribution

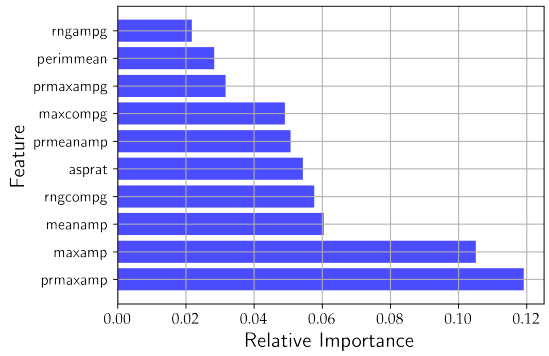


(Q) Dataset 7 Droplets

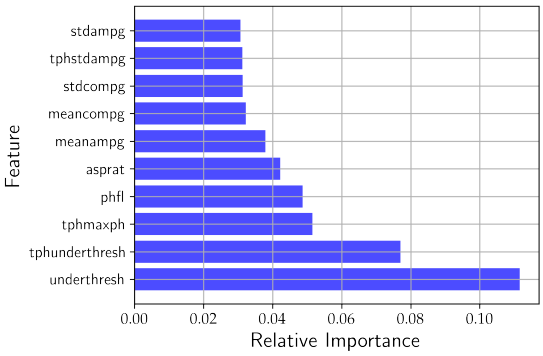


(R) Dataset 7 Artifacts

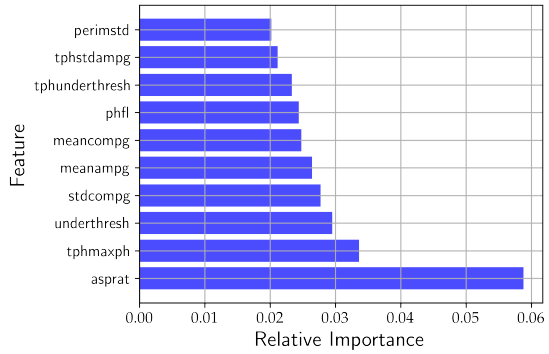
FIGURE 4: Particle size binned metrics, datasets 3-7. The figures show that, performances for particle sizes are variable between datasets but often worst in particle size ranges smaller than 12.5 µm.



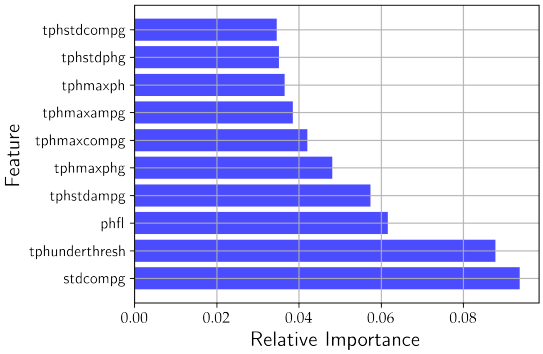
(A) Dataset 3



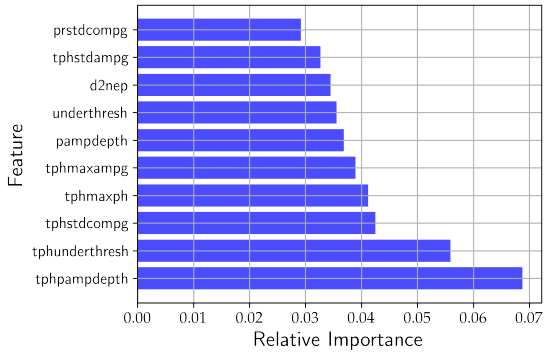
(B) Dataset 4



(C) Dataset 5

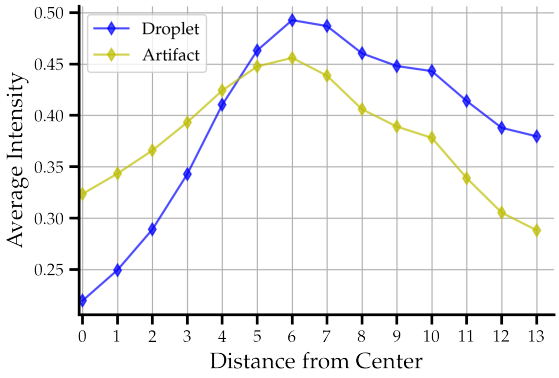


(D) Dataset 6

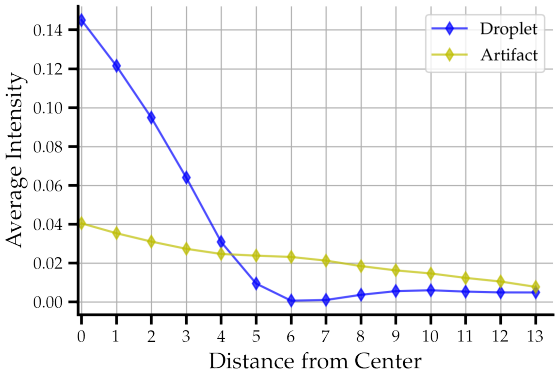


(E) Dataset 7

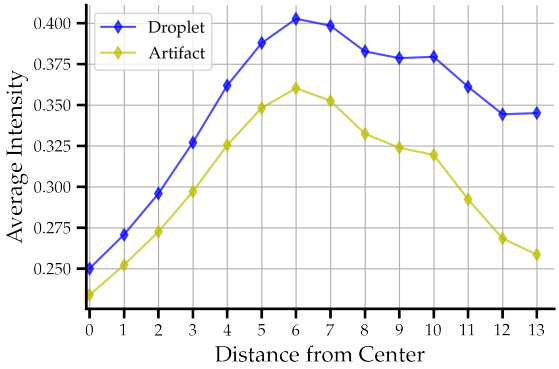
FIGURE 5: Feature importance, datasets 3-7. The figures show the top ten most important features for every dataset. Full explanations of the features can be found in Schlenczek, 2018



