Question3

This paper focuses on the problem of doppelganger effects in biomedical data in machine learning. Machine learning is widely used in biomedical data research, such as drug discovery and biomedical imaging, which increases the efficiency of finding the potential target. To evaluate a machine learning model, cross-validation is a common way. However, although the independence of the training and validation set is satisfied, the evaluation could still be unreliable because of the doppelganger effects. The Doppelganger effect means that the independent training and validation data are highly similar, which always makes the good performance of the model regardless of the actual effect. And this kind of data doppelganger is called the functional doppelganger.

There are several methods developed for the identification of data doppelgangers, such as ordination and embedding methods. Although not perfect, the pairwise Pearson's correlation coefficient (PPCC) is relatively reasonable and could be used to identify potential functional doppelgangers. And using the renal cell carcinoma (RCC) proteomics data, a data doppelganger has been tested by this method. How severely the PPCC data doppelganger affects the validation accuracy of machine learning model performance was also explored. If data doppelgangers exist in both training and validation sets, the performance of the model is exaggerated. The more pairs of data doppelgangers, the more severe the performance is inflated. Moreover, different machine learning models show different influences by data doppelgangers.

Doppelganger effects can be avoided in the practice and development of machine learning models for health and medical science in the following ways. The author put forward three recommendations to reduce the effects.

The first recommendation is to perform a careful cross-check using meta-data as a guide. By identifying data doppelgangers, they can be put all on training or validation set to prevent the doppelganger effects between two sets. There still exist some problems: putting them on testing data will make the model less generalized since fewer features could be learned. Putting on validation data lets the doppelgangers either be predicted correctly or wrongly.

The second recommendation is to perform data stratification by stratifying data into strata of different similarities. Since the proportion is known, it is still possible to consider the real-world performance in specific strata.

The third recommendation is to perform extremely robust independent validation checks involving as many data sets as possible, which is a way to inform on the objectivity and generalizability of the model.

I think doppelganger effects are not unique to biomedical data. According to the definition of the functional doppelganger, data are similar in both training and validation sets and confound the machine learning model outcomes. Not only is the data in the biomedical area satisfactory. If the data collected have relatively similar features both in the training and validation set, no matter how the models are trained, the performance of the model will be excellent.

Reference

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://doi.org/10.1016/j.drudis.2021.10.017