**Question 3**

The paper highlights several benefits of incorporating machine learning (ML) in the process of drug discovery. One key advantage is the acceleration of drug development while simultaneously reducing the associated costs. Notably, ML has been successfully used to identify potential drugs for COVID-19 treatment, which have advanced to clinical trials. In ML, a classifier needs to be trained to recognise patterns in data and make predictions for unseen data based on those patterns (Fumo, 2017). Therefore, it’s important for classifier to accurately categorize new data based on their distinctive features. One challenge that could occur in datasets is referred as the “doppelganger effect (DEs).” It’s a situation where classifier falsely performs well because of the presence of data doppelgangers, which are samples in a validation set that are similar to the samples in the training set but are independently derived, not replicates. Data doppelgangers that cause DEs are called functional doppelgangers (FDs), which can create confusion in ML task and compromise the accuracy of model performance (Wang et al., 2022). For instance, this paper highlighted the protein function prediction models assume that structurally similar proteins have similar functions, but this may not always be true, leading to inaccurate prediction.

The issue of data doppelgangers is particularly prevalent in biomedical data because of the nature of the data itself, including the complexity and variability of the biological processes (Yang et al., 2020). In genomics, there are many gene families with similar sequences. biological data contains information on the structure and function of biological molecules. However, since many molecules or compounds have similar structures or properties, making it difficult to accurately identify the relevant features or distinguish between different molecules or compounds. Another overwhelming feature of biomedical data is it consists of tens of thousands of features, but many of which may be weakly related or completely irrelevant to the disease (Zhang & Cao, 2019).

While data doppelgangers are frequently encountered in biomedical data, the doppelganger effect is not unique to biomedical data and this phenomenon can arise in any dataset. In general, presence of FDs can lead to erroneous conclusions in data analysis and result in inaccurate predictions and incorrect decision-making. If FDs are not accurately identified and addressed, there will be significant consequences especially in the biomedical field. It may lead to wrong diagnosis or treatment, which could be detrimental to the patients.

The paper suggested that to avoid the unfavourable effects of data doppelgangers, potential doppelgangers should be identified before performing the training-validation split. This approach can help to ensure that the trained machine learning model is more robust and reliable. However, despite increasing awareness of these issues, procedures for eliminating similarity between test and training date before splitting training and validation data are not standard practice for biomedical datasets.

Since data doppelgangers can have undesirable impacts, preventing or minimizing their presence in datasets is an important step in ensuring the accuracy and reliability of ML algorithms. Here are a few recommended strategies that can be implemented to mitigate the doppelganger effects:

1. Data processing: Removing duplicates from the dataset can help to remove both “true duplicates” and “false duplicates”. The objective of is to reduce the risk of data doppelgangers in downstream analysis. DoppelgangR in R package can be used to identify duplicate samples between and within data sets.
2. Clustering classification: It can be useful to group similar data points together. This can facilitate the identification and removal of potential data doppelgangers. In short, it’s to ensure that training and test samples are not duplicates or samples of high similarity.

Mentioned strategies can be useful if data sets are large. Since completely removing duplicates can result in loss of valuable information. However, in real situation, it’s hard to obtain enough biological samples for research. Firstly, the collection and preparation of biological samples for sequencing can be expensive and time-consuming. Secondly, the availability of patients who are willing to provide biological samples can be limited. Finally, laboratory errors can occur, which can result in the wastage of samples. In return, biological sample size can be utilized for ML may be insufficient. Having enough data sets and removing potential data doppelgangers can be one way to avoid the DEs in practice. Thus, taking appropriate precautions and fostering collaboration between researchers and experts in relevant field to ensure best practices are followed can be crucial. Academic researchers who have expertise in machine learning can collaborate with more hospitals or companies that specialized in data collection and processing or establish collaborations across borders. By pooling their resources and expertise, more useful datasets are obtainable to develop a robust trained model for health and medical science.

A recent study from Wang et al., 2022 reported that a newly developed R package called *doppelgangerIdentifier* was developed to identify PPCC DDs. These are data doppelgangers identified through pairwise Pearson’s correlation coefficient. The study used microarray gene expression and RNA-Seq gene expression data to demonstrate the presence of DEs and successfully identified FDs using the PPCC DD identification method. This can help in preventing deletion of true duplicates and retaining meaningful datasets to improve the model’s robustness.

Perhaps, understanding how doppelganger effects emerge can help us in searching for a solution too. The principle of ML is to find patterns in data using algorithms and statistical models, and then use these patterns to make predictions or decisions about new data (Fumo, 2017), Once the patterns are identified, the ML model can then be used to classify new data or make prediction about future events based on the patterns it has learned. While data can represent a wide range of things and is useful in identifying relationships, but these predictions are all based on the algorithm behind. The algorithm may be overfitting to the training set, learning to identify the FDs rather than the true underlying patterns in the data (Ying, 2019). When new and independent data are fed into not well-trained ML model, the algorithm may struggle to apply its learning to these new samples and fail to perform the real expectation.

If we take human recognition technology as an example, image recognition using ML involves feeding a large dataset of labelled examples to an algorithm, which then learns the patterns of pixel associated with each label, identifying correlations and features in the pixel patterns, and allowing it to recognize similar patterns in new images. According to Rathgeb et al.,2022 it reported a significant increase of false match rates because of high similarity scores are produced by doppelganger image pairs. Figure 1 illustrated the example of doppelganger image pairs from the HAD Doppelganger database. It’s important to note that in this instance, the distinction may be more apparent to the human eye than to a machine.

A collage of people

Description automatically generated with low confidence

Figure 1. Doppelganger image pairs from the HAD Doppelganger database (Rathgeb et al., 2022)

In general, machine learning is still in development, but it is increasingly being utilized in various applications to aid in human progress and development. While models can be improved through practice, achieving a perfect model may not be possible. For instance, a model previously trained on biological samples from mice may not be entirely accurate when applied to other species. As such, it is crucial to rigorously test a model before utilizing it to ensure that the results are reliable. In the future, advancements in artificial intelligence may enable machines to bridge the gap to possess complex reasoning and critical thinking as humans and bring significant benefits to all mankind.

Reference：

1.Fumo, J. (2017, August 17). *Types of machine learning algorithms you should know*. Medium. Retrieved February 26, 2023, from https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

2.Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020). Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.01032

3.Wang, L.R., Choy, X.Y. and Goh, W.W. (2022) “Doppelgänger spotting in biomedical gene expression data,” *iScience*, 25(8), p. 104788. Available at: https://doi.org/10.1016/j.isci.2022.104788.

4.Zhang, B. and Cao, P. (2019) “Classification of high dimensional biomedical data based on feature selection using redundant removal,” *PLOS ONE*, 14(4). Available at: https://doi.org/10.1371/journal.pone.0214406.

5.Rathgeb, C. *et al.* (2022) “Reliable detection of doppelgängers based on Deep Face Representations,” *IET Biometrics*, 11(3), pp. 215–224. Available at: https://doi.org/10.1049/bme2.12072.

6.Ying, X. (2019) “An overview of overfitting and its solutions,” *Journal of Physics: Conference Series*, 1168, p. 022022. Available at: https://doi.org/10.1088/1742-6596/1168/2/022022.

7.Wang, L. R., Wong, L., & Goh, W. W. (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