

## Machine-learning-based detection of spin structures


Isaac Labrie-Boulay<sup>1,§</sup>, Thomas Brian Winkler<sup>1,\*</sup>, Daniel Franzen<sup>2</sup>, Alena Romanova<sup>1</sup>,  
Hans Fangohr<sup>3,4</sup> and Mathias Kläui<sup>1,†,‡</sup>

<sup>1</sup>*Institute for Physics, Johannes Gutenberg University, Mainz 55099, Germany*

<sup>2</sup>*Institute of Computer Science, Johannes Gutenberg University, Mainz 55099, Germany*

<sup>3</sup>*Max Planck Institute for the Structure and Dynamics of Matter, Hamburg 22761, Germany*

<sup>4</sup>*Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, SO17 1BJ, Hampshire, United Kingdom*

 (Received 26 May 2023; revised 11 August 2023; accepted 16 November 2023; published 10 January 2024)

One of the most important magnetic spin structures is the topologically stabilized skyrmion quasiparticle. Its interesting physical properties make it a candidate for memory and efficient neuromorphic computation schemes. For device operation, the detection of the position, shape, and size of skyrmions is required and magnetic imaging is typically employed. A frequently used technique is magneto-optical Kerr microscopy, in which, depending on the sample's material composition, temperature, material growing procedures, etc., the measurements suffer from noise, low contrast, intensity gradients, or other optical artifacts. Conventional image analysis packages require manual treatment, and a more automatic solution is required. We report a convolutional neural network specifically designed for segmentation problems to detect the position and shape of skyrmions in our measurements. The network is tuned using selected techniques to optimize predictions and, in particular, the number of detected classes is found to govern the performance. The results of this study show that a well-trained network is a viable method of automating data preprocessing in magnetic microscopy. The approach is easily extendable to other spin structures and other magnetic imaging methods.

DOI: [10.1103/PhysRevApplied.21.014014](https://doi.org/10.1103/PhysRevApplied.21.014014)

### I. INTRODUCTION

In the last decade, magnetic quasiparticles [1,2] have raised the attention of researchers due to their interesting properties and their potential applicability in next-generation memory and neuromorphic devices [2,3]. In particular, skyrmions in magnetic thin films are under investigation because of their topological stability, their small size, and their ease of manipulation by spin currents [4,5]. However, depending on the magnetic properties of the thin film, the skyrmion size might vary from a few nanometers [6] up to a few micrometers [7]. Furthermore, skyrmions can be circular, elliptical, or other shapes, and a typical spin structure of a skyrmion with a Bloch domain wall is shown in the inset of Fig. 1(a). Skyrmions have been suggested and experimentally demonstrated for a wide range of nonconventional computing schemes, such as stochastic computing [8,9], token-based Brownian computing [10,11], and reservoir computing [12–14]. Common to realizing all these approaches is the reliable

detection of the skyrmions. In particular, magneto-optical Kerr effect (MOKE) microscopy is used to study magnetic skyrmions with diameters up to the micrometer scale [5,9]. The obtained grayscale contrast image in a Kerr microscope corresponds to the out-of-plane magnetization of the sample. For the analysis of skyrmions and their static and dynamic properties one needs to identify the position of all skyrmions as well as their shape [9,15]. This has been very labor intensive so far as it has typically been done manually, precluding large-scale systematic analysis. Additionally, manual data preprocessing is often necessary when using image analysis software to be able to evaluate the data correctly. Challenges are thermal and electrical noise or low contrast, which might arise from an imperfect alignment of the polarization filters, stray light, or the use of nonmagnetic capping layers required to avoid the oxidation of the sample, while partially absorbing the light. Furthermore, an intensity gradient might be visible in the data if the sample is not perfectly aligned or if it is inhomogeneous. Structural defects and/or dust particles might also impede data evaluation, as can sample edges. In Figs. 1(a), 1(d), and 1(g) we show a collection of typical MOKE microscopy data.

The study of skyrmions typically involves evaluating the size and position, and tracing the motion of skyrmions.

