Doppelgänger Effects in Biomedical Machine Learning

1. Introduction

Data Doppelgängers are a common occurrence in biomedicine (Wang et al., 2021), resulting from chance or other factors. When highly similar but independent data sets exist, they can be considered Doppelgängers. However, these data sets should not have a Doppelgänger effect unless the model performs well on the training and validation sets but fails to generalize to new data. When Doppelgängers generate an effect, it is known as functional Doppelgängers.

In recent years, the phenomenon of the Doppelgänger effect has garnered considerable attention due to its far-reaching influence on the application of machine learning (ML) techniques, as noted by Whalen et al. (2022). While this effect has been recognized as a potent force that shapes decision-making processes, its impact is not always beneficial, as highlighted by Eid et al. (2021). Of particular concern is the potential for the Doppelgänger effect to bias model performance. Failure to identify and remove such biases can result in a misguided and overly simplistic training process, where the model primarily learns from biases specific to the training dataset, thereby failing to generalize effectively across new datasets. This can have serious consequences for the reliability and accuracy of the model's output, as well as for the broader application of ML techniques. It is imperative to make efforts to avoid Doppelgänger effect, as doing so is of paramount importance.

This report explores potential methods to mitigate the Doppelgänger effect in the three stages of training: data preprocessing, model selection, and validation. We will also discuss how the Doppelgänger effect compares to other challenges in machine learning, such as data leakage, and the importance of addressing this issue in future research.

2. Literature Review

The field of biomedical data, which encompasses diverse data types such as clinical, genomic, imaging, proteomic, metabolomic, electronic health, and wearable device data, has witnessed significant advances in recent years with the application of ML techniques (Baldi, 2018). The performance of ML models is typically evaluated by various metrics, with an emphasis on their ability to generalize to novel data (Zou et al., 2019; Rajkomar et.al, 2018). However, achieving high performance on such metrics, such as accuracy, may not necessarily be desirable in the context of biomedical data, as it can be subject to a phenomenon known as the Doppelgänger effect. This effect refers to the potential for consistent excellent model performance on training and validation datasets that have high similarity, but poor generalization to new data (Wang et al., 2021).

Wang et al. (2021) proposed three related concepts to describe the Doppelgänger effect: data doppelgänger, doppelgänger effects, and functional doppelgänger. Data doppelgänger refers to data with high similarity that can potentially generate the Doppelgänger effect, rendering it a candidate for functional doppelgänger. However, not all data is susceptible to this effect. This raises the important question of whether high model performance on biomedical data can be deemed reliable.

In the field of predicting enhancer-promoter interaction, Cao et al. (2019) utilized ML methods and achieved high evaluation metrics, with an F_1 score value of 0.9. However, re-checking the dataset revealed a strong dependency between the training and validation sets, which was also observed in other types of datasets such as consecutive images of brain tumors obtained by optical coherence tomography (OCT). Ignoring the properties of these data and the high overlap of training and validation sets resulted in performance inflation, which can be seen as an instance of the Doppelgänger effect (Irmark, 2021; Sadad et al., 2021). These experiments highlight the drawbacks of the Doppelgänger effect, which can impede objective analysis of results and lead to the weakening of interpretability and overly optimistic estimates of the potential applications of discoveries.

3. Potential Methods for Avoiding Doppelgänger Effects

To mitigate the negative influences of Doppelgänger effects, it is necessary to identify potentially effective strategies to circumvent them. The pipeline for ML based on health and medical science data generally includes three parts: data preparation, model training, and validation. The potential methods for weakening or estimating doppelgänger effects can be carried out from these three aspects.

Data preparation. Identifying data doppelgänger with strong prior knowledge and carefully adopting data splitting methods facilitate the achievement of the goal in this step. As suggested by Whalen et al. (2021), it is necessary to use the graphical technique to visualize the correlations or interactions. Specifically, nodes can represent different biological entities, such as genes, proteins, and chemical materials, while edges can present relationships. A dendrogram is also probably another effective tool for analysis to visualize proteins from the same ancestor. In addition, correct data splitting should be the same as the Figure. 1, which refers to taking full consideration one single sample including itself and relative products and relationships.

Model training. Selecting the proper model to describe more information for biomedical data and using training strategies are two feasible approaches. Wang et al. (2021) outlined the performance of four models for classification tasks on the renal cell carcinoma data by comparing different numbers of data doppelgänger. However, the accuracy of randomly deriving feature set centered at 0.5 without and with all data doppelgänger, which probably reflects the weak ability to classify cancer tissue only by mining shallow relationships of data through the model had been trained fully. The situation implied that a more effective model, such as deep learning models to mine deep information, should be included. Therefore, selecting a proper model based on data complexity probably can avoid the doppelgänger effect by exploring different levels of information. Increasing or decreasing the weight in the constraint as the training strategy is also a considerable method, which probably has potential for artificially mitigating negative effects but existing hidden dangers with poor generalization and biological interpretability (Whalen et al., 2021).

