Content

I. Introduction

Due to the one-and-a-half-year epidemic, global vaccine research and development attracts the attention of each country. Different methods for developing vaccines have been tested in these few years. However, all the vaccines aim are to make the human body produces antigen-specific antibodies. Then how to find out the antigen becomes the most significant part. The subregion of antigen proteins is called epitope regions, which helps B-cell inducing antigen-specific immune responses in vivo. By recognizing the epitope regions of antigen proteins – the part of an antigen molecule to which antibody attaches itself, they produce large amounts of antigen-specific antibodies that inhibit antigen protein function.

Therefore, the computer-based automated prediction of epitope regions is extremely beneficial for developing vaccines to induce antigen-specific antibody production. It is currently one of the key research topics in vaccine development that will be widely useful for future medical data analysis. In conventional, the method to predict epitope cost huge labor, money, and time, like conducting three-dimensional structures and analyzing the antigen by X-ray or nuclear magnetic resonance (NMR) spectroscopy and focused only on the target sequence in the amino acid sequences of an entire antigen protein and have not thoroughly considered its sequence and features as a whole.

In this report, we will show several prediction methods of how to predict B-cell epitope and SARS to consider the characteristics of a whole antigen protein in addition to the target sequence, using Logistic Regression, KNN, Decision Tree, and SVM. We use CNN to apply the deep learning method for feature engineering. In addition, we will also show a deep learning method based on short-term memory with an attention mechanism and compare the accuracies and results of each model.

II. Dataset

A. Data Source

We used the existing dataset, which combines information extracted from both the Immune Epitope Database (IEDB), a free epitope database and prediction resource, as well as Universal Protein Resource (UniProt) – a comprehensive collection of protein sequences and their annotations. As a result, there are two data files, one for B-Cell, and one for SARS.

B. Data Introduction

The B-Cell and SARS dataset consist of an activity label which indicates – based on the number of antigen-binding sites (0 or 1) – whether a peptide exhibits antibody-inducing activity, as well as other non-class attributes: parent protein ID, parent protein sequence, start/end positions of the peptide, peptide sequence, 4 peptide features: the possibility of β turn is calculated by chou_fasman method, which helps to analyze the relative frequency of amino acid; use emini to get relative surface accessibility that represents the surface area of a biomolecule that is enabled to access to a solvent; Antigenicity, refers to an interaction between antigenic determinants (epitopes) and antibodies or specific T cell receptor for antigen; Hydrophobicity, means insoluble in water, do not across and maximum the contact the water, the opposite of hydrophilic. Also, there are 4 protein features, one is the isoelectric point, which represents the pH when the molecule carries no net electrical charge; Aromaticity allows the electrons in the molecule to be delocalized around the ring, increasing the molecule's stability; Hydrophobicity is the same as peptide feature; Stability, defined as the quality of maintaining a constant character in the presence of forces that resistance to change. The total number of records is 14387 for the B cell (training) dataset and 520 for the SARS (test) dataset. Presented antibody proteins were restricted to Immunoglobulin G (IgG), the most recorded antibody type in IEDB, and records that

represent different quantitative measures of antibody activity for the same peptide had been excluded for convenience.

C. Data Analysis and Visualization

Figures 3, 4, 6 below are the comparison of the distribution of the peptide properties between B-Cell and SARS. Almost all of them follow the normal distribution. As a result, we do not need more standardization or normalization to process the data. In both B-Cell and SARS datasets, one-third of data targets are 1, the most target is 0 (Figure 1), which shows the target shares the same distribution. The peptide character count distribution is not so clear, in B-Cell the 8,10 and 15 are top 3 counts and others are around or lower than 1000. In SARS, the peptide character count is quite abnormal because one specific length of peptide is way more than others, nearly 200. Others are all below 25(Figure 2). For code frequency, all characters are counted separately. Both B-Cell's and SARS's most characters is "L" and the least is "W"(Figure 3). chou_fasman is an empirical technique for predicting the secondary structure of proteins. This method is based on the analysis of the relative frequency of each amino acid in the alpha helix, beta sheet and transition based on the known protein structure solved by X-ray crystallography. This method can reach up to 50-60% accuracy in identifying the correct secondary structure, which is far lower than the GOR method or technology based on modern machine learning (Figure 4). kolaskar_tongaonkar and parker are peptide feature, representing the antigenicity and hydrophobicity

