# Doppelganger effects in Machine Learning

## I. Introduction

Nowadays, with the continuous development of machine learning, machine learning models have been widely used in the biomedical field. However, the presence of doppelganger effects affects the evaluation of the effectiveness of machine learning models.

Doppelganger effects refer to the phenomenon that two or more individuals have similar or identical features that may be mistaken for each other [1]. In machine learning models, this can lead to machine learning models performing surprisingly well in the training model but poorly in the test.

## II. Doppelganger effects in data

Doppelganger effects do not only occur in biomedical data and may also occur in other data types. This section will give some examples of doppelganger effects on other data sets.

### A. Doppelganger effects in the facial recognition system

Doppelganger effects usually yield an increased probability of false matches in a facial recognition system, as opposed to random face image pairs selected for non-mated comparison trials. Apart from demographic attributes, doppelgangers often share facial properties such as facial shape (Figure 1) [2].

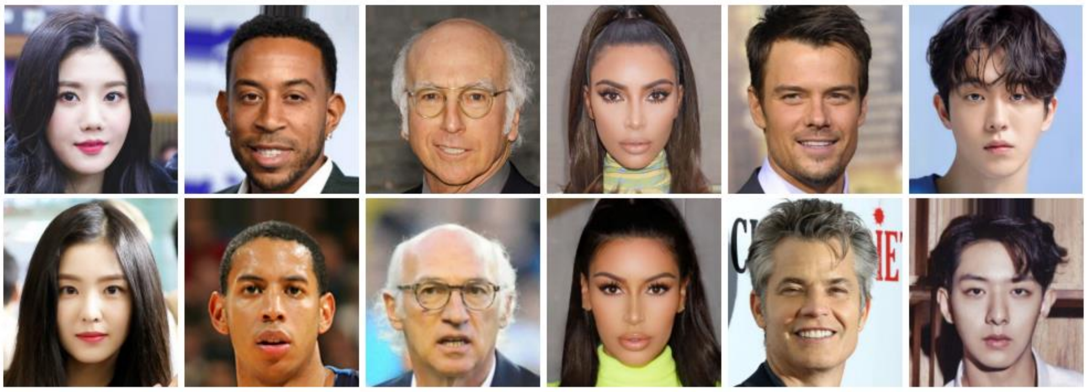

Figure 1: Example of doppelgangers image pairs

### B. Doppelganger effects in genomic data

Sufficient germ-line sequence markers provide a “fingerprint” that can be matched uniquely in a database of genotypes. Publicly available human genomic data is therefore normally summarized at a level that cannot be identified uniquely to protect patient privacy. Cancer transcriptomes undergo alterations that are highly distinctive but much more difficult to identify uniquely in summarized form. Re-use of tissue specimens is widespread in clinical genomic studies, creating a “doppelgänger effect” in publicly available datasets: hidden duplicates that, if left undetected, can inflate statistical significance or apparent accuracy of genomic models when combining data from different studies [3].

### C. Doppelganger effects in text data

This refers to the phenomenon where two or more pieces of text appear to be very similar, or even identical, but were produced by different authors or sources. Examples of doppelganger effects in text data include:

- Plagiarism, where one person copies the work of another without giving credit
- Paraphrasing, where one person rewrites the work of another in their own words
- Duplicate content on the internet, where the same content appears on multiple websites
- Automated text generation, where a machine learning model generates text that is similar to existing text

In all of these cases, the doppelganger text is similar or identical to the original text, but it is not an exact copy.

## III. Quantitative analysis of doppelganger effects

In quantitative terms, doppelganger effects can be measured by comparing the model's performance on the training dataset with its performance on the test dataset. If the model performs significantly lower than in training, this may indicate that the model is subject to a duality effect.

Moreover, doppelganger effects can also be measured by looking at the prediction accuracy of the model on different subgroups of the target population. If the error rate or accuracy of the model varies significantly between subgroups, this may indicate that the model is biased. It is likely because the model was trained on a dataset not representative of the target population. Then it may make inaccurate predictions when applied to individuals not adequately represented in the training data. In this case, the model obtained from such learning does not generalize well to all members of the population.

