Question 3 Report

This report is a reflection after reading the article. The report will be divided into the following three parts, first summarizing the findings of the given paper *How doppelganger effects in biomedical data confound machine learning*, second giving my thoughts about doppelganger effects, and the third purpose the potential solutions to reduce the doppelganger effects.

Literature Summary

Data doppelgängers occur when the training data and validation data are very similar to each other because of chance or otherwise. It causes models to perform well regardless of how they are trained. Functional doppelgangers are sample pairs that, when split across training and validation data results in the inflation of model performance. Both of them form doppelganger effects. The paper mentioned the phenomenon that data doppelgangers are abundant in biological studies, identify it with the example of PPCC measurement such that sample pairs with high PPCCs are also referred to as PPCC data doppelgangers, and also purposed possible methods to mitigate the doppelganger effect by identifying it before the training-validation split.

The study also noted that not all models are equally affected by this phenomenon. In particular, KNN and naive Bayes models tend to show a more direct relationship between the amount of doppelgangers and the level of performance inflation, which might be useful in data doppelgangers detection. The paper also offers three suggestions for minimizing the impact of data doppelgänger effects:

1. Carefully verify the data using meta-data as a reference.
2. Divide the data into stratified groups.
3. Perform thorough independent validation checks using a variety of data sets (divergent validation)

My Thoughts about Doppelganger Effect

The doppelganger effect is a universal problem in machine learning because data doppelgangers might exist in any kind of data. Because there exists a situation that the data are quite similar to each other. It might occur more frequently in biological data modeling because one species shares a large proportion of similarity in genes as for humans, only 0.1% of DNAs are different. The high similarity will lead to data doppelgangers. Here are some examples.

1. Imaging data: In medical imaging, doppelganger effects can occur when two individuals have similar anatomies or when an individual has a rare or unique anatomical feature that is shared with another individual. For example, two individuals with extremely similar brain structures may be misclassified as the same person by a machine learning model trained on brain imaging data.
2. Gene sequencing data: Doppelganger effects can also occur in gene sequencing data when two individuals have very similar genetic profiles. This can lead to confusion when analyzing genetic data for disease risk or ancestry. Or there already happened that whole-genome analysis of cancer samples is a common practice, and researchers often share or reuse these specimens in subsequent studies. However, the presence of duplicate expression profiles in public databases

can have an impact on the accuracy of reanalysis. [1]

3. Metabonomics data: When two people have identical metabolic profiles, there may be Doppelganger effects in metabonomics data, which studies tiny molecules in biological systems. When examining data on metabolism or drug metabolism, this may cause confusion.

I encountered this problem when building my own image classifier at that time I did not know the term doppelganger effect. Since the image dataset used for training and validation of the CNN model is a standard dataset, the images here are clear white background images with the target object in the center of the image. After training, the training accuracy and validation accuracy of this model was above 98%. However, when I used my own pictures (sight different from stander pictures) as input, the accuracy of the model decreased significantly.

Also, I think the doppelganger effect is a little bit similar to the overfitting problem in machine learning which occurs when a model is overly complex and has learned patterns in the training data that do not generalize well to new data. For the model that encounters overfitting problems, it will perform well on training data but poor in validation or test data. It indicates that the model is not robust enough. For the doppelganger phenomenon, the validation set is too similar to the training set so validation cannot indicate the model performance like validating it with the original training data. The model might perform badly on new or unseen data. There are some overlapping features between overfitting and doppelganger since both of the models already learned enough from the training data. Maybe the possible solution will be similar to handling the overfitting problem.

Potential Solutions

The doppelganger detection based on deep face representations [2] proposed a machine learning detection system for doppelganger images in facial data. They trained the classifier with generated doppelganger image pairs utilizing face morphing techniques by differential detection. To extract differences between facial features. Inspired by this paper, we can train a classifier to detect the data doppelganger in biological data. But this method needs extra prior knowledge and benchmark datasets to train a model fitting special needs.

Doppelganger effects can be avoided in the practice and development of machine learning models by following methods:

1. Ensuring that data used to train machine learning models are representative and diverse can help reduce the likelihood of doppelganger effects occurring in the data. For example, using an independent training and validation data set to train the model.
2. Using data augmentation techniques: Data augmentation involves creating synthetic data that is similar to the real data, but slightly different. This can help reduce the impact of doppelganger effects on machine learning models. The Generative adversarial Network is suitable for this task to generate fake data that is similar to real data.
3. Incorporating additional information into the machine learning model: For example, in the case of medical imaging data, incorporating information on patient demographics or medical history can help distinguish between individuals with similar anatomies.

4. Using multiple machine learning models: Training and evaluating multiple machine learning models on the same data can help identify and mitigate the impact of doppelganger effects. Like the study using KNN, naïve Bayes, decision tree… By comparing different model performances, the doppelganger can be detected. [3]

To sum up, doppelganger effects can appear in any kind of data and are not just found in biomedical data. In order to prevent doppelganger effects in the use and development of machine learning models for health and medical science, it is crucial to apply data augmentation techniques, incorporate extra information into the model, and use several machine learning models.

References:

[1] New Cancer Research Findings from Massachusetts General Hospital Outlined (The Doppelganger Effect: Hidden Duplicates in Databases of Transcriptome Profiles). (2017). Obesity, Fitness, & Wellness Week, 1659–.

[2] Rathgeb, C., Fischer, D., Drozdowski, P., & Busch, C. (2022). Reliable detection of doppelgängers based on deep face representations. IET Biometrics, 11(3), 215–224. https://doi.org/10.1049/bme2.12072

[3] Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022). Doppelgänger spotting in biomedical gene expression data. iScience, 25(8), 104788–104788. https://doi.org/10.1016/j.isci.2022.104788

[4] Wang, L. R., Fan, X., & Goh, W. W. B. (2022). Protocol to identify functional doppelgängers and verify biomedical gene expression data using doppelgangerIdentifier. STAR Protocols, 3(4), 101783–101783. https://doi.org/10.1016/j.xpro.2022.101783