In this section, I cover the results achieved from my interpretation and analysis of the paper on doppelgangerIdentifier (DI) which is a R package and reproduce the code to detect and verify doppelganger data [1]. In my code demonstration, out of the four ready-to use datasets and their metadata, I have leveraged only on the renal carcinoma microarray dataset and the breast cancer RNA Sequence dataset from [2] while invoking the following four functions as described below.

1) getPPCCDD - This function detects data doppelgangers between 2 batches or within a batch by calculating pairwise pearson correlation coefficient between every sample in one dataset against every sample in the other dataset. It fulfils all the steps as detailed in Figure 1 flow of the paper (pg 3). In my replication, I have only catered for the pairwise pearson correlation metric and not the Spearman or Kendrall metric in the function parameter.

2) visualizePPCCDD - This function visualises doppelganger detection results in univariate plot.

3) verifiyDD - This function establishes baseline for model evaluation by training and validating (train-validation sets) KNN classifier models with random feature sets.

4) visualizeVerificationResults - This function visualises the validated KNN classifier models.

Dataset	**Description**	**Citation / Source (Original paper where dataset were first derived from)**
rc	**Renal Cell Carcinoma Proteomics DataSet**	**Guo et al**
dmd	Duchenne Muscular Dystrophy (DMD) Microarray DataSet	Haslett et al., Pescatori et al
lk	Leukemia Microarray DataSet	Golub et al., Armstrong et al
bc	**Breast cancer dataset**	**https://github.com/lr98769/doppelgangerIdentifier/tree/main/tutorial/dataset**

Table 1

All datasets used in the DI package and for my code demonstration purpose it is strictly highlighted in blue.

Untitled1* | ppccDoppelgangerResults$PPCC_df | ppccDoppelgangerResults$PPCC_matrix | ppccDoppelgangerResults$Processed_...

Filter

	Sample1	Sample2	PPCC	ClassPatient	DoppelgangerLabel
1	normal_cc_patient_1_rep1	normal_cc_patient_1_rep2	0.9623107	Same Class Same Patient	Not Doppelganger
2	normal_cc_patient_1_rep1	tumor_cc_patient_1_rep2	0.8539736	Different Class Same Patient	Not Doppelganger
3	normal_cc_patient_1_rep1	normal_cc_patient_2_rep2	0.9415945	Same Class Different Patient	Doppelganger
4	normal_cc_patient_1_rep1	tumor_cc_patient_2_rep2	0.6420951	Different Class Different Patient	Not Doppelganger
5	normal_cc_patient_1_rep1	normal_cc_patient_3_rep2	0.9417308	Same Class Different Patient	Doppelganger
6	normal_cc_patient_1_rep1	tumor_cc_patient_3_rep2	0.9225526	Different Class Different Patient	Not Doppelganger
7	normal_cc_patient_1_rep1	normal_p_patient_4_rep2	0.8820859	Same Class Different Patient	Not Doppelganger
8	normal_cc_patient_1_rep1	tumor_p_patient_4_rep2	0.8670501	Different Class Different Patient	Not Doppelganger
9	normal_cc_patient_1_rep1	normal_ch_patient_5_rep2	0.9552486	Same Class Different Patient	Doppelganger
10	normal_cc_patient_1_rep1	tumor_ch_patient_5_rep2	0.8170501	Different Class Different Patient	Not Doppelganger
11	normal_cc_patient_1_rep1	normal_cc_patient_6_rep2	0.9174800	Same Class Different Patient	Not Doppelganger
12	normal_cc_patient_1_rep1	tumor_cc_patient_6_rep2	0.8355978	Different Class Different Patient	Not Doppelganger
13	normal_cc_patient_1_rep1	normal_cc_patient_7_rep2	0.8734413	Same Class Different Patient	Not Doppelganger
14	normal_cc_patient_1_rep1	tumor_cc_patient_7_rep2	0.8697289	Different Class Different Patient	Not Doppelganger
15	normal_cc_patient_1_rep1	normal_cc_patient_8_rep2	0.9736597	Same Class Different Patient	Doppelganger
16	normal_cc_patient_1_rep1	tumor_cc_patient_8_rep2	0.7181924	Different Class Different Patient	Not Doppelganger
17	normal_cc_patient_1_rep1	normal_p_patient_9_rep2	0.8627765	Same Class Different Patient	Not Doppelganger
18	normal_cc_patient_1_rep1	tumor_p_patient_9_rep2	0.8955285	Different Class Different Patient	Not Doppelganger

Image 1

R List with object ppcc_df: This is an output from running the function getPPCCDD. It is a correlation dataframe where each row in the dataframe describes information about the sample pair.