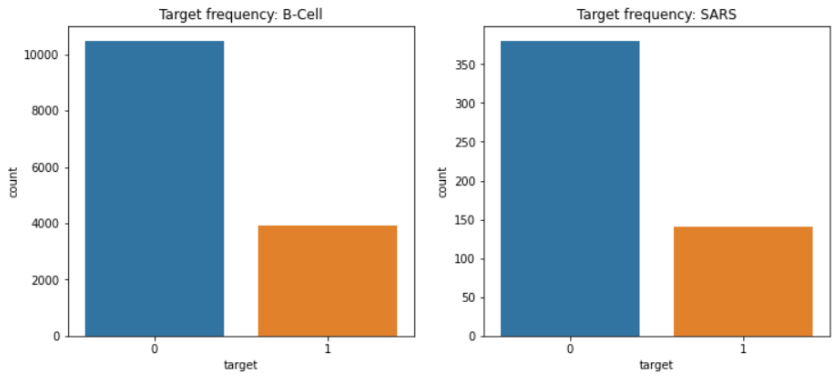

Figure 1. Total_frequency

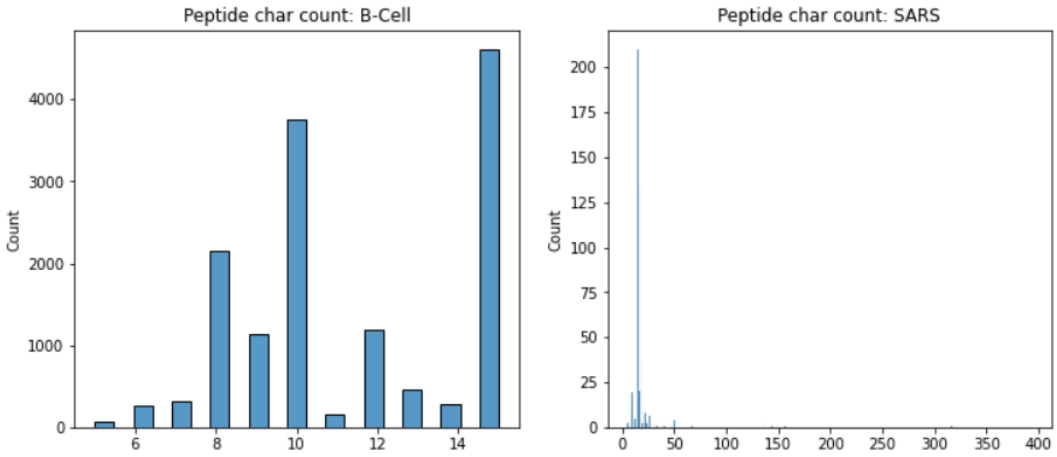

Figure 2. Peptide_count

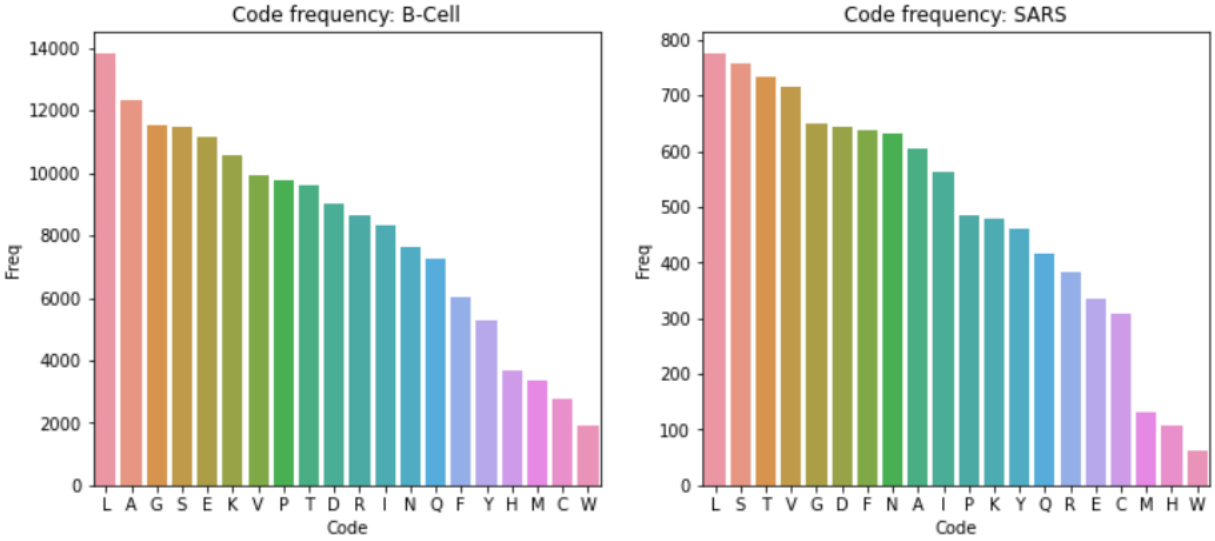

Figure 3. Code_frequency

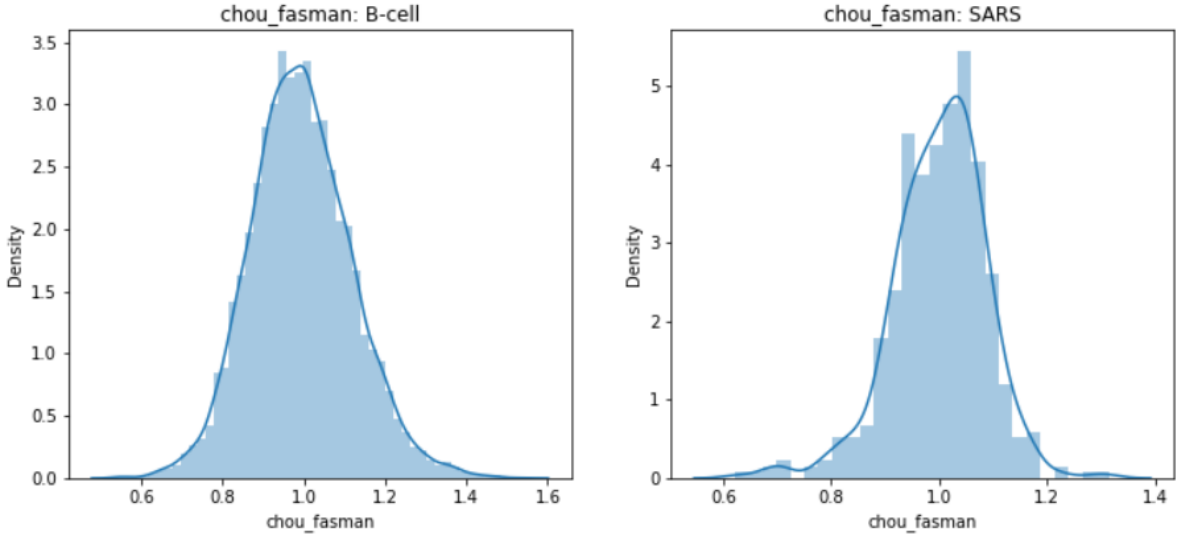

Figure 4. Chou_fasman

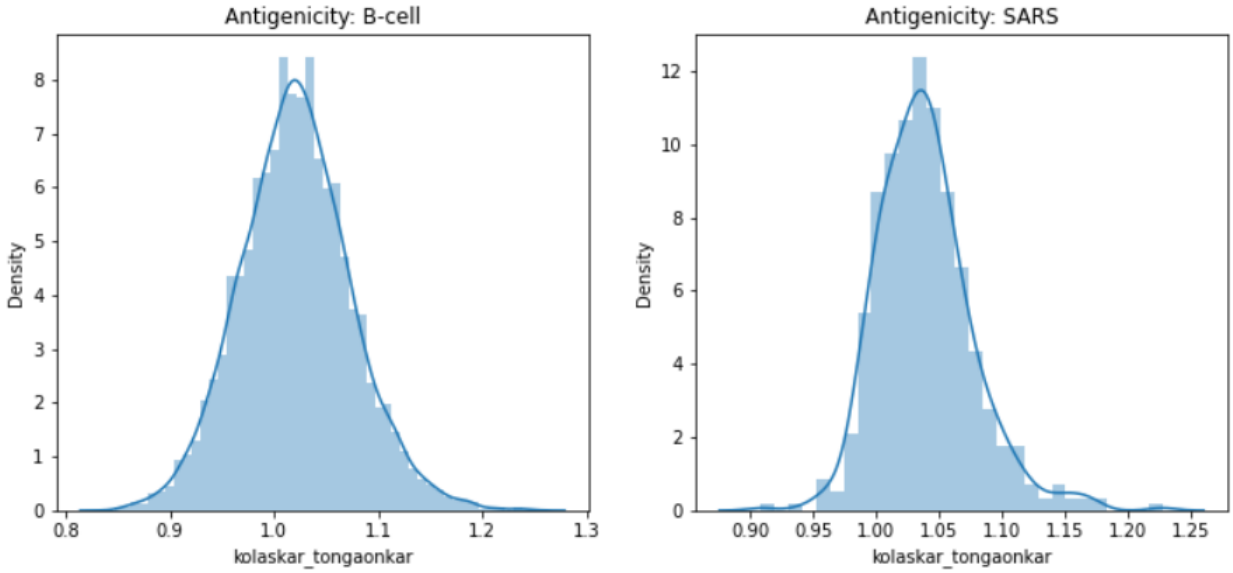

Figure 5. Antigenicity

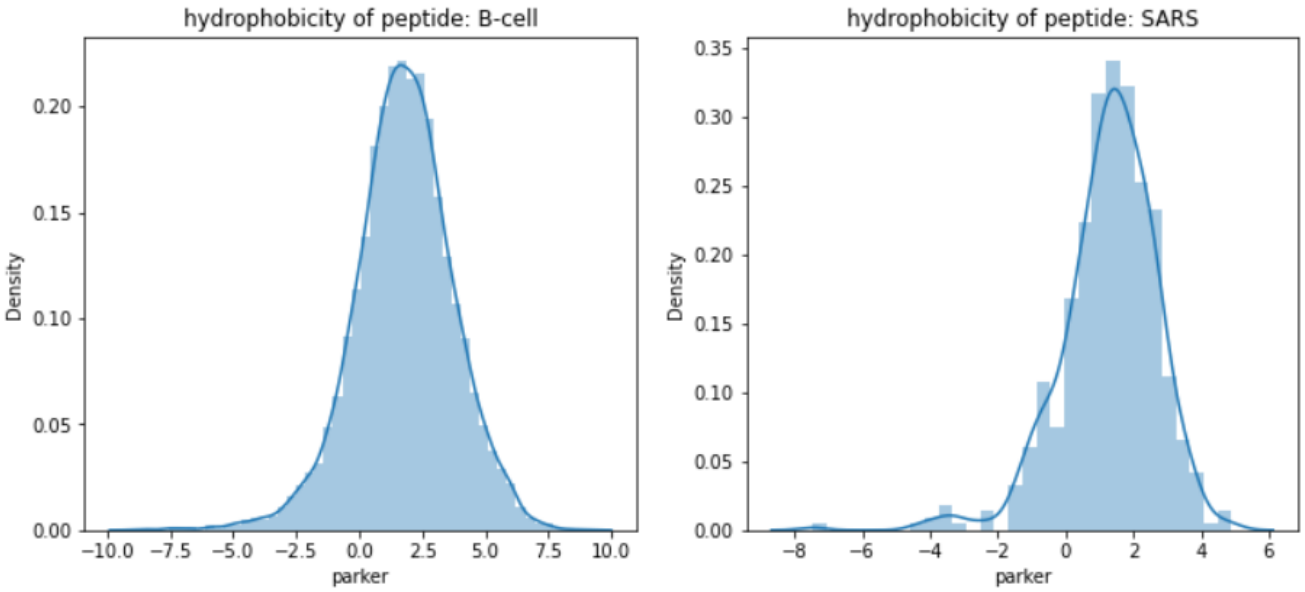

Figure 6. Hydrophobicity

D. Feature Selection and Engineering

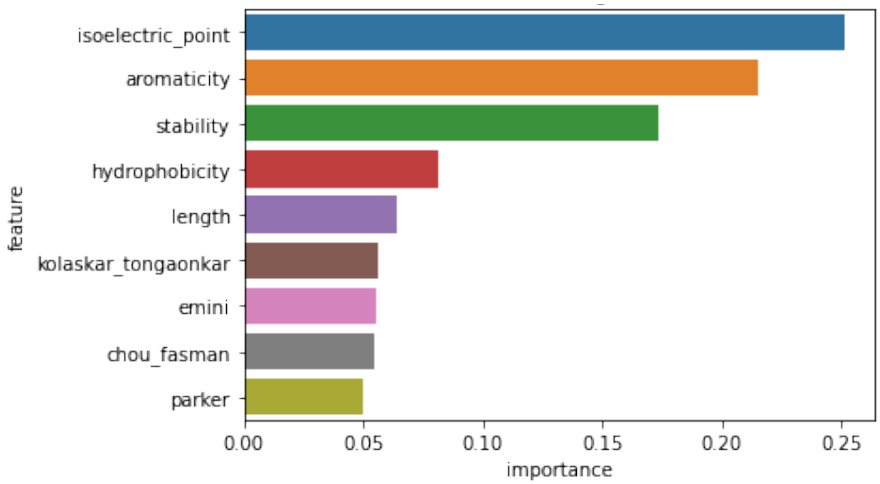

Figure 7. Feature Importance

To begin with, in order to explain the data feature, we decided to introduce the Explainable AI technique. We used a python package called Lime, where Lime stands for Local interpretable model agnostic explanations, which tries to give a local linear approximation of the model's behavior by creating local surrogate models which are trained to mimic the ML model's predictions locally. We built a combination deep-learning model of Lightgbm

