Question 3: Doppelgänger effects

Introduction

Data doppelgängers are defined when two sets of separately-generated data look remarkably similar to one another. Machine learning (ML) models are generally exhibiting the doppelgänger effect (DE) when they show promising results on a validation set, independent of the manner in which they were trained (Wang et al., 2022a). As DE may cause inflated results when testing the machine learning model on real data, it should be minimized. Moreover, DE becomes problematic in model selection methods, which rely exclusively on validation accuracy, and as a result, this may make it more difficult to proceed. As a result, it is essential for ML professionals to be conscious of the possibility of duplicates prior to conducting model validation. From this background, the article aims to discuss the doppelgänger effect on biomedical data and ways to prevent it. Thus, to prove this, the impact of the doppelgänger effect in different biological settings will be discussed. After that, the emergence of doppelgänger effects from a quantitative angle will be discussed. Furthermore, useful ways of avoiding doppelgänger effects will be discussed.

Impact of Doppelgänger effects in a different biological setting

Machine learning is complicated by doppelgänger effects in biological data in a way that it isn't in other fields (ML). The use of machine learning (ML) techniques has grown in the pharmaceutical industry in recent years as a means of accelerating the discovery of new therapeutic targets. Thus, in this regard, cross-validation methods are frequently employed in these machine-learning processes to assess the accuracy of these models. However, the existence of data doppelgängers has the potential to raise concerns about the accuracy of such validation systems. There are two types of doppelgängers; one is data doppelgängers (DDs) and another one is functional doppelgängers (FDs) (Wang et al., 2022a). DDs are pairs of samples that are highly correlated or similar to one another. In contrast, FDs are pairs of samples that produce artificially high ML performance when divided into training and validation sets (Wang et al., 2022b). Hence, doppelgängers can have a substantial effect on biomedical data when applied to machine learning.

In the field of contemporary bioinformatics, data doppelgängers have also been discovered. In this regard, Cao and Fullwood (2019) extensively analyzed state of the art in chromatin interaction prediction tools. Their investigation showed that the reported performance of these chromatin interaction prediction tools was inflated due to flaws in the assessment methods used at the time (Cao and Fullwood, 2019). In particular, the effectiveness of these systems was measured using test data that was highly analogous to the training data, and this implies that doppelgängers also affect the way chromatin interacts in living systems. Consistent with this research, Goh and Wong (2019) found evidence of data doppelgängers, in which particular validation data were predicted to perform well given specific training data. Thus, it can be said that doppelgängers have an effect on the process of genetic control.

Well-established areas of bioinformatics, such as protein function prediction, also showed evidence of data doppelgängers. Regarding protein function prediction, when two proteins have an identical function but different sequences, it is reasonable to assume that they are descendants of the same ancestor protein. Most data doppelgängers are mainly seen in this type of protein function prediction, which gives the wrong illusion that two proteins have similar functions (Wang et al., 2022b). However, on greater inspection, it has been inferred that though doppelgänger impacts protein function, it cannot be correctly predicted, where sequence similarity was quite low but showed similar functions.

Quantitative angle of emergence of doppelganger effects

The quantitative angle of the doppelgänger effect can be stated through the statistical aspect of clinical genomic studies. Human genomic data that is available to the public is usually summarized at a level that makes it impossible to identify an individual patient, which has been done to protect medical confidentiality. The "doppelgänger effect" occurs when researchers use the same tissue samples for many studies, which is common in clinical genomics. Thus, when combining the findings of many studies, the presence of these undiscovered duplicates has the potential to exaggerate the statistical significance of the findings or the apparent correctness of genomic models (Waldron et al., 2016). Thus, based on the findings of Waldron et al. (2016), it has been seen that due to the presence of hidden duplicates, predictive and prognostic frameworks become more accurate than the original. In this study, a serous ovarian cancer model was trained and subsequently validated in two investigations, where duplicates were found through analysis with doppelgänger R. In this case, after the progressive removal of duplicates, the hazard ratio (HR) was determined, which showed an improvement from 1.1 to 1.7 for every 30% of duplication (Waldron et al., 2016). In this way, through statistical approach quantitative angle of doppelganger effects has been highlighted.

