Learning Channel Importance for High Content Imaging with Interpretable Deep Input Channel Mixing

Daniel Siegismund, Mario Wieser, Stephan Heyse, and Stephan Steigele

Genedata AG, Basel, Switzerland

Abstract. Uncovering novel drug candidates for treating complex diseases remain one of the most challenging tasks in early discovery research. To tackle this challenge, biopharma research established a standardized high content imaging protocol that tags different cellular compartments per image channel. In order to judge the experimental outcome, the scientist requires knowledge about the channel importance with respect to a certain phenotype for decoding the underlying biology. In contrast to traditional image analysis approaches, such experiments are nowadays preferably analyzed by deep learning based approaches which, however, lack crucial information about the channel importance. To overcome this limitation, we present a novel approach which utilizes multi-spectral information of high content images to interpret a certain aspect of cellular biology. To this end, we base our method on image blending concepts with alpha compositing for an arbitrary number of channels. More specifically, we introduce DCMIX, a lightweight, scaleable and end-to-end trainable mixing layer which enables interpretable predictions in high content imaging while retaining the benefits of deep learning based methods. We employ an extensive set of experiments on both MNIST and RXRX1 datasets, demonstrating that DCMIX learns the biologically relevant channel importance without scarifying prediction performance.

Keywords: Biomedical Imaging · Interpretable Machine Learning · Explainable AI · Image Channel Importance.

1 Introduction

High-Content Imaging (HCI) has developed to one of the main driving factors in biopharma early discovery research to reveal novel drug candidates for sophisticated treatment strategies such as cancer immunotherapies [21]. HCI is based on a standardized experimental protocol that allow for the systematic acquisition of multi-spectral images, e.g., in form of a cell painting assay protocol that requires a high number of channels with the benefit of a highly generalizable assay [3]. Here, high-content images are recorded by automated instruments on microtiter plates which allow for large-scale drug candidate testing and an automatic analysis procedure to assess the mechanics of a drug candidate for a certain disease. When running such HCI experiments, scientists prepare typically a set of 4 to 15

channels [23,31] with a specific fluorophore that tags a certain cellular protein or compartment. Subsequently, the scientist aims to analyze the experimental outcome with respect to the importance of the fluorescence channels to validate the findings or refine the experiment and, therefore, requires a fast and easy-to-use analysis workflow. This is particularly important as the specific functional or mechanistic knowledge is encoded via the specific staining per image channel [3] and hence required for decoding the underlying biology.

However, to analyze such complex multi-channel cell-painting assays, the scientist requires the ability of sophisticated image analysis to distill the information from the multi-spectral information. In biopharma research, the traditional analysis [5] is gradually replaced by deep learning based approaches [40,14,48,39]. Despite the superior performance of such models in comparison to conventional segmentation based analysis [5], the scientist lacks informative insights in terms of understanding about which fluorescence channel influenced the decision [7].

In the past, various approaches have been proposed to extract the most relevant information from high-dimensional datasets. The most basic approach to determine the most relevant channels is a preprocessing step by applying an unsupervised dimensionality reduction method such as Principal Component Analysis (PCA) [17]. However, employing such a preprocessing step does not guarantee for phenotype-specific channels as the method only optimizes for the directions with the highest variance and not necessarily for the highest phenotypic information. More recently, attention-based approaches have been introduced for image channel selection [4,15,24,32] which suffer from high computational costs and poor scalability. In addition, there are model-agnostic approaches such as Shapely values [36,19] which, however, can suffer from sampling variability [27] and be time consuming in terms of highly complex models [6].