and Lime with 5-fold cross-validation B-cell locally, to visualize the feature importance. Meanwhile, we also calculate each observation's peptide sequence length, using them as a feature as well in the prediction of SARS.

As for the peptide sequences column, we first used one-hot encoding and the K-mer sequence pre-processing method to convert the sequences into categorical variables. Then we built a Convolutional Neural Network to extract the feature map from the sequences in order to identify the local information of proteins. The constructed CNN module consists of two filter layers where has the size of 64 × 20 × 1 and 32 × 10 × 1 respectively; a max-pooling layer with size 2 × 2 is introduced between the two hidden layers, and then three dense layers are used afterward to flatten the feature map. Figure 8 illustrates the layout of the constructed CNN module.

Figure 8. CNN Model Layout

The extracted feature map of the peptide sequences was joint with the property data columns with high feature importance and passed for future training.

III. Machine Learning Models

A. Logistic Regression

Logistic regression is a statistic supervised Machine Learning model that predicts the categorical values, especially the binary classification. In this case, the individuals are assigned to "Positive", "Negative". If the Linear Regression aims to calculate the line in

fitting space, the Logistic Regression uses a sigmoid function to transfer the linear output to discrete values. For example, in order to check spam, the sigmoid function can map the result between 0 and 1. If the result is larger than 0.5, then it considers spam. If the result is smaller than 0.5, then it considers not the spam, the same in this model.

B. K-Nearest Neighbors

K-Nearest Neighbors is also a supervised learning method used for classification. K means the number of clusters. Before applying this method, K needs to be defined. The principle of this algorithm is when predicting a new value x, determining which category the x belongs to according to the category of the nearest K points. Here we use the sklearn-GridSearchCV, exhaustive search over specified parameter values for an estimator automatically. The parameter "cv" is set as 10 which means 10 folder cross-validation, refer to the previous study. Also, we set k ranges from 1 to 31. The figure below shows that when k equals to 5, the model will have the highest score.