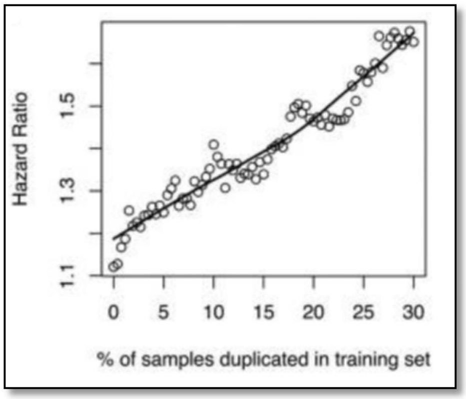


Fig 2: Statistical aspect of doppelgänger effect (Source: Waldron et al., 2016).

Examining or avoiding doppelgänger effects

When there are several doppelgänger effects (DEs) in a data set, it might complicate cross-validation procedures such as evaluating the model, tuning hyperparameters and selecting the feature. Thus, avoiding the randomization of DEs across both the validation and training sets is the most reliable strategy for mitigating the detrimental effects of these consequences. Before beginning the training model, it is required to identify the DEs that could potentially be affected by this mitigation strategy. Hence, Wang et al. (2022b) introduced the doppelgänger Identifier, which is a software package for the verification and identification of doppelgängers. This is done in an effort to limit the impact of the problem. When the Doppelgänger Identifier is applied across a broad spectrum of illnesses and different data, it is possible to demonstrate the ubiquitous presence of DEs in biomedical gene expression information. Earlier, Wang et al. (2022c) presented the problem of DE in biological data sets and widely offered a technique for recognizing probable doppelgängers within and among data sets. However, that approach could only prove the presence of DE within a dataset of single proteomics; but this does not demonstrate the pervasiveness of DE when considering other types of data, such as high-throughput data on gene expression. Thus, to address this knowledge gap, Wang et al. (2022b) have created a new R package called doppelgänger Identifier that provides straightforward tools for identifying pairwise Pearson's correlation coefficient data doppelgängers (PPCC DDs) and verifying the inflationary effects of DDs. This new incorporation has been done with the expectation that this will result in improved data science methods among the community. The pairwise Pearson's correlation coefficient, also known as the PPCC, measures the degree to which two distinct data sets are related to one another (Waldron et al., 2016). The presence of PPCC data doppelgängers in a pair of samples is indicated by a Pearson's correlation coefficient (PPC) value that is significantly higher than expected (Fig 3). Using the doppelgänger Identifier R package, users may quickly and effortlessly discover PPCC DDs between the data sets and validate the effects of these PPCC DDs on the accuracy of ML model validation (Wang et al., 2022b). It offers four different computation and visualization features, including the detection of PPCC DDs, as well as its visualization and verification, and lastly, the visualization of the verification results (Fig. 4). The final visualize verification results are shown through a scatter-violin plot, where a positive relationship between the number of PPCC DDs and validation accuracies in the scatter-violin plot helps to identify functional doppelgängers (FDs) (Wang et al., 2022a). Thus, through this method, PPCC can check the accuracy and detect DDs.

