Since the advent of computer, the rise of digital transformation has transformed the traditional ways of collecting and processing data at an unprecedented speed and scale. For example, in the biomedical domain, the electronic health records (EHRs) system documents patient data and the new generation of sequencing technologies enable the processing of billions of DNA sequence data in the laboratory.[1] Furthermore, one of the main enablers of digital transformation – artificial intelligence which is driven by the data has been applied in almost every sector to analyze data trends thus improving the decision making of the problem. Machine learning applications in the biomedical domain such as personalized medicine and cancer detection, have been increasingly used to assist biomedical scientists and medical professionals by summarizing and identifying patterns from a large amount of data in a shorter time. Several ML-identified drugs for coronavirus 2019 (COVID-19) treatment have also advanced into clinical trials in the drug development of fewer than two years time span. [2]

Despite the advances of machine learning in the biomedical domain, data quality concerns arise after the rapid growth and widespread use of databases. Data doppelgangers occur when independently derived training and test set are the same as each other. [2] It is one of the critical factors to train machine learning models, causing the models to perform well nevertheless the condition they are trained, so-called data doppelganger effect. It is quite difficult to detect in a large amount of biomedical data. Although biomedical professionals have been paying more attention to data doppelganger effects in these few years, there are no standard practices before model training are constituted for diminishing the similarity between test and training data.

To identify the presence of data doppelgangers, several logical approaches and measures are proposed such as ordination method, embedding methods coupled with scatterplots, dupChecker and pairwise Pearson's correlation coefficient (PPCC). [2] The basic design of PPCC is reasonable methodologically compared with other identification methods. It captures relations between sample pairs of different sets and validates the sample pair with a cut off threshold as shown in Figure 1. A pair of samples constitutes doppelgangers normally having extreme high PPCC value.

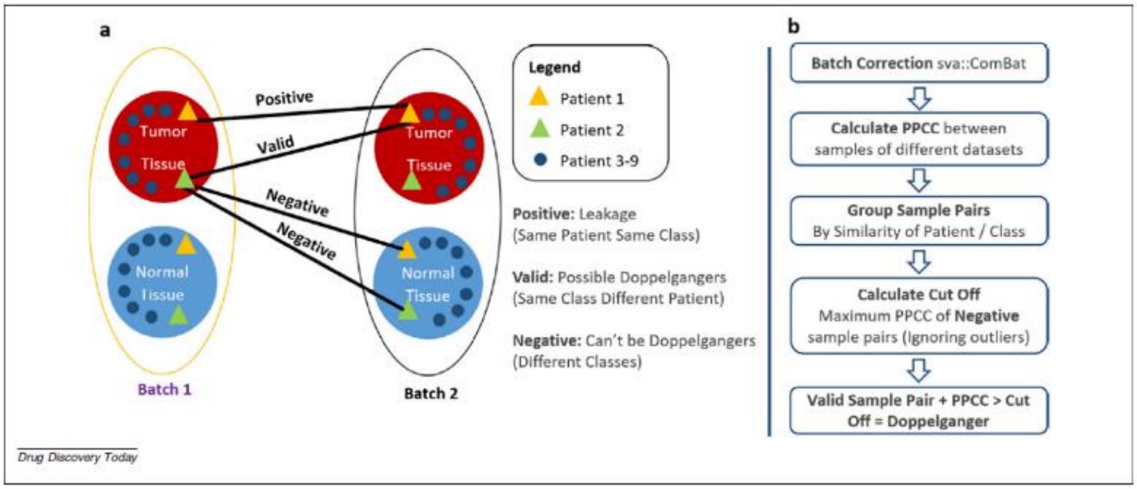
a
Positive
Valid
Negative
Negative
Tumor Tissue
Normal Tissue
Tumor Tissue
Normal Tissue
Batch 1
Batch 2
Legend
Patient 1
Patient 2
Patient 3-9
Positive: Leakage (Same Patient Same Class)
Valid: Possible Doppelgangers (Same Class Different Patient)
Negative: Can't be Doppelgangers (Different Classes)
b
Batch Correction sva::ComBat
Calculate PPCC between samples of different datasets
Group Sample Pairs By Similarity of Patient / Class
Calculate Cut Off Maximum PPCC of Negative sample pairs (Ignoring outliers)
Valid Sample Pair + PPCC > Cut Off = Doppelganger
Drug Discovery Today