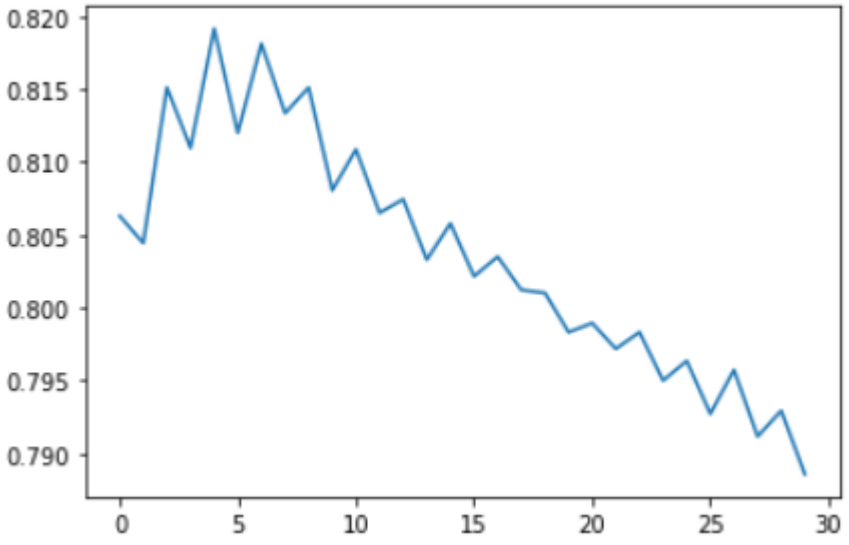

Figure 9. K Estimation

C. Support Vector Machines

SVM is a binary classification algorithm, which supports linear classification and nonlinear classification also regression as well. There are two important advantages of SVM: efficient in high space dimensional spaces; still effective when several dimensions are more than samples. However, there are too many samples' features, SVM may easily cause the overfitting.

D. Decision Tree

A Decision Tree is an unsupervised learning method, the goal is to create a model that learns simple decision rules from data characteristics to predict the value of a target variable. The advantage is easy to understand and the tree can be visualized. Also, it can handle both continuous data and discrete data, and no need to standardize the distribution. The disadvantage is the model may be too complex due to large data, and the result may not be the global optimal solution.

IV. Deep Learning Models

In this study, we also propose to use a deep learning model based on short-term memory with an attention mechanism. The RNN module is applied in the prediction model. However, the original RNN encounters gradient vanishing and gradient explosion problems. Therefore, we use LSTM, which is normally augmented by recurrent gates, which include input gates, output gates, and forget gates, to prevent back-propagating errors from gradient vanishing and gradient explosion. Furthermore, to better calculate the relationship, the bidirectional LSTM (BLSTM) is used in the prediction model, which combines forward LSTM and backward LSTM to preserve the upstream and downstream information of the peptide sequence by combining the two hidden states. In this model, the RNN module analyses the training input using 64 + 20 LSTM units in both directions within two layers.

V. Results & Comparison

A. Performance Measurements

In order to estimate the model's performance, we use the Precision, Recall, F1-score, Accuracy, receiver operating characteristic curve(ROC), and Area Under Curve(AUC). TP represents the number of True Positive, TN represents the number of True Negatives, FN represents the False Negative, N is the total samples in the dataset.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{N} \tag{4}$$

F1-score is also called F-score, the harmonic means of precision and recall. The highest value of the F1-score is 1.0 with perfect precision and recall and the lowest value is 0. ROC curve is a graph that shows the performance of a classification model at each threshold. The roc_curve() function needs two parameters, the True value, and the Predicted value. And it returns the False Positive rate and True Positive Rate, which represent the x-axis and y-axis respectively.

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

AUC measures the total area under each curve in a two-dimensional coordinate. The range of AUC can be between 0 and 1, the best case is TPR equals one and FPR equals zero. In general, the larger the area is, the better the performance.