\*twinkler@uni-mainz.de

†mathias.klaui@klaui-lab.de

‡klaui@uni-mainz.de

§Both authors contributed equally to the work

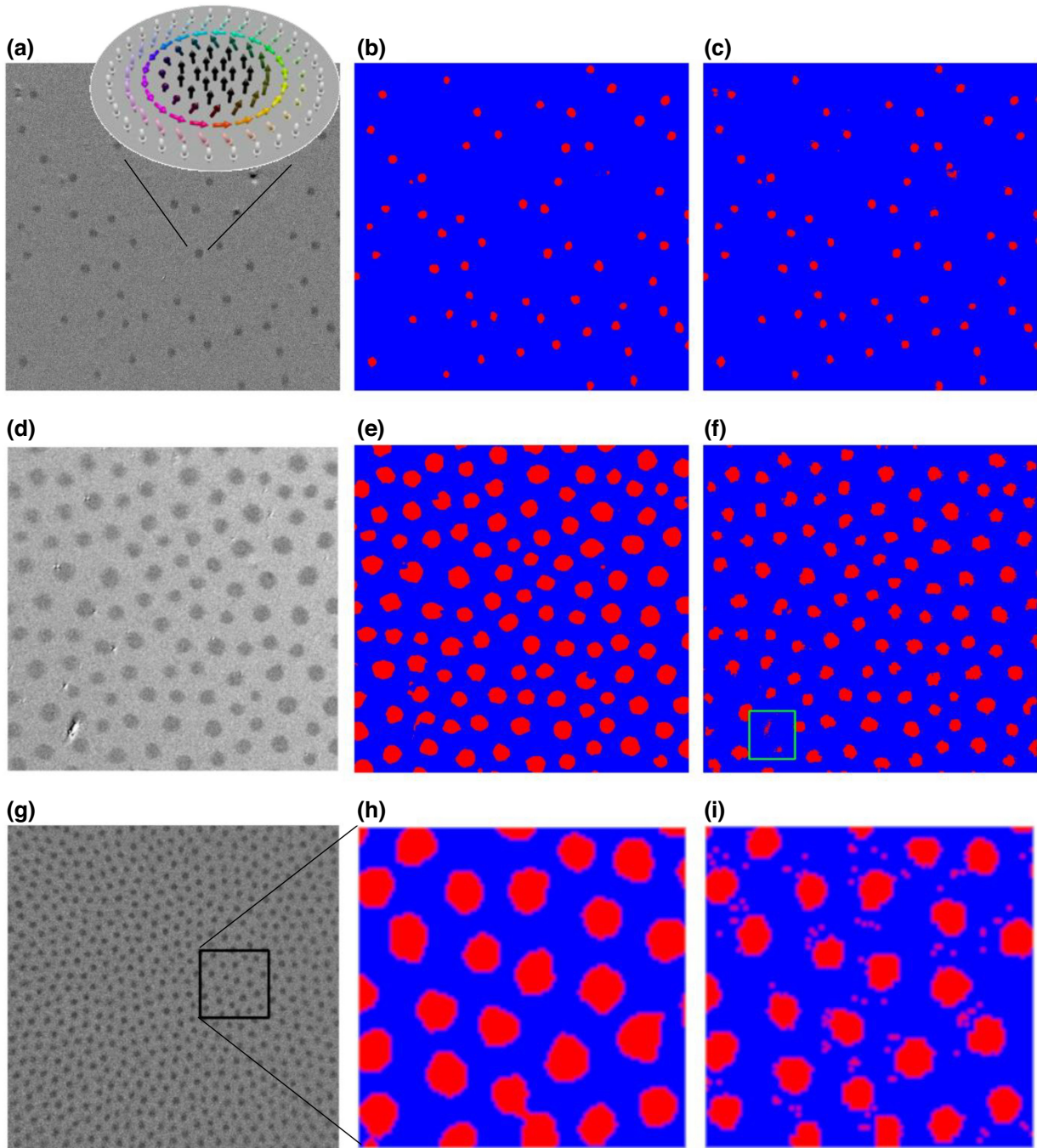


FIG. 1. Two-class model predictions. The pictures on the left show MOKE measurements, which are difficult to mask using traditional methods; the middle column shows the labels of these measurements; and the right column shows predictions made by these two-class models. (a) A sample with few defects. The inset visualizes the spin structure of a skyrmion. The centre of the skyrmion with the magnetization “down” leads to black contrast and the surrounding area with the magnetization “up” leads to light gray contrast. Note that the colored Bloch domain wall in the inset is not visible in the experimental image due to the limited spatial resolution. (b) Label of (a). (c) Prediction of (a). (d) This sample has defects and a slight brightness gradient. (e) Label of (d). (f) Prediction of (d) where the green square indicates an example of how these models predict defects. (g) This sample is particularly noisy and the black square indicates an enlarged area. (h) A 4 times enlargement of the label of (g). (i) A 4 times enlargement of a prediction of (g) (note the small speckles indicating incorrect detection). The inset skyrmion sketch in (a) is adapted from Ref. [1].

This is often done by first creating a binary mask of measurements and then subjecting it to standard data analysis techniques. Binary mask creation through image processing software requires manually performing a series of manipulations that are highly dependent on the measurement to convert the raw data to masks [16]. Some of these manipulations include denoising, contrast enhancement, the application of a Gaussian filter, thresholding, and more, depending on the measurement. Furthermore, no classical filter can reliably distinguish skyrmions or different magnetic structures, such as stripe domains, vortices, or domain walls. Figures 1(d) and 1(g) are examples of measurements that require significant effort to preprocess using conventional methods.

Neural networks can provide a powerful means of automating the evaluation process since image segmentation is one of the key abilities of convolutional neural networks (CNNs). Huge progress has been made in the last decade due to the extended use of general-purpose graphical processing units [17]. Since machine learning has previously been successfully applied to magnetic problems [18–21], we here tackle the key problem of detecting the most important spin structure, the skyrmion, to automate the analysis of magnetic microscopy measurements to enable applications such as unconventional computing.

We are aiming for a segmentation of the MOKE microscopy image, i.e., a pixel-wise classification of the data. The thresholded output will be the same size as the input data and deliver the most likely prediction for every pixel (see Fig. 1 for examples). With that approach, we have direct access to (i) a size-shape analysis, and (ii) a direct separation between magnetic structures (skyrmions) and optical defects (scratches, dust particles, imperfections, sample drift artifacts, etc.). While the first may also be achieved with smart “classical” algorithms [21], the latter really needs manual treatment in every video or even image. A well-trained neural network is able to deliver both in one step. However, if the quality of the data is too poor, e.g., because of a very low signal-to-noise ratio, a neural approach might not help. In Supplemental Material Appendix A [33] we analyze the limits of our approach in detail.

In this study, we explore the ability of a CNN to segment our MOKE microscopy data and obtain an automatic detection of skyrmions. We compare a two-class and a three-class network including defects and we tune the CNN to optimize the skyrmion and defect segmentation that exceeds the performance of conventional pattern detection. So far, no machine-learning-based approach has been proposed to tackle the specific task of segmentation of magnetic thin-film spin structure data; however, CNNs are a possible choice that must be explored, as in the following.

## II. METHODS

The labels of the datasets were one-hot encoded for two classes. The model was trained on 1901 training images sampled from 19 separate Kerr microscope videos. The labels were manually created using conventional mask creation techniques with the image processing software IMAGEJ [16]. The validation dataset was composed of 200 images taken from five different videos. Different CNNs can be used. Here we have chosen a U-Net derivative as the framework is sufficiently flexible for the detection of different physical entities in images [22]. After training, the model generated a prediction on the validation set and from there one could extract its performance metric based on a confusion matrix. The performance metric serves as a method for comparing models when one attempts to optimize a neural network for a specific task. After all optimizations have been applied, the test set is used for a final benchmarking. In our case, the test set was composed of 200 images from six different videos taken on different samples. It is important to note that both the validation set and the test set included images of skyrmions of different sizes (from around 30 up to 1000 pixels) We also refer to Supplemental Material Appendix B [33] for more details about the material stacks, the true skyrmion sizes, and the data acquisition, but want to emphasize that the true skyrmion size is not relevant for the prediction as our network is size agnostic. Our data included images with different degrees of noise, and images with different defects. An analysis on the dataset composition by evaluating the labels resulted in the following numbers: The training set had a total of 1 466 843 skyrmions (counting the clusters of pixels labeled as skyrmions), which made up 18.9% of all pixels in the set. The validation set had 17.8% and the test set 9.9% of its pixels labeled as skyrmions. The dataset was therefore considered to be imbalanced.

