### *Data doppelgänger analysis, demonstration in machine learning, and discussion of application scenario*

Data doppelgänger refers to the phenomenon that data from different taxa are too similar in the selected characteristics and become separate from each other (Wang, L.R et al., 2022). In machine learning datasets, we often use multi-dimensional features to describe the data and help classify and satisfy more complex algorithmic models, thus avoiding the widespread occurrence of this phenomenon. In biomedical data, there is often a large similarity between the same character data from different groups. Within the same individual, the transcriptome of neighboring cells in the same tissue site tends to be the same, which may be due to the same microenvironment in which they are located (Rabinowitz, J. S. et al., 2017). Even cells from the same site in different individuals may have similar transcriptome data. However, when we classify the training and test sets according to the traditional method and test the model, the resulting high-performance model is likely to be affected by this doppelgänger data, giving an inflated accuracy. In different situations, we want the mo to learn how to extract features, rather than simple, lazy learning and training. Focusing on health and medical science, we discuss the effects of this effect in categories.

| Cell/Gene | Gene 1 | Gene 2 | Gene 3 | … | Gene m | Type |
|---|---|---|---|---|---|---|
| Cell 1 | | | | | | A |
| Cell 2 | | | | | | A |
| Cell 3 | | | | | | C |
| … | | | | | | B |
| Cell n | | | | | | E |

Table 1: Transcriptomics data form with the amount of expression of each gene in each cell

First, at the data level, doppelgänger effects are not unique to biomedical data. For any classification task, data with the same label and similar feature values are separated from each other and will bring different degrees of separation effect. For example, in the example below, we took about 300 photos of different leaves that all suffer from different kinds of diseases, and there are three diseases. We hope that the image recognition model can obtain the ability to reclassify after training. At this point, we do some processing on the original image to simulate the doppelgänger effect in various omics and expression profiles. The traditional CNN model uses a convolution kernel to extract the features of the image through convolution operation and finally classifies it through the fully connected layer (Singh, A. K. et al., 2022). Convolution, activation, maximum pooling, regression, and other operations are all done based on the most original image pixels. In transcriptome analysis, we obtain the expression of each gene in different kinds of cells and can classify them by reducing them to low dimensions by PCA (Duan, Y. et al., 2021). Then each gene becomes a feature. For computer vision, each pixel is its corresponding feature. Let's take a photo of a leaf and use it to generate many similar photos at the same time. By adding noise, we gentile avatar photos. For humans, we can see at a glance that although the distribution of these scattered dots is very different, the most crucial leaves in the picture are exactly the same. For computers, the whole picture is a huge matrix of different RGB values (Mingfeng, W. et al., 2021), and all the information is represented in numbers. By adding noise, we ensure that the characteristic values of the leaf photos classified as 0, that is, the first kind of disease, are not the same in the dataset. In this way, every time the model extracts feature and processes them, it is also unaware of potential data twins.

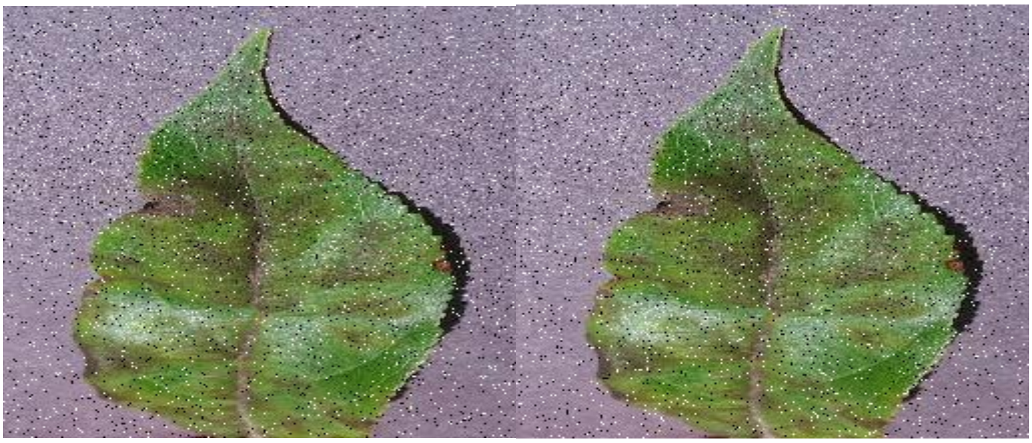
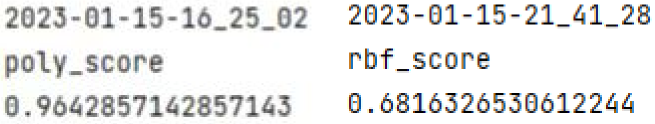