Figure 1: Diagram illustrating the pairwise Pearson's correlation coefficient (PPCC) data doppelgänger identification method. [2]

After identifying the data doppelganger in the biomedical dataset, it is time to have an understanding of the confounding effect of data doppelganger. Besides the observed doppelganger effect where the classifier falsely performs well during training, the functional doppelganger effect is what we are most concerned about. It inflates the machine learning model performance in both the training set and validation set. Moreover, the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance. In the journal, the effects on the validation accuracy across various machine learning (kNN, Naïve Bayers, Decision Tree and Logistic Regression) were studied. KNN model was selected as it showed a linear relationship between non-doppelganger datasets and doppelganger datasets. As shown in Figure 2, the accuracy of the kNN model with data doppelgangers showed better performance than the kNN without doppelgangers and the inflationary effect is almost similar with the data leakage when the machine learning has 8 duplicates. The doppelganger effect was also observed by other researchers in 2016, where validation set hazard ratio was calculated with duplicates and it showed that 30% of duplication inflates the hazard ratio from 1.1 to 1.7. [3]

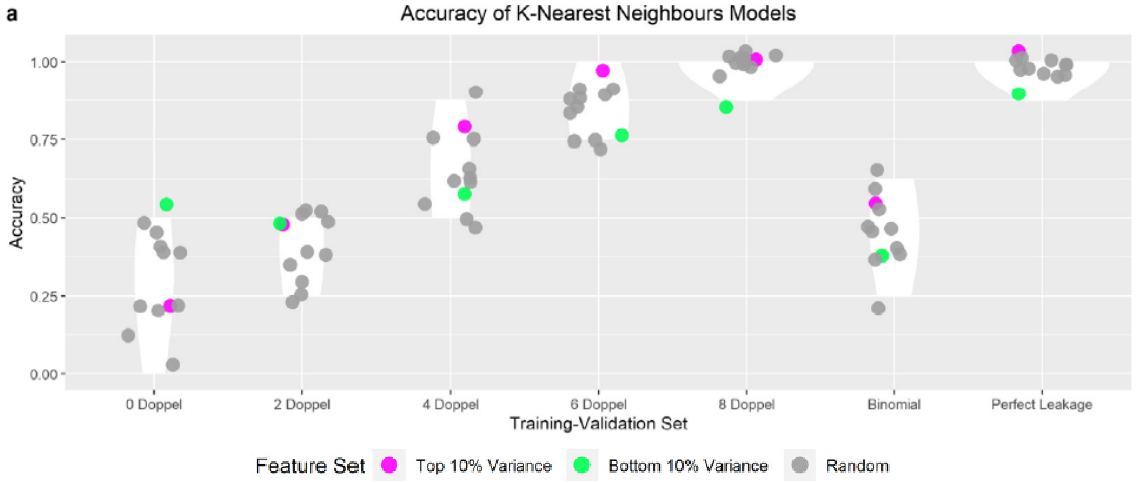

Figure 2: Accuracy of K-Nearest Neighbours Model on pairs of training-validation sets with varying numbers of PPCC doppelgangers in the validation sets. [2]

