REPORT

INTRODUCTION

Machine Learning Algorithms become more accurate in identifying and predicting outcomes the more they are fed with training and validation data from a dataset. Hence, it becomes advantageous to opt for larger datasets. These larger datasets however have a more common problem than previously identified, which is that of a DD (Doppelganger Data). DD refers to data which are similar in features but are independently derived. The training and testing datasets should be independently derived for a classifier algorithm to give realistic prediction of future inputs[1]. But the possibility of independent data to be doppelganger data is more common than previously thought of, especially in the field of biomedical data. DD being present in both training and validation data can misrepresent the accuracy of the algorithm being trained by giving false positive results. This can lead to many negative outcomes, such as monetary loss and in terms of the biomedical field, possible loss of human life.

EMERGENCE OF DOPPELGANGER EFFECTS FROM A QUANTITATIVE ANGLE

DD in datasets may give rise to Doppelganger effects. An example which can illustrate the emergence of doppelganger effect is from a paper discussing the example of video surveillance [4]. The video surveillance can capture images of thousands of people in a day and store images of individuals in the form of tracklets. A tracklet is a sequence of chips that have been cropped to an individual that was tracked across the full-frame video. Let's assume we supply two tracklets with similar images: 1) Tracklet A and 2) Tracklet B. Tracklet A is supplied to the training dataset and Tracklet B is supplied to test the dataset on a k-NN classifier algorithm. Each Tracklet can represent duplicated images of a person, hence during cleanup, a reference image from A and B is used respectively for running of the algorithm. If the individual in the images in Tracklet A and Tracklet B are lookalikes, but not two images are of the same person, they will be considered doppelgangers and it will cause the algorithm to overperform.

For an illustrative example, let's assume Tracklet A and Tracklet B have 1000 images each generated from video surveillance. One image is to be selected from each tracklet after cleanup

of pixelated/ blurred images. Say Reference image A and Reference image B are selected. Reference Image A is supplied as part of the training dataset and Reference Image B is part of the testing dataset.

Further, let each image have 10 features *(f1, f2,.....f10)* which are used for identification. If reference images A and B have 8 matching features *(f1, f2……f8)* the two images are considered to be doppelganger images. If PCA identifies features from *(f1, f2…..f5)* to be of importance, it may judge images A and B to be a match.

In summary, Reference Image A is supplied to k-NN algorithm to generate output, which accidentally matches test image i.e. Reference Image B,creating doppelganger effect resulting in inflated accuracy. Figure1 represents the above sequence.

