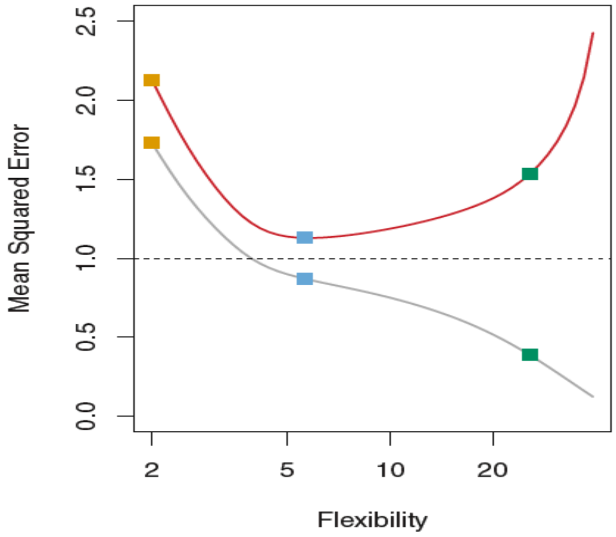
Group 13 Mengdi Gao, Joria Wang, Junyu Zhang

Table:

	Accuracy on training set	Accuracy on validation set
KNN	0.9351	0.9633
CNN	0.9858	0.9760
CNN with improvement	0.9780	0.9841

K-Nearest Neighbors Classifier:

K-Nearest Neighbors Classifier classify the data to different class by identifies the K neighbors in the training data that are closest to our test data point x1, then estimate the conditional probability for each class and classifies the test observation to the class with highest probability. The choice of K has a drastic effect on the KNN classifier obtained. Here are three KNN fits to the simulated data with 3 difference choices of K. When K = 1, the decision boundary is too flexible and overfitted the data, so we have very low train error rate but high-test error rate. There is a variance-bias trade off as K grows. As a result, we choose K=3 to get lowest train error rate. Before starting KNN method, we choose the PCA to reduce the dimension in order to reduce the running time of KNN. Moreover, it is efficient and accuracy by the PCA.



Convolutional Neural Network:

We split the training set in two part, the validation set and test set with a ratio of 1:9. A large validation set help us to obtain more information of the data to ensure the accuracy of the model.

The first layer we choose is the convolutional(conv2D) layer, so the CNN model can isolate the features that are meaningful from the transformed data. Each filter in the conv2D layer is a matrix of numbers. The matrix corresponds to a pattern or feature that the filter is looking for. In the image below, the filter is looking for a curved line .We set the filter in the conv2D layer as 32 and 64 respectively. Each layer is a matrix of numbers and the matrix corresponds to a pattern or feature that the filter is looking for. The second layer is the pooling layer which down sampling the data so we can reduce some computation cost. There is also a flatten layer so we can combine the features we got.

We choose RMSprop as our optimizers and choose 5 hidden layers and 30 epochs. The accuracy is about 97, but it takes hours to run. Then, we decrease the numbers of hidden layers and epochs and still receive a decent accuracy with better efficiency.

Convolutional Neural Network with Improvement:

To improve the CNN method, we expand the dataset with data augmentation technique to avoid overfitting. the improvement raise the accuracy form 0.9760 to 0.9841