**Generating a unified train test split**

To better evaluate and improve model performance, we used a unified train test split for all models. Then each model can be trained with the same training set and test on the same test set, so that we can better benchmark and test different model architectures. To do this, we randomly shuffled and divided the dataset into 10 folds with stratification, using 0 as the random seed. Each fold has 90% of the dataset as training set and 10% as test set. We made the folds by preserving the percentage of samples for each class in both the training sets and test sets.