# ML models and doppelganger effects

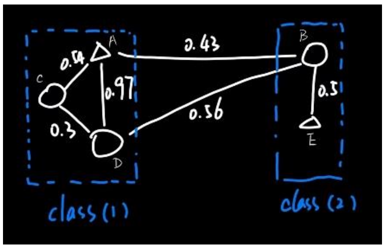
## Abstract

As the AI technique develops, larger and larger data size could be well handled, and the problems of data qualification are taking more and more attention. In this report, I would like to discuss why doppelganger effects are not unique to biomedical data, and explain my idea of using inner product similarity (cosine similarity) to better detect data doppelgangers and some data argumentation tricks which could maybe avoid this in practice. Some interesting image data samples are provided to better explain my idea.

## Related work

A doppelganger is a double or a look-alike of a person. The term "doppelganger effect" refers to the phenomenon of seeing someone who looks like a familiar person, but is not actually that person [1]. The doppelganger effects are first introduced to the data science area in 2016 [5], and another work gave a further discussion in 2021 [2], discussing the effects that could be caused due to the presence of data doppelgangers. Data doppelgangers mean that there are some independently derived data that are very similar to each other, and therefore many ML models are generally overstated – most of which are actually having a similar performance as the model totally trained on random signatures when removing all the data doppelgangers in the test set.

The earlier studies used MD5 fingerprints of the CEL files to identify the data doppelgangers [3], but since the limitation of the MD5 technology [4], this can only identify the duplicate samples but not true data doppelgangers. Another study used the pairwise Pearson's correlation coefficient (PPCC) to capture the similarity between sample pairs [5]. And the later work conclusively made a link between PPCC data doppelgangers and their impact on confounding ML tasks [2] – PPCC data doppelgangers are defined as valid sample pairs with PPCC values greater than all negative sample pairs in this work. Taking an example I draw on my pad, assuming that we have 5 samples in 2 classes (the float number is the similarity; different shape means from different patients; each sample is an image of tissue cell):

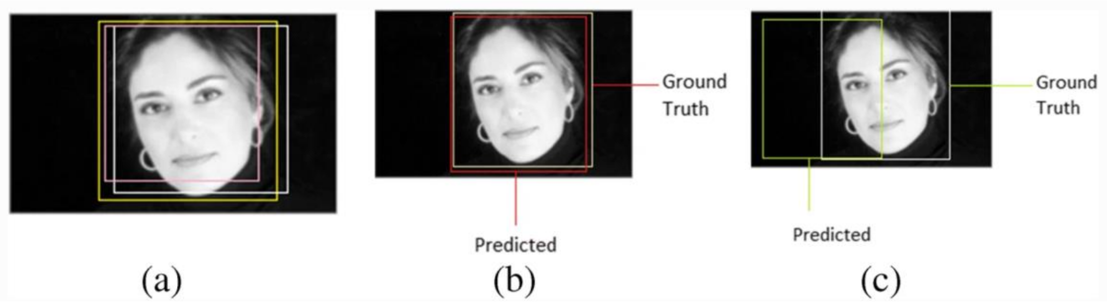


It’s easy to know that sample A and D is the pair that could cause data doppelgangers since $0.97 > 0.56 > 0.43$. When we use A to train and D to test the performance, if the images have $32 \times 32$ pixels, the similarity of 0.97 means 97% of the value of pixels in images A and D are the same, only 30 pixels are different (1024 pixels in total). In this case, running a ML model like logistic regression which has 1024 parameters, if the parameters value are roughly the same, then even if drop these 30 pixels out, the model could still predict class 1 as a result – this data pair could even considered as the same sample with different resolution. So if we put A in training set and only D in test set, the model performance will be dramatically high.

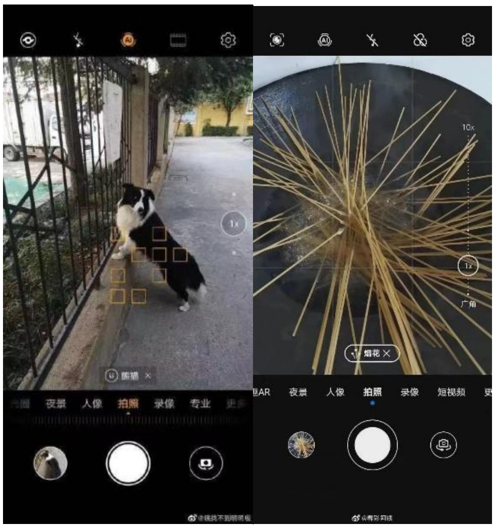
There are already some solutions for PPCC data doppelgangers: one is splitting training and test sets based on individual chromosomes, which could be helpful but really difficult to do practically because it relies on the prior knowledge and good-quality contextual data; another is removing the PPCC data doppelgangers using PPCC outlier detection package, doppelgangR [6][7], which is also getting troublesome when work with small data sets. There are also some recommendations provided by Wang, L. R., Wong Limsoon and Goh, W. W. B., pointing that it’s possible to use metadata to perform the cross-checks, stratify data into different similarities, and perform extremely robust independent validation checks as many as possible to avoid performance inflation.