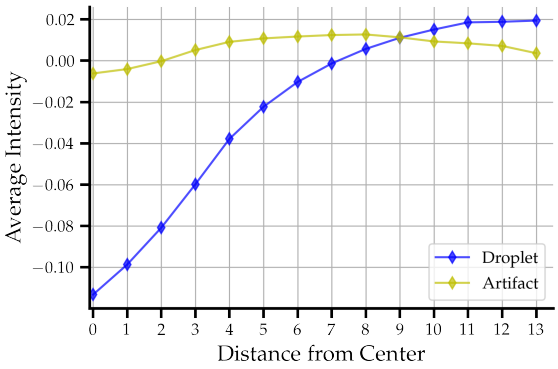
(A) Dataset 1 Amplitude



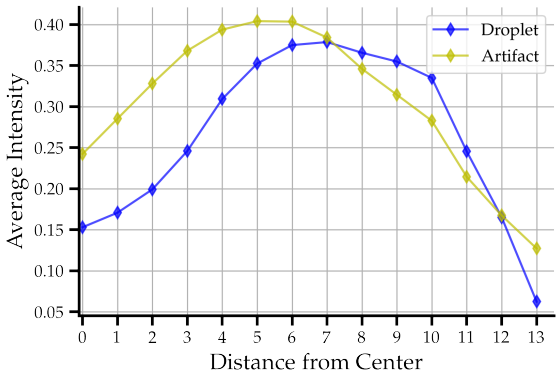
(B) Dataset 1 Phase



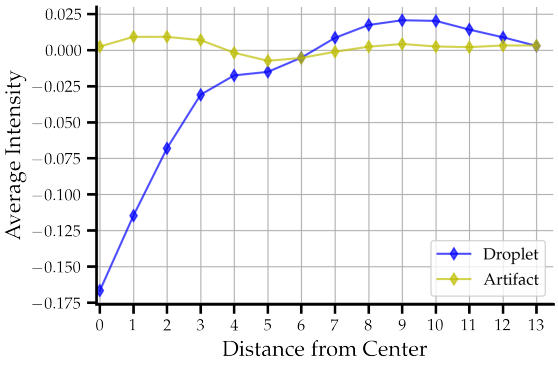
(C) Dataset 2 Amplitude



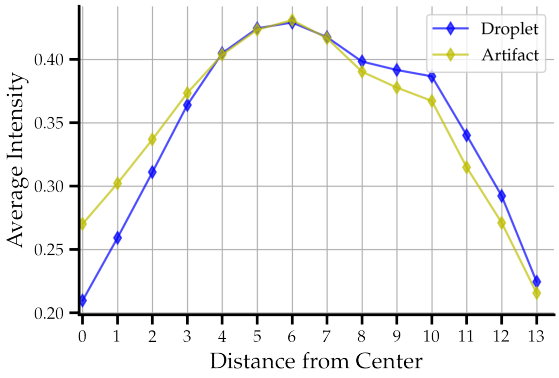
(D) Dataset 2 Phase



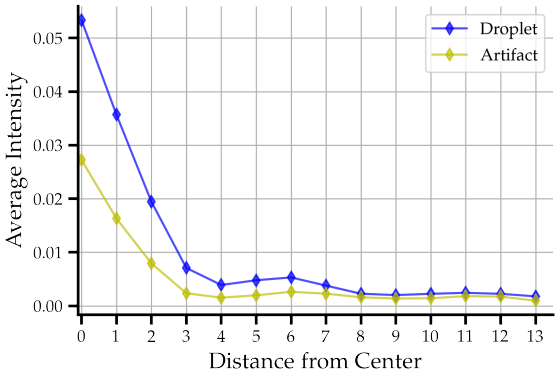
(E) Dataset 3 Amplitude



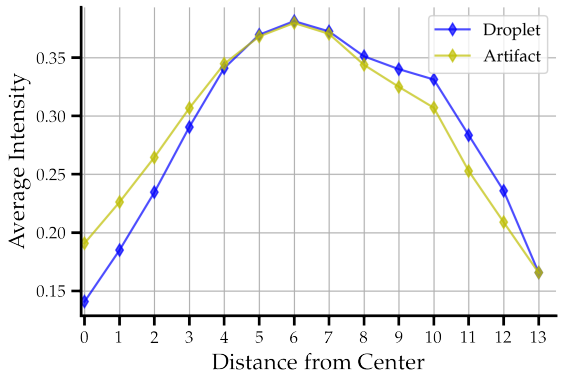
(F) Dataset 3 Phase



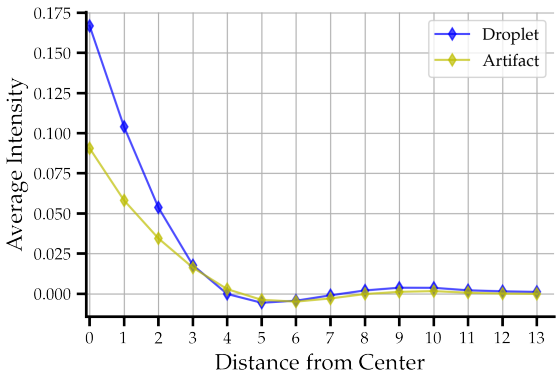
(G) Dataset 4 Amplitude



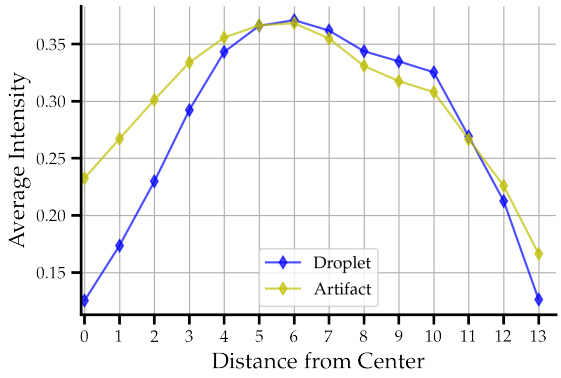
(H) Dataset 4 Phase



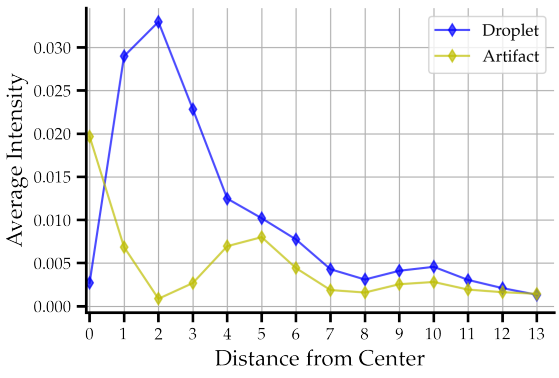
(I) Dataset 5 Amplitude



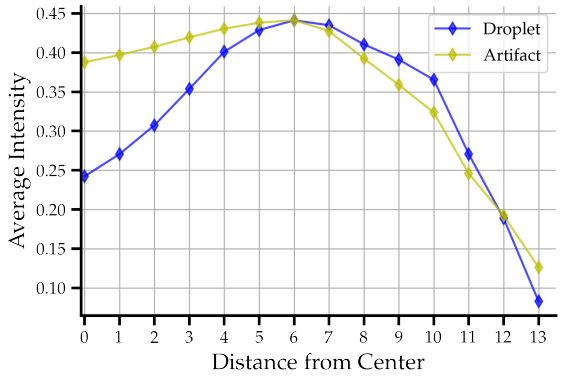
(J) Dataset 5 Phase



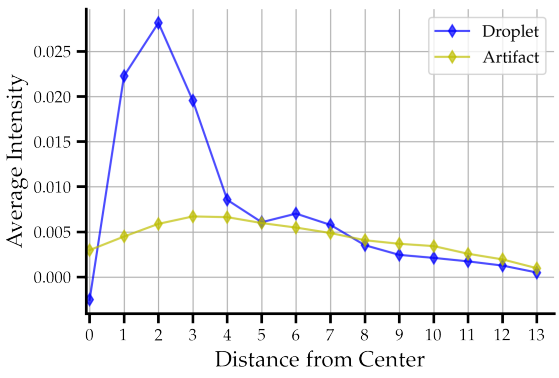
(K) Dataset 6 Amplitude



(L) Dataset 6 Phase

