Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks

Sanchari Dhar[a], Lior Shamir[a,*]

[a]*Kansas State University, Manhattan, KS 66506, USA*

Abstract

In the past decade, deep neural networks, and specifically convolutional neural networks (CNNs), have been becoming a primary tool in the field of biomedical image analysis, and are used intensively in other fields such as object or face recognition. CNNs have a clear advantage in their ability to provide superior performance, yet without the requirement to fully understand the image elements that reflect the biomedical problem at hand, and without designing specific algorithms for that task. The availability of easy-to-use libraries and their non-parametric nature make CNN the most common solution to problems that require automatic biomedical image analysis. But while CNNs have many advantages, they also have certain downsides. The features determined by CNNs are complex and unintuitive, and therefore CNNs often work as a "Black Box". Additionally, CNNs learn from any piece of information in the pixel data that can provide discriminative signal, making it more difficult to control what the CNN actually learns. Here we follow common practices to test whether CNNs can classify biomedical image datasets, but instead of using the entire image we use merely parts of the images that do not have biomedical content. The experiments show that CNNs can provide high classification accuracy even when they are trained with datasets that do not contain any biomedical information, or can be systematically biased by irrelevant information in the image data. The presence of such consistent irrelevant data is difficult to identify, and can therefore lead to biased experimental results. Possible solutions to this downside of CNNs can be control experiments, as well as other protective practices to validate the results and avoid biased conclusions based on CNN-generated annotations.

Keywords: Convolutional neural networks, data acquisition bias, deep learning, experimental design.

1. Introduction

Enabled by the increasing availability of digital imaging and large storage devices, the ability to analyze large databases of images has become pivotal in discovery from data in a broad range of fields. In particular, convolutional neural networks (CNNs) are widely popular in biomedical research, and are used for a very large number of applications within the biomedical domain (Litjens et al., 2017; Min et al., 2017; Shen et al., 2017; Cao et al., 2018; Wainberg et al., 2018), and obviously also other domains such as object recognition, face recognition, and more. The use of machine learning has also reinforced the need for performance analysis models (Liu et al., 2017).

In the past decade, the rapid advancement in digital imaging and storage devices have enabled the collection of very large datasets of biomedical images. For instance, microscopes with robotic stages are capable of collecting thousands of microscopy images within a few hours of operation (Abraham et al., 2004; Zanella et al., 2010; Shamir et al., 2010; Singh et al., 2014). Digital radiography has generated a high number of radiographs, allowing to automate the data analysis to make new discover-

*Corresponding author: Tel.: +1-785-532-4809;
Email address: lshamir@mtu.edu (Lior Shamir)

ies or improve healthcare practices through tasks such as automatic image-based diagnostics (Hu et al., 2018; Kermany et al., 2018; Bychkov et al., 2018; Aina et al., 2019; Thomsen et al., 2020).

The availability of large databases of biomedical images has reinforced the need for methodology that can analyze these images and turn them into scientific discoveries or new healthcare practices. Such datasets also allows the AI community to develop AI-based solutions to problems within the biomedical domain. Once the datasets become public, the AI community can use them as benchmarks, develop algorithm that can analyze them, and compare the performance of different algorithms to identify the optimal solutions.

In the past decade, deep learning, and specifically convolutional neural networks have become the most common AI approaches for biomedical image analysis (Anwar et al., 2018; Chen et al., 2019; Zhang et al., 2019). CNNs can be applied to a broad range of image data without the need to tailor specific algorithms, and can achieve superior performance. With the availability of easy-to-use libraries, CNNs have become very common also among researchers who are not necessarily computer scientists. However, while CNNs have substantial advantages, their prevalence also requires studying their disadvantages in the context of biomedical images, and the common practices in the application of CNN to biomedical image datasets. CNNs identify features automatically from the image pixels, leading to non-intuitive features that often act as a "black box". These features are determined automatically by their ability to discriminate between the different classes, and therefore might also be driven by signal that does not necessarily reflect the biomedical problem at hand.

While deep learning is clearly an emerging trend in virtually all aspects of biomedical image analysis, and that trend is bound to continue, it is also important to carefully analyze its weaknesses. Unlike traditional "shallow learning" methods designed to measure specific aspects of the image data, CNNs are designed to learn automatically from the pixel data without the need to design task-specific features. The ability of deep neural networks to identify complex features automatically is a substantial advantage that makes deep learning highly prevalent, and much easier to use by eliminating the need for deep knowledge in image processing. However, these automatically defined data-driven features are determined based on their ability to separate between the different image classes. That, however, reinforces a careful experimental design that might be as critical for testing "shallow learning" algorithms that work by task-specific features. For instance, pixels might have different values based on subtle differences in the lighting conditions at the time of imaging. CNNs might be able to capture these differences in the case they are not normalized between the different classes. Slight changes in the position of the imaging device or subject, differences in the temperature of the CCD when the images are taken, and even different technicians can lead to certain differences that are difficult to sense by eye, but can have strong impact on CNNs. A CNN can make predictions based on any information in the training set, regardless of whether it is of biomedical meaning. Because the features are non-intuitive, it is difficult for the experimentalist to identify situations in which the predictions are driven by background noise or artifacts rather than their biomedical meaning. Such situations can lead to biased results published in scientific papers, while the experimentalist is not aware of the bias and believes they report on accurate findings.

In many cases the visual features leading to the bias are too subtle to notice by eye, leading the experimentalist to believe that the predictions made by the CNN reflect the ability of the CNN to identify differences between the visual content. Consequently, the experimentalist might reach certain conclusions regarding the presence of differences between the image classes, and the ability of the CNN to identify these differences. As a simple example, the ability of a CNN to identify a certain disease automatically by analyzing radiographs can be estimated by the benchmark dataset differently than the ability of the same CNN to identify the same disease in a real-world setting.

Here we study possible dataset bias using several different datasets acquired in a controlled process, which is typical to the acquisition of biomedical image datasets. We show that in many cases the application of the CNNs to the datasets leads to results driven by dataset bias related to the data acquisition process, and can therefore lead to biased results. To avoid biased results, we propose simple control experiments. These experiments can be performed at the time of applying a CNN to annotate the data, and can assist in identifying situations in which the annotation is driven by bias.