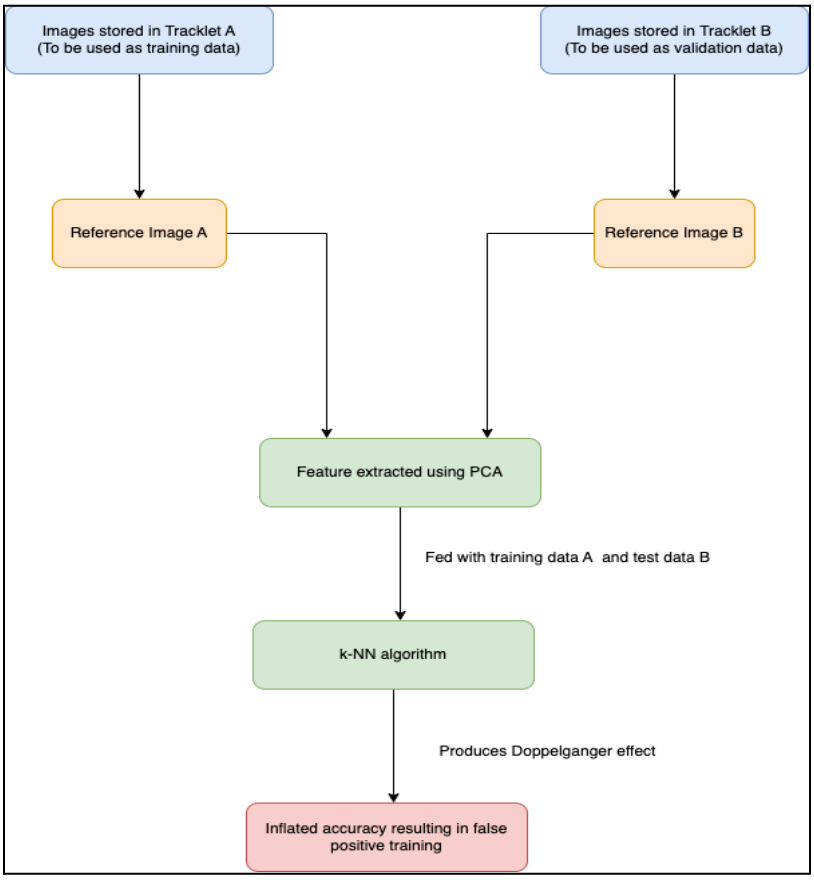

Figure 1. How doppelgangers can cause doppelganger effect in imaging dataset

EXAMPLES OF DOPPELGANGER DATA IN OTHER FIELDS

DD is not limited to biomedical datasets, but also in various imaging datasets. For example, the "VGG Face Dataset" [2] uses doppelganger data as part of 2.6 million mined face images from the internet, and has enabled Felipe Cunha (https://towardsdatascience.com/celebrity-doppelganger-finder-using-vgg-face-dataset-dlib-and-opencv-30ea1806200) to create a Celebrity Doppelganger Finder by developing a facial recognition algorithm using Dlib to recognize facial landmarks in an input image and present a closely matched celebrity face i.e doppelganger, based on those facial landmarks. Here we see an application where doppelganger data is used to achieve an end rather than hindering the machine learning process.

Similarly, Wang et al. [3] created RNA-Sequence dataset of 'lymph_lung' and 'large_upper' and tested a newly created R package called 'doppelgangerIdentifier' to identify PPCC Data Doppelgangers within and between datasets. The focus of their paper was to illustrate how doppelganger effects in biomedical data confound machine learning.

Another example of DD which has adverse effects on humans is in the facial recognition of surveillance systems as described in the illustration in the previous section. Modern surveillance systems are heavily dependent on AI for facial recognition to provide actionable information for real time decision making [4]. But relying on these systems poses an ethical dilemma when Re-identification (ReID) of similar looking individuals needs to be done. Potential misidentification of people can result in the wrong people being accused of committing crimes, for which a person may need to have to pay heavily in the court of justice, which can potentially taint their reputation along with the social injustice brought about by the situation. This consequence is similar to that of training datasets containing DDs when it comes to shortlisting candidates for drug testing[1], who may be falsely identified as suitable for the test and suffer adverse effects if they were picked up as a false positive.

TO CHECK FOR AND AVOID DOPPELGANGER EFFECTS

Doppelganger effect in biomedical data arises mostly due to DD present in both the training and the testing dataset which are passed to an algorithm. The doppelganger effect is responsible for giving a false positive result by increasing the accuracy of trained models.
A few suggested ways from Wang et al [1] are:

i) To avoid doppelganger effect is to perform careful cross-checks from meta-data and assort the doppelganger data into either training or validation dataset.
ii) Perform stratification of types of DD, namely pairwise Pearson's correlation coefficient (PPCC) and non-PPCC and observe performance of each training model on stratum separately. This model will help in identifying any gaps with the classifier algorithm, although not entirely eliminate the doppelganger effect.
iii) Perform robust validation on a large number of datasets, which can be helpful in informing objectivity of classifiers, rather than eliminating the doppelganger effect.

'doppelgangerIdentifier' is an R package [3] which can be efficiently used to easily identify PPCC DDs between and within datasets and verify the impacts of these detected PPCC DDs on ML model validation accuracy.

B. RichwardWebster et al. in their paper on doppelganger saliency [4] discuss an interesting method in the example of image recognition as cited above. Using the Explainable AI Toolkit (XAITK) they generate saliency maps which help highlight the differences between images that may be doppelgangers. Perhaps future biomedical data scientists will be able to apply this toolkit in the same way to identify doppelgangers in biomedical data.

CITATIONS

[1] Wang LR, Wong L, Goh WWB. How doppelgänger effects in biomedical data confound machine learning. Drug Discov Today. 2022 Mar;27(3):678-685. doi: 10.1016/j.drudis.2021.10.017. Epub 2021 Oct 28. PMID: 34743902.
[2] O. M. Parkhi, A. Vedaldi, A. Zisserman Deep Face Recognition British Machine Vision Conference, 2015.

[3] Wang, Li & Choy, Xin & Goh, Wilson. (2022). Doppelgänger Spotting in Biomedical Gene Expression Data. iScience. 25. 104788. 10.1016/j.isci.2022.104788.
[4] B. RichardWebster, B. Hu, K. Fieldhouse and A. Hoogs, "Doppelgänger Saliency: Towards More Ethical Person Re-Identification," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 2846-2856, doi: 10.1109/CVPRW56347.2022.00322.