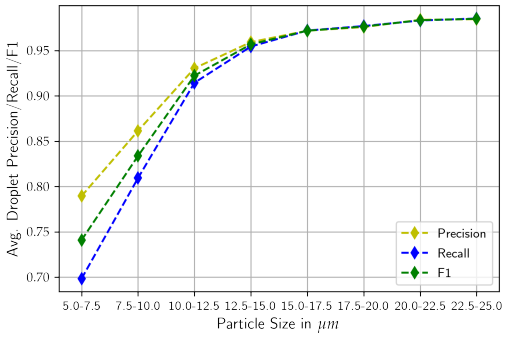


(M) Dataset 7 Amplitude

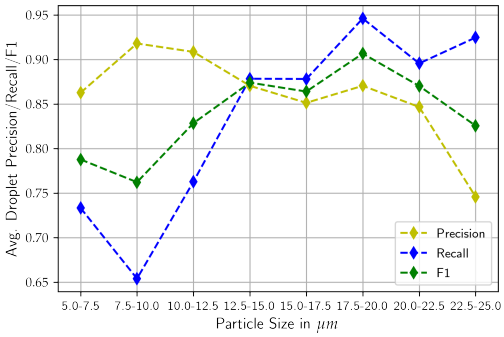


(N) Dataset 7 Phase

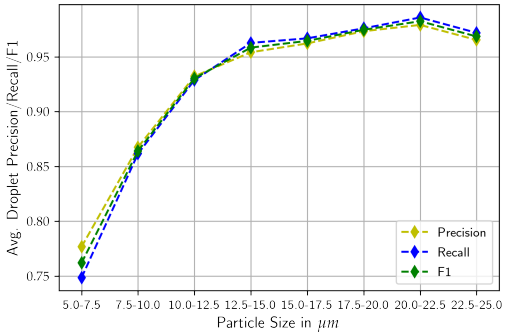
FIGURE 6: Average radial intensity for amplitude and phase. The figures show that except for dataset 2, pixel intensities at the center of amplitude images are lower for droplets than for artifacts. The datasets can be parted in three groups regarding the phase image intensities. Droplets for datasets 1, 4 and 5 have their maximum average intensities at the center. Droplet intensities for datasets 6 and 7 start out low and peak at a distance of two and droplet intensities for dataset 2 is an outlier with negative intensities for droplets in the center.



