preface

Machine learning models are increasingly being used in biomedical data models. Cross-validation techniques are often used to evaluate these models. However, the reliability of such validation methods can be affected by data confounding. Data Doppelgänger occurs when independently derived data are very similar to each other, resulting in models that perform well no matter how they are trained (i.e., the Doppelgänger effect). The Doppelgänger effect is not unique to biomedical data, but occurs when samples exhibit similarity, and similar data are not only unique to biomedical data, but also exist in other disciplines, such as the common Face recognition systems, for example, have a large number of face databases that may bring about the Doppelgänger effect and lead to recognition failure. However, the Doppelgänger effect is now prevalent in the field of biomedical data analysis. Therefore it is necessary to investigate the nature of data Doppelgänger and propose an improved Doppelgänger recognition method.

method

Given the potential for confusion due to doppelganger effects, it is critical to be able to identify whether data doppelgangers exist between the training and validation sets prior to validation. A logical approach to data doppelgänger identification is to use ranking methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE) coupled with scatter plots to see how the samples are distributed in the reduced dimensional space. However, the data Doppelgänger are not necessarily distinguishable in the reduced dimensional space(Figure 1)

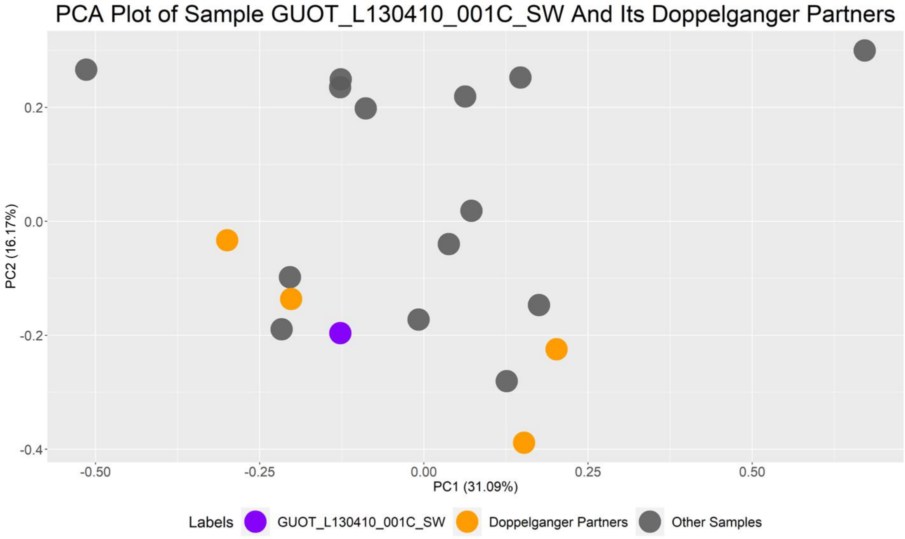

Figure 1. Plot of PC1 against PC2 for GUOT_L130410_001C_SW and its PPCC data doppelgänger partners. GUOT_L130410_001C_SW is represented by the purple dot. The PPCC data doppelgänger partners of GUOT_L130410_001C_SW are represented by orange dots. The samples in dark grey represent other samples of different batch that are not PPCC data doppelgängers with GUOT_L130410_001C_SW. No visible distinction

between PPCC data doppelgängers and non-PPCC data doppelgängers can be observed from the PCA plots above.

A method by which dupChecker identifies duplicate samples by comparing the MD5 fingerprints of their CEL files, where identical MD5 fingerprints indicate that the samples are duplicates (essentially duplicates and therefore indicate a leakage problem). Thus, dupChecker does not detect true data splits, which are independently derived samples that are incidentally similar. Another measure is the paired Pearson correlation coefficient (PPCC), which captures the relationship between sample pairs from different datasets. An unusually high PPCC value indicates that a pair of samples constitutes a PPCC data split. Thus Wang et al. used the renal cell carcinoma (RCC) proteomics data from Guo et al. and determined PPCC data splits based on the PPCC distributions of valid scenarios versus negative and positive scenarios, ultimately observing a high percentage of PPCC