The score used for benchmarking our models was the Matthews correlation coefficient (MCC) [23]. Studies have shown that the MCC gives a strongly indicative performance score for both positively biased and negatively biased predictions for every kind of dataset [23]. Generally speaking, a positively biased prediction accurately predicts positive values but does poorly when predicting negative ones. The MCC has the ability to properly score both positively and negatively biased predictions from positively imbalanced datasets, negatively imbalanced datasets, and balanced datasets [23]. As such, this metric is used as a reference performance measure for imbalanced data in different fields of research. In our dataset, skyrmions were classified as positive and the background as negative. Our datasets were negatively imbalanced as there were more background pixels than skyrmion pixels in the images. Therefore, the MCC’s ability to account for class imbalance is very important in our case.

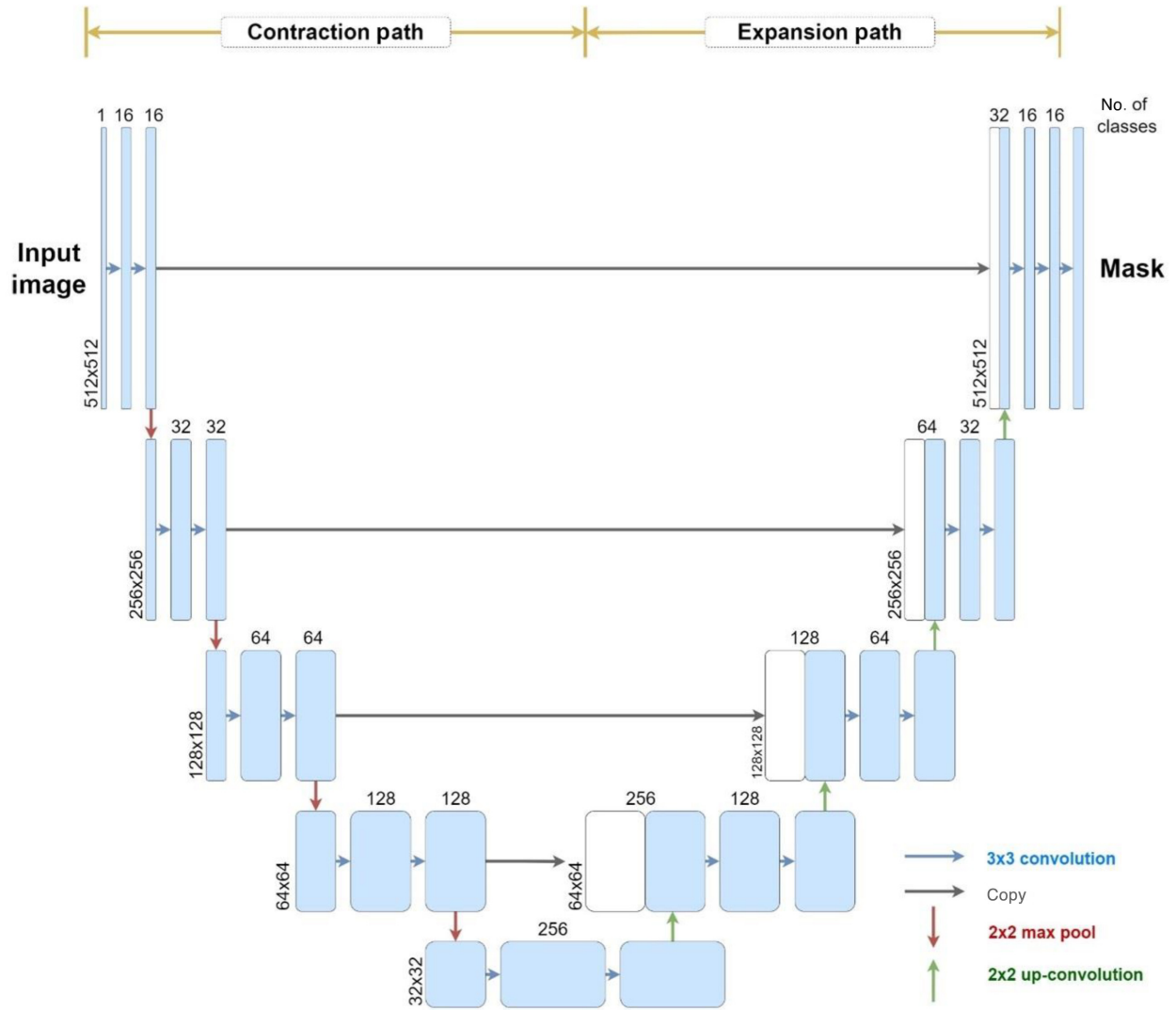


FIG. 2. Network topology for the CNN used in this study. The left side is the contraction path and the right side is the expansion path. The gray arrows connecting both paths indicate the skip connections. Figure is adapted from Ref. [22].

The MCC for a binary classifier is given by [23],

$$m_{\text{MCC}} = \frac{N_{\text{TP}} \times N_{\text{TN}} - N_{\text{FP}} \times N_{\text{FN}}}{\sqrt{(N_{\text{TP}} + N_{\text{FP}})(N_{\text{TP}} + N_{\text{FN}})(N_{\text{TN}} + N_{\text{FP}})(N_{\text{TN}} + N_{\text{FN}})}}$$

with TP being the true positive classified pixel in the prediction, TN the true negative, FP the false positive, and FN the false negative ones. The MCC only provides a high score if all the confusion matrix categories perform well, making it a superior metric compared to an often-used metric like accuracy or the  $F1$  score.

The training was run 5 times for every model. The MCC was calculated for every run. The mean and standard deviation of each model's performance were then calculated. This method of analyzing the performance of a model gives a more realistic idea of the performance. Note that,

unlike other standard measures, the  $m_{\text{MCC}} \in [-1, 1]$ , where  $-1$  refers to a totally incorrect classification,  $0$  means perfectly inconclusive, and  $1$  indicates a perfect classification. The MCC, however, only measures the prediction performance in a pixel-wise approach, and does not provide a physical interpretation of the data.

The models were written in PYTHON using the TENSORFLOW 2.9 [24] machine-learning library with KERAS [25] running on top providing the high-level application programming interface (API). This study's software also utilized functionalities from NUMPY [26], SCIKIT-LEARN [27], and PIL [28]. All models were trained on an Nvidia RTX 3060 GPU [17].

