Thesis Reading Report of How Doppelgänger Effects in Biomedical Data Confound Machine Learning

A. The basic information of this paper:

Li Rong Wang, Limsoon Wong and Wilson Wen Bin Goh. *How Doppelgänger Effects in Biomedical data Confound Machine Learning.* Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.017.

B. The question that this paper wants to solve:

When using ML, training and test data sets should be exported independently when evaluating the performance of a classifier. However, separate training and test sets can still produce unreliable validation results. When classifiers incorrectly perform well due to the presence of data doppelgänger, we say that there is an observed doppelgänger effect. So this paper wants to investigate the properties of doppelgänger data and propose an improved method to identify doppelgänger.

C. What solution does this paper put forward?

The paper provides a method that use the Pairwise Pearson's Correlation Coefficient (PPCC) to identify potential functional doppelgänger from constructed base scenarios. And the author used the renal cell carcinoma (RCC) genomic data obtained by Guo et al from NetProt software library to construct the baseline scenario. Based on the PPCC distribution of the effective scenario, the authors identified the PPCC data doppelgänger by

comparing the PPCC distribution of the negative and positive scenarios. The authors next examined the PPCC distribution between the same and different tissue pairs (Figure 2b). After identifying PPCC data doppelgänger in RCC, the authors also investigated their impact on the validation accuracy of different random training classifiers (training classifiers are ML models that "learn" from training data). Finally, different machine learning (ML) models were used to predict the performance of the training validation set with different amounts of Pearson correlation coefficient (PPCC) data doppelgangers. The ML models evaluated included k-Nearest Neighbor (kNN) (a), Naive Bayes (B), decision tree (C), and logistic regression (D) models.

D. The motivation that this paper proposes the solution:

The paper is motivated to propose this solution for the following reasons:

1. Although the biomedical data community appears to be increasingly aware of the problem of this data doppelgänger, procedures for eliminating or minimizing similarities between tests and studies are unproven.

2. Earlier studies of similar problems cannot solve the problem.

E. The disadvantage of the proposed solution:

Doppelgänger effect is eliminated when all PPCC data doppelgängers are placed in the training set. This provides a possible way to avoid the doppelgänger effect. However, binding PPCC data doppelgängers to a training set or validation set is a suboptimal solution. In the former, when

the size of the training set is fixed (thus, each contained data results in a less similar sample being excluded from the training set), the model does not generalize well due to lack of knowledge. In the latter case, spectacular winner-takes-all scenarios can occur (doppelgängers are either be predicted correctly or wrongly).

F. Previous methods to solve the problem and shortcomings:

1.One logical approach to data recognition by doppelgänger is to use either a sorting method (e.g., principal component analysis) or an embedding method (e.g., T-SNE), along with a scatter plot, to see how the sample is distributed in a dimensional-reduction space. However, we find that this approach is not feasible because the data doppelgängers are not necessarily distinguishable in dimensionality reduction space.

2.DupChecker, identifies duplicate samples by comparing their CEL files' MD5 fingerprints. The same MD5 fingerprints indicate that the samples are duplicated (essentially duplicates, thus indicating a leak problem). Therefore, dupChecker does not detect the real data doppelgängers, which are independently exported samples that are accidentally similar.

3.Another measure, the pairwise Pearson correlation coefficient (PPCC), captures the relationship between sample pairs of different data sets. An unusually high PPCC value indicates that a pair of samples constitutes the PPCC data doppelgängers (note that it is impossible to determine which of the pair of samples is original). Although this original PPCC paper was

sound and intuitive, its main limitation was that it never ultimately linked PPCC data doppelgängers to their ability to confound ML tasks (i.e., to have functional effects and thus act as functional doppelgängers). When reanalyzing their data, we also realized that the doppelgängers they reported were actually the result of leaks (between sample duplicates) and therefore did not constitute true data doppelgängers.

4.In studies using PPCC outlier detection kit (doppelgangR) to identify doppelgängers, PPCC data doppelgängers can be deleted to mitigate its impact. However, this approach is not suitable for small data sets with a high proportion of PPCC data doppelgängers, such as RCC, because deleting PPCC data doppelgängers reduces the data to an unusable size.

F. Methods to guard against doppelgänger effects:

While deleting data doppelgängers directly from data has proved difficult to implement, we still need to guard against the doppelgängers effect. There are three methods that this paper proposed:

First method is to carefully perform cross-checking using metadata as a guide. The second one is to implement data layering. The third one is to perform very robust independent validation checks involving as many data sets as possible (divergent validation).

G. Own understanding of this paper:

Firstly, for the doppelgänger effect, I think it is not a unique effect only existing in biomedical data. Because when you are using ML models in any area and when assessing the performance of a classifier, separate training and test sets can still produce unreliable validation results. Therefore, there is an observed doppelgänger effect. I found an example that Lookalikes, a.k.a. doppelgängers, increase the probability of false matches in a facial recognition system, in contrast to random face image pairs selected for non-mated comparison trials.[1]

After reading this paper, I think to avoid doppelgänger effect in using machine learning models for health and medical science is make sure we use the three methods which are used to guard against that effect mentioned in this paper. The situation that the method PPCC that was proposed in this paper cannot work is there is a small data sets with a high proportion of PPCC data doppelgängers. So, if we can gather more data and scale up the data set, then the proportion of data doppelgängers could be decreased, and we can apply the PPCC method to delete data doppelgängers and will not reduce the data to a unusable size.

H. Reference:

[1]: https://ieeexplore.ieee.org/document/9548306.