2. Related work

Benchmark datasets are compiled for developing machine vision algorithms, and testing and comparing the performance of different algorithms to identify the most effective solution to a given biomedical image analysis problem. These datasets have enabled substantial research, and their use is a common standard practice in machine learning. However, benchmark datasets can also be biased for different reasons. For instance, in the context of benchmarks for object recognition, the perception of the people annotating the samples by their ground truth or selecting the samples for the dataset can lead to bias, especially when the dataset is collected from the web (Torralba and Efros, 2011; Khosla et al., 2012; Tommasi et al., 2017; Kortylewski et al., 2019). That bias can also be shown experimentally by the fact that training an algorithm with one benchmark dataset and testing it with another dataset leads to weaker results compared to training and testing using the same benchmark (Torralba and Efros, 2011). That difference in performance is not expected given that larger training sets are expected to provide stronger or equal performance to smaller training sets, and therefore the weaker performance can be considered evidence of dataset bias (Torralba and Efros, 2011).

In some cases, images that look visually identical can be classified differently by a machine learning algorithm due to certain information in the image that is too subtle to notice by eye, but can have critical impact on the classification process. That is known as *adversarial machine learning* (Huang et al., 2011). Such images can be used to attack machine learning system, and can specifically impact the performance of artificial neural networks (Goodfellow et al., 2014). The effect of adversarial samples can also impact video data (Zhang et al., 2020). The ability to attack neural networks by using data that seem visually indifferent to the human eye demonstrates that artificial neural network can be sensitive to bias originated from the way the neural networks operate.

One of the solutions to the problem of dataset bias is to increase the variability in the datasets. That can be done by using data augmentation (McLaughlin et al., 2015; Jaipuria et al., 2020), combining different datasets (Khosla et al., 2012), or synthetically change the variability of the dataset (Khosla et al., 2012). **Another proposed approach is to weight the different features by the ability of a certain feature set to classify them, and penalize samples that are easy to classify (Li et al., 2019). The weighted dataset can then be used to reduce the bias in the results.**

Benchmark datasets used in the domain of machine learning aim at representing the real world as reliably as possible (Torralba and Efros, 2011), allowing to develop and compare the performance of different algorithms that solve general common problems such as automatic object recognition or face recognition. In the biomedical domain, benchmark datasets are used to develop and compared algorithms for different biomedical image analysis problems. In other cases biomedical datasets are collected for one single experiment, and used by a single research team. While substantial work has been done to analyze bias in datasets that were collected from the web, collecting images from the web is much less frequent in the preparation of biomedical image datasets. Here we focus on biases in datasets collected in controlled environment and well-defined data acquisition processes, as often is the case in image analysis in the biomedical domain.

Assuming no bias in the dataset, the classification accuracy shown by a convolutional neural network can be trusted as an indication of differences between the different classes, and the ability of the algorithm to provide an automatic solution to the problem reflected by that dataset. The application of the trained CNN to large datasets can be used to annotate a large number of samples and make discoveries in the data, or automate annotation as done in tasks such as image-based diagnostics. However, if the dataset is biased in a certain way, that bias could lead to signal driven by the bias rather than by the biomedical information. That misleading situation can lead to differences between the performance achieved by the neural network when using the benchmark dataset, and the performance of the neural network when applied to real-world medical images.

A controlled data acquisition process does not necessarily guarantee unbiased datasets. For instance, a medical dataset for automatic image-based diagnostics can be acquired at more than one clinic. If the positive cases are not distributed equally across the different clinics, a CNN can learn the features that characterize a certain clinic, and in fact develop an algorithm for "clinic classification" rather than identification of the actual medical condition. Because different clinics can use different hardware, dif-

ferent settings, and different technicians, it is difficult to guarantee that the image acquisition process in all clinics is identical.

An example of such bias was demonstrated using microscopy data, where an algorithm could predict the treatment applied to cells in microscopy images (Zanella et al., 2010; Singh et al., 2014). But the ability to classify the cells was also driven by the imaging session rather than the morphology of the cells (Shamir, 2011). That was shown by the consistency of the results regardless of the presence of cells in the images, demonstrating that the signal was driven by the background noise rather than the cells (Shamir, 2011). That is, even when the cells were completely removed from the images, the classification accuracy was nearly identical to the classification accuracy of the original dataset, when the cells were present (Shamir, 2011).

3. Biomedical image datasets

Several datasets of biomedical images were tested, as shown in Table 1. The first dataset that was used was *COVID-CT* (Khan et al., 2020). The *COVID-CT* dataset was compiled in order to test whether COVID-19 can be diagnosed through automatic analysis of chest X-rays. The dataset was used for CoroNet (Khan et al., 2020), a deep convolution neural network that can identify COVID-19 infection from chest X-ray radiographs.

Figure 1 shows examples of original chest x-rays images from the dataset, and a 20×20 pixels sub-images from the top left corner of the original image. It is clear that by using the human eye alone it is not possible to identify differences between the different classes based on the cropped sub-images, as these are blank background areas that do not contain visual information of any part of the body. Table 1 shows the classification accuracy when classifying the blank 20×20 sub-images using a LeNet-5 convolutional neural network (LeCun et al., 1998; Sultana et al., 2018). The activation function used in most layers of the convolutional neural network is Rectified Linear Unit (ReLU), except for the output layer, where the sigmoid activation function is used. During training the model used the Adam (Adaptive Moment estimation) optimizer (Kingma and Ba, 2014) with an adaptive learning rate, and the binary cross entropy is used as the loss function because of binary classification. The number of training epochs was 120.

The experiments were done by training the network with the blank sub-images of the training set, and then testing with the test blank sub-images in the test set. For comparison, the same experiment was done by also training and testing with the original images. In any case, in all experiments the training and test images were of the same type, and no attempt was done to classify the 20×20 sub-images with a CNN trained with the original images, or vice versa. Obviously, training images were not used for testing. As the table shows, despite the absence of information related to COVID-19 in the seemingly blank background sub-images, the CNN was able to achieve classification accuracy of 67.14%, much higher than mere chance accuracy of 50%.