To overcome the aforementioned limitations, we present a simple yet effective method to estimate channel importance for HCI images. More specifically, we introduce a lightweight, easy to use mixing layer that is composed of a generalized image blending mechanism with alpha compositing [50,1] which converts a d-dimensional channel image into a 2D image retaining all phenotype relevant information. This allows not only to incorporate an arbitrary number of channels in a highly scalable fashion but also leads to a reduced network size with faster inference times while being able to facilitate the use of transfer learning of pretrained networks. To summarize, we make the following contributions:

- We extend the imaging blending concepts of [50,1] and apply these to images with an arbitrary number of channels.
- We encapsulate the generalized image blending into a lightweight, scalable and end-to-end trainable mixing layer, called DCMIX, to estimate channel importance for multi-spectral HCI data.
- Experiments on MNIST as well as on the challenging multi-channel real-world imaging data set RXRX1 [42] with 31 different cell phenotype classes demonstrate that the proposed method learns the correct channel importance without sacrificing its model performance.

2   Related Work

In this section, we review related work on interpretable and explainable machine learning [30]. Broadly spoken, we can distinguish between interpretable models that are interpretable by design and explainable models that try to explain existing models post-hoc [30].

Interpretable Machine Learning Methods can be separated into the following model classes: score-based [44], rule-based [9], sparse [43] and neural networks [12], among others [30]. In this review, we focus more closely on sparsity inducing and attention-based interpretable methods. Sparsity-based approaches introduce a sparsity constraint on the model coefficients to determine the feature importance. One of the most basic approaches is the least absolute shrinkage and selection operator (LASSO) introduced by [43] which is employing the L_1-norm to ensure feature sparsity. This approach has subsequently been extended to various lines of research, including dealing with grouped features [49,33], estimating network graphs [13,34] or learning sparse representations in neural networks [26,47]. Most closely related to our work is LassoNet [22] which employs a group lasso constraint based on the feature channels that are obtained from a pretrained feature extraction network. In contrast, our approach is end-to-end trainable and hence does not require a two step approach of feature extraction and importance estimation. More recently, attention-based approaches [45] have emerged in the context of interpretable machine learning. [8] introduced an attention-based model for the analysis of electronic health records and [37] learns important features with an attentive mixture of experts approach. Moreover, attention is used in the context of hyper spherical band/channel selection [4,15,24,32]. In contrast, our approach works on image blending and alpha compositing and hence reducing high computational costs.

Explainable Machine Learning Methods denote approaches that aim to explain decisions of an already trained machine learning model post-hoc by learning a surrogate model [30]. In summary, we distinguish between attribution methods that try to quantify the attribution of a feature to the prediction [41], concept-based explanations trying to explain predictions with high-level concepts [20], symbolic metamodels employing symbolic regression as a surrogate [2] and counterfactual explanations [46]. In the context of our work, we focus on attribution models. [35] learns a surrogate classifier to explain an arbitrary black-box model based on submodular optimization. [38] introduced DeepLIFT to decompose the input contributions on the model prediction. In addition, Shapley values gained a wide adoption in the machine learning domain mainly for feature selection and model explain ability [36,19]. As a result, Lundberg & Lee [28] introduced Shapely additive explanations (SHAP) to explain model predictions based on Shapely regression values. Finally, Shapley values have been used in the context of HCI channel importance estimation [42]. More specifically, the authors adopt Shapely values to explain the channel importance of HCI images from a

pretrained black-box model. Opposed to our approach, this method requires the training of two separate models and hence does not allow for end-to-end training.
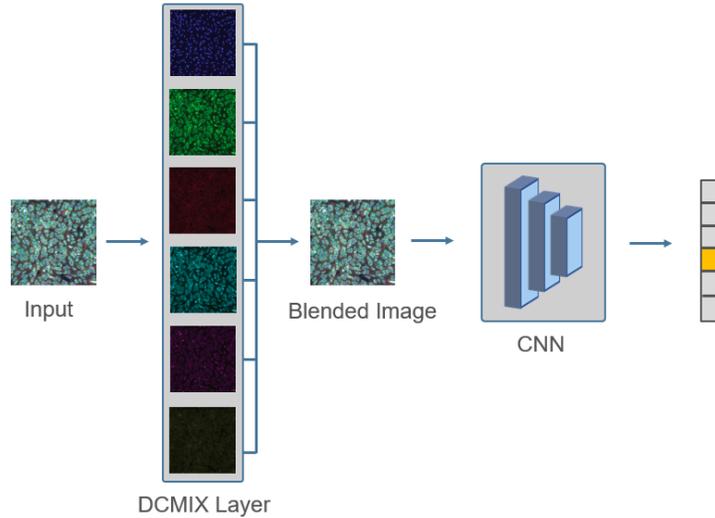

Fig. 1. Blue arrows denote steps and gray boxes actions in our workflow, respectively. In the first step (1.), we take a multi-channel cellular image and split it into single channels. Subsequently, we mix the channel within our DCMIX layer to obtain the most important part of each channel. In the second step (2.), we take the blended image into our classification network.