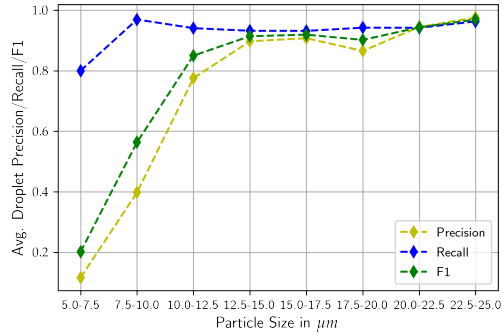
(A) Dataset 2 Validation



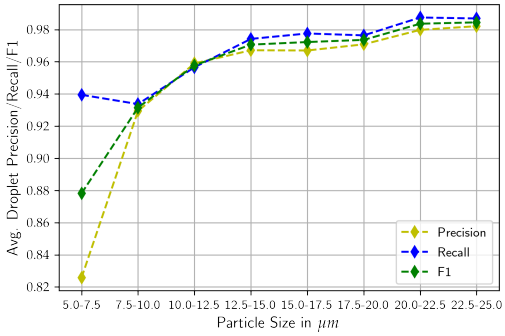
(B) Dataset 2 Test



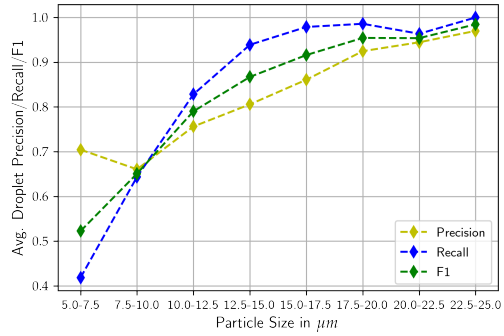
(C) Dataset 3 Validation



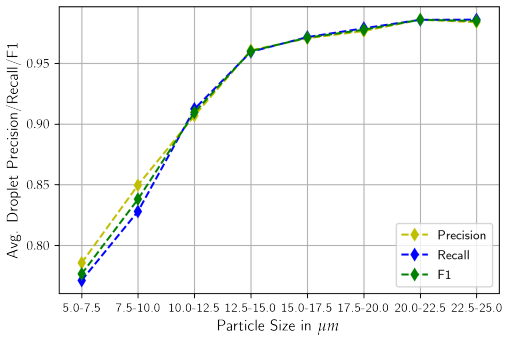
(D) Dataset 3 Test



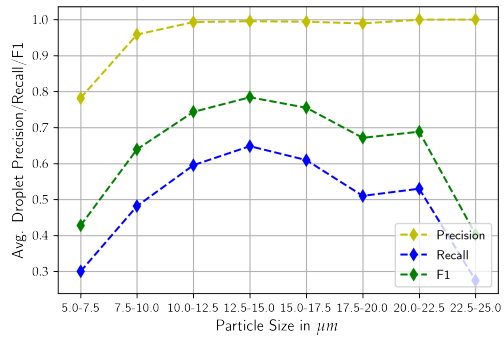
(E) Dataset 4 Validation



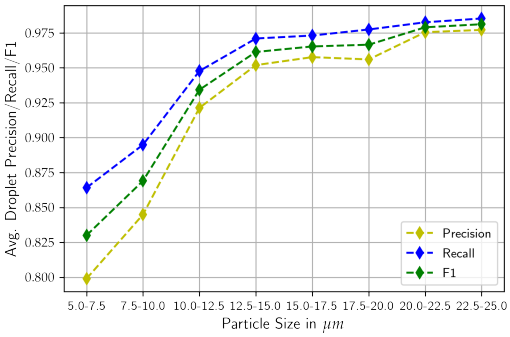
(F) Dataset 4 Test



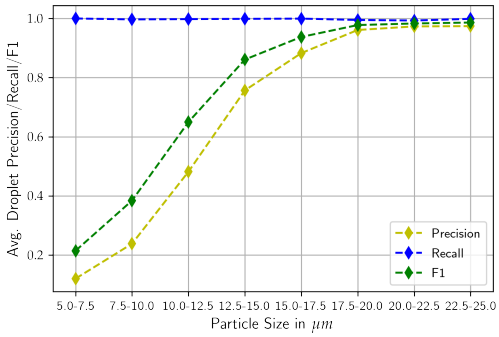
(G) Dataset 5 Validation



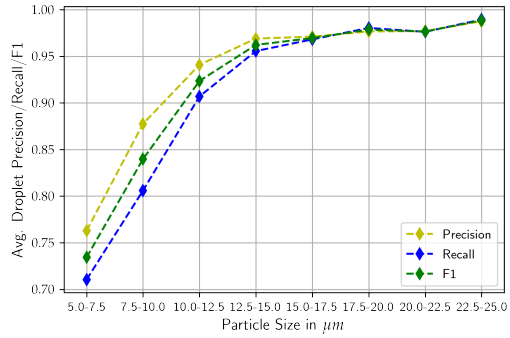
(H) Dataset 5 Test



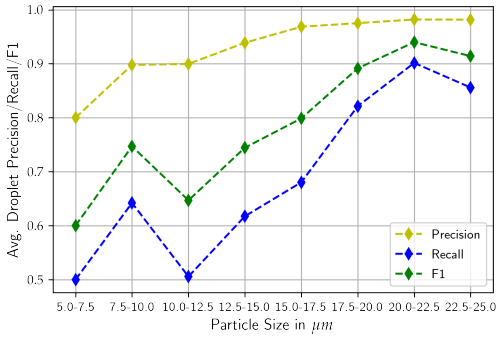
(I) Dataset 6 Validation



(J) Dataset 6 Test

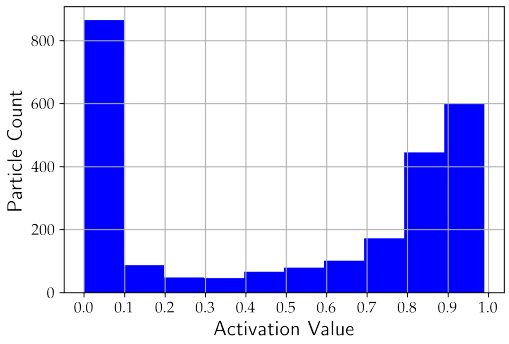


(K) Dataset 7 Validation