Classification accuracy of ~62.5% was observed with the original dataset for *COVID-CT*, when classifying the images into COVID-19 or not COVID-19 using LeNet-5 architecture. This shows that the dataset of the sub-images provided better prediction accuracy than the original dataset. That surprising observation can be due to the signal from the differences in the imaging process being stronger than the signal from the medical condition reflected by the images. A dataset of sub-images is more consistent, allowing the CNN to learn the subtle but consistent differences between the images originated from the imaging process.

Another biomedical dataset that was tested is a dataset with four X-ray classes (Khan et al., 2020). In this dataset the chest X-rays were separated into the classes *COVID*, *Normal*, *Pneumonia bacterial* and *Pneumonia viral*.

Figure 1 shows examples of the original images and the corresponding cropped top left corner of the original images. As the figure shows, the cropped images are similar to each other, and cannot be classified easily by the naked eye. They only contain background areas, and not any part of the body. A convolutional neural network used to distinguish the classes based on the cropped images alone provided classification accuracy of ~41.25%, which is far higher than the expected random chance classification of ~25%. That is, even when no biomedical information was present in the image, the CNN was able to identify COVID-19 cases with accuracy far greater than mere chance. The classification accuracy when applying the same LeNet-5 architecture to the original dataset pro-

No	Dataset	Classes	# Training images	# Test images	Image size	Accuracy (%)
1	COVID-19 (two classes)	2	558	140	20×20 pixels	67.14
2	COVID-19 (four classes)	4	960	240	20×20 pixels	41.25
3	Kvasir	8	3200	800	20×20 pixels	30.75

Table 1: Medical image datasets used in this study, the size of the separated seemingly blank background area, and the classification accuracy in each dataset achieved when classifying the blank sub-images with a LeNet-5 convolutional neural network.

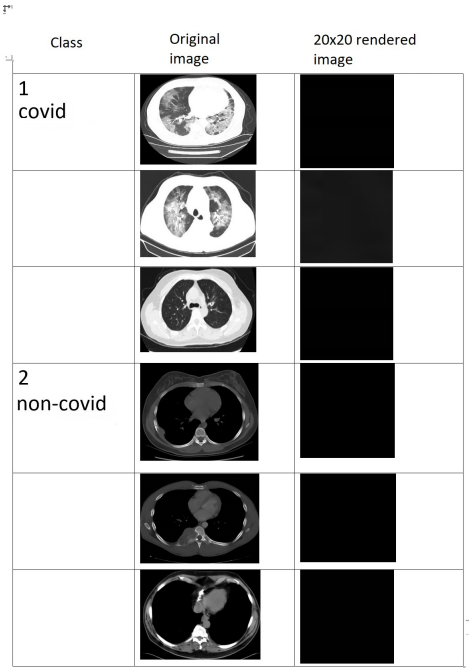

Figure 1: Example images from COVID-CT and a 20×20 portion of the top left corner separated from the original images. Only the sub-images were used for the classification.

vided accuracy of ~77.50%. That shows that the CNN trained and tested with the full images provided a higher classification accuracy compared to the CNN trained and tested using the blank sub-images. However, the dataset made of the small blank sub-images also provided prediction accuracy far higher than mere chance.

The Kvasir dataset (Pogorelov et al., 2017) is a biomedical dataset that contains images of endoscopic examinations. Figure 2 shows the original images and the blank sub-images of size 20×20 pixels that were cropped from the topmost left corner of the original dataset. As the figure shows, it is virtually impossible to classify the images into their respective labels by using the visual information in the cropped sub-images alone. However, LeNet-5 can perform this distinction with an accuracy of ~30.75%, even when the images do not contain any information that can allow a person identify the class. As before, the accuracy achieved with the blank sub-images is much higher than mere chance accuracy, which is ~12.5% for the eight classes.

When applying the LeNet-5 to the original images, the classification accuracy was ~73.75%. That accuracy is higher than the accuracy when using the blank sub-images. That difference can be attributed to the ability of the CNN to identify differences between the classes based on the visual content. But since the classification of the blank sub-images also shows accuracy higher than mere chance, it can be assumed that background visual features related to the image acquisition process might have a certain impact on the results.

Another dataset that was used was a dataset of 200 microscopy images of drosophila (D. melanogaster) cells separated into 10 classes, taken from the public benchmark of (Shamir et al., 2008b). Each class is a different gene being masked using mRNA interception, and the cells are stained with DAPI (4',6-diamidino-2-phenylindole). The images were acquired using a DeltaVision light microscope with a robotic stage and a 60× objective as described in (Shamir et al., 2009). Fig-

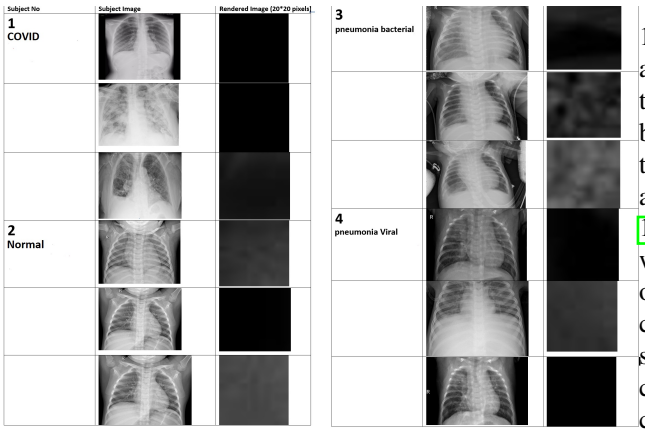

Figure 2: Example images from the COVID-19 dataset with four classes, and the 20×20 portion of the top left corner separated from the original images. When only the sub-images were used the classification accuracy was ~41%.

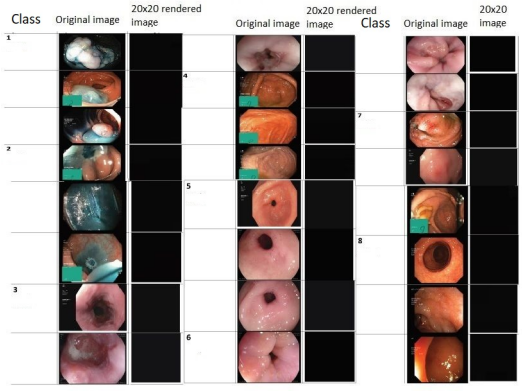

Figure 3: Example original images from *KVASIR* and the 20×20 portion of the top left corner separated from the original images. Only the sub images were used for the classification.

ure 4 shows example images from the dataset for each of the 10 genes being masked.

