Question 3: Report on Paper "How Doppelganger Effects in Biomedical Data Confound Machine Learning"

1. Understanding of Data Doppelgangers Effects

Data doppelgangers are described to be samples that appear to be similar in terms of their features. When they appear in both the training and validation datasets, they can cause the machine learning model to appear to perform well through an inflated validation result.

An example from the paper of such an effect is in prediction of protein function. Such a model would predict a protein's function based on the structural sequence, where similar structural sequences would have similar functions. However, such a model will be unable to identify the proteins with similar functions that do not have similar structural sequences. As the "rule" of similar structure proteins having similar functions hold true for most of the proteins, this will cause the doppelganger effect of validation result inflation.

As stated in the paper, due to the abundance of data doppelgangers in biomedical and biology data, there is a need for more robust methods of detecting and mitigating doppelganger effects when validating a machine learning model.

2. Doppelganger Effects in Other Data Types

Based on my understanding on doppelganger effects from the paper, a possible example of data doppelganger that came to mind is in image classification of animal species that look very different during their baby and adult stage. Below are some examples of such animals:

Figure 1 The difference in physical appearances between a baby panda (left) and an adult panda (right).

Figure 2 The difference in physical appearances between a baby flamingo and an adult flamingo.

Being able to classify the species of these baby animals will probably require an expert with domain knowledge in animals. As most of these animals grow to resemble their adult forms within a short period of time, and spend most of their lives in their adult forms, the amount of images of their adult form would be far greater than their baby forms. For example, if the training and validation dataset both contain many images of adult pandas but few images of baby pandas, the classifier performance will experience inflation due to the doppelganger effect.

3. Suggested Methods of Reducing Doppelganger Effects

The paper reported the effectiveness of the pairwise Pearson's correlation coefficient (PPCC) method in detecting data doppelgangers, but also described the challenges faced by the various stated methods on reducing data doppelganger's inflationary effects on machine learning. Some of these methods discussed in the paper include:

1. Removing data doppelgangers

2. Splitting training and validation datasets based on prior knowledge of structural similarities of data doppelgangers

3. Data trimming by only removing variables that contribute much to the doppelganger effect

Taking the example of protein function prediction, I would like to suggest the following method:

1. Upon identification of data doppelgangers (proteins that have similar sequence structures and functions), identify also the proteins with the similar function that do not have similar structures.

2. Assign different data labels to these two groups.

3. As there will be far more data doppelgangers, class weights should be assigned during model training to mitigate the effects of class imbalance

4. For validation, use a stratified approach so as to get a more accurate representation of the model performance.

4. Conclusion

In this report, I have attempted to explain my understanding of data doppelgangers and the doppelganger effects in affecting machine learning models in biomedical and biology applications. From my understanding on this topic, I have suggested an example of a possible doppelganger effect in image classification and also proposed a method to mitigate the doppelganger effect using the example of protein function prediction.

The suggested method for mitigating the doppelganger effect would still require prior knowledge on the nature of proteins, their structure and functions. Therefore, domain knowledge in the field that machine learning is applied to is important to ensure doppelganger effects can be mitigated.