3   Model

As illustrated in Figure 1, we utilize a two step approach for estimating channel importance in multi-spectral bioimage classification settings by introducing a lightweight, easy to use and end-to-end trainable mixing layer. To do so, we propose a blending layer which combines the most important parts of the distinct channels into a new 2D image. After, we perform a classification based on the blended image.

3.1   Conception of the Image Blending Layer

We start with an input image $I \in \mathbb{R}^{h \times w \times c}$ where h denotes the height, w the width and c the number of channels in the multi-spectral image. Subsequently, the image I is split into its distinct channels and processed in the DCMIX layer. The DCMIX layer is inspired by simple image blending and alpha compositing

[50,1]. More specifically, the idea behind Alpha blending is to combine two images as follows:

$$C = \alpha_1 \cdot A_1 + (1 - \alpha_1) \cdot A_2, \tag{1}$$

where $A_1 \in \mathbb{R}^{h \times w \times c}$ and $A_2 \in \mathbb{R}^{h \times w \times c}$ are the corresponding image matrices to blend and $C \in \mathbb{R}^{h \times w \times c}$ the blended image matrix. The trainable parameter α_1 determines the transparency of each channel.

In this work, we take advantage of the ideas proposed in [50,1] and generalize the idea by employing the trainable alpha values as weights for each channel that has to be blended with:

$$C = \sum_{i}^{n} \alpha_i \cdot A_i \text{ where: } \alpha_i \geq 0, \tag{2}$$

where α_i is multiplied with each channel A_i. The parameter n defines the number of channels and C is the blended image which will be subsequently used for the further analysis.

3.2   Classifying Genetic Perturbations based on DCMIX-blended Images

Our goal is to learn a classification model $F_\theta(y \mid C)$ of the blended image C for distinct classes of genetic perturbations y^c where c is the number of genetic perturbations to be predicted. In this work, our model F is a Deep Convolutional Neural Network which extracts a cascade of feature maps M^l where l denotes the current layer. The last feature map is used as an input to the multi-class classification head that predicts the genetic perturbation vector y^c using a softmax output.

3.3   End-to-End Training Algorithm

The model training is described in Algorithm 1. As an input, we use the multi-spectral images X and the genetic perturbation labels y. Subsequently, we draw minibatches from the training data X, y (line 1). For each of the minibatches, we obtain the blended images c_i as well as the corresponding mixing factors α_i. The blended images c_i are fed in the neural network F_θ (line 3) and the corresponding predictions $\hat{y}_i$ are used to calculate the loss in line 5. Finally, we update the parameters θ and α based on the loss by using gradient descent (line 7).

4   Experiments

A description of setups and additional hyperparameters can be found in the supplementary materials.

Algorithm 1 DCMIX training algorithm

INPUT: X images, y labels
OUTPUT: The prediction $\hat{y}$, mixing factors α

1: **for** minibatch x_i, y_i from X, y **do**
2:     $c_i, \alpha_i \leftarrow \text{DCMIX}(x_i)$
3:     $\hat{y_i} \leftarrow F_\theta(c_i)$
4:
5:     loss $\leftarrow$ crossentropy$(\hat{y_i}, y_i)$
6:
7:     update θ, α using gradient descent
8: **end for**

4.1   MNIST

Dataset. To demonstrate the efficacy of DCMIX for estimating channel importance, we generate an artificial dataset based on MNIST [10]. MNIST consists of 70000 samples with images $x \in \mathbb{R}^{28x28x1}$ and labels y that represent numbers from 0 to 9. For our dataset, we randomly select a subset of 10000 samples from MNIST. In order to assess the channel importance, we extend the MNIST images with two additional noise channels. Therefore, we draw two noise matrices with shape $28x28$ from a uniform distribution defined on $[0, 255]$. Subsequently, we add the previously generated noise channels to the input image such that we obtain a three channel input image $x \in \mathbb{R}^{28x28x3}$ where the first denotes the most important channel. For training, we split the data into a 70 percent training and a 30 percent hold-out set. The training set is further split into a 80 percent training and 20 percent validation set, respectively.