The dimensions of each microscopy image is 1024×1024 pixels, and each image contains multiple cells as can be seen in Figure 4. The distribution of the cells in the images is expected to be random, and the differences between the classes are expected to be reflected through the cells. Therefore, the cells were separated from the images by applying a simple Otsu binary threshold (Otsu, 1979), and objects with more than 40 neighboring pixels were identified. The 60×60 subimage around each such object was separated to create a dataset of 2000 images of cells, and another dataset of 2000 subimages of the same size taken from background parts of the image, where cells were not present. The Table shows that while the cell images are classified with much higher accuracy, the background subimages are classified in accuracy much higher than the expected 10% mere chance.

Objects	# Training images	# Test images	Image size (pixels)	Accuracy (%)
Cells	1500	500	60×60	81.11
Background	1500	500	60×60	63.45

Table 2: The classification accuracy when using the cell images and when using subimages taken from the background.

3.1. Face and object recognition

The source of the bias shown with the biomedical datasets above could be the image acquisition process, as will be discussed in Section 5. Some of the most commonly used datasets in the machine learning domain are datasets of images downloaded from the web such as ImageNet, and in these cases the process of image acquisition is not controlled. In the biomedical domain, however, datasets are normally not acquired by downloading different images from the web, but through a controlled process. That process can be compared to the the process of preparation of some face recognition datasets and object recognition datasets.

For controlled face recognition datasets we tested the *Yale Faces A* and the *Yale Faces B* face recognition benchmark datasets. The *Yale Faces A* dataset has 15 subjects, where each subject has 11 face images. The *Yale Faces B* dataset has 28 subjects, where each subject has 585 images.

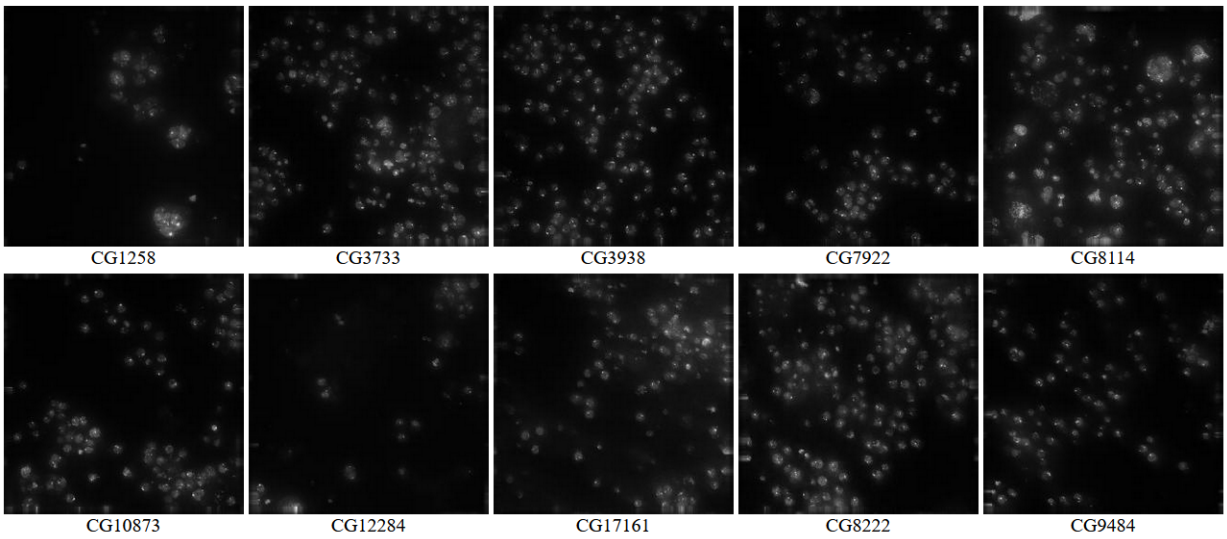

Figure 4: Example images and gene IDs from the RNAi dataset. Each class is the result of masking a different gene.

The *Yale Faces B* was transformed into a dataset of the same number of images, where each image in the original dataset was transformed into an image containing the 27×20 pixels of the top left corner in the original image. That part of the image contained just the background, which was visually identical in all images. Figure 5 shows the first five images of the first five subjects in *Yale Faces B*. As the figure shows, the sub-images separated from the original images of the different subjects are very difficult to distinguish when using the unaided human eye.

In the *Yale Faces A* dataset the background was removed from the image, leading to an artificially blank background. Therefore, in the case of the *Yale A* dataset, each image was transformed such that the 22×29 pixels from the forehead of each subject were used. Unlike *Yale B* dataset, in which no pixel containing any feature of the face or hair was used, in the *Yale A* dataset the small images contained pixels representing the skin of the person. However, the images did not contain information that allows to identify the face by visually looking at the image, or to even identify that the image is a face or any other part of a person's body. Figure 6 shows examples of the original face images and the smaller images that were used for classification by the CNN.

As with the biomedical datasets, the classification accuracy of the dataset was measured by using the common LeNet-5 CNN architecture. The number of training and test images in each dataset and the classification accuracy of each dataset are shown in Table 3.

No	Dataset	classes	# training images	# test images	Image size (pixels)	Accuracy (%)
1	Yale Faces A	15	132	33	22×29	54.6
2	Yale Faces B	28	13104	3276	27×20	87.8

Table 3: The size of the datasets and the classification accuracy when using LeNet-5 for classifying the cropped small sub-images from the face recognition datasets.