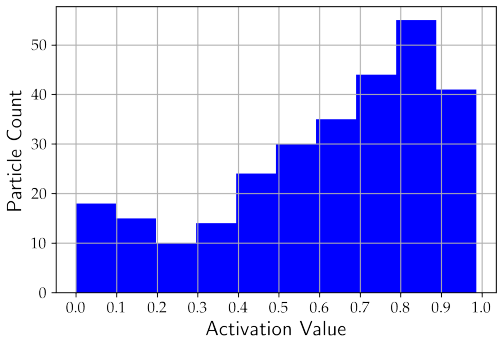


(L) Dataset 7 Test

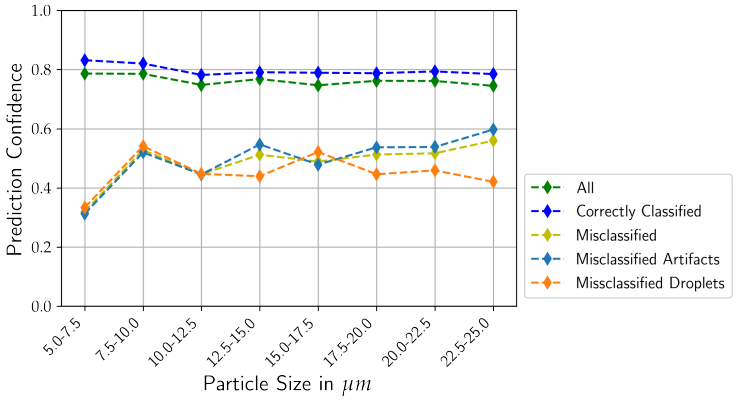
FIGURE 7: Individually centered images train and test performance by particle size, datasets 3-7. The validation performance figures show, that particles below 12.5 µm are classified less reliably than larger particles and that this effect is not due to generalization. This can also be seen in the test performances but there are larger variances depending on the dataset.



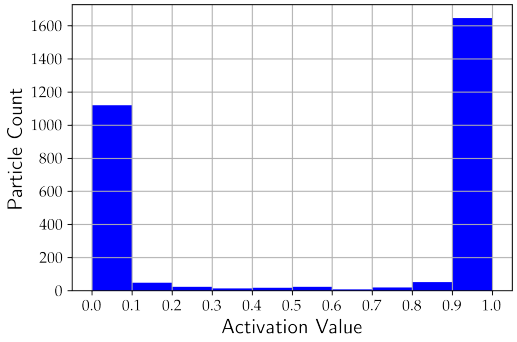
(A) Dataset 2 All Particles



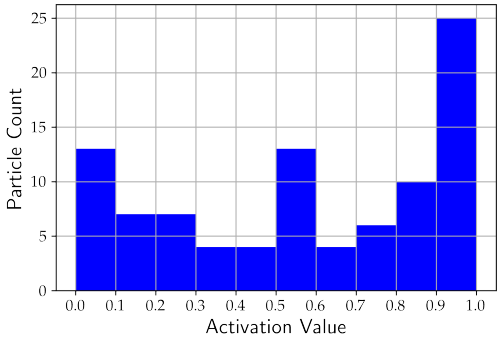
(B) Dataset 2 Missclassified Particles



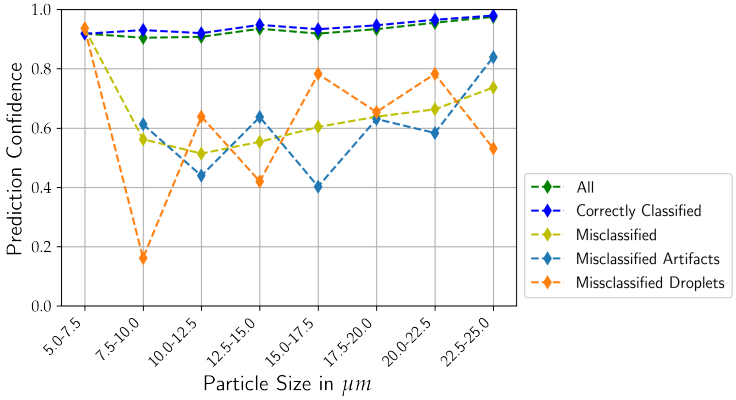
(C) Dataset 2 Confidence by Particle Size