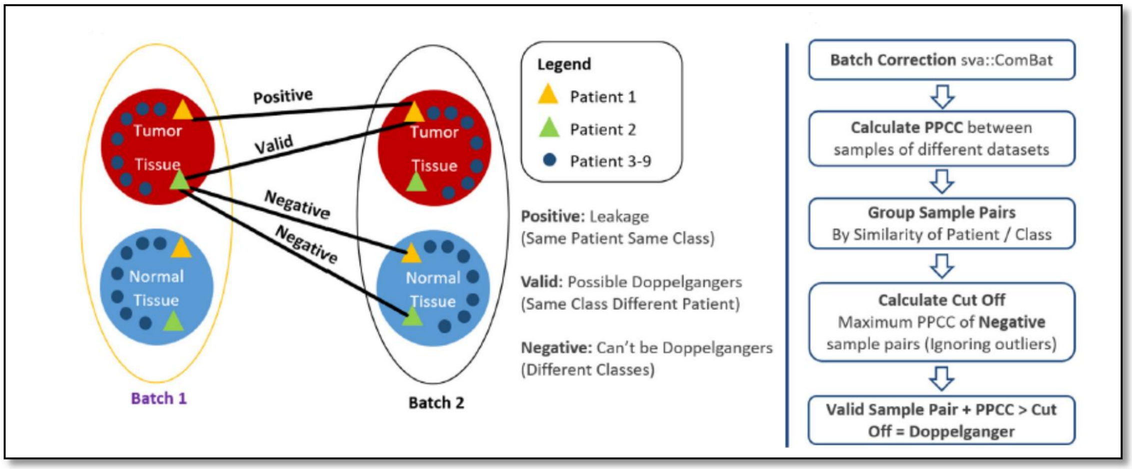
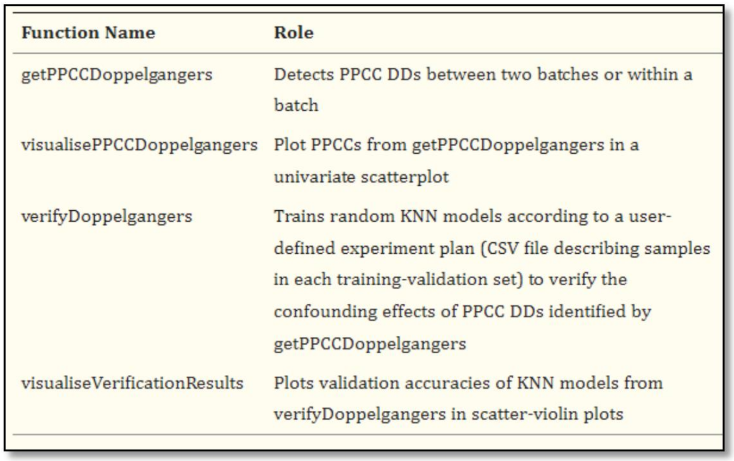

Tumor Tissue
Normal Tissue
Batch 1
Tumor Tissue
Normal Tissue
Batch 2
Positive
Valid
Negative
Negative
Legend
Patient 1
Patient 2
Patient 3-9
Positive: Leakage (Same Patient Same Class)
Valid: Possible Doppelgangers (Same Class Different Patient)
Negative: Can't be Doppelgangers (Different Classes)
Batch 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

Fig 3: Data doppelgänger identification through pairwise PPCC (Source: Wang et al., 2022a).



Function Name	Role
getPPCCDoppelgangers	Detects PPCC DDs between two batches or within a batch
visualisePPCCDoppelgangers	Plot PPCCs from getPPCCDoppelgangers in a univariate scatterplot
verifyDoppelgangers	Trains random KNN models according to a user-defined experiment plan (CSV file describing samples in each training-validation set) to verify the confounding effects of PPCC DDs identified by getPPCCDoppelgangers
visualiseVerificationResults	Plots validation accuracies of KNN models from verifyDoppelgangers in scatter-violin plots

Fig 4: Functions in the R package "doppelgänger Identifier" (Source: Wang et al., 2022a).

In addition to Pearson's correlation coefficient, the Spearman Rank correlation coefficient can additionally be employed to find DDs in a correlation matrix. Thus, in order to determine the Spearman rank correlation coefficient, the values should be arranged in one rank in each sample; after that, the sorted variables are subjected to an analysis using Pearson's coefficient of correlation (Lin et al., 2019). Spearman rank correlation coefficient is a measurement of the consistent association between the samples, and this is a more comprehensive approach than Pearson's correlation coefficient, which can only quantify linear relationships between two variables (Wang et al., 2022b). However, though lots of advances have occurred to minimize the doppelgänger effects, this is quite difficult to resolve this problem analytically. Thus, it is vital to check for possible doppelgängers in data before assortment in training and validation information to prevent performance inflation.

Conclusion

From the above analysis, it became evident that there is a wide range of activities involving doppelgängers in biomedical activities. Moreover, within the body, the evidence of its occurrence has also been highlighted, where the similarity in function has been shown in genomics, chromatin interactions, and protein functions. Thus, in this regard, the biomedical data science community is more aware of these data doppelgänger. In this regard, they developed the R package of doppelganger identifiers because they prioritize the detection of these effects first. However, complete removal of doppelgängers is not quite possible; thus, it is quite important to check for any possible doppelgängers in the data before collecting any validation data.