Although all images are visually similar to each other, the CNN was able to classify the images with accuracy far higher than mere chance. With 15 subjects, the expected mere chance accuracy of *Yale Faces A* is ~7%, while the mere chance accuracy expected for the *Yale Faces B* dataset is ~3%. The dramatically higher classification accuracy as shown in Table 3 shows that the CNN identifies discriminating features that are not necessarily related to the faces, and therefore not related to the machine learning problem at hand. That shows that even of the CNN achieves classification accuracy higher than mere chance, it does not necessarily mean that the CNN is indeed able to identify faces, but could possibly identify features of the dataset that allows discrimination between the differ-

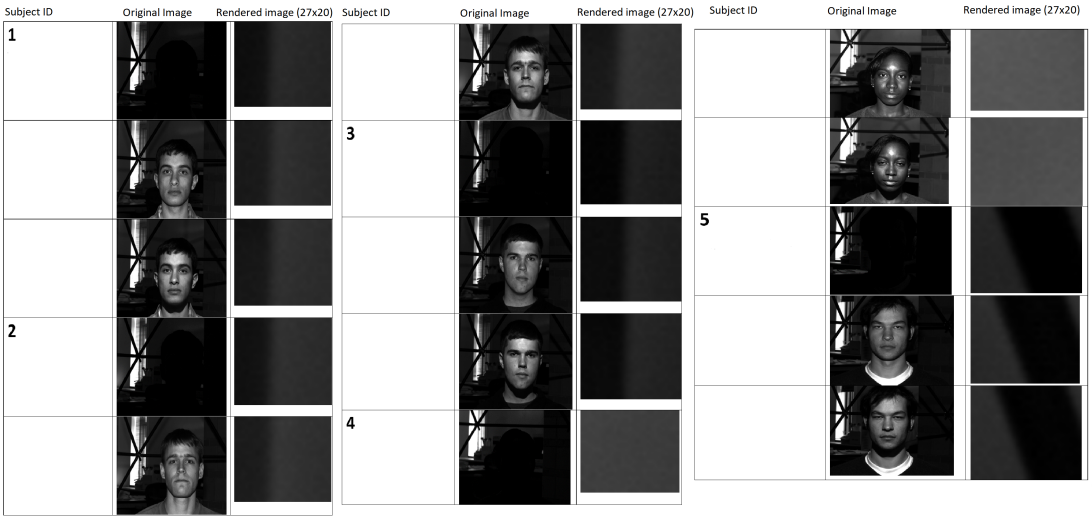

Figure 5: Example images from *Yale Faces B* and the small sub-images from the topmost left corner separated from the original images. Only the sub images were used for the classification.

ent subjects.

When applying the CNN to the original *Yale Faces A* dataset with the same number of training and test images specified in Table 3, the classification accuracy is ~96.97%. That is clearly higher accuracy than the ~54.55% accuracy when the transformed dataset of cropped sub-images are used. That shows that the images have more information than what can be identified in the sub-images that only contain the forehead, but the forehead information identified by the CNN can be used to identify the subject with accuracy is still far higher than random chance accuracy.

When applying the CNN to the original *Yale Faces B* dataset, also with the same number of training and test images specified in Table 3, the classification accuracy was ~99.97%. As with *Yale Faces A*, there is a significant drop in the classification accuracy when using just the small blank background sub-images. But the accuracy with the seemingly uninformative small parts of the background is still far higher than mere chance accuracy. In the case of *Yale B*, the small sub-images are taken from the background of the images, and do not contain any part of the face, hair, clothes, or anything else that might allow the identification of the person in the image. Sine the subject can be identified without any feature of the face or body, the only explanation is that the imaging process led to information present in the images, and allows the CNN to identify the subject by identifying the session in which the image was taken.

Object recognition benchmark datasets are in many cases collected from the world wide web. Commonly used benchmarks include ImageNet, MS COCO, PASCAL, or CIFAR. That method of collecting the data and preparing benchmark datasets is different from collecting images in the biomedical domain, where the process is normally a controlled process. To test for the possible bias, we used object recognition datasets that were collected in a controlled imaging process. We used two datasets: *COIL-20* and the *COIL-100*. *COIL-20* contains 20 object classes, and each object has 72 image samples in the dataset. *COIL-100* has 100 subjects, each contains 72 images(Nene et al., 1996b) (Nene et al., 1996a).

A separate dataset of sub-images was created from the original *COIL-100* and *COIL-20* datasets. The new datasets were equal in the number of image to the original datasets, but each image was replaced with the 128×128 sub-image, cropped from the top right corner of the original image. Figures 7 and 8 show examples of the original images, and the sub-images that were separated from the original images to form the new datasets. As the fig-

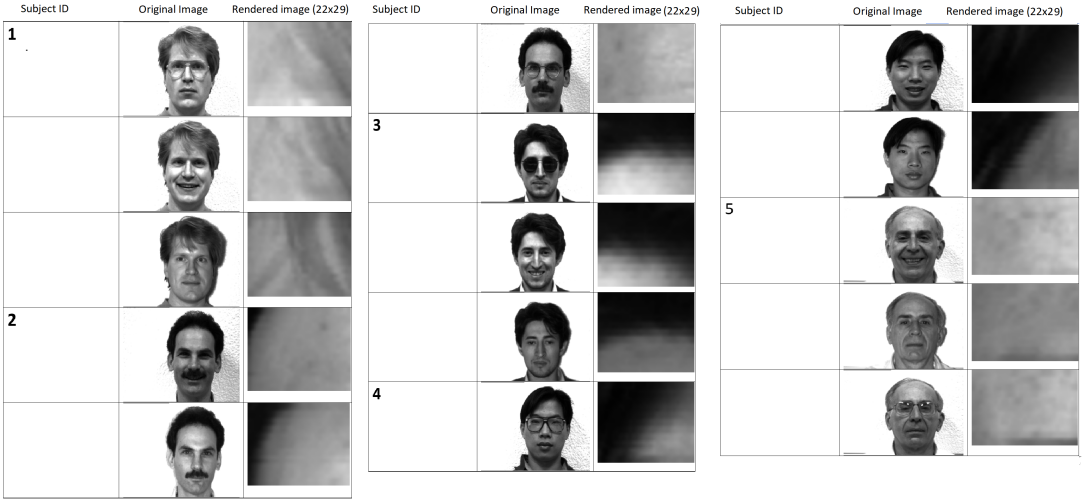

Figure 6: Example images from Yale Faces A and the sub-images of the forehead separated from the original images. Only the forehead sub-images were used for the classification.

ures show, the new datasets are made of seemingly blank images that contain no features that can allow a person identify the class.