## Data doppelgangers are not unique to biomedical data

Data doppelgangers mean there are many similar data samples appearing which could cause performance inflation, especially when many data doppelgangers appear in the test dataset. And this not only happens in biomedical data, it also happens in computer vision (CV) area. For example, according to the YOLOv4 paper [7], its accuracy is over 94% but sometimes you can still get the following results [6] – it can’t provide a correct anchor box for the objects.



And the recognition function of the camera app in Huawei phones, sometimes works wrong, either [8]:



I think although the YOLO already performs very well in object detection, the test result could be overstated because of some data doppelgangers in their test set – I found that it could easily recognize the shadow as a person at night when I was in an internship doing a project about object detection, which is not supposed to happen with such high accuracy.

## Methods to detect data doppelgangers

Although there are already many methods trying to detect data – for example, ordination methods (e.g., PCA), and embedding methods (e.g., t-SNE), which project the sample into a reduced-dimensional space and do the detection. These methods are unfeasible because data doppelgangers are not necessarily distinguishable in a lower-dimensional space [2]. PPCC method uses Person’s correlation coefficient, which formula is shown as below:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

Although it's widely used to detect data doppelgangers, it can only detect linear similarity, and since it's a pairwise calculation, which is very computationally expensive. I think maybe there are some better ways to calculate the similarity without reducing the dimension of space, like cosine similarity (formula (2)) (or we say "inner product similarity").

$$sim(a,b) = \frac{A \cdot B}{|A||B|} \tag{2}$$

The latter one is widely used in the NLP area, and the most famous method related to this is the attention mechanism [9], which computes the similarity considering all the features in the original high-dimensional space. In this way, we can detect the non-linear similarity more accurately owning the advantages of matrix computation.

To avoid doppelganger effects, we can first filter the data in the same class from different patients according to the metadata; then calculate the similarity using PPCC, cosine similarity or inner product similarity, and remove the data doppelganger in the test set if the dataset size is big enough; otherwise, don't put those data whose similarity with any training data greater equal to $\min\{maximum\_similarity\_in\_negative\_data \times 1.2,\ 1\}$ (the hyperparameter 1.2 here means we don't expect the highest negative sample's PPCC value higher than 0.83), so that we don't remove any useful data and also evaluate more objectively. We can also consider the n-fold cross-validation for evaluating the model performance. If the dataset size is still too small, we can consider some data argumentation methods (e.g., pooling, random erasing, flipping, crop, Mosaic data argumentation, et.). But if still cannot solve the problem, the only way is to get more data.

## Conclusion

Doppelganger effects appear in many different areas like object detection, face recognition, biomedical data area, and so on. As the data-driven model exists, there could be doppelganger effects existing. In this case, data qualification and size are really important. To get more data, we can use BERT, U-Net or some other deep learning models to assist us according to the specific cases. To avoid doppelganger effects, it's better to carefully split the train, validation and test sets – don't put those data doppelgangers in validation or test sets. Last but not least, take validation sets as many as possible, testing multiple times and finally take the average as the report result.

## Reference

[1] O'Meara, James Stephen. "The Doppelgänger Effect." Astronomy.com, 24 Aug. 2015, https://astronomy.com/magazine/stephen-omeara/2015/08/the-doppelganger-effect.

[2] Wang, L. R., Wong, L. & Goh, W. W. B. (2022). How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today, 27(3), 678-685. https://dx.doi.org/10.1016/j.drudis.2021.10.017.

[3] Q. Sheng, Y. Shyr, X. Chen, DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis, BMC Bioinform 15(2014) 323.

[4] Shacklett, Mary E., and Peter Loshin. "What Is MD5 (MD5 Message-Digest Algorithm)?" Security, 1 Aug. 2021, www.techtarget.com/searchsecurity/definition/MD5.

[5] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, M. Birrer, The Doppelganger effect: hidden duplicates in databases of transcriptome profiles, J Natl Cancer Inst 108 (2016) djw146.

[6] Diwan, T., Anirudh, G. & Tembhurne, J.V. Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tools Appl (2022). https://doi.org/10.1007/s11042-022-13644-y

[7] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[8] 泡沫酱. "那些沙雕的 AI 智能识别……" 微信公众平台, 8 Feb. 2021, https://mp.weixin.qq.com/s?__biz=MzU4NTE0ODYwNQ==&mid=2247490972&idx=1&sn=d4e50760ed4009deaf399eb73c6b0f35&chksm=fd8fac4ecaf8255852e153baffba39eafb8100321168c43aa38eab6cf2d32f4ace7ab1910059#rd.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.