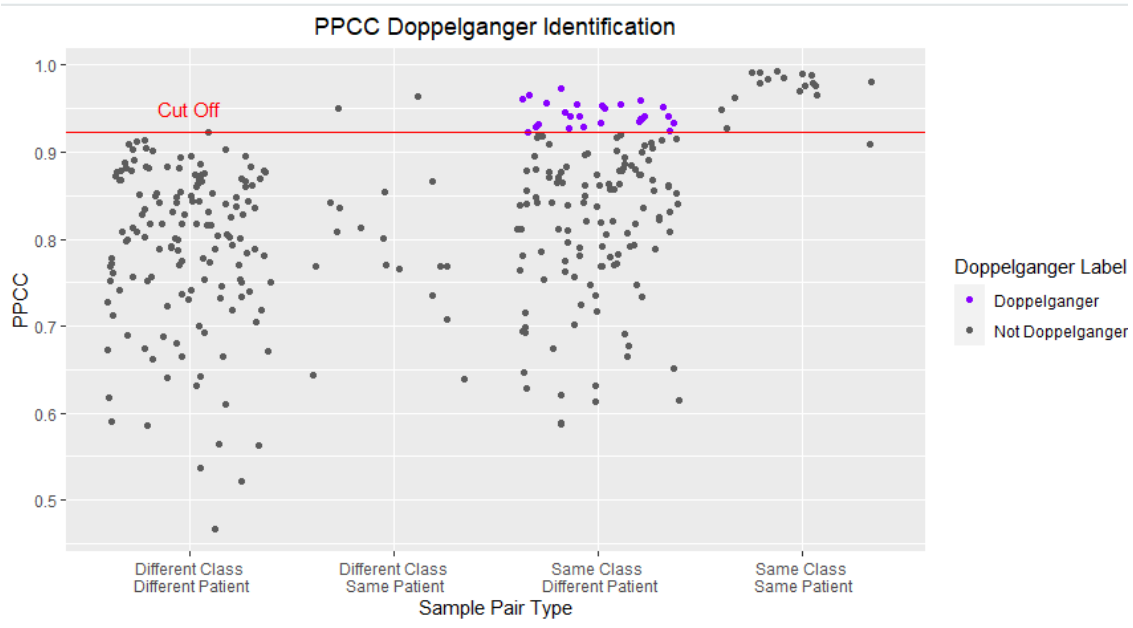

Image 2

Identifying doppelgangers from the microarray dataset - renal carcinoma was based on the criteria of selecting sample pairs from the same class but different patients. In the scatterplot below, X axis shows the ppcc of sample pairs with varying similarities in class and patient and the Y axis shows the pcc value of each sample pair. Red line is the identification cut-off

threshold. Data doppelgangers are labelled as purple as they are the sample pairs in the same class and different patient columns where ppcc values are greater than the threshold.