### III. RESULTS AND DISCUSSION

CNNs used for optical pattern recognition use both a down-sampling structure called the contraction path and

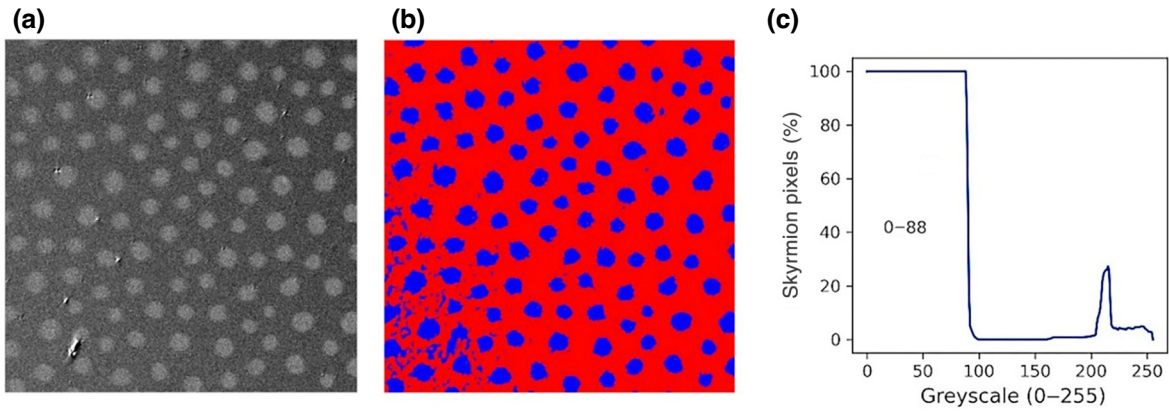


FIG. 3. Inverted data predictions. (a) Color-inverted sample. (b) The prediction. This figure shows that any dark pixel is nearly always automatically classified as a skyrmion for the two-class case. The splash in the bottom left corner is likely caused by the fact that the sample has a color gradient where the image gets progressively darker from the bottom left to the top right corner. (c) Grayscale pixels ranging from 0 to 88 are predicted to be skyrmions 100% of the time.

an up-sampling structure called the expansion path. The contraction path consists of two  $3 \times 3$  convolutions followed by a rectified linear unit (ReLU) activation layer, and the down-sampling is finally achieved using a  $2 \times 2$  maximum pooling layer with a stride of two [22] in our case. For every down-sampling block, the number of feature channels is doubled. The expansion path concatenates the corresponding depth feature map from the contraction path, applies two  $3 \times 3$  convolutions and a ReLU activation layer, and up-samples the feature map via a  $2 \times 2$  up-convolution [22]. The concatenation of high-resolution features from the left side and up-sampled output from the right side is required for localization. These are called shortcuts. It should also be noted that the frequently used U-Net implementation employs unpadded convolutions, which has the effect of creating masks with reduced size relative to the input image. Our implementation used consistent padding to keep the dimensions of the original image at the output. An illustration of the CNN used in this study can be seen in Fig. 2. This two-class CNN had a dropout rate of 5% between its  $3 \times 3$  convolution blocks. Dropout is a regularization technique that makes it so that nodes in a layer are randomly ignored throughout training [29,30]. Details of the network and the analysis are provided in Sec. II.

To train our CNN, we carried out five trainings over 100 epochs. We used a few `TENSORFLOW callback` functions to prevent overfitting (EarlyStopping, ReduceLROnPlateau, ModelCheckpoint [24]). The average MCC of the validation data prediction was 0.7520 with a standard deviation of 0.032. Examples of predictions for this benchmark model can be seen in Figs. 1(c), 1(f), and 1(i).

We note that the model had difficulties when classifying defects, see Fig. 1(f) for an example. Often these defects have a wide mixture of very different contrasts caused by sample drift and background subtraction when taking

MOKE microscopy measurements. This causes part of the defect to be classified as a skyrmion and the other as background. Even though defects are quite different in shape and contrast to the skyrmions, the models could not easily classify the two as two distinct objects.

Furthermore, in the case of noisy images, the masks created by the models were predicted to have tiny skyrmion speckles. Figure 1(i) shows a prediction done on one such sample. This is problematic from an applied physics viewpoint. An experimentalist's analysis might, for example, involve finding and tracking the average size of skyrmions in a set of measurements. One can see how such small speckles might create difficulties for reliable detection.

In the spirit of interpretable artificial intelligence (AI), it was found that the cause for this problem is that the models are mostly predicting based on intensity values. We found that grayscale image analysis could reveal this: We tasked one of these models with making predictions on 255 uniformly colored 8-bit grayscale images whose color ranged from 0 to 255. This experiment found that any pixel having a value smaller than 88 was classified as skyrmion and any value above was predicted as background. The result can be seen in Fig. 3(c). This provides clear evidence for this conjecture as to why the speckles appear in noisy images. Their noise produces a wide range of pixel brightness values. Random higher brightness pixels are classified as skyrmions.

Taking this analysis further, the model was tasked with predicting the inverted sample in Fig. 3(a), where skyrmion pixels have high grayscale values and background pixels have low grayscale values. The model's prediction can be seen in Fig. 3(b). As expected, one can see that the model automatically classifies any dark object as a skyrmion (white), which results in a nearly inverted prediction.

