INTRODUCTION:

Machine learning improves drug discovery efficiency in a variety of ways. For example, it can shortlist better drug candidates faster, lowering the amount of time spent on discovery and testing. In machine learning, it is widely accepted that the training and test data sets should be generated independently. However, validation findings based on separately obtained training and test sets may be inaccurate. Models trained and verified on data doppelgängers, for example, may perform well independent of training quality.

DATA DOPPELGANGERS AND FUNCTIONAL DOPPELGANGERS:

When a classifier falsely performs well because of data doppelgängers, we say that there is an observed doppelgänger effect. When samples appear to be identical throughout their measurements, this is known as data doppelgängers. However, there is no certainty that this will result in a doppelgänger effect. Functional doppelgängers are data doppelgängers that provide a doppelgänger effect.

DOPPELGANGERS IN BIOMEDICAL DATA:

The prevalence of functional doppelgängers may be demonstrated using a renal cell carcinoma as the data set and correctly constructed controls. Protein function prediction uses data doppelgängers to infer that proteins with similar sequences are descended from the same ancestor protein and inherit that ancestor's function. On the other hand, this method would be unable to predict activities for proteins with less identical sequences reliably but similar functionalities and enzymes with distinct sequences overall but similar active site residues.

Testing their performance on comparable compounds with different activities can help distinguish badly trained models from their well-trained counterparts (SAR paradox). If minor structural differences determine the outcome, a well-trained model should perform well in these situations since it is trained on relevant structural features and capable of detecting these little variations. On the other hand, an inadequately trained model might miss the actual biological action. Although various strategies for detecting data doppelgängers have been presented, most of them are neither generalizable nor robust enough.

IDENTIFYING THE DATA DOPPELGANGER

Using ordination methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE) in combination with scatterplots to observe how samples are distributed in reduced-dimensional space is one logical way to data doppelgänger detection. However, because data doppelgängers are not always discernible in reduced-dimensional space, we discovered that such a strategy was not practical.

Duplicate samples are identified using dupChecker, which compares the MD5 fingerprints of their CEL files. The presence of 18 identical MD5 fingerprints suggests that the samples are duplicates. A PPCC result that is unusually high suggests that a pair of samples are PPCC data doppelgängers. The original PPCC paper's main flaw was that it never definitively tied PPCC data doppelgängers to their capacity to confuse ML tasks.

Proteomics of renal cell cancer (RCC) was chosen because of its efficacy in creating clear-cut scenarios: (i) valid circumstances, in which doppelgängers are acceptable by generating sample pairs assigned to the same class label but from separate samples; (ii) negative instances, in which doppelgängers are nonpermissible by constructing sample pairs with different class labels. The PPCC

data doppelgängers are found by comparing the valid scenario's PPCC distribution to the negative and positive scenarios. Surprisingly, a large number of PPCC data doppelgängers were discovered. This means that outlier identification will be insufficiently sensitive.

CONFOUNDING EFFECTS:

Even if the features are chosen at random, the presence of PPCC data doppelgängers in both training and validation data inflates ML performance. Furthermore, the higher the number of doppelgänger pairings in the training and validation sets, the better the ML performance. This finding supports the idea that PPCC data doppelgängers behave like functional doppelgängers, causing inflationary consequences comparable to data leaking. The k-nearest neighbour parallels between doppelgänger effects and leakage are striking.

The doppelgänger effect is eliminated when all PPCC data doppelgängers are grouped together in the training set. This may give a solution to avoid the doppelgänger effect. The PPCC data doppelgängers, however, are restricted to either the training or validation sets due to a poor solution.

To lessen the impact of PPCC data doppelgängers, they might be eliminated. However, in small data sets with a large fraction of PPCC data doppelgängers, such as RCC, this strategy does not work because removing PPCC data doppelgängers will decrease the data to an undesirable size.

CAUSES OF DOPPELGANGER:

Doppelgängers may be found in a variety of biological data sets, including as genetic sequences. A sufficient number of germ-line sequence markers creates a "fingerprint" that can be uniquely matched in a genotyping database. To safeguard patient privacy, publicly available human genetic data is generally summarised to a level that cannot be recognised uniquely. Cancer transcriptomes change in a precise way, but they're considerably more difficult to recognise in a summary form. In clinical genomic research, tissue specimen reuse is common, resulting in the "doppelganger effect" in publically available datasets: hidden duplicates that, if left undiscovered, might inflate statistical significance or apparent accuracy of genomic models when data from several studies are combined.[1]

UNIQUENESS OF DOPPELGANGER:

Not only in biomedical data sets, but doppelgangers also prevail in various real-time data sets. For example, it has caused problems in the face recognition systems. Lookalikes, a.k.a. doppelgangers, increase the probability of false matches in a facial recognition system, in contrast to random face image pairs selected for non-mated comparison trials. Hence it is proved that doppelgangers are not unique to biomedical datasets.[2]

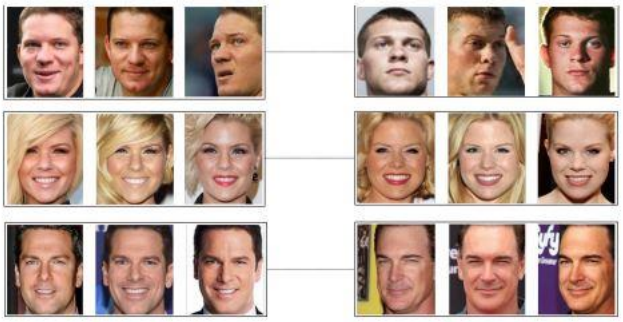

Figure1. Examples of the doppelganger identities from the Lowshot face recogntion challenge dataset.