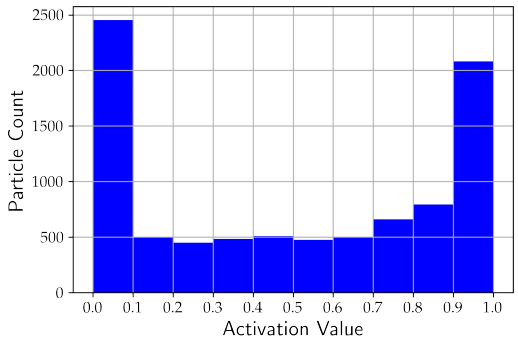
(D) Dataset 3 All Particles



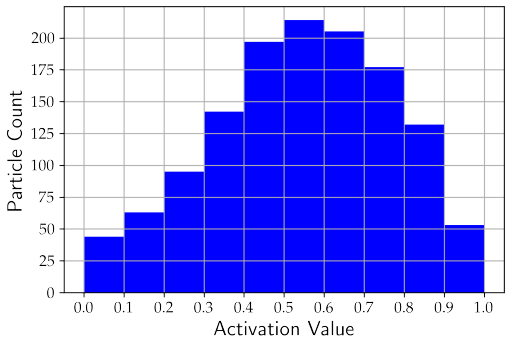
(E) Dataset 3 Missclassified Particles



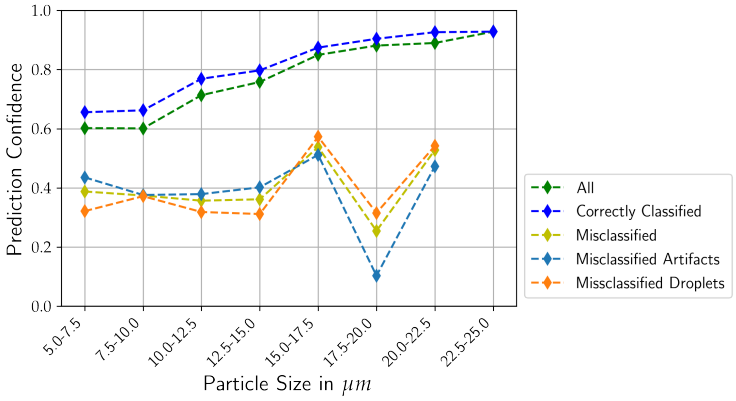
(F) Dataset 3 Confidence by Particle Size



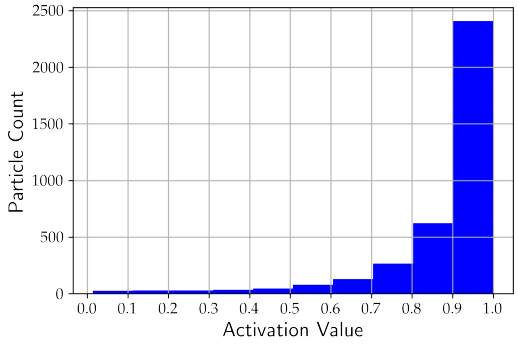
(G) Dataset 4 All Particles



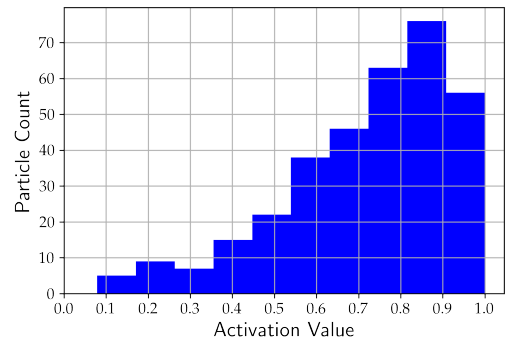
(H) Dataset 4 Missclassified Particles



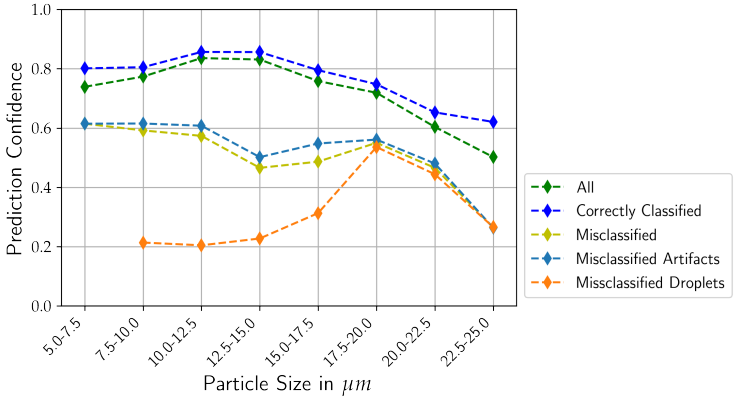
(I) Dataset 4 Confidence by Particle Size



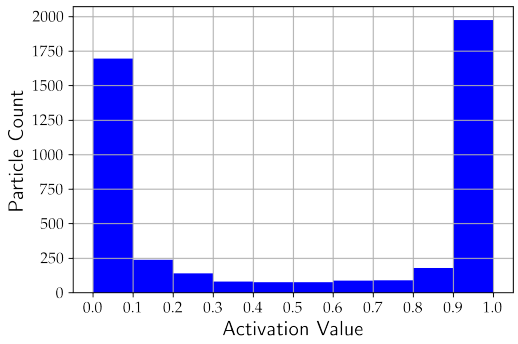
(J) Dataset 5 All Particles



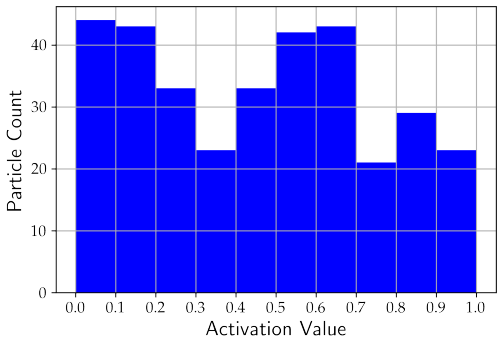
(K) Dataset 5 Missclassified Particles



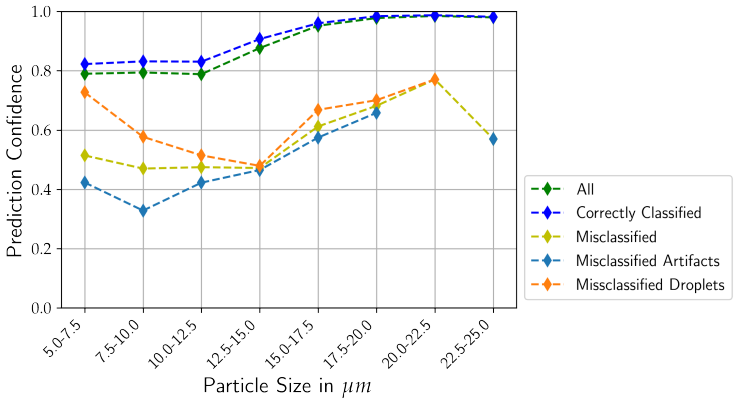
(L) Dataset 5 Confidence by Particle Size