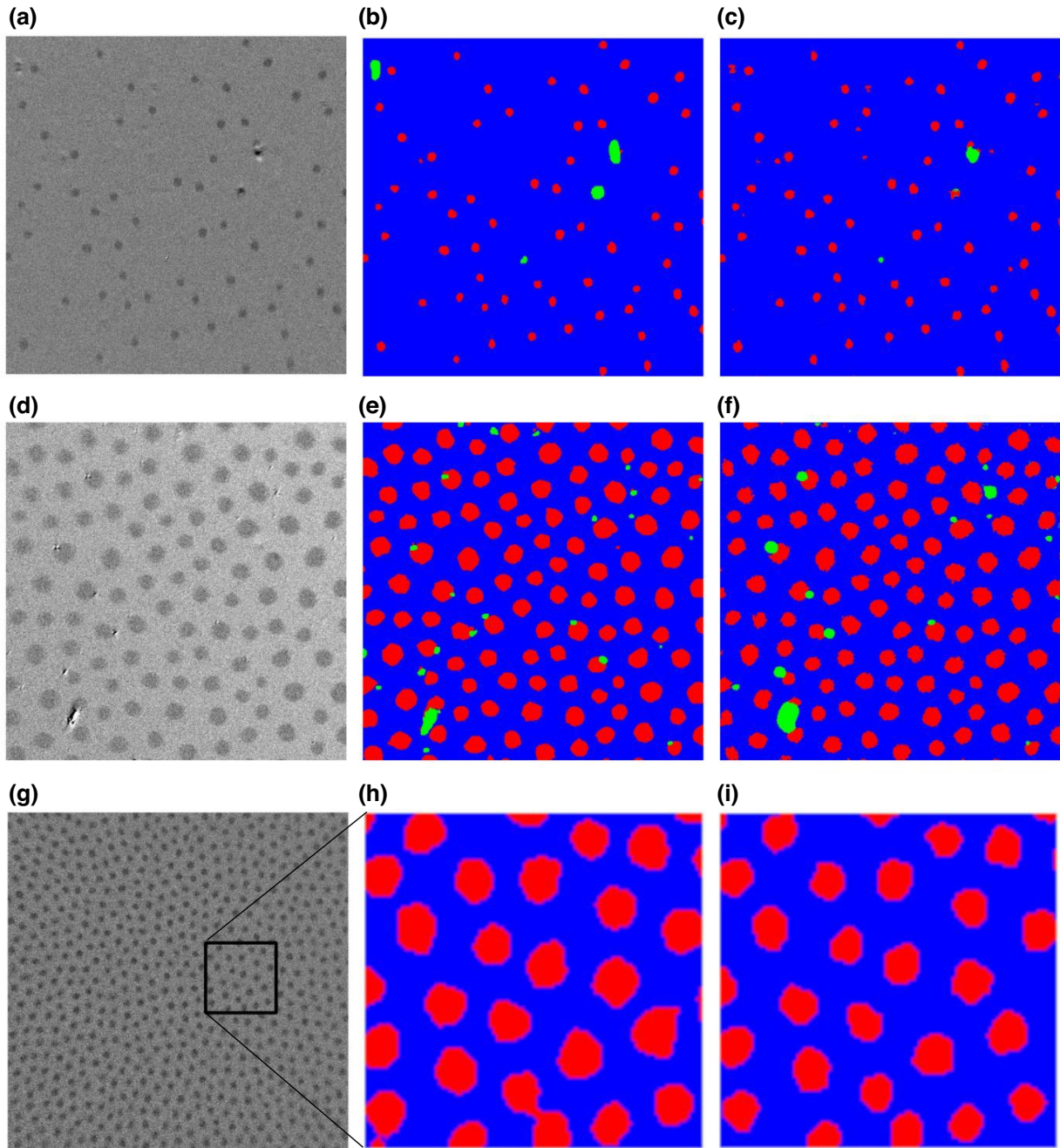


FIG. 4. Three-class model predictions. The pictures on the left (a),(d),(g) show MOKE measurements; those in the middle column (b),(e),(h) show the labels (note the defects labeled in green); and those on the right (c),(f),(i) show predictions made by these three-class models. (c),(d) show that this model architecture can tell the difference between skyrmions and defects. As can be noted in (i), this kind of model no longer predicts small skyrmion speckles in noisy images.

To remove this shortcoming and create a model that is better suited for discriminating against defects and random noise, a third-class was added to represent defects. By forcing the model to differentiate between the two different objects, the network is forced to learn to predict based on shapes and other factors beyond simply color intensity. This new multiclass model (skyrmion,

defect, background) was supervised using three class labels.

The training data for this work comprised data from previous publications, where areas with few defects were studied. Therefore, defects in these data were rare. Here we added 100 new measurements deliberately containing many defects that had been properly labeled by hand and a

TABLE I. Performance of three-class model optimizations. This table shows the MCC value for each optimization. The tested activation functions were parametric rectified linear unit function (PReLU), a hyperbolic tangent function (tanh), and “mish,” mathematically defined as  $f(x) = x \tanh(\text{softplus}(x))$ . All are common activation functions used in neural networks. For label smoothing, we explored with  $\alpha = 0.3$  and  $\alpha = 0.4$ . The default model uses  $\alpha = 0.2$ . We also tried training the model using different amounts of dropout following the  $3 \times 3$  convolutions (10% and 15%).

Optimization type	Optimization	Mean and standard deviation (SD) of MCC
Two-class model	None	0.752 (SD = 0.064)
Three-class initial model	None	0.896 (SD = 0.002)
Data augmentation	Without inversion	0.898 (SD = 0.003)
Activation function	With inversion PReLU	0.884 (SD = 0.012)
	tanh	0.869 (SD = 0.014)
	mish	0.891 (SD = 0.006)
	30%	0.875 (SD = 0.033)
Label smoothing	40%	0.884 (SD = 0.005)
	0%	0.896 (SD = 0.002)
	10%	0.900 (SD = 0.006)
Dropout	15%	0.892 (SD = 0.004)

model was then trained on these images to label defects on the original 1901 images. The predicted labels were post-processed so that the three-class model had 2001 images and is said to be weakly labeled. The three-class model also had a validation set and a testing set, which both contain 200 images with respective handmade labels.

The MCC was computed by keeping skyrmions as the positive class and combining defects and backgrounds as the negative class. Just like the original network, the three-class CNN was trained 5 times over 100 epochs. The average MCC of the three-class predictions on the validation set was 0.849 with a standard deviation of 0.0075, which reveals a significant increase in predictive capability. Example predictions from this model can be seen in Figs. 4(c), 4(f), and 4(i). It should be noted that now the small skyrmion speckles are all gone in the model.

Since the three-class model clearly performed better, the next step was to test additional optimizations to the three-class neural network to see if they could improve performance further. One modification was made at a time and the means and standard deviations of the MCC were recorded. In the end, all the modifications showing improved performance were combined to generate a master model. The master model was then tested against the

TABLE II. Data augmentation used. This table summarizes the data transformations used in the data-augmented models. Image inversion is marked with an asterisk to signify that only a subset of models was subjected to this augmentation both in training and at test time.

Data augmentation techniques
Random 90° rotation
Gaussian noise
Random shift
Random scaling (up to a maximum factor of 0.2)
Random contrast change with a (−0.3, 0.3) limit
Random brightness change with a (−0.3, 3) limit
Image inversion*

\*Only applies to entries marked “with inversion” in Table I.

test data to prevent bias towards the validation set. For this study, the training epochs were reduced and restricted to 15. The callback function hyperparameters were also adjusted accordingly.