B. Result Analysis

We used two approaches for every model, the first approach is to separate the B-cell data into a training set and test set, and see the result mainly in B-cell data; the second approach is to use B-cell as a training set, and SARS as test set, to observe the model performance beyond B-cell. These results show the performance in the same model between B-Cell and SARS. Support means the number of observations. In Logistic Regression, both scores of 0 predictions are higher than 1. We can see the recall and F1-score in B-Cell prediction has a huge difference, which are 0.97 and 0.11. And recall in SARS is 0.98 and 0.17 respectively.

	precision	recall	f1-score	support
0	0.75	0.97	0.84	3460
1	0.55	0.11	0.19	1288

	precision	recall	f1-score	support
0	0.76	0.98	0.86	380
1	0.73	0.17	0.28	140

Figure 10. B-Cell and SARS Results for Logistic Regression

In the KNN model, the gap has narrowed a little bit. Unlike Logistic Regression, the difference between 0 and 1 of precision and F1-score in both B-Cell and SARS are larger than Logistic Regression. The recall in B-Cell is 0.92 and 0.57, in SARS they are 0.88 and 0.21.

	precision	recall	f1-score	support
0	0.85	0.92	0.89	3460
1	0.73	0.57	0.64	1288

	precision	recall	f1-score	support
0	0.75	0.88	0.81	380
1	0.41	0.21	0.28	140

Figure 11. B-Cell and SARS Results for KNN

In the SVM model, we found that it somehow cannot process the dataset when the target equals 1. This leads to a model which only predicts negative. And it is very weird to see that the score of B-Cell and SARS are all the same. Since we do not find out why, so we may discuss it in the future.

	precision	recall	f1-score	support
0	0.73	1.00	0.84	3460
1	0.00	0.00	0.00	1288

	precision	recall	f1-score	support
0	0.73	1.00	0.84	380
1	0.00	0.00	0.00	140

Figure 12. Results for SVM

In the Decision Tree model, the performance is also has a huge variation between B-Cell and SARS. The score of B-Cell is all above 0.5, especially for predicting Negative, they are almost near 90%. However, in the SARS dataset, the scores are generally low. Only the value of predicting the Positive recall has over 0.5.

	precision	recall	f1-score	support
0	0.86	0.87	0.87	3460
1	0.64	0.62	0.63	1288

	precision	recall	f1-score	support
0	0.38	0.11	0.17	380
1	0.18	0.52	0.26	140

Figure 13. Results for Decision Tree

C. Comparison

TABLE 1. Comparison of the performances of different models on B-Cell training dataset for predicting B-Cell

Method	Precision	Recall	F1-score	Accuracy
Logistic Regression	70%	74%	67%	74%
KNN	**82%**	**83%**	**82%**	**83%**
SVM	53%	73%	61%	73%
Decision Tree	80%	80%	80%	80%
RNN with BLSTM	8%	-	4%	75.27%

Method	Precision	Recall	F1-score	Accuracy
Logistic Regression	**77%**	**77%**	**72%**	**77%**
KNN	68%	72%	69%	72%
SVM	53%	73%	62%	73%
Decision Tree	40%	26%	24%	26%
RNN with BLSTM	16.93%	-	0.9%	77.99%

The performance metrics of different methods of B-Cell are shown in Table 1. All the values are weighted average. KNN has the highest score in precision, recall, F1-score, and accuracy. The RNN with BLSTM is ranked second, which reached 75.27%. However, it has very low at other measurements except for accuracy.

The performance metrics of SARS are shown in Table 2. RNN with BLSTM performs best according to accuracy, reached 77.99%, however, the F1-score almost equals 0, and the precision is also much lower than average. Logistic Regression's accuracy is very close to the RNN module, which is 77%. And each of its values is higher than other models, with 77%, 77%, and 72% respectively. KNN and SVM have similar recall and accuracy, but SVM's precision and F1-score are relatively lower than KNN. The worst model is the decision tree, in which the performance distribution is 40%, 26%, 24% 26% respectively, far away from the other models. Compare Table 1 and Table 2, it is easy to observe that the KNN and the Decision Tree models are over fittings, which suggests they are not suited for these data inputs.