Models. In order to demonstrate the effectiveness of our approach, we benchmark DCMIX against a plain LCNet050, LassoNet [22] as well as on an attention-based [29,25] LCNet050.

Quantitative Evaluation. Channel Importance In this experiment, we evaluate the channel importance on the validation set, and the results are reported in Table 1. As we can observe in the channel importance ranking, DCMIX can effectively learn the most important channel one and is in line with the more complex LassoNet and attention-based LCNet050. At the same time, DCMIX requires only a fraction of GFLOPS and model parameters. More specifically, DCMIX requires solely 5.9271 GFLOPS compared to 17.809 GFLOPS for the Attention-LCNet050. In addition, DCMIX need three times less parameters (0.2789 million) in contrast to Attention-LCNet050 (0.9281 million) and requires only the same amount of GFLOPS and parameters as the plain LCNet050.

Quantitative Evaluation. Model Performance Despite the fact, that the aim of this method is not to improve the model performance but rather learn

Table 1. Results of the MNIST channel importance and model size. Channel importance ranking denotes the rank of the weights depicted in the second column. The model size is evaluated on GFLOPS and the number of model parameters where lower is better.

Method	Channel importance ranking	Channel weights	GFLOPS	# Parameters (million)
LCNet050	-	-	5.9269	0.2789
LassoNet [22]	1,3,2	120259, 51003, 52318	-	-
Attention[29,25]-LCNet050	1,3,2	$1, 3.24 \times 10^{-11}, 2.33 \times 10^{-6}$	17.809	0.9281
DCMIX-LCNet050	1,3,2	0.82,0.21,0.22	5.9271	0.2789

the most important channel to gain biological insights for a drug discovery experiment, we want to ensure that DCMIX archives competitive performance to state-of-the art approaches. To do so, we compared DCMIX to a plain LCNet050, LassoNet and Attention-LCNet050 in Table 2. Here, we observe that DCMIX obtains competitive results compared to both LCNet050 and Attention-LCNet050 and outperforms LassoNet on accuracy, precision, recall and f1-score measures.

Table 2. Results of model performance for the MNIST dataset on the hold-out dataset. We assess the model performance on four different metrics: accuracy, precision, recall and f1-score where higher is better. Values in brackets denote the standard deviation.

Method	Accuracy	Precision	Recall	F1-Score
LCNet050	0.992 (0.0008)	0.991 (0.002)	0.991 (0.002)	0.991 (0.002)
LassoNet [22]	0.963 (0.012)	0.888 (0.002)	0.888 (0.002)	0.887 (0.002)
Attention[29,25]-LCNet050	0.992 (0.002)	0.991 (0.001)	0.991 (0.001)	0.991 (0.001)
DCMIX-LCNet050	0.991 (0.002)	0.990 (0.002)	0.990 (0.002)	0.990 (0.002)

4.2   RXRX1

Dataset. For our real world experiment, we employ the RXRX1 dataset[42] which consists of 125510 512x512 px fluorescence microscopy images (6 channels) of four different human cell lines that are perturbed with 1138 genetic perturbations (including 30 different positive control perturbations). In this study, we used as the training data 30 positive control siRNAs plus the non-active control which lead to 31 classes in total. All images were normalized using the 1 and 99 percent percentile and after, we extract image patches with a size of 192x192 px and an offset of 96 px. This step leads to 32776 image patches. For training,

we split the data into a 70 percent training and a 30 percent hold-out set. The training set is further split into a 80 percent training and 20 percent validation set, respectively.