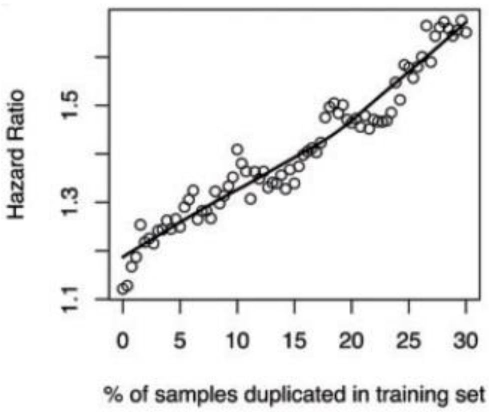

Figure 3: The doppelganger effect: hidden duplicates can inflate the apparent accuracy of predictive and prognostic models. [3]

Data doppelganger effects are not unique to biomedical data where this is a common problem affecting the machine learning model in every domain. Before further elaborating, some studies would address data doppelganger as duplicate data or data duplication and the definition of it is mostly context-dependent.[4] In general, the model does not generalize well when training with data duplication and might cause overfit problem. An overfitting model has a similar outcome as observed doppelganger effects where the model falsely performs well on the training data no matter the condition of the training set. However, its performance in production is anomalously poorly. There are certain conditions that data duplication helps model in training, but this is not under this discussion. Besides data duplication, data leakage happens when your training data contains information about the target, but similar data will not be available when the model is used for prediction. This leads to high performance on the training set and possibly even the validation data. The model will in turn predict poorly in production too. In other words, leakage causes a model to look accurate until you start making decisions with the model, and then the model becomes very inaccurate.

As the doppelganger effects provide an illusion of inflationary effect on the machine learning model, removing the data doppelganger in the dataset might be the way to diminish the effect. Theoretically, it is possible, but most of the data doppelgangers are uncharacterized and not well understood. So far, the method to dimmish the doppelganger effects are not generalizable and not robust to tackle in each condition. DoppelgangerR, a doppelganger detection package was used for identifying duplicates in the dataset. But it is not feasible to a small dataset with high portions of data doppelganger such as RCC. It is because the dataset would be reduced to a usable size. So, we should be aware of such data doppelgangers effects before the training validation split.

Although diminishing the data doppelganger effect has proven elusive, but we must do whatsoever to protect the model against them. There are three recommendations proposed in this journal. Firstly, performing careful cross-checks using meta-data in RCC to construct positive and negative cases. The score ranges of PPCC can be determined whether the data doppelganger exists in the data. Secondly, performing data stratification into strata of different similarities instead of evaluating model performance on test data. Hence, we can evaluate model performance on each stratum separately. Thirdly, performing comprehensive independent validation checks including divergent data sets as possible. [2]

Besides drug development, data doppelgangers exist in other data types like imaging and gene sequencing. In the imaging field, duplicate images can be found in the publication under different experimental conditions. The image duplication in biomedical research publications was observed by Bik, Casadevall and Fang. They visually screened through images from a total of 20621 papers in 40 scientific journals from 1995 to 2004. Based on their analysis, 3.8% of published papers contained duplicated images, with at least half features suggestive of deliberate manipulation.[5] This scenario is worsened in the publications containing photographic image data, especially in the Western blotting technique. 1 of the 25 publications was demonstrated duplicated images.

Those duplicated images under different experimental conditions are misleading future researchers when reading the previous studies. Moreover, publications with inaccurate data influence the quality and integrity of the scientific literature. And, it also reduces the efficiency of the scientific institutions by directing additional investigators to check for false leads or construct unsupportable hypotheses.

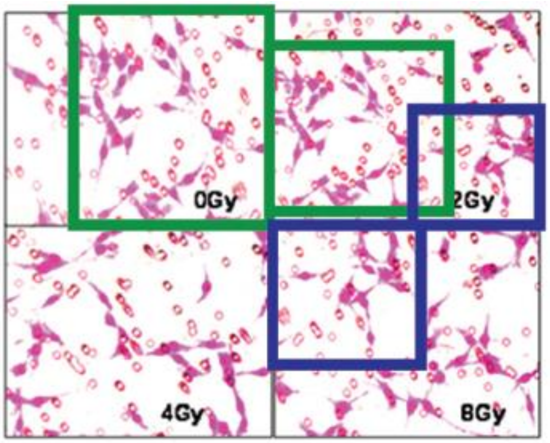

Figure 4: Although the panels represent four different experimental conditions, three of the four panels appear to show a region of overlap (green and blue boxes), suggesting that these photographs were obtained from the same specimen. [5]

In the gene sequencing field, duplication arises when two or more database records represent the same biological entity, a problem that has been known for over 20 years. In terms of definitions, data duplication in gene sequencing is defined by at least three relevant aspects of context; Different biological databases, different biological methods, and different biological tasks.