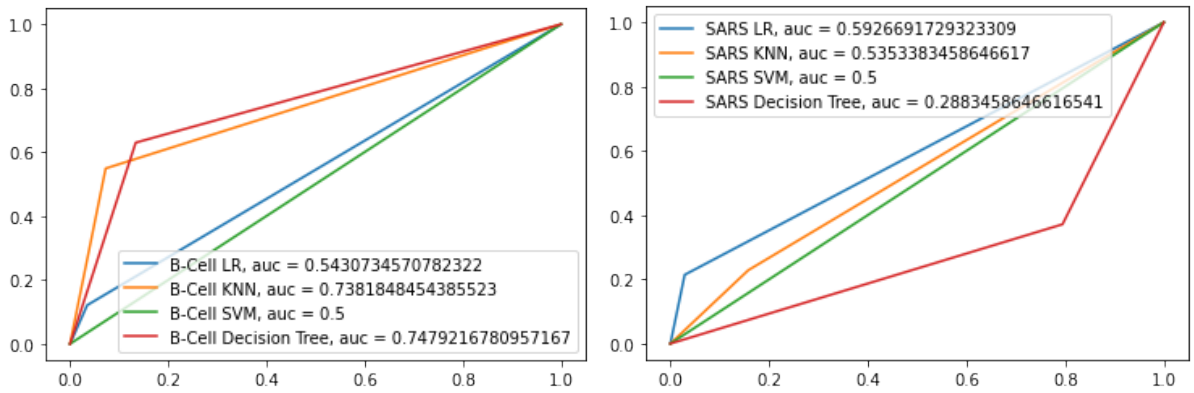

Figure 14. roc_curve and AUC

The Roc curves are plotted in Fig.14, which shows KNN increases dramatically in the early stage and AUC is much higher than other models. Since SVM can only predict one target, so maybe that is the reason AUC equals 0.5. For the SARS roc curve, the Logistic Regression boost in haste and the true positive rate of Decision Tree ranges from 0 to 0.4, and the false-positive rate ranges from 0 to 0.8 in the early stage, which shows that this model is not so well-fitted.

VI. Discussion & Future Work

In this project, there are several points and difficulties we have not figured out yet, such as why Deep Learning has such a lower score in precision and recall score, and why the SVM model cannot predict specific targets, and how to improve the Decision Tree. However, based

on the current comparison of various models' results, it appears RNN with BLSTM and Logistic Regression, combines with the CNN Feature Engineering, achieves the highest classification accuracy with 78% and 77%. For future work, we hope we can build a more robust model for peptide sequence pre-processing based on the amino acid's property instead of one-hot encoding in order to increase the accuracy. Moreover, we want to use this model to not only predict SARS with the epitope region but also the Covid-19 as well and apply it to various viruses in the future.

VII. Reference

[1]. Rux JJ, Burnett RM. Type-specific epitope locations revealed by X-ray crystallographic study of adenovirus type 5 hexon. Mol Ther. 2000 Jan;1(1):18-30. doi: 10.1006/mthe.1999.0001. PMID: 10933908.

[2]. Mayer M, Meyer B. Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. J Am Chem Soc. 2001 Jun 27;123(25):6108-17. doi: 10.1021/ja0100120. PMID: 11414845.

[3]. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997 Nov 15;9(8):1735-80. doi: 10.1162/neco.1997.9.8.1735. PMID: 9377276.

[4]. www.iedb.org

[5]. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019 Jan 8;47(D1):D339-D343. doi: 10.1093/nar/gky1006. PMID: 30357391; PMCID: PMC6324067.

[6]. http://www.thinkpeptides.com/bcell.html

[7]. Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986). https://doi.org/10.1038/323533a0

[8]. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol. 1978;47:45-148. doi: 10.1002/9780470122921.ch2. PMID: 364941.

[9]. Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol. 1985 Sep;55(3):836-9. doi: 10.1128/JVI.55.3.836-839.1985. PMID: 2991600; PMCID: PMC255070.

[10]. https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

[11]. https://towardsdatascience.com/protein-sequence-classification-99c80d0ad2df

[12]. https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/

[13]. https://towardsdatascience.com/explainable-ai-xai-a-guide-to-7-packages-in-python-to-explain-your-models-932967f0634b