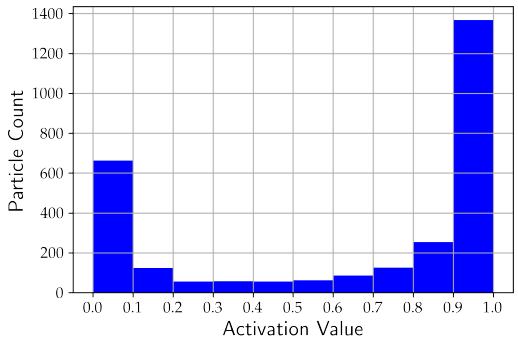
(M) Dataset 6 All Particles



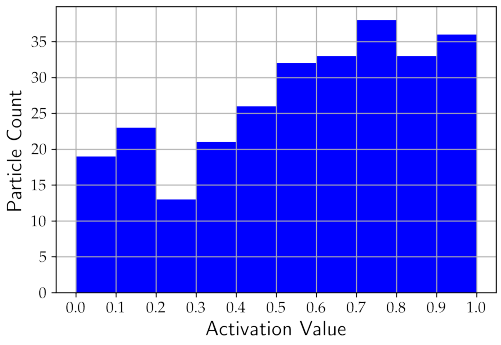
(N) Dataset 6 Missclassified Particles



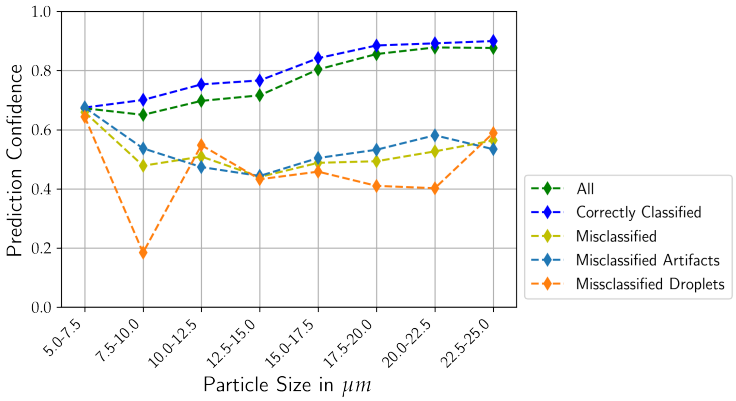
(O) Dataset 6 Confidence by Particle Size



(P) Dataset 7 All Particles



(Q) Dataset 7 Missclassified Particles



(R) Dataset 7 Confidence by Particle Size

FIGURE 8: Output neuron activation and confidence by size, datasets 2-7. The figures show that the neural networks confidences correlate to a predictions' correctness for all datasets and that this is true accross particle sizes.

Bibliography

Batista, Gustavo E. A. P. A., Ronaldo C. Prati, and Maria Carolina Monard (June 1, 2004). "sci". In: *ACM SIGKDD Explorations Newsletter* 6.1, pp. 20–29. DOI: 10.1145/1007730.1007735.

Baumgardner, D., J. L. Brenguier, A. Bucholtz, H. Coe, P. DeMott, T. J. Garrett, J. F. Gayet, M. Hermann, A. Heymsfield, A. Korolev, M. Krämer, A. Petzold, W. Strapp, P. Pilewskie, J. Taylor, C. Twohy, M. Wendisch, W. Bachalo, and P. Chuang (Oct. 1, 2011). "Airborne instruments to measure atmospheric aerosol particles, clouds and radiation: A cook's tour of mature and emerging technology". In: *Atmospheric Research* 102.1, pp. 10–29. DOI: 10.1016/j.atmosres.2011.06.021.

Burman, Prabir (Sept. 1, 1989). "A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods". In: *Biometrika* 76, pp. 503–514. DOI: 10.1093/biomet/76.3.503.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (June 1, 2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. DOI: 10.1613/jair.953.

Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz (June 1, 2004). "Editorial: special issue on learning from imbalanced data sets". In: *ACM SIGKDD Explorations Newsletter* 6.1, pp. 1–6. DOI: 10.1145/1007730.1007733.

Corfield, David (2009). "Projection and Projectability".

Fan, W. and I. Davidson (2007). "On Sample Selection Bias and Its Efficient Correction via Model Averaging and Unlabeled Examples". In: *SDM*. DOI: 10.1137/1.9781611972771.29.

Fugal, J. P. and R. A. Shaw (June 17, 2009). "Cloud particle size distributions measured with an airborne digital in-line holographic instrument". In: *Atmospheric Measurement Techniques* 2.1. Publisher: Copernicus GmbH, pp. 259–271. DOI: https://doi.org/10.5194/amt-2-259-2009.

Glorot, Xavier and Y. Bengio (Jan. 1, 2010a). "Understanding the difficulty of training deep feedforward neural networks". In: *Journal of Machine Learning Research - Proceedings Track* 9, pp. 249–256.

Glorot, Xavier, Antoine Bordes, and Y. Bengio (Jan. 1, 2010b). "Deep Sparse Rectifier Neural Networks". In: Journal of Machine Learning Research. Vol. 15.

Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing (May 1, 2017). "Learning from class-imbalanced data: Review of methods and applications". In: *Expert Systems with Applications* 73, pp. 220–239. DOI: 10.1016/j.eswa.2016.12.035.