The optimizations tested on our network and their resulting performance can be seen in Table I. It should be noted that both the two-class and three-class benchmark models were also trained over 15 epochs to compare them with the optimized models. We explored different activation functions, fine-tuning the dropout percentage, different levels of label smoothing, and two different data augmentation schemes.

Data augmentation refers to a technique that performs random transformation on the data during training. In that way, the amount of training data is artificially increased, as previously proposed [31]. Test time augmentation refers to a data augmentation that happens at the prediction stage (after training). It entails pooling predictions from several transformed versions of a given test input to obtain a “smoothed” prediction [32]. In our case, a prediction was made on three differently transformed versions of the image, and the final prediction was averaged over the three.

The data transformations used in the data augmentation can be seen in Table II. Applying random 90° rotations, random lateral shifts, or a random scaling simply artificially expands the number of measurement images seen. The maximum scaling factor of 0.2 was chosen so the model was not trained on images containing skyrmions that were unnecessarily big. Some images were randomly injected with Gaussian noise to give the model access to more noisy data. Random changes in brightness and contrast were also applied to images. Because of the nature of the MOKE microscopy measurements, some of the defects contained pixels that were clipped or close to being clipped by the 255 limit of the 8-bit grayscale images. The high brightness of defects was caused by the continuous background subtraction from the Kerr microscope measurements. These background subtractions are essential to obtain magnetic data but after a certain time sample drift

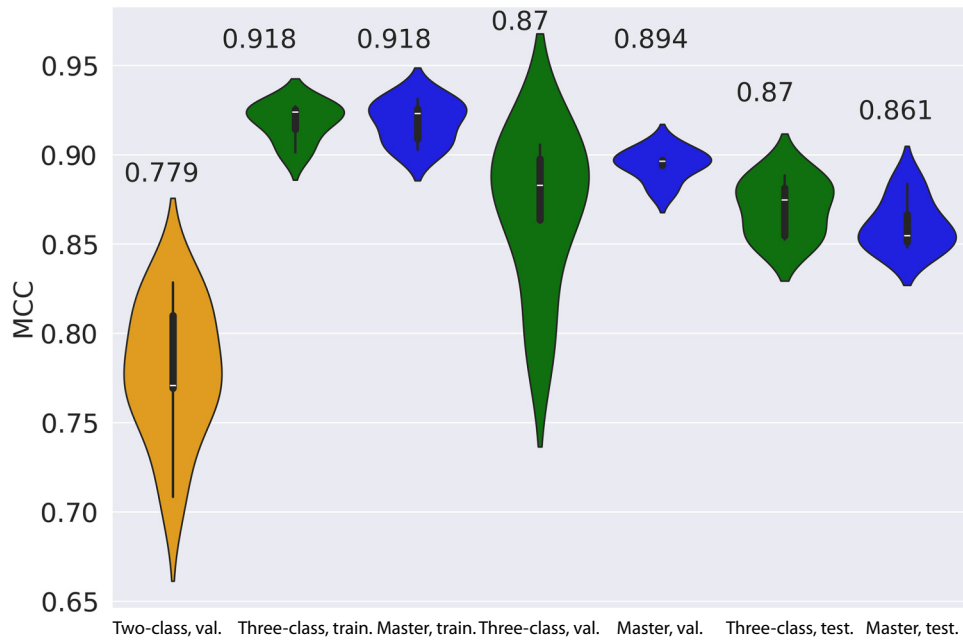


FIG. 5. Performance statistics of the benchmark and master models on the three sets. Plot comparing the average MCC and MCC standard deviation of the three-class benchmark model and the master model on all three datasets. The average value is written next to the respective distribution.

and the shifting of the microscope illumination intensity becomes so significant that it creates shadow-like shapes around very bright defects. To prevent the model from registering objects that were uniformly composed of clipped pixels, the brightness and contrast were only allowed to be amplified by a factor of  $\leq 0.2$ , as seen again in Table I.

A subset of models was trained using both the same data augmentation techniques and random image inversion. The idea was to force the network to distinguish skyrmions not only by relative contrast differences but by shape in general. It should be noted that an inverted magnetic field creates negative images in MOKE measurements, making this potential feature very valuable.

Considering the standard deviations shown in Table II, none of the optimizations show very large improvements relative to the three-class initial benchmark. Most resulted in worse performance. Our default model's activation function (ReLU) and label smoothing (20%) performed best. Data augmentation improved performance by an MCC increase of 0.002. Data augmentation with image inversion most likely results in the models having to solve the much more sophisticated task of perfectly recognizing the shape of skyrmions and defects, hence the reduction of performance. Furthermore, increasing the dropout rate to 10% seemed to increase the MCC by 0.004. Taking into account these slight improvements, the master network was created.

Finally, the master model was trained for 15 epochs and tasked with predicting the test set. The prediction results for all datasets can be seen in Fig. 5. The validation and

training set performances can also be seen in this figure. The master models show a nearly identical performance to the benchmark models when predicting the test set. The generated masks are all of reasonable quality.

The MCC metric is only a pixel-wise quality measure for the prediction. Since we are interested in whole objects, we compared the true size of the skyrmions in the test set to the predicted size of the skyrmions. This way, the investigation is semantic rather than at the pixel-level. A histogram comparing the true size of the skyrmions to the master model's predictions of the sizes is shown in Fig. 6. As one can see, the prediction's distribution matches the true size distribution. However, a second smaller peak at around 150 pixels can be seen. This being said, because of the difficulty of estimating the real size of skyrmions when labeling the data and as a result of the weak re-labeling of the three-class labels, the size of skyrmions in the labels should not be confused with the true size of the skyrmions. In other words, the labels themselves do not perfectly represent the magnetization profile. The Supplemental Material provides additional performance analysis of our models [33].