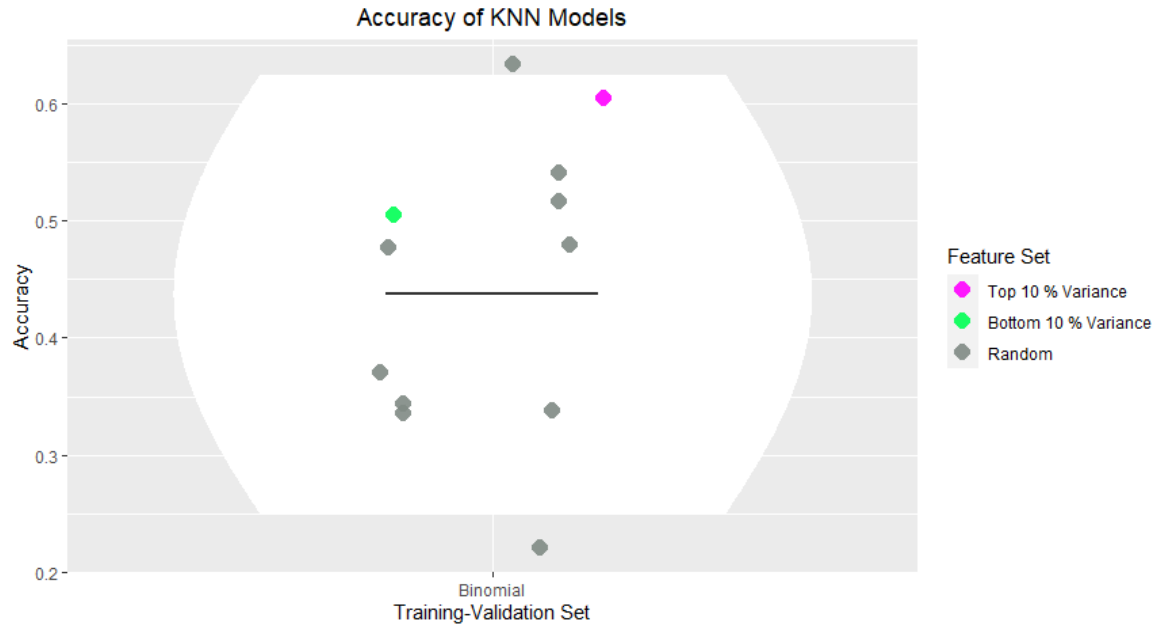

Image 3

Visualising the functional doppelgangers reveal the grey dots that represent random feature sets while pink and green represent the highest and lowest bounds of model accuracy for train-validation sets. Y-axis represents the validation accuracy and it is observed that 60% of validation samples were correctly classified. The model average seems to be at 0.45 as depicted by the black crossbar.

Next, I also attempted to identify doppelgangers in the RNA-Sequence dataset - breast cancer from tutorial [3] after replicating results from the paper. The steps and interpretation are similar to the above. (Image 4-6)