Han, Jun and Claudio Moraga (1995). "The influence of the sigmoid function parameters on the speed of backpropagation learning". In: *From Natural to Artificial Neural Computation*. Ed. by José Mira and Francisco Sandoval. Red. by Gerhard Goos, Juris Hartmanis, and Jan Leeuwen. Vol. 930. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 195–201. DOI: 10.1007/3-540-59497-3_175.

Henneberger, J., J. P. Fugal, O. Stetzer, and U. Lohmann (Nov. 6, 2013). "HOLIMO II: a digital holographic instrument for ground-based in situ observations of microphysical properties of mixed-phase clouds". In: *Atmospheric Measurement Techniques* 6.11. Publisher: Copernicus GmbH, pp. 2975–2987. DOI: https://doi.org/10.5194/amt-6-2975-2013.

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov (July 3, 2012). "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv:1207.0580 [cs]*.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey Hinton (Jan. 1, 2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Neural Information Processing Systems* 25. DOI: 10.1145/3065386.

LeCun, Yann and Yoshua Bengio (1995). "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553. Publisher: Nature Publishing Group, pp. 436–444.

Liu, X., J. Wu, and Z. Zhou (Apr. 2009). "Exploratory Undersampling for Class-Imbalance Learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), pp. 539–550. DOI: 10.1109/TSMCB.2008.2007853.

López, Victoria, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera (Nov. 20, 2013). "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". In: *Information Sciences* 250, pp. 113–141. DOI: 10.1016/j.ins.2013.07.007.

López, Victoria, Alberto Fernández, Jose G. Moreno-Torres, and Francisco Herrera (June 1, 2012). "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics". In: *Expert Systems with Applications* 39.7, pp. 6585–6608. DOI: 10.1016/j.eswa.2011.12.043.

Loyola-González, Octavio, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and Milton García-Borroto (Jan. 29, 2016). "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases". In: *Neurocomputing* 175, pp. 935–947. DOI: 10.1016/j.neucom.2015.04.120.

Nair, Vinod and Geoffrey Hinton (June 16, 2010). "Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair". In: Proceedings of ICML. Vol. 27, pp. 807–814.

Peng, Yonghong, Peter Flach, Pavel Brazdil, and Carlos Soares (July 1, 2008). "Decision Tree-Based Data Characterization for Meta-Learning". In:

Ramachandran, Prajit, Barret Zoph, and Quoc V. Le (Oct. 27, 2017). "Searching for Activation Functions". In: *arXiv:1710.05941 [cs]*.

Ramelli, Fabiola, Alexander Beck, Jan Henneberger, and Ulrike Lohmann (Feb. 27, 2020). "Using a holographic imager on a tethered balloon system for microphysical observations of boundary layer clouds". In: *Atmospheric Measurement Techniques* 13.2. Publisher: Copernicus GmbH, pp. 925–939. DOI: https://doi.org/10.5194/amt-13-925-2020.

Schlenczek, Oliver (2018). "Airborne and ground-based holographic measurement of hydrometeors in liquid-phase, mixed-phase and ice clouds". PhD thesis. Johannes Gutenberg-Universität Mainz. DOI: 10.25358/OPENSCIENCE-4124.

Silverman, B. W. and M. C. Jones (1989). "E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)". In: *International Statistical Review / Revue Internationale de Statistique* 57.3. Publisher: [Wiley, International Statistical Institute (ISI)], pp. 233–238. DOI: 10.2307/1403796.

Storkey, Amos (Jan. 1, 2009). "When Training and Test Sets Are Different: Characterizing Learning Transfer". In: *Dataset Shift in Machine Learning*. Journal Abbreviation: Dataset Shift in Machine Learning, pp. 3–28. DOI: 10.7551/mitpress/9780262170055.003.0001.

Sugumaran, V., V. Muralidharan, and K. I. Ramachandran (Feb. 1, 2007). "Feature selection using Decision Tree and classification through Proximal Support

Vector Machine for fault diagnostics of roller bearing". In: *Mechanical Systems and Signal Processing* 21.2, pp. 930–942. DOI: 10.1016/j.ymssp.2006.05.004.

Tomek, Ivan (1976). "TWO MODIFICATIONS OF CNN." In: *IEEE Transactions on Systems, Man and Cybernetics* SMC-6.11, pp. 769–772. DOI: 10.1109/TSMC.1976.4309452.

Touloupas, Georgios, Annika Lauber, Henneberger Jan, Alexander Beck, and Aurélien Lucchi (May 8, 2020). "A convolutional neural network for classifying cloud particles recorded by imaging probes". In: *Atmospheric Measurement Techniques* 13.5, pp. 2219–2239. DOI: 10.5194/amt-13-2219-2020.

Yamashita, Rikiya, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi (Aug. 2018). "Convolutional neural networks: an overview and application in radiology". In: *Insights into Imaging* 9.4. Number: 4 Publisher: SpringerOpen, pp. 611–629. DOI: 10.1007/s13244-018-0639-9.

Zadrozny, Bianca (Sept. 20, 2004). "Learning and Evaluating Classifiers under Sample Selection Bias". In: *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004* 2004. DOI: 10.1145/1015330.1015425.

Zadrozny, Bianca, J. Langford, and Naoki Abe (Dec. 19, 2003). "Cost-Sensitive Learning by Cost-Proportionate Example Weighting". In: Proceedings - IEEE International Conference on Data Mining, ICDM, pp. 435–442. DOI: 10.1109/ICDM.2003.1250950.

Declaration of Authorship

I hereby confirm that this thesis constitutes my own work, produced without aid and support from persons and/or materials other than the ones listed.Quotation marks indicate direct language from another author. Appropriate credit is given where I have used ideas, expressions or text from another public or non-public source.The paper in this or similar form has never been submitted as an assessed piece of work in or outside of Germany. It also has not yet been published.

City, Date:

_____

Authors' Signature:

_____