## IV. Avoiding doppelganger effects in the practice

### A. Cross-checks using meta-data

With this information from the meta-data, we are able to identify potential doppelgängers and assort them all into either training or validation sets, effectively preventing doppelgänger effects, and allowing a relatively more objective evaluation of ML performance. In a similar vein, technical replicates arising from the same sample should also be dealt with similarly.

### B. Data stratification

Instead of evaluating model performance on whole test data, we can stratify data into strata of different similarities. Assuming each stratum coincides with a known proportion of real-world population, we are still able to appreciate the real-world performance of the classifier by considering the real-world prevalence of a stratum when interpreting the performance at that stratum. More importantly, strata with poor model performance pinpoint gaps in the classifier.

### C. Robust and independent validation checks involving many data sets

Divergent validation techniques can inform the objectivity of the classifier. It also informs on the generalizability of the model (in terms of real-world usage) despite the possible presence of data doppelgängers in the training set.

## D. Explainable AI

Applying interpretable machine learning techniques (Figure 2) [5] such as LIME (Local Interpretable Model-Agnostic Explanations) can help to understand what the model is looking at when it makes a prediction. This can be useful for identifying the parts of the image that the model relies on for the prediction, which can help to detect doppelganger effects.

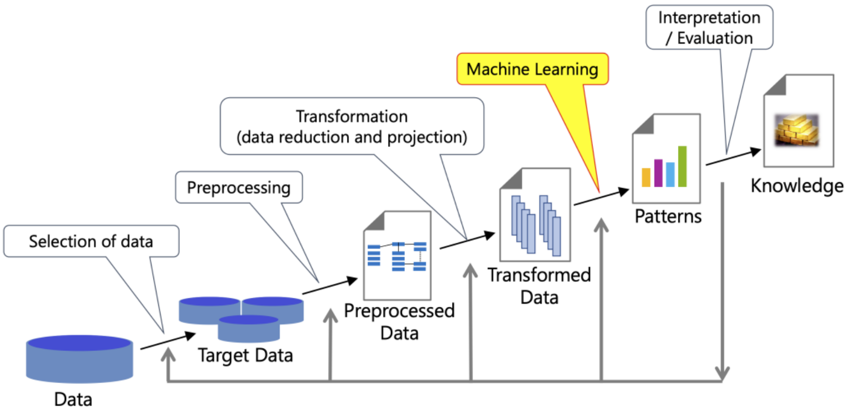

Figure 2: Knowledge Discovery Process (Figure adapted from Fayyad et al., 1996)

## E. Adversarial training

Adversarial training is another way to avoid doppelganger effects. This involves training the model with a dataset that includes images of individuals who are similar to each other, but not identical. This can help the model to learn to recognize subtle differences and variations, which can reduce the likelihood of doppelganger effects (Figure 3) [6].

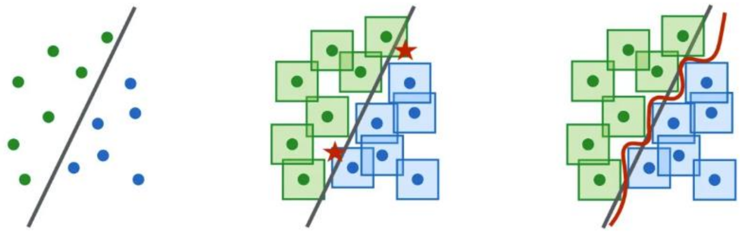

Figure 3: A conceptual illustration of standard vs. adversarial decision boundaries. Left: A set of points that can be easily separated with a simple (in the case, linear) decision boundary. Middle: The simple decision boundary does not separate $l_\infty$-balls (here, squares) around the data points. Hence there are adversarial examples (the red stars) that will be misclassified. Right: Separating the $l_\infty$-balls requires a significantly more complicated decision boundary. The resulting classifier is robust to adversarial examples with bounded $l_\infty$-norm perturbations.

Adversarial training can be seen as a special kind of Regularization, where the learned features are "purified" (Figure 4) [7].

