The report of How doppelgänger effects in biomedical data confound machine learning

This paper introduces us to the phenomenon "doppelgänger", which researchers have discovered when conducting scientific research on biomedical data, which refers to the fact that when the samples in the training and test sets of machine learning are highly similar, the models always perform better in training. This phenomenon still occurs even if the samples from these datasets are all generated independently. This phenomenon ultimately leads to excellent test results even when the actual training and construction of the model is not perfect, which is detrimental to the subsequent improvement of the model and affects the usefulness of the model in the face of widely varying data samples.

I don't think doppelganger effects are specific to biomedical data and as a senior student I am currently working on my final project. Proudly my final project is also about machine learning and deep learning, he is a convolutional neural network based sound classification system that has many similarities to some of the cases presented in that article. After reading the article, I tried to explore the existence of doppelganger effects on my final project. To my surprise, the doppelganger effects were also present in my graduation design. When my test set was more similar

to the training set (e.g. taken from the same sound source, even if the sound slices were longer spaced in the time dimension), my model always showed very good performance (up to 94% correct, as shown in Fig. 1), but once the data samples in the test set were Once the data samples for the test set were taken from open—world sound samples, the overall model performance dropped significantly to around 60% (as shown in Fig. 2).

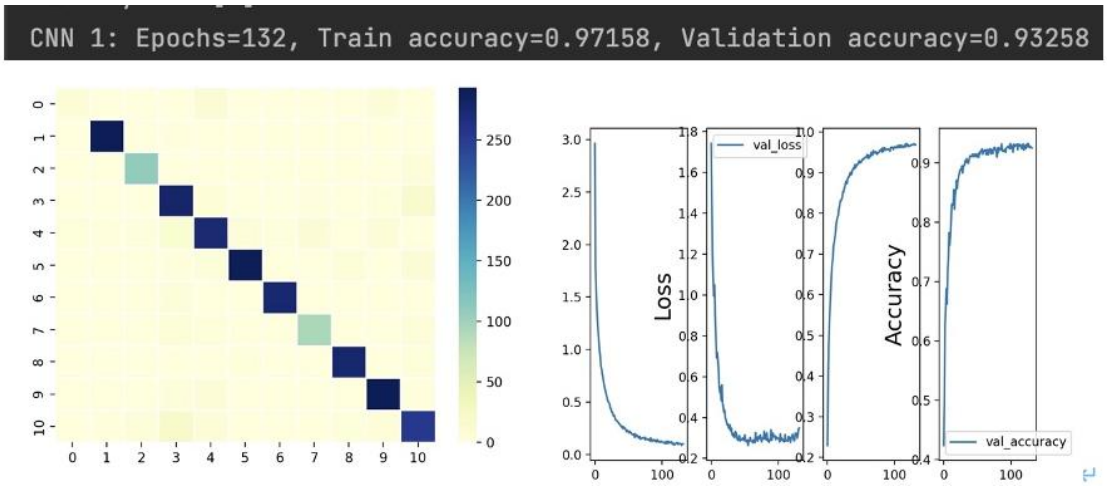

Fig.1

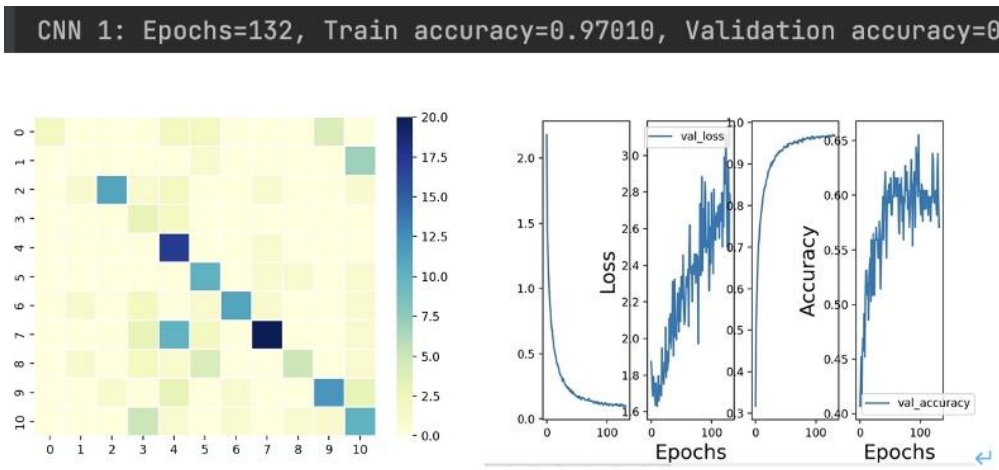

Fig.2