Table 4 shows the number of images in the training and test sets, as well as the classification accuracy achieved when using the cropped sub-images that contain no intelligible image content. For *COIL-20* and *COIL-100*, the mere chance accuracy is ~5% and ~1%, respectively. The much higher classification accuracy achieved by the CNN clearly illustrates that the model is able to capture the hidden compounded relationship between the segmented image an the target label. That can be explained by the ability of the CNN to recover information from the pixels of the background, and that information can distinguish between the object classes. Since no objects are present in these sub-images, the identification is driven by the imaging process, and the CNN can identify the imaging session rather than the object in the image.

When applying the CNN to the original images the classification accuracy of *COIL-20* dataset is ~98.61%, which is far higher compared to the classification accuracy when using the cropped sub-images of ~35.42%. Similarly, the classification accuracy when applying the CNN to the original *COIL-100* dataset is ~96.46%, far higher than the ~27.48% accuracy achieved when applying the CNN to the cropped sub-images.

4. Proposed solutions to CNN classification bias

One of the primary advantages of convolution neural networks (CNNs) is their innate power to select a feature map automatically when supplied with training images. However, the downside of that nature might in some cases lead to potential weaknesses. The automated process of feature map selection without human interference may lead to the use of features that are not necessarily a reflection of the image analysis problem at hand. The CNN is designed to select the most discriminating features automatically, and if such features exist in the dataset the classification accuracy provided by the CNN can be misleading.

Several practices can be used to avoid misleading results due to classification bias driven by discriminating yet irrelevant features. Firstly, the background of an image can provide substantial information about the soundness of the image acquisition process. By separating small seemingly blank sub-images of the background we can create a control dataset made with just background information. The ability of a CNN to identify the correct class based on the background alone can alert on the existence

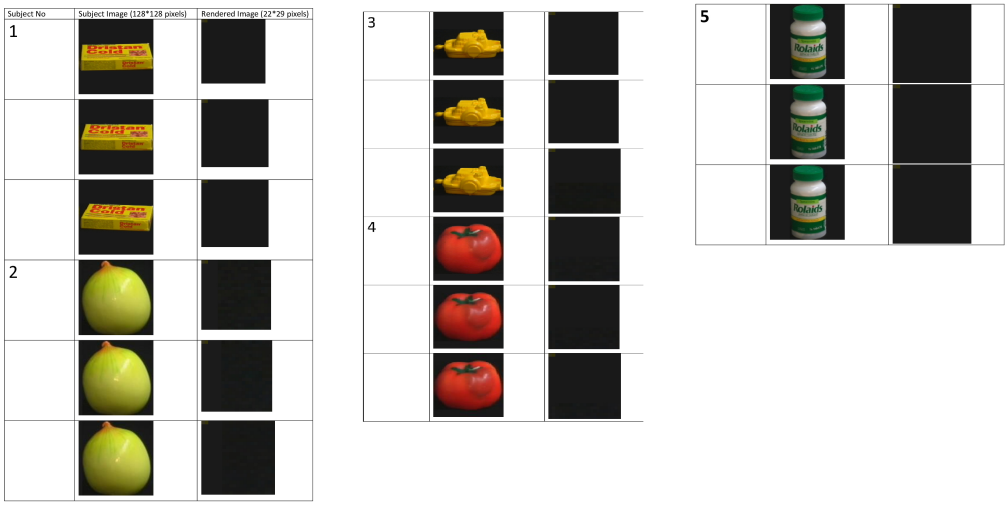

Figure 7: Example images from COIL-100 and the seemingly blank images of the background separated from the original images. Each sub image is the 128×128 sub-image separated from the top right part of the image. Because that part of the image is background, the sub-images seem to the unaided eye as black squares, and do not seem to contain meaningful information. Only the blank sub-images were used for the classification.

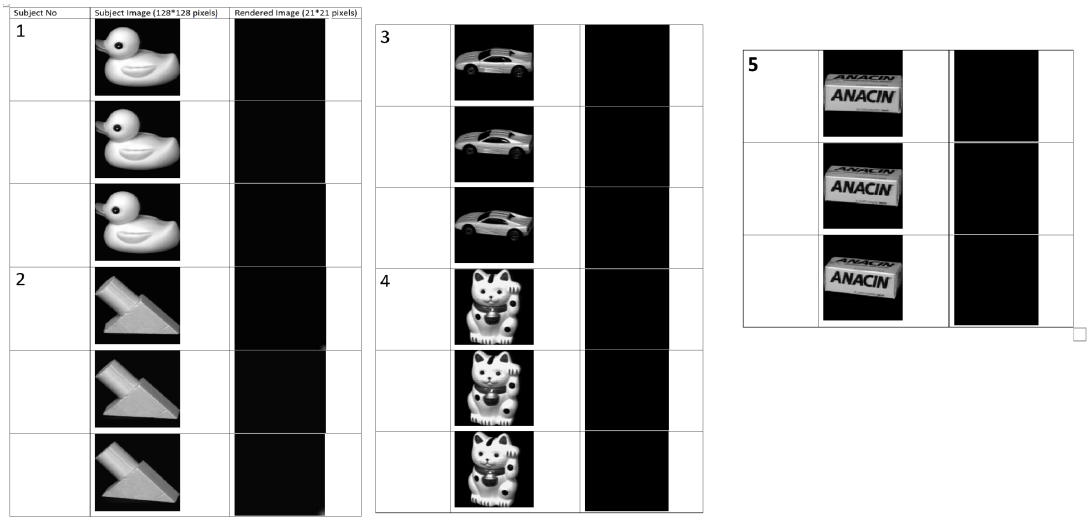

Figure 8: Example images from COIL-20 and the seemingly blank sub-images of the background separated from the original images. The background sub-images are the 128×128 sub-image separated from the top right part of each original image. That part of the images contain black background only, and therefore the separated sub-images are visually black squares. Only the blank sub-images were used for the classification.

No	Dataset	# classes	# training images	# test images	Image size	Accuracy (%)
1	COIL-20	20	1152	288	21×21 pixels	35.42
2	COIL-100	100	5760	1440	21×21 pixels	27.85

Table 4: Datasets used for testing the classification of object recognition benchmarks using deep neural networks.

of certain anomalies in the data acquisition process. These anomalies are difficult to detect, but CNNs can use them to make a classification with accuracy higher than its actual ability to classify these images when anomalies are not present. That is, if a CNN can predict the class of an image based on its background with accuracy higher than mere chance, the overall classification accuracy achieved by that CNN on the entire dataset might be biased, and therefore no strong assumptions can be made on the ability to use that CNN as a valid solution.

