Systematic biases in machine learning and their impact on astronomy research

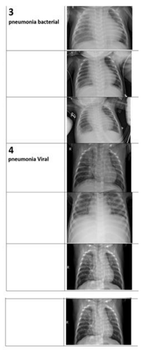
Lior Shamir

Department of Computer Science, Kansas State University

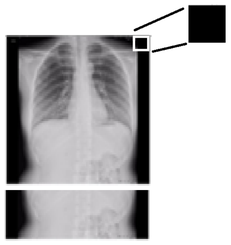
November, 2023

Thousands of false or biased (published) discoveries using ML/DNN

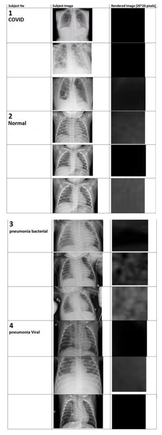
Example 1: experiment that shows that covid-19 can be identified through CNN classification of chest x-rays by CNN (Khan et al, 2020). When applying a CNN to just small seemingly blank background Areas taken from the top right corner of each image, the classification accuracy is nearly the same as with the entire chest x-rays. That shows that covid-19 can be identified by background areas of the chest x-rays that contain no part of the body. That is obviously impossible, unless there is some hidden information in the data that allows the Classification regardless of coivd-19.



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



Dhar, S., Shamir, L., 2021, *Visual Informatics*, 5(3), 92-101



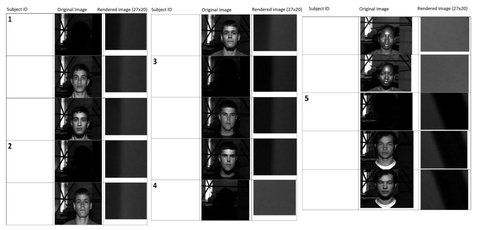
CNN classification accuracy:

Original images – 67%
Blank background images – 62%
Mere chance accuracy – 25%

The analysis with the blank images was done by just using the images on the right, that contain no part of the body.

Many other examples: The experiment shown above was repeated with very many datasets from the biomedical domain, but also Common and very commonly used benchmark datasets of face recognition, object recognition, and many more, as explained in detail In (Dhar & Shamir, 2021). In all cases, repeating the same experiment with just "blank" background of the images led to very similar Results, showing that the algorithm can provide good classification accuracy even with no relevant information (Dhar & Shamir, 2021).

Automatic face recognition

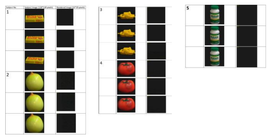


CNN classification accuracy:

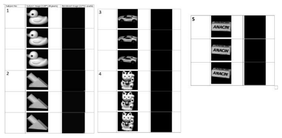
Original images – 99%
Blank background images – 87%
Mere chance accuracy – 4%

This experiment shows that face recognition can be done with no faces. The algorithm uses background information that identifies the imaging conditions rather than the face.

Automatic object recognition



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



There are many more examples in many other fields as shown in (Dhar & Shamir, 2021). Other relevant information about face, object, and microcopy images analysis can be found in previous papers (Shamir, 2008, 2011; Model & 2015), and audio data (Bock & Shamir, 20016). The main challenge of these biases is that they are very difficult to notice, and many, even experienced, researchers have them in their data without being aware of them, ultimately leading to biased results and conclusions.

The implications of systematic machine learning and deep learning bias in Digital Sky Surveys

Like other disciplines as shown here, astronomy is also prone to systematic machine learning bias. The most vulnerable systems are digital sky surveys, where machine learning is used to handle the vast pipelines of data. The following example shows an experiment with two digital sky surveys: SDSS and Pan-STARRS. The experiment tested bias in the broad annotation of galaxies into "late-type" and "early-type". An experiment using a DNN was done, and the confusion matrix was generated. Then, another experiment was done such that the training set was made of elliptical galaxies from one part of the sky, and spiral galaxies from another part of the sky. When changing the training set, the confusion matrix showed a statistically significant different portion of spiral galaxies I the sky (Dhar & Shamir, 2022). That is, although the test set was identical, using different training sets based on the part o the sky they were taken from provided a significantly different conclusion about the Universe.

Galaxy image annotation using CNNs

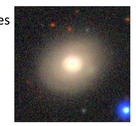
SDSS

Training set and test set from the same part of the sky

Classification to spiral galaxies and elliptical galaxies

When the test set and training set are from the same part of the sky, the CNN shows a different Universe then when the training and test images come from different parts of the sky.

The results show that the location of the training galaxies in the sky completely change the results, even when the test set is identical. That bias is unexpected and very difficult to notice.



Pan-STARRS

Training set and test set from the same part of the sky

	Elliptical	Spiral
Elliptical	2891	109
Spiral	85	2915

Training set and test set from different parts of the sky. Training spiral galaxies are taken from the same part of the sky were the test set is taken.

	Elliptical	Spiral
Elliptical	2704	296
Spiral	31	2969

	Elliptical	Spiral
Elliptical	7850	150
Spiral	756	7244

Training set and test set from the same part of the sky. Training spiral galaxies are taken from the same part of the sky were the test set is taken.