References

Allesina, S., & Tang, S. (2015). The stability–complexity relationship at age 40: a random matrix perspective. *Population Ecology*, *57*(1), 63-75. https://doi.org/10.1007/s10144-014-0471-0.

Bianconi, F., Antonini, C., Tomassoni, L., & Valigi, P. (2019). CRA toolbox: software package for conditional robustness analysis of cancer systems biology models in MATLAB. *BMC bioinformatics*, *20*(1), 1-19. https://doi.org/10.1186/s12859-019-2933-z.

Bianconi, F., Baldelli, E., Luovini, V., Petricoin, E. F., Crinò, L., & Valigi, P. (2015). Conditional robustness analysis for fragility discovery and target identification in biochemical networks and in cancer systems biology. *BMC systems biology*, *9*, 1-18. https://doi.org/10.1186/s12918-015-0216-5.

Bramanti, D., & Nanetti, S. (2022). Fragility or Frailty? The Stories of Five Women's Transition to Old Age. *Italian Sociological Review*, *12*(6S), 0_1-320.

Cao, F., & Fullwood, M. J. (2019). Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nature genetics*, *51*(8), 1196-1198. https://doi.org/10.1038/s41588-019-0434-7.

Gatenby, R. A., & Brown, J. S. (2020). Integrating evolutionary dynamics into cancer therapy. *Nature reviews Clinical oncology*, *17*(11), 675-686. https://doi.org/10.1038/s41571-020-0411-1.

Goh, W. W. B., & Wong, L. (2019). Turning straw into gold: building robustness into gene signature inference. *Drug discovery today*, *24*(1), 31-36. https://doi.org/10.1016/j.drudis.2018.08.002.

Kaneko, T., & Kikuchi, M. (2022). Evolution enhances mutational robustness and suppresses the emergence of a new phenotype: A new computational approach for studying evolution. *PLOS Computational Biology*, *18*(1), e1009796. 10.1371/journal.pcbi.1009796.

Kim, H., Muñoz, S., Osuna, P., & Gershenson, C. (2020). Antifragility predicts the robustness and evolvability of biological networks through multi-class classification with a convolutional neural network. *Entropy*, *22*(9), 986. https://doi.org/10.3390/e22090986.

Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2019). Generating high-fidelity, synthetic time series datasets with doppelganger. *arXiv preprint arXiv:1909.13403*.

Pasqualetti, F., Zhao, S., Favaretto, C., & Zampieri, S. (2020). Fragility limits performance in complex networks. *Scientific Reports*, *10*(1), 1774. https://doi.org/10.1038/s41598-020-58440-6.

Radványi, Á., & Kun, Á. (2021). The Mutational Robustness of the Genetic Code and Codon Usage in Environmental Context: A Non-Extremophilic Preference?. *Life*, *11*(8), 773. https://doi.org/10.3390/life11080773.

Sakthivel, R., Sakthivel, R., Selvaraj, P., Alzahrani, F., & Marshal Anthoni, S. (2021). Robust non-fragile memory feedback control for multi-weighted complex dynamical networks with randomly occurring gain fluctuations. *International Journal of Systems Science*, *52*(12), 2597-2616. https://doi.org/10.1080/00207721.2021.1892861.

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The doppelgänger effect: hidden duplicates in databases of transcriptome profiles. *JNCI: Journal of the National Cancer Institute*, *108*(11). https://doi.org/10.1093/jnci/djw146.

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

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

Wang, L. R., Wong, L., & Goh, W. W. B. (2022c). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*, *27*(3), 678-685.

Zhang, L., Zhang, K., Zhang, J., Zhu, J., Xi, Q., Wang, H., Zhang, Z., Cheng, Y., Yang, G., Liu, H. & Zhang, R. (2021). Loss of fragile site-associated tumor suppressor promotes antitumor immunity via macrophage polarization. *Nature Communications*, *12*(1), 4300. https://doi.org/10.1038/s41467-021-24610-x.