To mitigate the data duplicate issues, data curation comes in but it involves experts filtering the dataset manually by applying their experience and intuition. The process is tedious, inefficient and also expensive to work out. So, before going into the data curation, a supervised learning duplicate-detection method for pair of genomic database records was proposed. It serves as one of the precision methods with high performance and efficiency compared with the simple heuristics detection method. For instance, machine learning techniques are widely used for finding duplicate records in general databases, but only a few have been proposed for bioinformatics.

A great duplicate detection method, however, must reflect such diversity, and its performance must be tested in data sets with different duplicate types derived from multiple sources, where the test data is independent of the method. Hence, the proposed supervised methods for duplicate detection in sequence databases show substantial promise. The features for meta-data, sequence similarity, and quality checks on alignments achieved the best results, especially in meta-data features. It has the potential to be used to identify and filter distinct records.

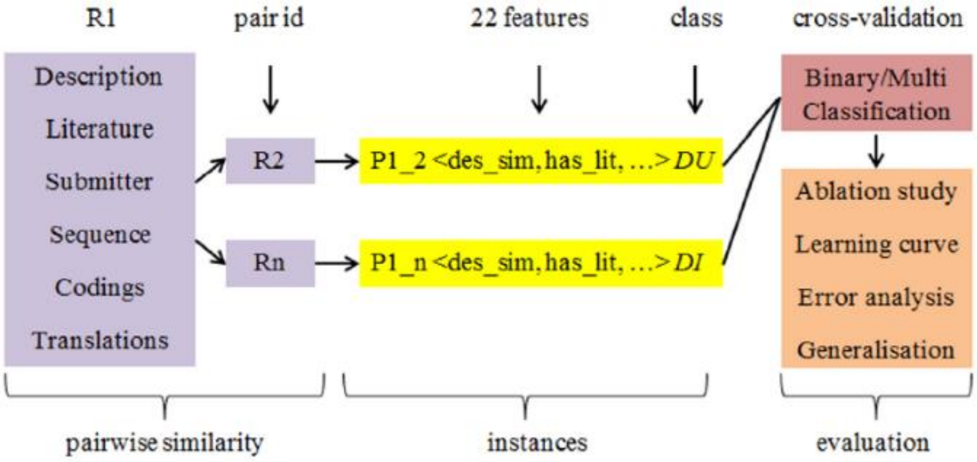

Figure 5: Diagram illustrating the Supervised Learning for Detection of Duplicates in Genomic Sequence [4]

References

1. J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, “Big Data Application in biomedical research and Health Care: A Literature Review,” *Biomedical Informatics Insights*, vol. 8, 2016.
2. L. R. Wang, L. Wong, and W. W. Goh, “How doppelgänger effects in biomedical data confound machine learning,” *Drug Discovery Today*, 2021.
3. L. Waldron, M. Riester, M. Ramos, G. Parmigiani, and M. Birrer, “The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles,” *Journal of the National Cancer Institute*, vol. 108, no. 11, 2016.
4. Q. Chen, J. Zobel, X. Zhang, and K. Verspoor, “Supervised learning for detection of duplicates in genomic sequence databases,” *PLOS ONE*, vol. 11, no. 8, 2016.
5. E. M. Bik, A. Casadevall, and F. C. Fang, “The prevalence of inappropriate image duplication in Biomedical Research Publications,” *mBio*, vol. 7, no. 3, 2016.