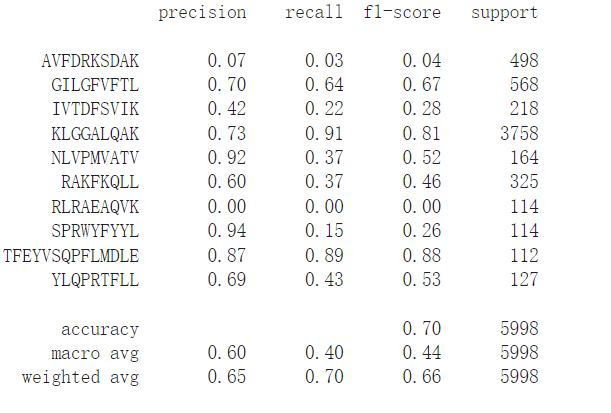
Results and Discussion

Reporting on the experiments with discussion on insights. Technical challenges are to be discussed here too.

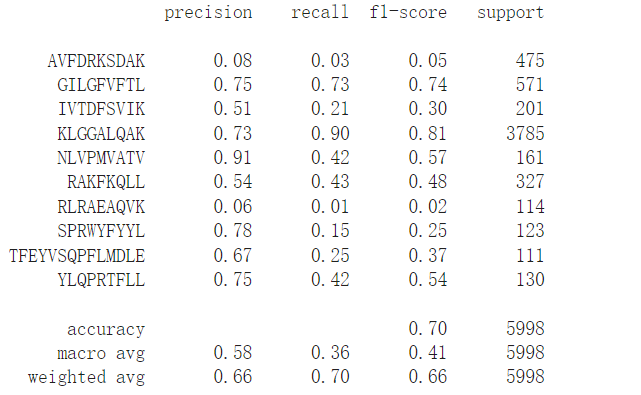
In task6, for the human dataset, K neighbors classifier is used to classify the 15000 data with the top 10 epitopes sampled hierarchically.

## Prediction of human epitopes (alpha chain)



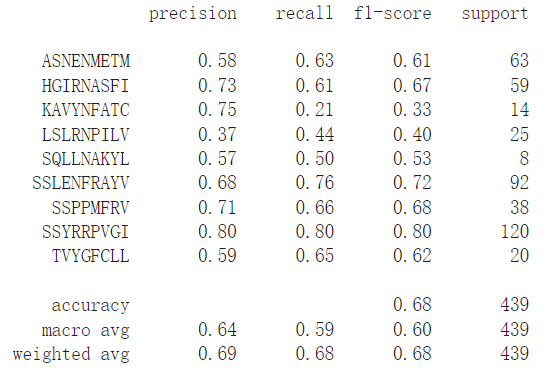
The overall accuracy of knn model is 0.7 and the macro average F1 score is 0.44. Notably, the F1 score shows significant difference in predicting different epitopes. For instance, F1 score of KLGGALQAK is the highest (0.81) and F1 score of RLRAEAQVK is 0. Obviously, a part of the lower f1 value reduces the average f1 value. The model performs well on some epitope categories and poorly on others. One potentially reason is related to the diversity of the TCR sequence corresponding to the epitopes. If the model can predict epitopes well, the TCRs sequence corresponding to them may have a higher similarity. This is reflected in the clustering plot by the more concentrated clustering of the same epitope and the smaller number of small clusters. On the contrary, if there are TCRs with a lot of categories and complex structure sequences that can specifically bind with a single epitope, the model prediction performance will be bad. In other word, TCRs’ specificity and cross-reactivity have a significant impact on model’s performance.

## Prediction of human epitopes (beta chain)

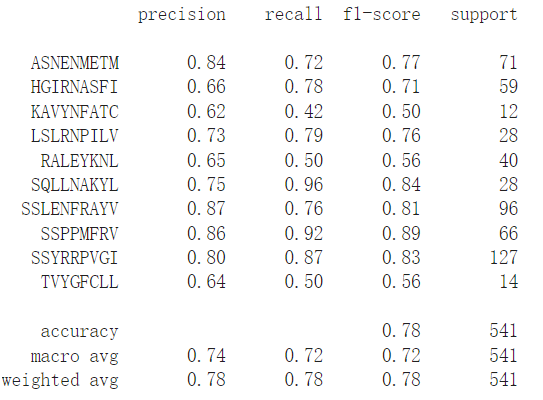


Comparing the prediction results of human alpha chains, the beta chain has an overall accuracy 0.70 and macro average F1 score 0.41 which is similar with the prediction evaluation of alpha chain. Similarly, the model performs poorly in predicting some specific epitopes such as AVFDRKSDAK and RLRAEAQVK. Generally, TCR specificity dictates that more than one TCR category can recognize the same epitope. That will reduce the predictive performance of some normal machine learning models.

## Prediction of mouse epitopes (alpha chain)



## Prediction of mouse epitopes (beta chain)



From the prediction of mouse epitopes, alpha chain and beta chain have overall accuracy 0.68, 0.78 and F1 score 0.6, 0.72 respectively. Noticed that the overall performance of mouse epitope prediction is greatly better than human. This may be related to the simpler structure of the mouse.

On the one hand, biological TCRs have a high complexity in feature space initially. On the other hand, the knn model is sensitive to the unbalanced data. Moreover, it has a high requirement for the feature space distance, which will have an impact on the final prediction results. Optimizing to a more complex and advanced model may be more helpful.