Models For the real-world RXRX1 experiment, we compare DCMIX to LassoNet [22] and the attention-based [29,25] LCNet050.

Table 3. RXRX1 channel importance evaluation for the HepG2 cell line. The importance ranking illustrates the most important channels form left to right based on the weights depicted in the second column. In addition, model statistics are measured in GFLOPS and the number of model parameters (lower is better).

Method	Importance ranking	Channel weights (in Channel order)	GFLOPS	# parameters (millions)
ViT-B16-Imagenet21k [11] + LassoNet [22]	6,4,1,5,2,3	73084, 52526, 31138, 87881, 55612, 107733	-	-
Attention-LCNet050	4,2,5,1,3,6	0.15, 0.17, 0.008, 0.48, 0.16, 0.007	35.61	1.75
DCMIX-LCNet050	4,2,3,5,1,6	0.30, 0.69, 0.38, 1.06, 0.36, 0.21	5.95	0.27

Quantitative Evaluation. Channel Importance Here, we describe the evaluation results on channel importance for the RXRX1 dataset which is illustrated in Table 3. To do so, we compare the results to the ground truth introduced in [42]. The experiment was manually designed by a scientist in the laboratory such that both channels four and two hold the most important biological information and channel 6 contains no important information for the phenotype. Keeping this information in mind, we assess the channel importance of DCMIX, LassoNet and Attention-LCNet050. Here, we can confirm that DCMIX learns the two most important channels four and two and the least important channel 6. These findings are also supported by Attention-LCNet050 which learned equivalent importance values. In contrast, LassoNet fails to uncover the correct channel importance by selecting the least important channel as the most important one. Despite finding the same important channels, DCMIX possess a 6-8 times higher speed and requires 6 times less parameters compared to the attention based networks and can be used in an end-to-end fashion which is not feasible for LassoNet.

Quantitative Evaluation. Model Performance In this experiment, we evaluate the model performance of DCMIX to LassoNet and Attention-LCNet050 and illustrate the results in Table 4. Here, we observe that DCMIX outperforms both LassoNet and Attention-LCNet050 in terms of accuracy by five and seven

percent, respectively. Furthermore, these finding are confirmed by precision, recall and f1-scores where DCMIX outperforms both competitors by approximately five and seven percent.

Table 4. Results of model performance for the RXRX1 dataset on the hold-out dataset. We asses the model performance on four different metrics: accuracy, precision, recall and f1-score where higher is better. Values in brackets denote the standard deviation.

Method	Accuracy	Precision	Recall	F1-Score
ViT-B16-Imagenet21k [11] + LassoNet [22]	0.695 (0.004)	0.705 (0.005)	0.705 (0.004)	0.704 (0.005)
Attention[29,25]-LCNet050	0.744 (0.019)	0.753 (0.014)	0.747 (0.014)	0.747 (0.013)
DCMIX-LCNet050	0.765 (0.004)	0.77 (0.037)	0.77 (0.042)	0.764 (0.043)

5   Discussion

DCMIX demonstrates state-of-the-art channel importance scores in fluorescence cellular imaging DCMIX employs image blending to estimate the importance of each image channel. In Figure 2, we provide an overview of the Spearman rank correlation of the channel importance estimates for all tested methods. The results are comparable for all methods (except of LassoNet) with a Spearman ρ always larger than 0.83. Especially the correlations of DCMIX and Attention-LCNet050 to the ground truth shapley values from [42] are evident with a Spearman ρ of 0.89. Both methods estimate the channel 2 and 4 as most important and channel 6 as least important which was the intentionally experimental design and furthermore shown via shapley values [42]. The authors explained their finding with a very large spectral overlap of the fluorescence signal from channel 2 and 4 to any other channel rendering them more important [42].
In contrast, LassoNet does not show any overlap with the rankings selected by all other methods (Figure 2) with a maximal Spearman ρ value of -0.08.