data splits

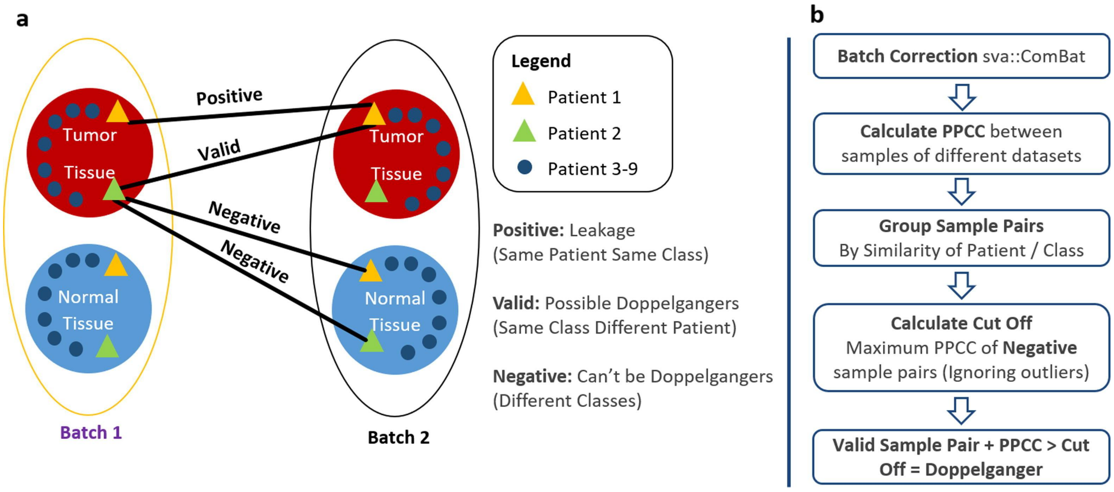
a
Positive
Valid
Negative
Negative
Tumor Tissue
Normal Tissue
Tumor Tissue
Normal Tissue
Batch 1
Batch 2
Legend
Patient 1
Patient 2
Patient 3-9
Positive: Leakage (Same Patient Same Class)
Valid: Possible Doppelgangers (Same Class Different Patient)
Negative: Can't be Doppelgangers (Different Classes)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

Figure 2. Diagram illustrating the pairwise Pearson's correlation coefficient (PPCC) data doppelgänger identification method. (a) Naming convention for different types of sample pair based on the similarities of their patient and class. (b) Process of PPCC data doppelgänger identification. PPCC data doppelgängers are defined as valid sample pairs with PPCC values greater than all negative sample pairs.

The effect of PPCC data doppelgangers on the validation accuracy of different randomly trained classifiers is then explored and it is found that the presence of PPCC data admixture in the training and validation data improves ML performance even if the features are randomly selected.The relationship between the number of PPCC data doppelgangers and the magnitude of the doppelganger effect is based on a dosage, i.e., the more doppelganger pairs appear in the training and validation sets, the more exaggerated the ML performance. However, if there are few similar examples (few data overlap), gaps in the model are exposed and thus the model tends to perform poorly