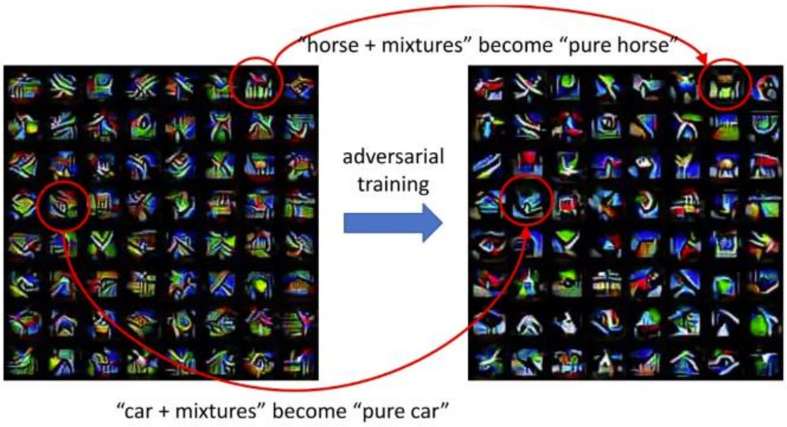
"horse + mixtures" become "pure horse"
adversarial training
"car + mixtures" become "pure car"

Figure 4: Experiments support the theory that adversarial training do purify dense mixtures, through visualizing some deep layer features of ResNet on CIFAR-10 data.

## F. Active learning

Utilizing active learning techniques can help to reduce doppelganger effects. This is done by using the model to identify and flag images that are likely to be doppelgangers, and then having human experts review and verify the flagged images (Figure 5) [8].

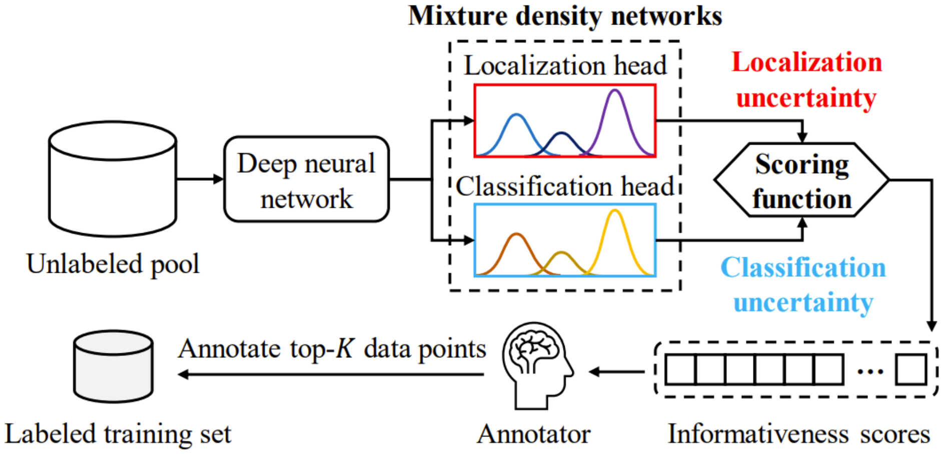

Figure 5: The Process of Active Learning (Figure adapted from Choi et al., 2021)

## References:

[1] Wang L R, Wong L, Goh W W B. How doppelgänger effects in biomedical data confound machine learning[J]. Drug Discovery Today, 2021.

[2] Hill K. The secretive company that might end privacy as we know it[M]//Ethics of Data and Analytics. Auerbach Publications, 2020: 170-177.

[3] Waldron L, Riester M, Ramos M, et al. The doppelgänger effect: hidden duplicates in databases of transcriptome profiles[J]. JNCI: Journal of the National Cancer Institute, 2016, 108(11). [4] Xue H T, Stanley-Baker M, Kong A W K, et al. Data considerations for predictive modeling applied to the discovery of bioactive natural products[J]. Drug discovery today, 2022.

[5] Fayyad U M, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework[C]//KDD. 1996, 96: 82-88.

[6] Hu W, Niu G, Sato I, et al. Does distributionally robust supervised learning give robust classifiers?[C]//International Conference on Machine Learning. PMLR, 2018: 2029-2037.

[7] Allen-Zhu Z, Li Y. Feature purification: How adversarial training performs robust deep learning[C]//2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS). IEEE, 2022: 977-988.

[8] Choi J, Elezi I, Lee H J, et al. Active learning for deep object detection via probabilistic modeling[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10264-10273.