DCMIX achieves state-of-the-art classification performance with lower model complexity Across all classification metrics DCMIX archives competitive results on MNIST and state-of-the-art performances on real-world RXRX1 compared to its competitors. Intuitively, we attribute the competitive results on MINST to the problem simplicity which is further supported by the high classification scores of 99% (Table 2). Concurrently, DCMIX requires merely a fraction of model parameters in all experiments (Tables 1,3) compared to the baselines.

Practical runtimes for DCMIX are 6-8 times faster than Attention-based approaches While DCMIX requires only 5.9271 GFLOPS and 5.95
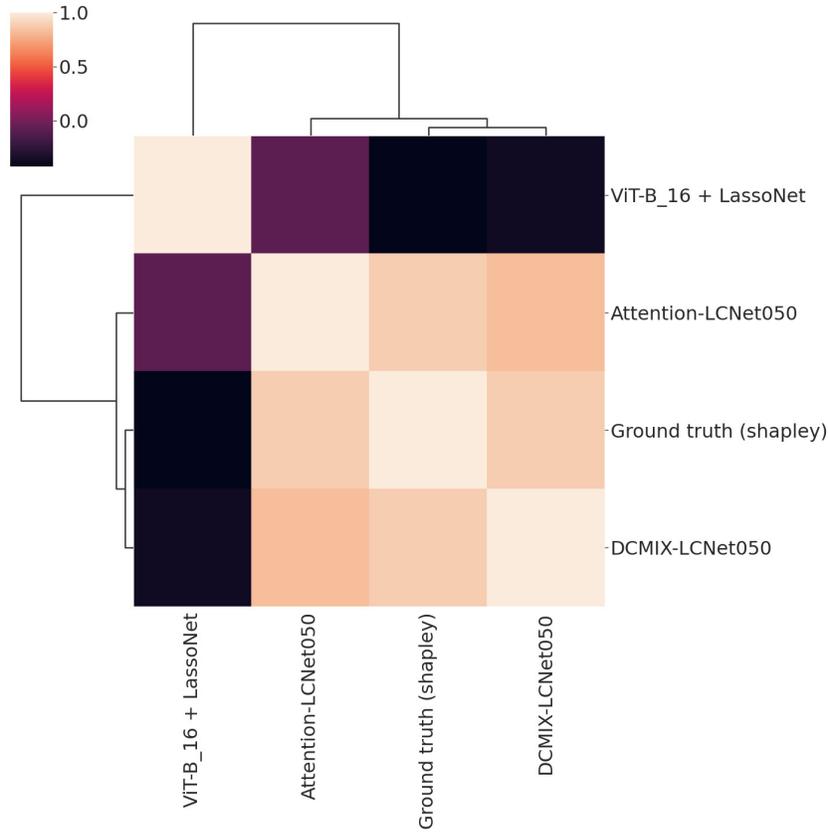
Fig. 2. Visualization of Spearman's rank correlation coefficient of the channel importance estimates for all different methods from Table 3. A value of -1 indicates maximal ranking difference between the channel importance estimates, 1 indicates no difference. Matrix has been sorted using average linkage hierarchical clustering with euclidean distance.

GFLOPS on RXRX1, achieving the same computational performance as plain LCNet050, Attention-LCNet050 needs 17.809 GFLOPS on MNIST and 35.614 GFLOPS on RXRX1, respectively (Table 1 and Table 3). Moreover, even post-hoc approaches such Shapely values that are trained on a black-box model require often more significant computation time. For example, the training time required for the Shapley value explanation are in the range of several minutes for the smaller CIFAR-10 dataset [16]. This demonstrates that the speed of DCMIX outperforms not only interpretable competitors but also explainable post-hoc approaches on a large scale.

DCMIX is applicable in real-world settings beyond biomedical imaging From an application standpoint we see an advantage of DCMIX over the

other tested methods, as the high scalability of DCMIX allows a model training workflow were channel importance is − per default − applied, such that a scientist gets immediate feedback about where the classification relevant information is coming from, and whether it correlates with the known understanding of the underlying biology.