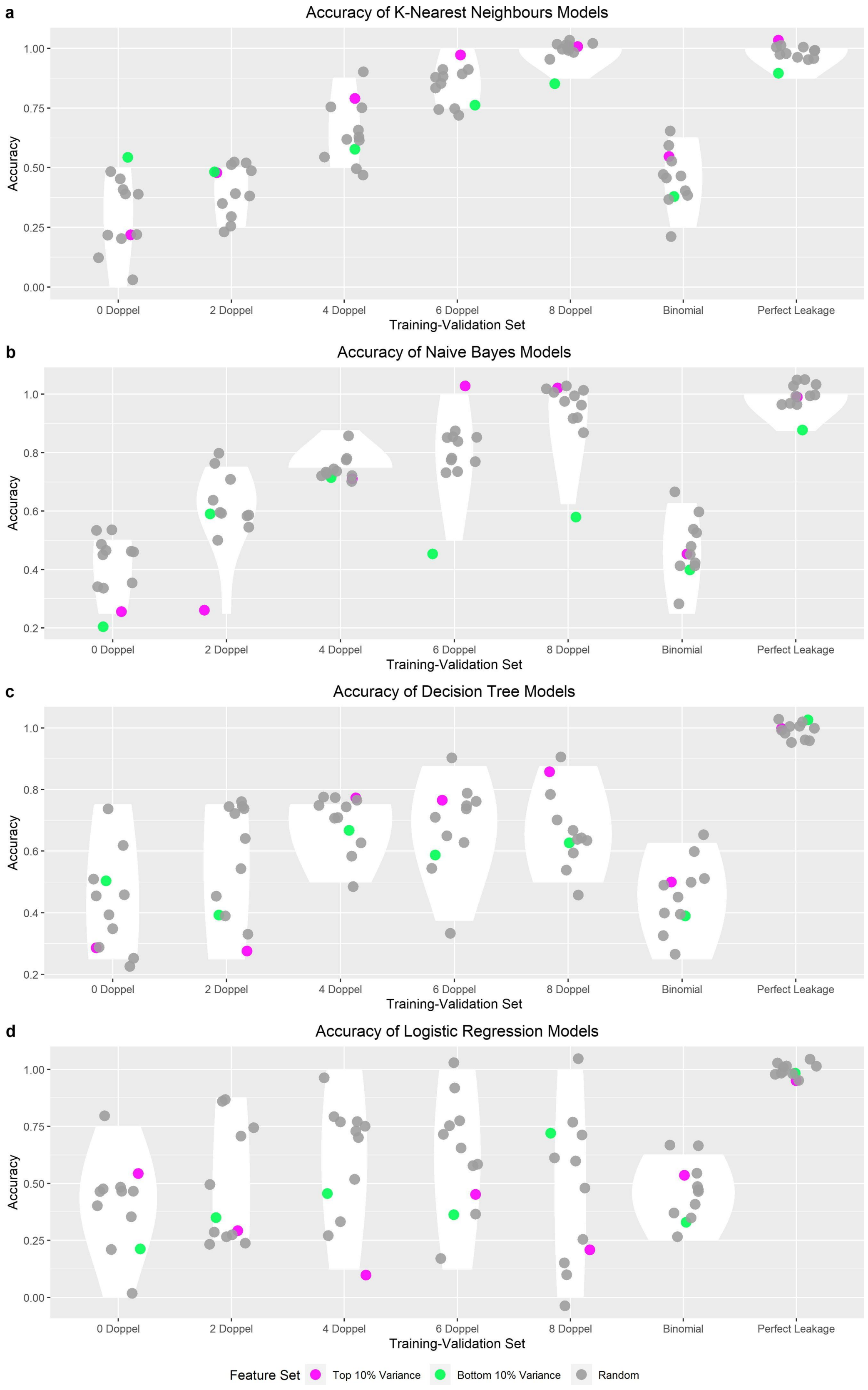

Figure 3. The prediction performance of different machine learning (ML) models on pairs of training-validation sets with varying numbers of pairwise Pearson's correlation coefficient (PPCC) data doppelgängers in the validation set. ML models assessed includes: k-nearest neighbor (kNN) (a), naïve bayes (b), decision tree (c), and logistic regression (d) models. X-axis indicates the type of validation set: 'i Doppel' refers to a validation set with i number of PPCC data doppelgängers in the training set (where i = 0, 2, 4, 6, and 8), 'Binomial' refers to the accuracies generated by 12 (number of feature sets) binomial distributions with N = 8 (because there are eight samples in the validation set) and P = 0.5 (probability of guessing the correct label for each validation sample) (negative control), 'Perfect Leakage' refers to a validation set with eight duplicates with the training set (positive control). Y-axis indicates the accuracy of ML models on a validation set of eight samples with the lowest accuracy being 0 and the highest accuracy being 1. 'Top 10% Variance' refers to the feature set comprising proteins of the highest variance (i.e., top 10% among the total number of proteins in the data set). 'Bottom 10% Variance' refers to the feature set comprising proteins of the lowest variance at 10% of the total number of proteins in the data set, 'Random' refers to the feature set comprising randomly select proteins at 10% of the total number of proteins in the data set.

This confirms that PPCC data doppelgängers (based on pairwise correlations) act as functional doppelgängers (confounding ML results), producing an inflationary effect similar to data leakage, so how to deal with the doppelgänger effect?

As mentioned earlier, we can use the pairwise Pearson correlation coefficient (PPCC) to identify data Doppelgänger , it is crucial to be able to identify the presence of data doppelgänger between the training and validation sets prior to validation, and PPCC it captures the relationship between sample pairs from different data sets. An unusually high PPCC value indicates that a pair of samples constitutes a PPCC Data Doppelgänger; in addition to the Pearson correlation coefficient, other correlations such as Spearman Rank correlation coefficient and Kendall Rank correlation coefficient can also be used to identify Data Doppelgänger. the Spearman Rank correlation coefficient is calculated by first ranking the values in each sample and then applying the Pearson correlation coefficient to the ranking variable. The Spearman Rank Correlation Coefficient measures the monotonic relationship between samples and is more general than the Pearson Correlation Coefficient, which measures only linear relationships. Like the Spearman Rank correlation coefficient, the Kendall Rank correlation coefficient also measures monotonic relationships between samples and is more general than the PPCC.

Splitting effects can also be mitigated using methods that do not result in a significant reduction in sample size or require large amounts of contextual data, e.g., we try to trim the data by removing variables that contribute strongly to the data splitting effect; furthermore careful cross-checking using metadata as a guide can also be done followed by data stratification. Instead of evaluating the model performance of the entire test data, we can stratify the data into layers of different similarities. Finally perform extremely powerful independent validation checks involving as many datasets as possible. It may be possible to use data from different database sources for validation, and different classification models can be used to validate the data, so check the data for Doppelgänger before model validation in order to recognize the interrelated patterns in the data and try to avoid Doppelgänger effects.