# Report of Question 3

## CONTENT

# 1. Definition and Uniqueness of doppelganger effect

Machine Learning models are currently becoming more popular for developing drugs as they are able to identify the possible objectives more quickly [1]. Before the model can be put into use, it must be validated first. However, whether the result of validation is reliable might be influenced by an effect called doppelganger [1], which means that if the similarity of the validation data set and training data set is very high, the model may perform well even if the quality of training is poor. Hence, doppelganger effect might have seriously bad impact on the reliability of system that use this ML model.

In fact, this effect is not unique to biomedical data used for drug development. It also exists in the field of facial recognition, which also negatively affects the reliability of the system. In a facial recognition system, a person can be possibly recognised as some other man by using his/her own biological features. This is called the *zero-effort impostor*. It has already been proved that related people, such as brothers or sisters in blood, can succeed in *zero-effort impostor* with a higher probability [3], [4].

In the figure below, it can be seen that these people have some of their facial characteristics in common with each other, even though they are not related in blood. As a result, it is also possible for them to make a successful *zero-effort impostor* attempt. Therefore, the doppelganger effect is at least not unique to the biomedical data of drug development. Whereas, although the data set used in facial recognition systems does not belong to medicine, it is still biological information.

Figure 1. Example of doppelganger effect in other type of data [2]

Theoretically, doppelganger effect will occur as long as the reliability of Machine Learning model can be potentially affected due to the similarity of data. Thus, it is reasonable and rational to argue that doppelganger effect can possibly exist in the systems that use Machine Learning models in other research fields.

## 2. Possible solutions

Typically, there are two ways to solve or avoid this problem.

The first solution is quite straightforward and easy to implement. To avoid the doppelganger effect, the validation data set should be enlarged so that this effect can be avoided with a higher probability. However, its drawbacks are obvious as well. It can be extremely time-consuming if the size of validation data set is too large, and it is hard

to decide the exact amount of data that is required. Therefore, this solution can be implemented easily but not efficient and effective.

The second way is to apply the regularisation method. One way to do so is to add a norm penalty item $\Omega(\theta)$ to the target function, which can decrease the capacity of the sample. For instance, if the target function is

$$f(\theta;\ X, y),$$

after regularisation it becomes

$$g(\theta;\ X, y) = f(\theta; X, y) +\ \alpha\Omega(\theta),$$

where $\alpha$ is a real number greater than or equal to 0, and it denotes the relative contribution of the norm penalty item. The larger $\alpha$ is, more regularisation it has. In addition to norm penalty, there are also some other regularisation methods such as L1 and L2 regularisation. By contrast with enlarging the validation data set, this approach seems more efficient and reliable.

To summarise, to tackle the doppelganger effect, either enlarge the validation data set or apply regularisation method.

# 3. Bibliography


[1] Wang, L.R., Wong, L. and Goh, W.W.B. (2021). How doppelgänger effects in biomedical data confound machine learning. Drug Discover Today. https://doi.org/10.1016/j.drudis.2021.10.017 [Accessed on Feb. 24th, 2022].

[2] Rathgeb, C., Fischer, D., and Drozdowski, P. et al. (2022). Reliable Detection of Doppelgangers based on Deep Face Representations.

https://deepai.org/publication/reliable-detection-of-doppelgangers-based-on-deep-face-representations [Accessed on Feb. 27th, 2022].

[3] Pruitt, M.T., Grant, J.M., and Paone, J.R. et al. (2011). Facial recognition of identical twins. In Int'l Joint Conf. on Biometrics. pp. 1-8.

[4] Phillips, P.J., Flynn, P.J., and Bowyer, K.W. et al. (2011). Distinguishing identical twins by face recognition. In Int'l Conf. on Automatic Face Gesture Rcognition. pp. 185-192.