DCMIX scales very well with the number of channels due to the addition of only one additional parameter per additional channel. This is particularly interesting for hyper spectral applications where hundreds of channels exist (e.g. in remote sensing) − a highly interesting application area for subsequent studies.

In addition, DCMIX allows for any arbitrary downstream network which can be fine-tuned / designed for other applications than fluorescence imaging.

DCMIX applies currently a simple addition channel mixing strategy to estimate channel importance without losing any classification performance (see model performance in Table 2 and 4). In principle several other channel blending methods exist, e.g. difference, multiplication or luminosity. Due to the flexibility of DCMIX these other mixing strategies can be easily integrated. Several studies already show the applicability of complex multi-spectral channel blending for visualization and classification in remote sensing [1,18].

6 Conclusion

In this work, we present a novel lightweight framework, DCMIX, which estimates channel importance of fluoresce images based on image blending. This empowers us to estimate phenotype-focused interpretations in a simple yet effective manner. Our experimental results demonstrate that the channel importance scores uncovered by DCMIX are both biologically supported and in line with competitive state-of-the-art approaches on MNIST and RXRX1 datasets. Concurrently, DCMIX is more effective in terms of runtime and scaleable to an arbitrary number of channels without scarifying the model performance.

Limitations. We discuss the limitations of our approach in the following two aspects. (1) The weights of DCMIX which determine the channel importance are solely a proxy and do not explain the absolute importance between channels. (2) DCMIX is based on image blending and hence only supporting image-based datasets. For future work, we plan to investigate how DCMIX can be extended to other data modalities.

References

1. Why Not a Single Image? Combining Visualizations to Facilitate Fieldwork and On-Screen Mapping. Remote Sensing (2019)
2. Alaa, A.M., van der Schaar, M.: Demystifying black-box models with symbolic metamodels. In: Advances in Neural Information Processing Systems (2019)
3. Bray, M.A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C., Carpenter, A.E.: Cell Painting,

a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature protocols **11**(9) (2016)

4. Cai, Y., Liu, X., Cai, Z.: BS-Nets: An End-to-End Framework for Band Selection of Hyperspectral Image. IEEE Transactions on Geoscience and Remote Sensing (2020)
5. Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., others: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology (2006)
6. Carrillo, A., Cantú, L.F., Noriega, A.: Individual explanations in machine learning models: A survey for practitioners. CoRR **abs/2104.04144** (2021)
7. Castelvecchi, D.: Can we open the black box of AI? Nature News (2016)
8. Choi, E., Bahadori, M.T., Kulas, J.A., Schuetz, A., Stewart, W.F., Sun, J.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in Neural Information Processing Systems (2016)
9. Cohen, W.W.: Fast effective rule induction. In: International Conference on Machine Learning (1995)
10. Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine (2012)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., others: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2020)
12. Feng, J., Simon, N.: Sparse-input neural networks for high-dimensional nonparametric regression and classification (2019)
13. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics (2008)
14. Godinez, W.J., Hossain, I., Lazic, S.E., Davies, J.W., Zhang, X.: A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics (Oxford, England) (2017)
15. He, K., Sun, W., Yang, G., Meng, X., Ren, K., Peng, J., Du, Q.: A Dual Global–Local Attention Network for Hyperspectral Band Selection. IEEE Transactions on Geoscience and Remote Sensing (2022)
16. Jethani, N., Sudarshan, M., Covert, I., Lee, S.I., Ranganath, R.: FastSHAP: Real-Time Shapley Value Estimation. In: International Conference on Learning Representations (2022)
17. Jolliffe, I.: Principal component analysis. Springer Verlag (1986)
18. Jordanova, G., Verbovšek, T.: Improved Automatic Classification of Litho-Geomorphological Units by Using Raster Image Blending, Vipava Valley (SW Slovenia). Remote Sensing (2023)
19. Jullum, M., Redelmeier, A., Aas, K.: groupShapley: Efficient prediction explanation with Shapley values for feature groups (2021), arXiv:2106.12228
20. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning (2018)
21. Kruger, S., Ilmer, M., Kobold, S., Loureiro Cadilha, B., Endres, S., Ormanns, S., Schuebbe, G., Renz, B., D'Haese, J., Schlößer, H., Heinemann, V., Subklewe, M., Boeck, S., Werner, J., von Bergwelt, M.: Advances in cancer immunotherapy 2019 - latest trends. Journal of Experimental and Clinical Cancer Research (2019)
22. Lemhadri, I., Ruan, F., Abraham, L., Tibshirani, R.: Lassonet: A neural network with feature sparsity. Journal of Machine Learning Research (2021)