I believe that the emergence of doppelganger effects in the field of sound recognition is well documented, and when my test set is taken from the same source, significant doppelganger effects are always observed, which has many similarities to doppelganger effects in biomedical data science. Samples of data from the same type of vocal source are always similar even if they are all generated completely independently, much like samples of medical image data from the same type of patient in biomedical data science.

Although I have only explored the field of sound recognition classification for my final project design, it is clear that doppelganger effects are not specific to biomedical data.

Regarding how to avoid this effect in machine learning models for health and medicine, I think it is possible to start experimenting in the pre—processing phase of the dataset. I think the main reason for the appearance of doppelganger effects is the high similarity between the features extracted from the training and validation sets. If we can explore a way to reduce the similarity between the training and validation sets without adversely affecting the model as much as possible, wouldn't it be possible to reduce doppelganger effects.

This will not only reduce the similarity between the training and validation

sets, but also improve the performance of the model, making it more generalisable and ensuring good performance in the face of new data samples.

In the case of my final project, after data augmentation, the overall model performance has improved somewhat in the face of open world data samples (Fig.3).

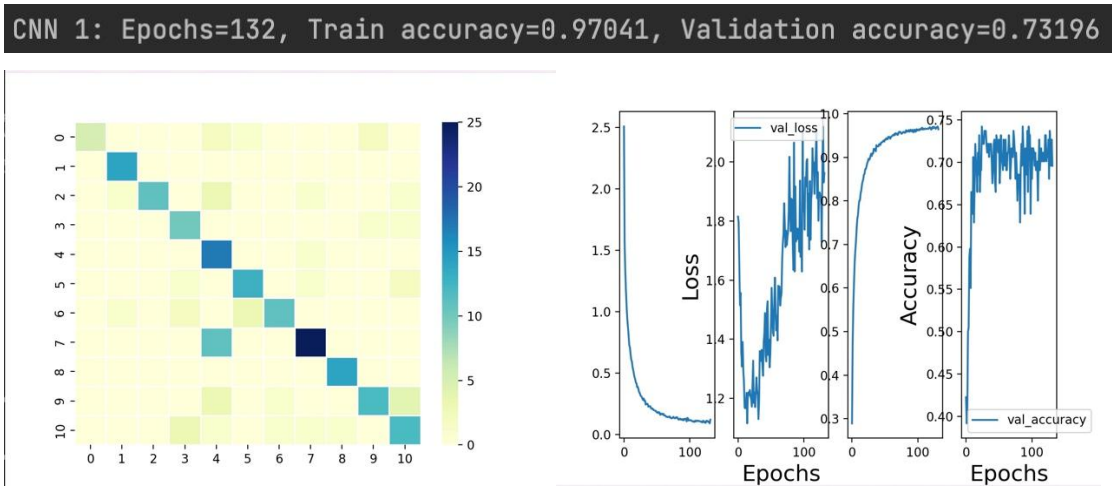

Fig.3

Secondly, the article suggests that doppelganger effects can be reduced by restricting the duality to the training or validation set. Instead of forcing the elimination of doppelganger effects, a different way of thinking would be to fuse doppelganger effects, and when all samples have similar features due to this new feature extraction method, doppelganger effects would lose their influence on the model because of their ubiquity.

These are just a few thoughts from my first encounter with doppelganger effects, so please bear with me if there are any errors or omissions.