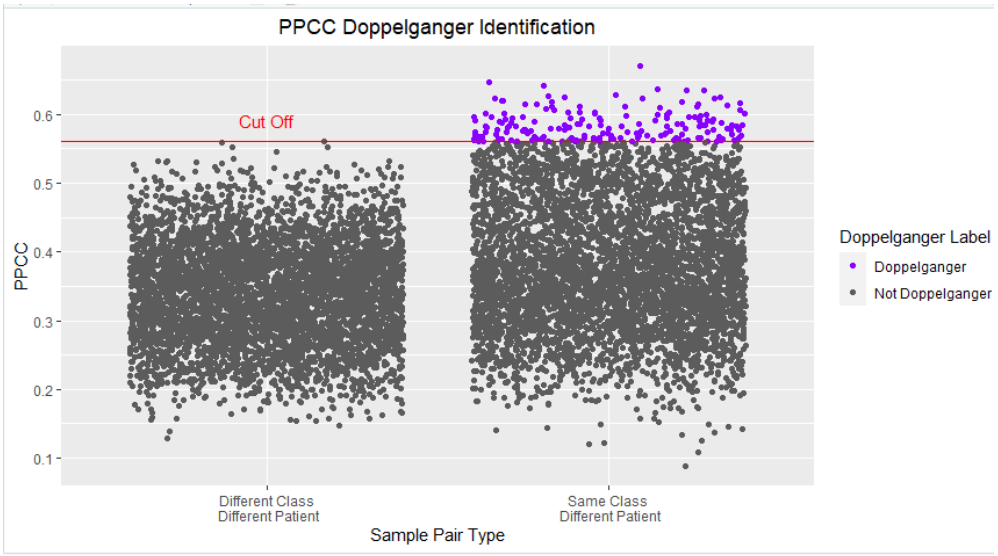

Image 4

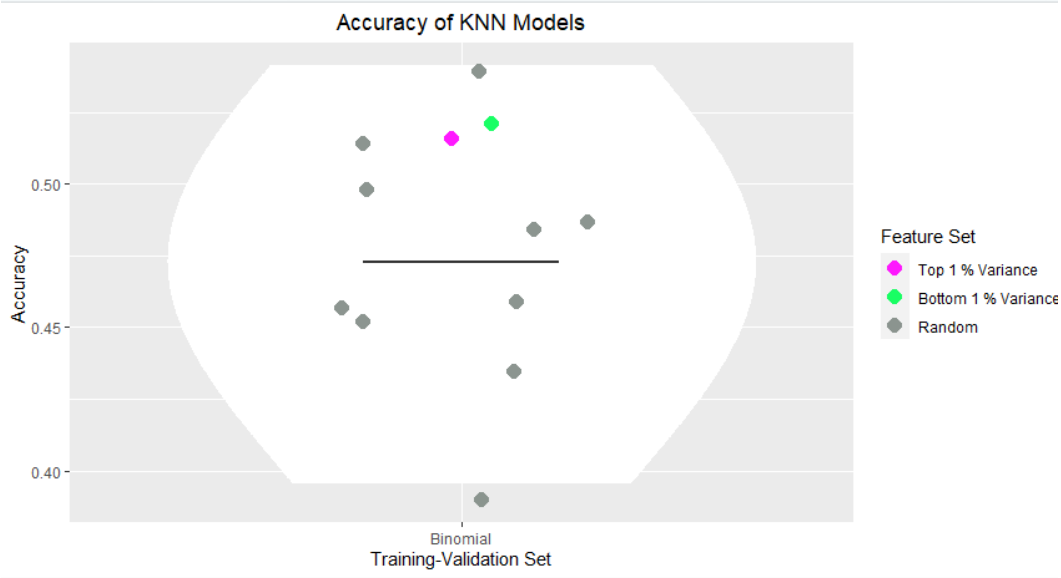

Image 5

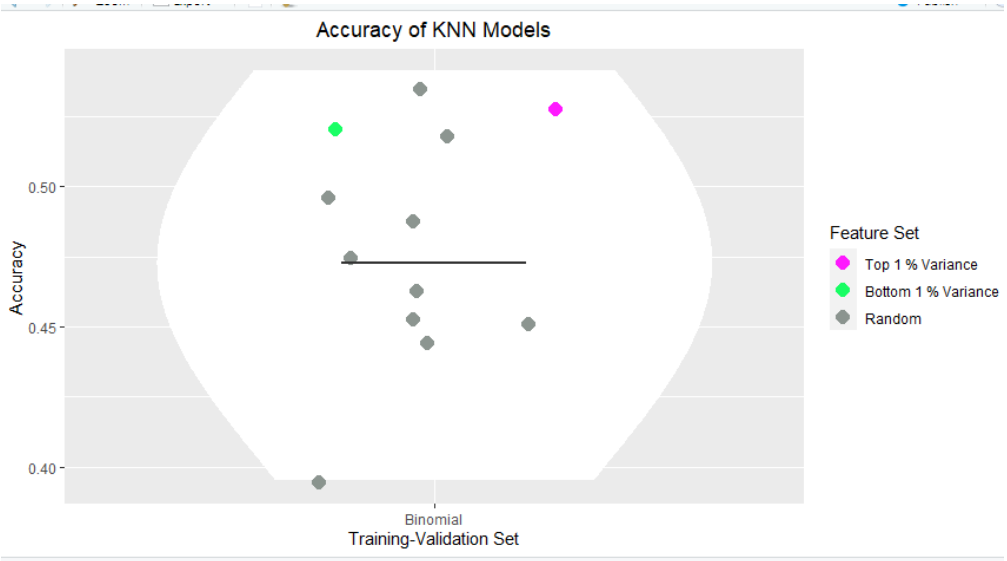

Image 6

Doppelganger effects are also found in building machine learning models for therapy and rehabilitation in response to neurological disorders that affect the human brain [4]. They are not, however, unique to the biomedical data domain, especially in examples of proteomics, neuroimaging and gene expression data as informed by the research paper [1]. From my preliminary investigations, I read that doppelganger effects also exist in other areas such as manufacturing and systems research and development through the creation of synthetic data models for time series data [5].

Although doppelgangers are a threat to the validity of inferences, they cannot be totally avoided from the health and biomedical field. Instead they can be checked and controlled through adjustment techniques such as applying statistical correction methods before combining multiple datasets from different sources.The presence of batch imbalance will otherwise affect the pairwise pearson correlation outcomes. An alternative method is to

adopt sensible data splitting strategy and randomization in data samples of the machine learning algorithm to avoid false positives.

Doppelganger spotting is not easy to identify in the biomedical domain because the presence of duplicate samples can artificially inflate the predictive accuracy performance of the machine learning model on real world datasets, which are often imbalanced. Referencing the paper [6], duplicate or replicate samples within or between datasets can be identified through expression profiles where expression doppelgangers have higher pairwise Pearson correlation coefficient than expected. More generally, another useful identification method is using duplicate outlier detection by obtaining outliers based on distance measures. These duplicates are then detected as outliers at the high end of batch corrected correlations [7].

References

[1] https://www.sciencedirect.com/science/article/pii/S2666166722006633

[2] https://github.com/lr98769/doppelgangerIdentifier

[3] https://github.com/lr98769/doppelgangerIdentifier/tree/main/tutorial/dataset

[4] https://stanfordvr.com/mm/2010/fox-ct-doppelgangers-health.pdf

[5] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01654-1

[6]https://bioconductor.org/packages/devel/bioc/vignettes/doppelgangR/inst/doc/doppelgangR.html

[7] https://ncbi.nlm.nih.gov/pmc/articles/PMC5241903/