Validation. Cross-validation is a widely recognized means of validation. However, due to the properties of structure-dependent biomedical data including temporary, space, group, and gene, block cross-validation can be used instead proposed by Roberts et al (2016). What’s more, external validation from multi-sources and multi-center data is essential, which assists to recheck the original data and evaluate the actual generalization ability to unseen data.

4. Discussion

The phenomenon of doppelgänger effects in biomedical data has not received sufficient attention in previous research, leading to potential biases and incorrect conclusions. It is crucial to recognize the negative impact of this phenomenon and take steps to prevent it.

The definition of data doppelgänger highlights two key attributes: independence and high similarity between datasets. Independence is particularly important in biomedical data as it is easy to draw optimistic conclusions without recognizing dependencies, which violates the assumptions of machine learning models based on independence. Therefore, doppelgänger effects can be considered a form of data leakage. Evaluation of similarity using methods such as the Pearson correlation coefficient and Euclidean distance is necessary but not sufficient to demonstrate doppelgänger effects, which requires cross-validation, checking the data using multiple calculation methods to account for independent model performance and external validation from multi centers.

Doppelgänger effects are not limited to genomics, but are present in other applications of machine learning, such as image analysis. Overcoming this obstacle requires identifying functional doppelgängers to derive more reliable results and prevent overly optimistic conclusions based on big data. Given the complexity of biomedical data, data scientists must possess a deeper knowledge of the field and apply multiple methods, including visualization and quantitative analysis, to cautiously check the data during training progress.

5. References

Baldi P: **Deep learning in biomedical data science**. *Annual review of biomedical data science* 2018, **1**:181-205.

Cao F, Fullwood MJ: **Inflated performance measures in enhancer-promoter interaction-prediction methods**. *Nat Genet* 2019, **51**(8):1196-1198.

Eid FE, Elmarakeby HA, Chan YA, Fornelos N, ElHefnawi M, Van Allen EM, Heath LS, Lage K: **Systematic auditing is essential to debiasing machine learning in biology**. *Commun Biol* 2021, **4**(1):183.

Irmak E: **Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework**. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* 2021, **45**(3):1015-1036.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M *et al*: **Scalable and accurate deep learning with electronic health records**. *NPJ Digit Med* 2018, **1**:18.

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W: **Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure**. *Ecography* 2017, **40**(8):913-929.

Sadad T, Rehman A, Munir A, Saba T, Tariq U, Ayesha N, Abbasi R: **Brain tumor detection and multi-classification using advanced deep learning techniques**. *Microsc Res Tech* 2021, **84**(6):1296-1308.

Wang LR, Wong L, Goh WWB: **How doppelganger effects in biomedical data confound machine learning**. *Drug Discov Today* 2022, **27**(3):678-685.

Whalen S, Schreiber J, Noble WS, Pollard KS: **Navigating the pitfalls of applying machine learning in genomics**. *Nat Rev Genet* 2022, **23**(3):169-181.

Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A: **A primer on deep learning in genomics**. *Nat Genet* 2019, **51**(1):12-18.

6. Appendices

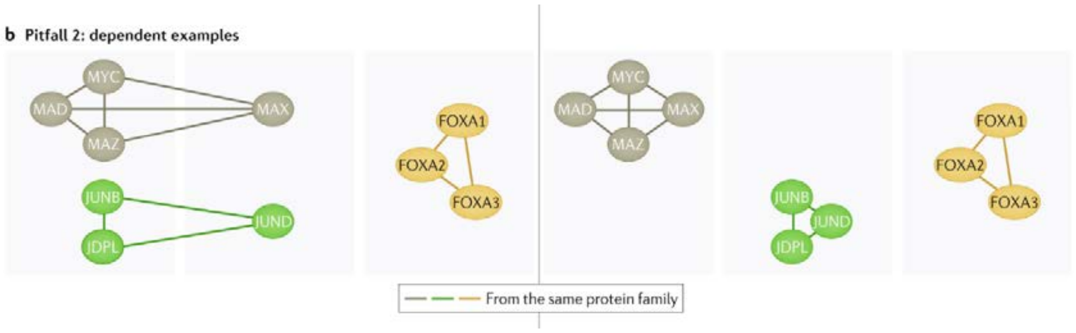

Figure 1. The correct way to split data suggested by Whalen et al. (2022) .