Figure 1: Diseased leaves, left and right are from the same photo, but treated with different noise

First, we train the model using the SVM algorithm. The training set data of the three classifications are the same, and the algorithm finally obtains a good classification effect by learning to fit the training set data continuously. On the test set, if we divide it by 20%, then about 20 photos in each category are used for testing. For the test data are classified as one, they all come from one photo. It's just that the addition of different noises causes their characteristic values to be different. Next, we test, and as expected, the effect is very good, because SVMs can be achieved optimally with different kernel methods in multi-classification (B., P. and Nagaraju, V., 2022).



```
2023-01-15-16_25_02    2023-01-15-21_41_28
poly_score             rbf_score
0.9642857142857143     0.6816326530612244
```

Figure 2: Left: Performance of SVM models on test sets from raw data partitioning. Right: Performance of the SVM model. After replacing the original training set with a picture from the dataset with a real classification of 0

For the photos with categories 1 and 2, the computer did learn how to classify them, the basis for each classification, and the underlying patterns between the photo pixels. For those with 0 disease type, we supplemented the test set with the original 20 data sheets to test whether the model could still perform well.

This time, we found that the classification effect of the machine is very poor, and it is even not as good as random classification (guess all categories). This shows that the machine is confused by these data avatars and has not learned how to extract features at all. There are two types of classification correctly, and the data products are evenly distributed, so the accuracy rate is also as we expected (about 66.7%).

It is worth noting that our classifier choice is the SVM model, which is representative of the discriminant model. For classification tasks, the discriminant model can only learn the boundary between each classification by remembering the previous data, and gradually strengthening this boundary to reach the extreme value of the loss function (Xie, Q. et al., 2022). That is, it does not describe what the data looks like. Generative models such as the Naive Bayes model can use association to summarize specific patterns for each classification from the data they have seen. Key features are given high weight and categorized according to probabilistic outputs. Therefore, SVM's poor performance in the face of real data is predictable, it can only learn from the photo that has seen the disease 0, and does not summarize the characteristics of the disease 0.

That is to say, in the face of the same data set, regardless of whether there is a doppelganger effect, no matter which cross-checking method is used, our model can only be trained and tested on these data. We can't see the performance of the model without the help of other data. A generative model, such as a deep neural network model with an encoder, might be trained on subtle data and be able to detect a certain amount of data. A lazy classifier may not be able to recognize data or classification at all. We expect classifiers to truly learn and associate, but for different tasks, whether the effect of the avatar effect is serious needs further discussion.

***Next, we analyze the practical application scenarios of the avatar effect in biomedical data and discuss how and whether to eliminate it.***

In the task of judging whether a patient's tissue is diseased through omics analysis, or in the task of determining whether there is a disease, we often find the most enriched molecules in the genome, transcriptome, proteome, and metabolome through omics analysis. Because most genes are expressed almost identically in cells, it is these different genes and abnormally occurring proteins, such as IL-17 (associated with mucosal immune responses), that have a critical impact on the outcome (Zhou, Y. et al., 2022). These data need to be solved using bioinformatics analysis and prior knowledge. In analyses to study the regeneration of amputated limbs in zebrafish, each cell needs to have a specific location memory to help them differentiate into new tissue. This classification task is characterized not by the equal assignment of transcriptomes to each of the genes, but by focusing on the genes that are enriched where the cells are located. If the PPCC method is used to calculate the similarity of data for each pair that may have a doppelgänger effect, then genes that are expressed the same but are not helpful for classification are likely to erase useful data (Wang, L.R et al., 2022). Therefore, in omics analysis, we only need to ensure that there is no information from cells from the same individual or the same location and ensure that the entire data set is large enough to reduce the impact of the avatar effect. In other words, the classification problem of diseases that are only plagued by the doppelganger effect only when the amount of data is so small that the proportion of repetitions is too high, and the expression of almost identical diseases from different patients are plagued by the doppelganger effect.

In the task of predicting molecular properties, we can choose different feature extraction methods to ensure that the model can extract features from the information of each atom and each chemical bond as much as possible. Instead of following the traditional QSAR method, it is a simple judgment based on functional groups (Barratt, M.D., 1998). Especially for some special molecules, such as the following two molecules tested by myself (Jin, W, 2022).