	Elliptical	Spiral
Elliptical	7699	301
Spiral	450	7550

Dhar, S., Shamir, L., 2022, *Astronomy and Computing*, 38, 100545

Photometry annotation using ML

Same as with galaxy images, bias with ML annotation can also be observed when using the photometry data for similar task (Goddard & Shamir, 2022). Photometry data taken from SDSS are separated into three different locations in the sky. Then, the training sets were taken such that spiral galaxies were taken from one part of the sky, and elliptical galaxies from a different part of the sky. Table 1 shows the binomial distribution P values for the difference between all combinations of spiral/elliptical galaxies taken from different parts of the sky.

Table 1. Two-tailed p-values of the binomial distribution.

Training Region		Evaluation Region					
		Virgo		Hercules		Cetus	
Ellipticals	Spirals	Elliptical	Spiral	Elliptical	Spiral	Elliptical	Spiral
Virgo	Virgo	—	—	$9.1 \cdot 10^{-2}$	$1.93 \cdot 10^{-1}$	$5.75 \cdot 10^{-1}$	$3.03 \cdot 10^{-1}$
Virgo	Hercules	$7.11 \cdot 10^{-1}$	$4.96 \cdot 10^{-1}$	$8.79 \cdot 10^{-3}$	$4.87 \cdot 10^{-2}$	$9.52 \cdot 10^{-1}$	$6.97 \cdot 10^{-1}$
Virgo	Cetus	$4.01 \cdot 10^{-1}$	$2.42 \cdot 10^{-1}$	$4.29 \cdot 10^{-1}$	$6.89 \cdot 10^{-1}$	$1.87 \cdot 10^{-1}$	$2.78 \cdot 10^{-2}$
Hercules	Virgo	$6.89 \cdot 10^{-1}$	$7.34 \cdot 10^{-1}$	$2.5 \cdot 10^{-1}$	$2.38 \cdot 10^{-1}$	$3.9 \cdot 10^{-1}$	$3.95 \cdot 10^{-1}$
Hercules	Hercules	$1.8 \cdot 10^{-1}$	$8.18 \cdot 10^{-2}$	—	—	$1.19 \cdot 10^{-1}$	$3.73 \cdot 10^{-3}$
Hercules	Cetus	$2.71 \cdot 10^{-1}$	$1.04 \cdot 10^{-2}$	$6.53 \cdot 10^{-1}$	$3.32 \cdot 10^{-1}$	$2.0 \cdot 10^{-1}$	$1.37 \cdot 10^{-3}$
Cetus	Virgo	$7.79 \cdot 10^{-1}$	$7.19 \cdot 10^{-1}$	$3.94 \cdot 10^{-2}$	$1.8 \cdot 10^{-1}$	$5.22 \cdot 10^{-1}$	$3.42 \cdot 10^{-1}$
Cetus	Hercules	$4.18 \cdot 10^{-1}$	$4.71 \cdot 10^{-1}$	$7.24 \cdot 10^{-4}$	$6.14 \cdot 10^{-2}$	$8.49 \cdot 10^{-1}$	$4.9 \cdot 10^{-1}$
Cetus	Cetus	$6.67 \cdot 10^{-1}$	$1.56 \cdot 10^{-1}$	$5.81 \cdot 10^{-4}$	$2.99 \cdot 10^{-2}$	—	—

Although most values show no statistical significance, some of these values show that in some cases the experiment show parts of the sky that have different distribution of spiral/elliptical galaxies compared to other parts of the sky.

Goddard, H., Shamir, L., Neural network bias in analysis of galaxy photometry data, *18th IEEE International Conference on eScience*, pp. 407-408, 2022.

References

Bock, B., Shamir, L., Assessing the efficacy of benchmarks for automatic speech accent recognition, *EAI Endorsed Transactions on Creative Technologies* , 15(4), e3. EAI, 2015.

Bock, B., Shamir, L., Assessing the efficacy of benchmarks for automatic speech accent recognition, *8th International Conference on Mobile Multimedia Communications (MOBIMEDIA)*, Chengdu, China. ACM, 2015.

Dhar, S., Shamir, L., Systematic biases when using deep neural networks for annotating large catalogs of astronomical images, *Astronomy and Computing*, 38, 100545, 2022.

Dhar, S., Shamir, L., Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks, *Visual Informatics*, 5(3), 92-101, 2021.

Goddard, H., Shamir, L., Neural network bias in analysis of galaxy photometry data, *18th IEEE International Conference on eScience*, pp. 407-408, 2022.

Model, I, Shamir, L., Comparison of dataset bias in object recognition benchmarks, *IEEE Access*, 3(1), 1953-1962, 2015

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

Shamir, L.; Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis, *Journal of Microscopy*, vol. 243, no. 3, pp. 284-292. Wiley-Blackwell, 2011.

Shamir, L., Evaluation of face datasets as tools for assessing the performance of face recognition methods, *International Journal of Computer Vision*, vol. 79(3), pp. 225-230. Springer, 2008.