Another approach that can be used is acquiring the training set and test set in two separate data acquisition sessions, or obtaining the training and test data from different clinics or other sources. That is, instead of acquiring the entire dataset in a single batch and then separating it to training and test sets, the training set is acquired in one session, and the test set is acquired in a different session. The common practice of acquiring the entire dataset in a single batch and then randomly splitting the data into training and test sets can allow CNNs to achieve stronger classification accuracy by using potential information from the imaging session. If all images of a certain class were acquired in a single session, a CNN might be able to select features that identify the session rather than the class. Separating the acquisition of the training and test sets ensures that no features that can identify the session in the training set can be used by the CNN to identify the images in the test set by their session. That is, if each class of images is acquired in a single imaging session, and then separated into training and test samples, a CNN can associate an image to its session to increase its ability to correctly classify the images into classes. If all test samples are acquired in a different session than the training images, the session information cannot be used to associate test samples with training samples of the same class.

The practice of separating the acquisition of the training and test sets can be tested with the microscopy images described in Section 3, and shown in Figure 4. The dataset was generated multiple times through the same process, leading to equivalent datasets with the same genes, but imaged in a different process and different slides. The same 10 genes were also used for different experiments (Shamir et al., 2008a, 2009). Table 5 shows the results of training with one slide imaged with one image acquisition batch, and testing with another slide imaged in a different image acquisition batch.

As the table shows, the classification accuracy of the background images when using different imaging sessions for training and testing dropped to ~12%, which is very close to random accuracy of 10%. The classification accuracy of the cells also dropped substantially from 81% when training and testing with the same batch, to 47% when training with images acquired in one batch and testing with images acquired in a different batch. That shows that training with one batch and testing with another batch reduces the accuracy, meaning that some of the signal used for classification was originated from the imaging session, in addition to the signal coming from the shape of the cell.

Avoiding acquisition of data in sessions can also improve the reliability of benchmark datasets commonly used to test the performance of CNNs. For instance, if each sample is acquired in a separate session, the CNN will not be able to use the subtle but significant information that reflect the imaging session. While imaging multiple images in one session is convenient and more efficient in terms of number of images that can be collected, it is also an unsound practice that can allow CNNs classify the imaging session (e.g., lighting conditions, temperature of the CCD, etc) rather than the subjects in the images. If each image is acquired separately, no session information will be present. Also, if the images are also acquired in a random order and not by imaging one class at a time, the class cannot be identified by the session its samples were acquired in.

To test whether imaging each sample in a different session, we used the *c. elegans* muscle age dataset, which

Objects	Classes	# Training images	# Test images	Accuracy (%)
Cells	10	1500	1500	47.27
Background	10	1500	1500	11.83

Table 5: The classification accuracy when using the cell images and when using subimages taken from the background. The training set was acquired with one slide and one imaging session, and the test set was acquired in a different slide and a different image acquisition session than the training set.

is also part of the benchmark dataset of (Shamir et al., 2008b). The dataset contains 252 microscopy images of the head of *c. elegans* nematodes, separated into four classes. Each class is of a different age, 1 day old, 2 days old, 4 days old, and 8 days old. Each image was acquired separately, through a long manual process of imaging the nematodes using a 20× objective light microscope. Table 6 shows the classification accuracy of the entire image, and the classification accuracy when separating the 20×20 top-left subimage from each image. Fifty images per class are used from training, and seven images per class are used for testing. Due to the relatively low number of test images, a 10-fold cross-validation is used.

Image part	Classes	# Training images	# Test images	Accuracy (%)
All image	4	50	7	51.16
20×20 top-left	4	50	7	24.2

Table 6: Classification accuracy of the *c. elegans* muscle age dataset when using the entire image, and when using the 20× seemingly blank part of the images.

As the table shows, when separating the 20×20 top-left subimage from each original image, the classification accuracy drops to very close to mere chance accuracy of 25%. When using the entire image, the classification accuracy is higher than random chance. That shows that when acquiring each image separately the classification accuracy is very close to the expected mere chance accuracy. That shows that if each image is acquired in a separate batch, the background information cannot be used to make associations between training and test images.

5. Conclusion

Dataset bias has been discussed in the computer vision literature in the context of the ability of benchmark datasets to reflect the real world. In the biomedical domain, CNNs applied to biomedical image datasets have been shown to provide in some cases better accuracy than the classification made by expert pathologists (Paul et al., 2021). Here we study biases that are not driven by the human selection of the samples or preferences in the annotation process, but driven by the image acquisition process. These biases are more relevant to the biomedical domain, where in many cases the medical images are acquired in a controlled process by a defined number of clinics. These biases are difficult to identify, and are sometimes not expected since controlled image acquisition is often assumed as a process that controls also for the possible biases.

Experiments show that when acquiring the training set in one batch, and the test set in a different batch, the ability of the CNN to make accurate classifications with non-informative background parts of the images drops to approximately mere chance accuracy. That is due to the absence of information that can associate the sessions. In that situation, the ability to make correct classifications of the foreground objects can be attributed to the real ability of the CNN to solve the image classification problem, rather than its ability to identify subtle patterns that are typical to certain image acquisition sessions. Acquiring each image in a different session might not always be practical, as it can require substantial labor. For instance, in microscopy images preparing a separate slide for each cell can be impractical when a single person attempts to produce a dataset of several thousand cells. Preparing one dataset for training and a separate dataset for testing can achieve the same results. That practice doubles the amount of work compared to the traditional approach of splitting the data to training and test sets, but that practice is still practical.

Being easy to use, powerful, and accessible through available open source libraries, CNNs have been becoming extremely popular in the biomedical domain, and the default solution to automatic image analysis problems. However, while CNNs are in many ways superior to pre-

vious approaches, they also have the downside of overfitting and uncontrolled learning. When a growing population of biomedical researchers who are not necessarily machine learning experts use CNNs, it is important to ensure all CNN users are informed also with the possible weaknesses of CNN. That will help and avoid experiments that might seem scientifically sound, but in fact provide biased or unreliable results.