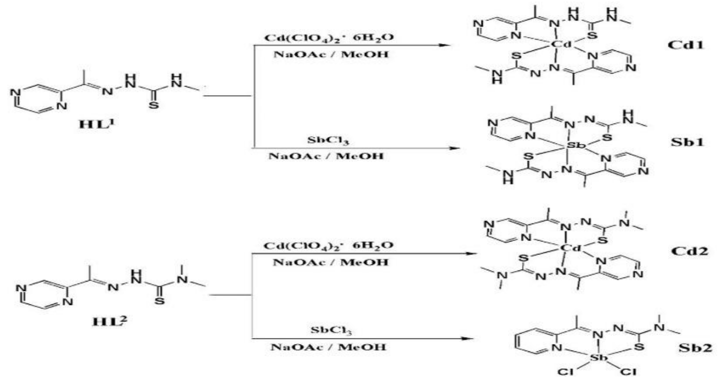


Figure 3: The metal chromium and antimony correspond to the structure of different Schiff base complexes

Just a methyl group replaces a hydrogen ion, and the corresponding complex structure is highly similar, and the bacteriostatic properties of its molecules differ by 10 times.
If our model has never seen changes in these sites, it will not be able to make the correct classification. Not only that, for models with simple attention, even if this particular molecule is taken to train and test the model, its generalization ability will not necessarily improve.

In the biological sciences, many similar structures bring similar functions, which is a commonality. The machine can classify by simply remembering this pattern, and the effect will not be too bad. For the machine to truly learn the effect of each atom on the entire molecule, it needs to disassemble the macromolecule into atoms, or even the characteristics of electrons to train the model. This makes training costs much higher. In addition, we can also improve the control of the model over the entire molecule by adding a super node (Li, S.( 1 ) et al., 2020).

In summary, the avatar effect appears in various machine learning datasets, but the impact on different tasks is related to the proportion of the dataset, data quality, and classification task type. For biomedical data, the effect of the avatar effect is even worse, often leading to the expansion of the model effect. Eliminating the avatar effect requires certain prior knowledge, and we can require the avatar effect data with different criteria for different tasks. Describing data with lower-level, lower-dimensional, more information-rich features can mitigate the impact. But rich, contextual data, more complex model generation, and training cost are the most effective factors to ensure real accuracy.

## Reference:

B., P. and Nagaraju, V. (2022) 'Preliminary Sensing of Wrong-Lane Accidents by Comparing Random Forest with Logistic Regression, Decision Tree and SVM Algorithm for Better Accuracy', *Journal of Pharmaceutical Negative Results*, 13, pp. 497–506. DOI: 10.47750/pnr.2022.13.S04.055.

Barratt, M.D. (1998) 'Integration of QSAR and in Vitro Toxicology', Environmental Health Perspectives, 106, pp. 459–465. doi:10.2307/3433795.

Duan, Y. et al. (2021) 'Low-complexity point cloud denoising for LiDAR by PCA-based dimension reduction', *Optics Communications*, 482. DOI: 10.1016/j.optcom.2020.126567.

Li, S.( 1 ) et al. (2020) 'MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities', Cell Systems, 10(4), p. 308–322.e11–322.e11. doi:10.1016/j.cels.2020.03.002.

Mingfeng, W. et al. (2021) 'Study on the quality detection method of biochar based on red–green–blue image recognition technology', *Biomass Conversion and Biorefinery: Processing of Biogenic Material for Energy and Chemistry*, pp. 1–9. DOI: 10.1007/s13399-021-01957-1.

Rabinowitz, J. S. et al. (2017) 'Transcriptomic, the proteomic, and metabolomic landscape of positional memory in the caudal fin of zebrafish', Proceedings of the National Academy of Sciences of the United States of America, 114(5), pp. 717–726. https://doi.org/10.1073/pnas.1620755114

Singh, A. K. et al. (2022) 'Multichannel CNN Model for Biomedical Entity Reorganization', *BioMed Research International*, pp. 1–11. DOI: 10.1155/2022/5765629.

Wang, J., Zhang, Z.-M. and Li, M.-X. (2022) 'Synthesis, characterization, and biological activity of cadmium (II) and antimony (III) complexes based on 2-acetyl pyrazine thiosemicarbazones', Inorganica Chimica Acta, 530. doi:10.1016/j.ica.2021.120671.

Wang, L.R., Wong, L. and Goh, W.W.B. (2022) 'How doppelgänger effects in biomedical data confound machine learning', Drug Discovery Today, 27(3), pp. 678–685. doi:10.1016/j.drudis.2021.10.017.

Xie, Q. et al. (2022) 'Semisupervised Training of Deep Generative Models for High-Dimensional Anomaly Detection', *IEEE Transactions on Neural Networks and Learning Systems, Neural Networks and Learning Systems, IEEE Transactions on, IEEE Trans. Neural Netw. Learning Syst*, 33(6), pp. 2444–2453. DOI: 10.1109/TNNLS.2021.3095150.

Zhou, Y. et al. (2022) 'Acinetobacter baumannii reinforces the pathogenesis by promoting IL-17 production in a mouse pneumonia model', Medical Microbiology and Immunology, pp. 1–9. doi:10.1007/s00430-022-00757-2.