DOPPELGANGERS BEING USEFUL IN VARIOUS FIELDS:

Although doppelgangers diminish the model's efficiency, they have proved beneficial in a variety of domains. Virtual doppelgangers are employed in biomechanics to improve voluntary imitations. Participants were shown four virtual avatars with varying degrees of realism and observer likeness in a within-subjects design, ranging from an abstract stick figure to a customised doppelganger avatar created from 3d scans of the observer. Participants were instructed to emulate the various trunk motions made by the characters. The functional ranges of motion (ROM) for spinal extension (bending backward, BB), lateral flexion (bending sideward, BS), and rotation in the horizontal plane (RH) based on shoulder marker trajectories as behavioral indicators of imitation.This shows that using a virtual doppelganger to maximise model-observer similarity might be useful in observational modelling, and that this could be utilised to improve maladaptive motor habits in patients with persistent back pain.[3]

Fig. 2. Virtual characters displayed in the experiment. Avatar number (AN) labels the equally distanced contrast for the different levels of character realism and personalization. The personalized character ("doppelganger," AN=4) was designed manually based on 3D photographs (Kinect sensor).

Artificial intelligence makes extensive use of it. Machines can anticipate human thought patterns and assist in the sorting of adverts and other data based on the preferences of the individual.[4]

Such use, however, can lead to cyber dangers, because people rely on AI to make tiny judgments. As a result, it's critical to keep the use of such doppelgangers to a minimum.[4]

DETECT AND AVOID DOPPELGANGERS:

Since the presence of it is unwanted in specific fields, it is essential to know how to detect and avoid such doppelgangers.

1. Approach

The goal is to identify multiple identities. For each pair of data A and B,the probability of A's information being attributed to B (P r(A → B)) and B's information being attributed to A (P r(B → A)) is calculated. A and B are the same if the combined probability is more significant than a threshold. To calculate the pairwise probabilities, for each data, Ai ∈ A used to train a model using all other data in A except Ai and test using Ai. This method is called Doppelgänger Finder.[5]

Procedure 1 *Doppelgänger Finder*

Input: Set of authors $\mathcal{A} = A_1, ..A_n$ and associated documents, D, and threshold t
Output: Set of multiple identities per authors, M
$F \Leftarrow$ Add weight k with every feature frequency (default k=10)
$F' \Leftarrow$ Features selected using PCA on F
▷ Calculate pairwise probabilities
for $A_i \in \mathcal{A}$ **do**
  n = Number of documents written by A_i
  $C \Leftarrow$ Train on all authors except A_i using F'
  $R \Leftarrow$ Test C on A_i (R contains the probability scores per author.)
  for $A_j \in R$ **do**
    $Pr(A_i \rightarrow A_j) = \frac{\sum_{x=1}^{n} Pr(A_{jx})}{n}$
  end for
end for
▷ Combine pairwise probabilities
for $(A_i, A_j) \in \mathcal{A}$ **do**
  $P = Combine(Pr(A_i \rightarrow A_j), Pr(A_j \rightarrow A_i))$
  if $P > t$ **then**
    $M.add(A_i, A_j, P)$
  end if
end for
return M

2. Feature extraction

To identify similarities between two data, the same features are used. After extracting all the features, the weight is added to the feature frequencies to increase distance among different data. This increases the distance between the present and not present features and gives better results.

Principal component analysis (PCA) is a widely used mathematical tool for high-dimensional data analysis. It uses the dependencies between the variables to represent the data in a more tractable, lower-dimensional form. PCA finds the variances and coefficients of a feature matrix by finding the eigenvalues and eigenvectors.

To perform PCA, the following steps are performed:

1) Calculate the covariance matrix of the feature matrix F. The covariance matrix measures how much the features vary from the mean concerning each other.

2) Calculate eigenvectors and eigenvalues of the covariance matrix. The eigenvector with the highest eigenvalue is the most dominant principal component of the dataset (PC1). It expresses the most significant relationship between the data dimensions. Principal components are calculated by multiplying each row of the eigenvectors with the sorted eigenvalues.

3) One of the reasons for using PCA is to reduce the number of features by finding the principal components of input data. The best low-dimensional space is defined as having the minimal error between the input dataset and the PCA[5]

3. Baseline

Two distance-based methods can be used to detect the doppelgangers:

1) Unsupervised: Calculate the Euclidean distance between any two dates. Choose a threshold. Two authors are the same if their distance is less than the threshold.

2) Supervised: Train a classifier using the euclidean distance between any two data in the training set. Test it using the euclidean distance between the data in the test set.

The same features and classifiers are used for both the method and the baseline method. The distance method might provide different results with different feature sets and classifiers.[5]

Reference:

1. (PDF) The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles (researchgate.net)

2. (PDF) Doppelganger Mining for Face Representation Learning (researchgate.net)

3. Exploring Virtual Doppelgangers as Movement Models to Enhance Voluntary Imitation | IEEE Journals & Magazine | IEEE Xplore

4. https://towardsdatascience.com/artificial-intelligence-might-just-get-you-your-doppelganger-6511be7a405b

5. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6956566

6. Impact of Doppelgängers on Face Recognition: Database and Evaluation | IEEE Conference Publication | IEEE Xplore