23. Levenson, R.M., Mansfield, J.R.: Multispectral imaging in biology and medicine: Slices of life. Cytometry Part A (2006)

24. Li, W., Chen, H., Liu, Q., Liu, H., Wang, Y., Gui, G.: Attention Mechanism and Depthwise Separable Convolution Aided 3DCNN for Hyperspectral Remote Sensing Image Classification. Remote Sensing (2022)

25. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A Structured Self-attentive Sentence Embedding. In: International Conference on Learning Representations (2017)

26. Louizos, C., Welling, M., Kingma, D.P.: Learning sparse neural networks through l_0 regularization. In: International Conference on Learning Representations (2018)

27. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence (2020)

28. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions

29. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Conference on Empirical Methods in Natural Language Processing (2015)

30. Marcinkevičs, R., Vogt, J.E.: Interpretable and explainable machine learning: A methods-centric overview with concrete examples. WIREs Data Mining and Knowledge Discovery (2023)

31. Nalepa, J.: Recent Advances in Multi- and Hyperspectral Image Analysis. Sensors (2021)

32. Nikzad, M., Gao, Y., Zhou, J.: An Attention-Based Lattice Network for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing (2022)

33. Park, T., Casella, G.: The Bayesian Lasso. Journal of the American Statistical Association (2008)

34. Prabhakaran, S., Metzner, K.J., Böhm, A., Roth, V.: Recovering networks from distance data. In: Asian Conference on Machine Learning (2012)

35. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)

36. Rozemberczki, B., Watson, L., Bayer, P., Yang, H.T., Kiss, O., Nilsson, S., Sarkar, R.: The Shapley Value in Machine Learning. In: International Joint Conference on Artificial Intelligence (2022)

37. Schwab, P., Miladinovic, D., Karlen, W.: Granger-causal Attentive Mixtures of Experts: Learning Important Features With Neural Networks. In: AAAI Conference on Artificial Intelligence (2019)

38. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning (2017)

39. Siegismund, D., Wieser, M., Heyse, S., Steigele, S.: Self-supervised representation learning for high-content screening. In: International Conference on Medical Imaging with Deep Learning (2022)

40. Steigele, S., Siegismund, D., Fassler, M., Kustec, M., Kappler, B., Hasaka, T., Yee, A., Brodte, A., Heyse, S.: Deep Learning-Based HCS Image Analysis for the Enterprise. SLAS DISCOVERY: Advancing the Science of Drug Discovery (2020)

41. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (2017)

42. Sypetkowski, M., Rezanejad, M., Saberian, S., Kraus, O., Urbanik, J., Taylor, J., Mabey, B., Victors, M., Yosinski, J., Sereshkeh, A.R., Haque, I., Earnshaw, B.: RxRx1: A Dataset for Evaluating Experimental Batch Correction Methods. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2023)
43. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B) (1996)
44. Ustun, B., Rudin, C.: Supersparse linear integer models for optimized medical scoring systems. Machine Learning (2015)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
46. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Cybersecurity (2017)
47. Wieczorek, A., Wieser, M., Murezzan, D., Roth, V.: Learning sparse latent representations with the deep copula information bottleneck. In: International Conference on Learning Representations (2018)
48. Wieser, M., Siegismund, D., Heyse, S., Steigele, S.: Vision transformers show improved robustness in high-content image analysis. In: Swiss Conference on Data Science (2022)
49. Yuan, M., Lin, Y.: Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society Series B: Statistical Methodology (2005)
50. Zhang, L., Wen, T., Shi, J.: Deep Image Blending. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2020)