Acknowledgments

The research was funded in part by NSF grant AST-1903823. We would like to thank the two knowledgeable anonymous reviewers for the insightful comments that helped to improve the manuscript. We would also like to thank Zhejiang University Press for the funding that makes the paper open access.

References

Abraham, V.C., Taylor, D.L., Haskins, J.R., 2004. High content screening applied to large-scale cell biology. Trends in Biotechnology 22, 15–22.

Aina, O.E., Adeshina, S.A., Aibinu, A., 2019. Deep learning for image-based cervical cancer detection and diagnosis—a survey, in: 2019 15th International Conference on Electronics, Computer and Computation, IEEE. pp. 1–7.

Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K., 2018. Medical image analysis using convolutional neural networks: a review. Journal of Medical Systems 42, 1–13.

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., Lundin, J., 2018. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific reports 8, 1–11.

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., Xie, Z., 2018. Deep learning and its applications in biomedicine. Genomics, proteomics & bioinformatics 16, 17–32.

Chen, Y.C., Hong, D.J.K., Wu, C.W., Mupparapu, M., et al., 2019. The use of deep convolutional neural networks in biomedical imaging: A review. Journal of Orofacial Sciences 11, 3.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 .

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q., 2018. Deep learning for image-based cancer detection and diagnosis- a survey. Pattern Recognition 83, 134–149.

Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D., 2011. Adversarial machine learning, in: Proceedings of the 4th ACM workshop on Security and artificial intelligence, pp. 43–58.

Jaipuria, N., Zhang, X., Bhasin, R., Arafa, M., Chakravarty, P., Shrivastava, S., Manglani, S., Murali, V.N., 2020. Deflating dataset bias using synthetic data augmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 772–773.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172, 1122–1131.

Khan, A.I., Shah, J.L., Bhat, M.M., 2020. Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. Computer Methods and Programs in Biomedicine 196, 105581.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A., 2012. Undoing the damage of dataset bias, in: European Conference on Computer Vision, Springer. pp. 158–171.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T., 2019. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data, in: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Li, Y., Vasconcelos, N.. REPAIR: Removing Representation Bias by Dataset Resampling, in: 2019 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 9572–9581.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Liu, S., Wang, X., Liu, M., Zhu, J., 2017. Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics 1, 48–56.

McLaughlin, N., Del Rincon, J.M., Miller, P., 2015. Data-augmentation for reducing dataset bias in person re-identification, in: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE. pp. 1–6.

Min, S., Lee, B., Yoon, S., 2017. Deep learning in bioinformatics. Briefings in Bioinformatics 18, 851–869.

Nene, S.A., Nayar, S.K., Murase, H., et al., 1996a. Columbia object image library (coil-100). Technical Report CUCS-005-96 .

Nene, S.A., Nayar, S.K., Murase, H., et al., 1996b. Columbia object image library (coil-20). Technical Report CUCS-005-96 .

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66.

Paul, H.Y., Singh, D., Harvey, S.C., Hager, G.D., Mullen, L.A., 2021. Deepcat: Deep computer-aided triage of screening mammography. Journal of Digital Imaging 34, 27–35.

Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., Riegler, M., Halvorsen, P., 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, in: Proceedings of the 8th ACM on Multimedia Systems Conference, ACM, New York, NY, USA. pp. 164–169. URL: http://doi.acm.org/10.1145/3083187.3083212, doi:10.1145/3083187.3083212.

Shamir, L., 2011. Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis. Journal of microscopy 243, 284–292.

Shamir, L., Delaney, J.D., Orlov, N., Eckley, D.M., Goldberg, I.G., 2010. Pattern recognition software and techniques for biological image analysis. PLoS Computational Biology 6, e1000974.

Shamir, L., Eckley, D.M., Delaney, J., Orlov, N., Goldberg, I.G., 2009. An image informatics method for automated quantitative analysis of phenotype visual similarities, in: 2009 IEEE/NIH Life Science Systems and Applications Workshop, IEEE. pp. 96–99.

Shamir, L., Orlov, N., Eckley, D.M., Macura, T., Johnston, J., Goldberg, I.G., 2008a. Wndchrm–an open source utility for biological image analysis. Source Code for Biology and Medicine 3, 1–13.

Shamir, L., Orlov, N., Eckley, D.M., Macura, T.J., Goldberg, I.G., 2008b. Iicbu 2008: a proposed benchmark suite for biological image analysis. Medical & Biological Engineering & Computing 46, 943–947.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Singh, S., Carpenter, A.E., Genovesio, A., 2014. Increasing the content of high-content screening: an overview. Journal of biomolecular screening 19, 640–650.

Sultana, F., Sufian, A., Dutta, P., 2018. Advancements in image classification using convolutional neural network, in: 2018 Fourth International Conference on Research in Computational Intelligence and Communi-

cation Networks (ICRCICN), pp. 122–129. doi:10.1109/ICRCICN.2018.8718718.

Thomsen, K., Christensen, A.L., Iversen, L., Lomholt, H.B., Winther, O., 2020. Deep learning for diagnostic binary classification of multiple-lesion skin diseases. Frontiers in Medicine 7, 604.

Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T., 2017. A Deeper Look at Dataset Bias. Springer International Publishing, Cham. pp. 37–55. URL: https://doi.org/10.1007/978-3-319-58347-1_2, doi:10.1007/978-3-319-58347-1_2.

Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias, in: CVPR 2011, IEEE. pp. 1521–1528.

Wainberg, M., Merico, D., Delong, A., Frey, B.J., 2018. Deep learning in biomedicine. Nature biotechnology 36, 829–838.

Zanella, F., Lorens, J.B., Link, W., 2010. High content screening: seeing is believing. Trends in Biotechnology 28, 237–245.

Zhang, H., Zhu, L., Zhu, Y., Yang, Y., 2020. Motion-excited sampler: Video adversarial attack with sparked prior, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, Springer. pp. 240–256.

Zhang, Y., Lin, H., Yang, Z., Wang, J., Sun, Y., Xu, B., Zhao, Z., 2019. Neural network-based approaches for biomedical relation classification: a review. Journal of Biomedical Informatics 99, 103294.