#### IV. CONCLUSION

In summary, it has been shown that our optimized CNN can produce machine-learning models that create accurate skyrmion detection masks from magnetic Kerr microscopy images. Binary classifiers had difficulties in differentiating between defects and skyrmions and did not perform very



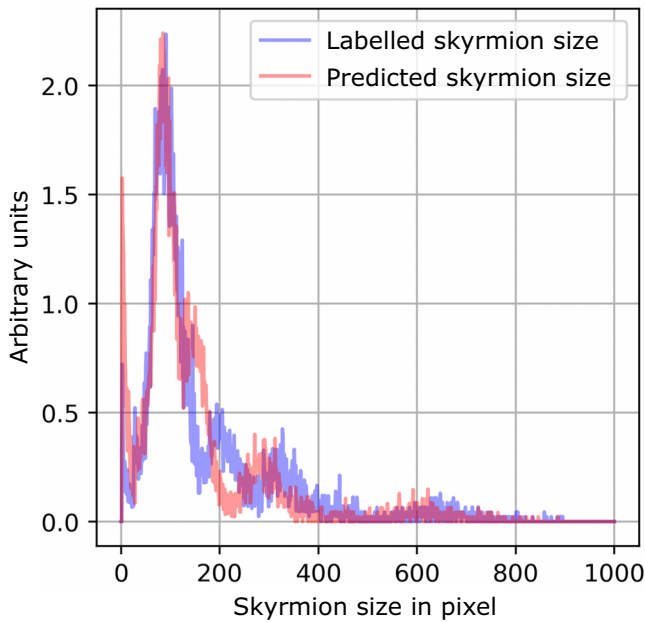


FIG. 6. Predictions on the size of the skyrmions in the test set, normalized per image to account for skyrmion densities varying strongly throughout the set images. This figure shows the skyrmion size distributions of the test set skyrmion labels and the master model predictions on this test set. The total numbers of labeled (45 707) and predicted (45 779) skyrmion objects coincide well.

well on noisy images. We found that a reason for this is that the two-class model relied too much on brightness and/or contrast when making predictions. Adding defects as a third class drastically improved the quality of the prediction. The average MCC significantly improved and the predictions were very close to ground truth. Datasets and networks have been uploaded for open access use [34].

The fine-tuning of the network to create a master CNN showed no strongly improved performance when tested against the test set. This being stated, none of the individual optimizations showed clear individual impact

for an improvement. The reason for this is that most of our three-class models can easily perform this task and that it may be extremely difficult to achieve an MCC greater than 0.90 given that most predictions had almost flawlessly passed a visual inspection test before the fine-tuning. This could be due to the training dataset being very strong or because the test set lacked more difficult samples. For this study, the dataset remained limited to data obtained from previous experiments. To challenge the network's predictive power beyond the test set, we tested a model trained with image inversion data augmentation on an inverted image, which also performed reasonably well, see Fig. 7. Still, the master model in general outperformed the models trained with the inversion augmentation. Maybe a less accurate but more modular model able to handle inverted images is in fact more desirable for daily use. Additionally, when magnetic contrast switches within a measurement, the ability to detect both configurations with only one inference may be convenient.

To conclude, we have developed a CNN to detect and characterize skyrmions in magnetic microscopy images. As preprocessing noisy Kerr microscopy skyrmion images for experimental physics is complex and time-consuming, automatization is important to increase the efficiency of research. This paper demonstrates that a CNN can be trained to create very accurate skyrmion masks given the appropriate training data. A possible future avenue for this research would be to see if a neural network can be trained to automate the preprocessing of other magnetic structures such as skyrmioniums [35], stripe domains [35], or even domain walls [35]. Our approach is also applicable to many magnetic microscopy techniques beyond Kerr microscopy, such as magnetic force microscopy, Lorentz microscopy, or scanning electron microscopy with polarization analysis, making our results broadly applicable. Furthermore, one could see if more powerful models can be created if trained with videos instead of images (with long short-term memory or echo-state networks) [36,37].

We provide datasets and trained models for open access use on Zenodo, see Ref. [34].

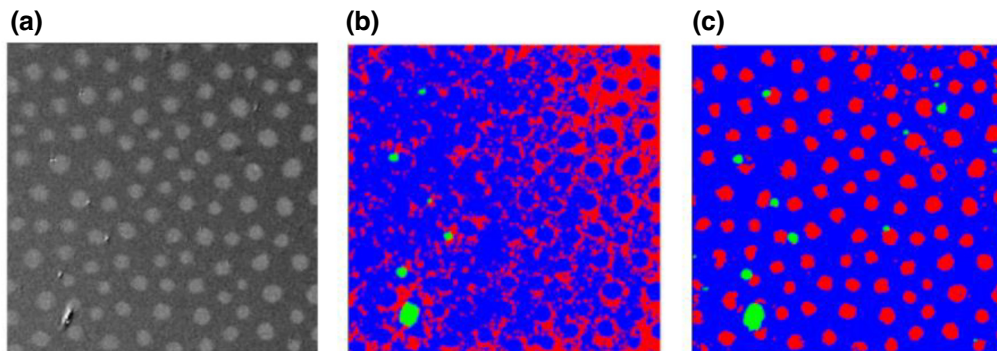


FIG. 7. Inverted data predictions from the master model and model trained on inversion data augmentations. (a) The inverted sample. (b) The master model prediction. (c) The inversion model predictions.

## ACKNOWLEDGMENTS

I.L.-B. gratefully acknowledges the German Academic Exchange Service (DAAD) and Mitacs for the scholarship that allowed him to complete his internship held at the Institut für Physik at the Johannes Gutenberg University of Mainz. He also thanks T.B.W. for providing him with technical assistance and supervising him throughout his internship. We thank Karin Everschor-Sitte for the fruitful discussions. D.F. was funded by DFG Project No. 233630050 (TRR 146). The work was further funded by the emergentAI center, funded itself by the Carl Zeiss Stiftung, further by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projects No. 403502522 (SPP 2137 Skyrmionics), No. 49741853, and No. 268565370 (SFB TRR173 Projects No. A01 and No. B02). The work is also supported by the Horizon 2020 Framework Program of the European Commission under FET-Open Grant Agreement No. 856538 (ERC-2019-SyG; 3D MAGiC) and the Horizon Europe Project No. 101070290 (NIMFEIA), which we acknowledge. We also want to acknowledge the colleagues that contributed to the datasets by fabricating the materials and devices, and carrying out the magnetic microscopy measurements.

- 
- [1] K. Everschor-Sitte, J. Masell, R. M. Reeve, and M. Kläui, Perspective: Magnetic skyrmions—Overview of recent progress in an active research field, *J. Appl. Phys.* **124**, 240901 (2018).
- [2] A. Fert, V. Cros, and J. Sampaio, Skyrmions on the track, *Nat. Nanotechnol.* **8**, 152 (2013).
- [3] J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, Neuromorphic spintronics, *Nat. Electron.* **3**, 360 (2020).
- [4] S. Woo, *et al.*, Observation of room-temperature magnetic skyrmions and their current-driven dynamics in ultrathin metallic ferromagnets, *Nat. Mater.* **15**, 501 (2016).
- [5] Wanjun Jiang, Pramey Upadhyaya, Wei Zhang, Guoqiang Yu, M. Benjamin Jungfleisch, Frank Y. Fradin, John E. Pearson, Yaroslav Tserkovnyak, Kang L. Wang, Olle Heinonen, Suzanne G. E. te Velthuis, and Axel Hoffmann, Blowing magnetic skyrmion bubbles, *Science* **349**, 283 (2015).
- [6] S. Meyer, M. Perini, S. von Malottki, A. Kubetzka, R. Wiesendanger, K. von Bergmann, and S. Heinze, Isolated zero field sub-10 nm skyrmions in ultrathin Co films, *Nat. Commun.* **10**, 3823 (2019).
- [7] J. Závorka, F. Dittrich, Y. Ge, N. Kerber, K. Raab, T. Winkler, K. Litzius, M. Veis, P. Virnau, and M. Kläui, Skyrmion lattice phases in thin film multilayer, *Adv. Funct. Mater.* **30**, 2004037 (2020).
- [8] D. Pinna, F. A. Araujo, J.-V. Kim, V. Cros, D. Querlioz, P. Bessiere, J. Droulez, and J. Grollier, Skyrmion Gas Manipulation for Probabilistic Computing, *ArXiv170107750 Cond-Mat Physicsphysics* (2017).
- [9] Jakub Závorka, Florian Jakobs, Daniel Heinze, Niklas Keil, Sascha Kromin, Samridh Jaiswal, Kai Litzius, Gerhard Jakob, Peter Virnau, Daniele Pinna, Karin Everschor-Sitte, Levente Rózsa, Andreas Donges, Ulrich Nowak, and Mathias Kläui, Thermal skyrmion diffusion used in a reshuffler device, *Nat. Nanotechnol.* **14**, 658 (2019).
- [10] M. A. Brems, M. Kläui, and P. Virnau, Circuits and excitations to enable Brownian token-based computing with skyrmions, *Appl. Phys. Lett.* **119**, 132405 (2021).
- [11] T. Nozaki, Y. Jibiki, M. Goto, E. Tamura, T. Nozaki, H. Kubota, A. Fukushima, S. Yuasa, and Y. Suzuki, Brownian motion of skyrmion bubbles and its control by voltage applications, *Appl. Phys. Lett.* **114**, 012402 (2019).
- [12] D. Prychynenko, M. Sitte, K. Litzius, B. Krüger, G. Bourianoff, M. Kläui, J. Sinova, and K. Everschor-Sitte, Magnetic skyrmion as a nonlinear resistive element: A potential building block for reservoir computing, *Phys. Rev. Appl.* **9**, 014034 (2018).
- [13] K. Raab, M. A. Brems, G. Beneke, T. Dohi, J. Rothörl, F. Kammerbauer, J. H. Mentink, and M. Kläui, Brownian reservoir computing realized using geometrically confined skyrmion dynamics, *Nat. Commun.* **13**, 1 (2022).
- [14] T. Yokouchi, Pattern Recognition with Neuromorphic Computing Using Magnetic-Field Induced Dynamics of Skyrmions, *Sci. Adv.* **8**, eabq5652 (2022).
- [15] N. Kerber, M. Weißenhofer, K. Raab, K. Litzius, J. Závorka, U. Nowak, and M. Kläui, Anisotropic skyrmion diffusion controlled by magnetic-field-induced symmetry breaking, *Phys. Rev. Appl.* **15**, 044029 (2021).
- [16] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, *Nat. Methods* **9**, 671 (2012).
- [17] J. Ghorpade, GPGPU processing in CUDA architecture, *Adv. Comput. Int. J.* **3**, 105 (2012).
- [18] G. Katsikas, C. Sarafidis, and J. Kioseoglou, Machine learning in magnetic materials, *Phys. Status Solidi B* **258**, 2000600 (2021).
- [19] A. Kovacs, J. Fischbacher, H. Oezelt, M. Gusenbauer, L. Exl, F. Bruckner, D. Suess, and T. Schrefl, Learning magnetization dynamics, *J. Magn. Magn. Mater.* **491**, 165548 (2019).
- [20] H. Y. Kwon, H. G. Yoon, C. Lee, G. Chen, K. Liu, A. K. Schmid, Y. Z. Wu, J. W. Choi, and C. Won, Magnetic Hamiltonian parameter estimation using deep learning techniques, *Sci. Adv.* **6**, eabb0872 (2020).
- [21] S. Gerber, L. Pospisil, M. Navandar, and I. Horenko, Low-cost scalable discretization, prediction, and feature selection for complex systems, *Sci. Adv.* **6**, eaaw0961 (2020).
- [22] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015).
- [23] D. Chicco and G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics* **21**, 6 (2020).
- [24] TensorFlow Developers, TensorFlow (2022).
- [25] F. Chollet, *et al.*, Keras. (2015).
- [26] C. R. Harris, *et al.*, Array programming with NumPy, *Nature* **585**, 357 (2020).

- [27] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, and Gaël Varoquaux, *API Design for Machine Learning Software: Experiences from the Scikit-Learn Project*. (2013).
- [28] Wiredfool, *Python-Pillow/Pillow: 4.0.0*. (2017).
- [29] J. Kukačka, V. Golkov, and D. Cremers, *Regularization for Deep Learning: A Taxonomy*. (2017).
- [30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, 1929 (2014).
- [31] C. Shorten and T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data* **6**, 60 (2019).
- [32] R. Müller, S. Kornblith, and G. Hinton, When Does Label Smoothing Help? (2019).
- [33] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevApplied.21.014014> for additional performance analysis.
- [34] T. B. Winkler, I. Labrie-Boulay, A. Romanova, H. Fangohr, M. Kläui, R. Gruber, F. Kammerbauer, K. Raab, and J. Zazvorka, MOKE-microscopy-Skyrmion-dataset (1.0) [Data set], Zenodo (2023), <https://doi.org/10.5281/zenodo.7636110>.
- [35] B. Göbel, I. Mertig, and O. A. Tretiakov, Beyond skyrmions: Review and perspectives of alternative magnetic quasiparticles, *Phys. Rep.* **895**, 1 (2021).
- [36] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.* **9**, 1735 (1997).
- [37] H. Jaeger, Echo state network, *Scholarpedia* **2**, 2330 (2007).