William Hou

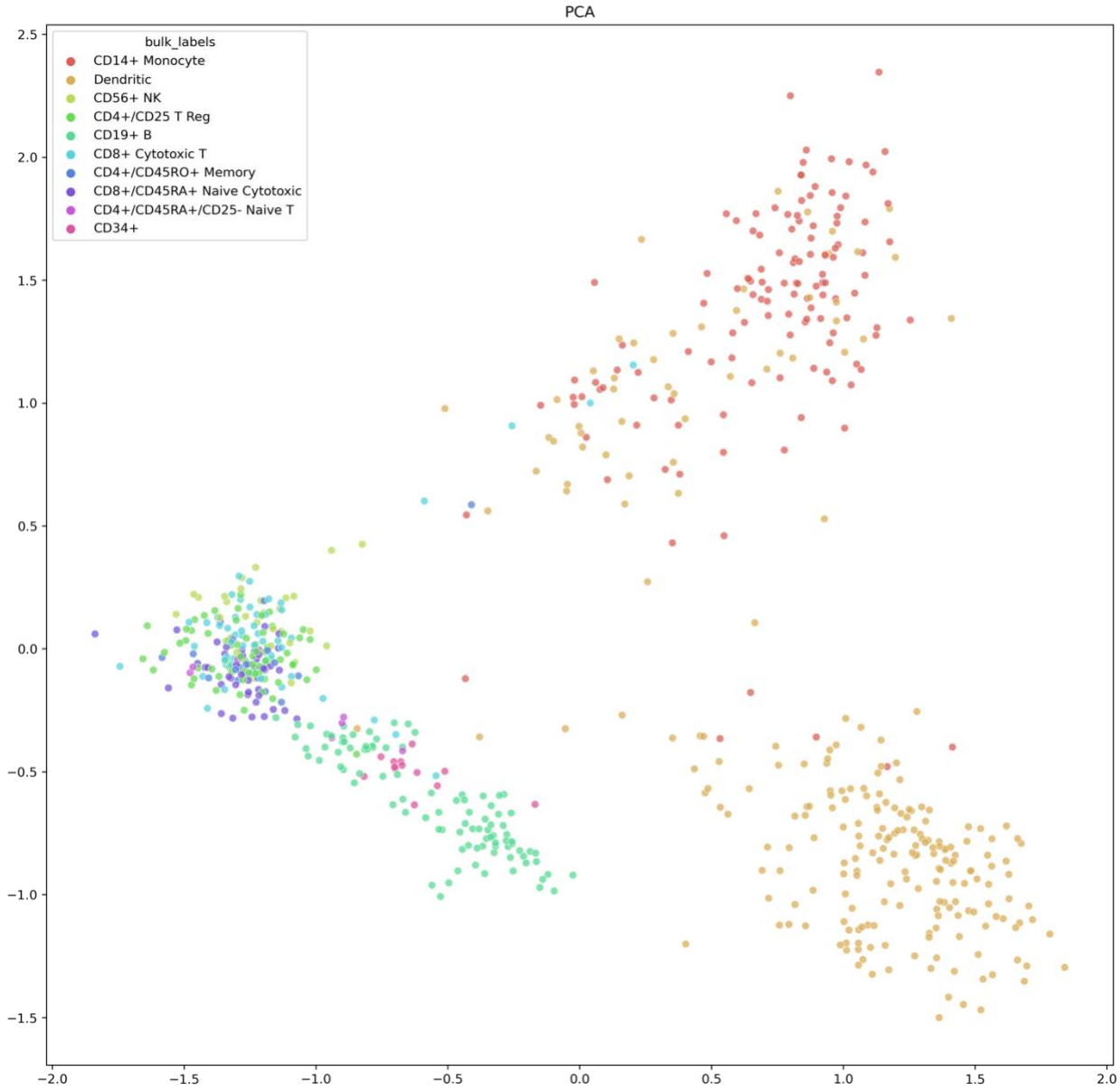
Prof. Liana Lareau

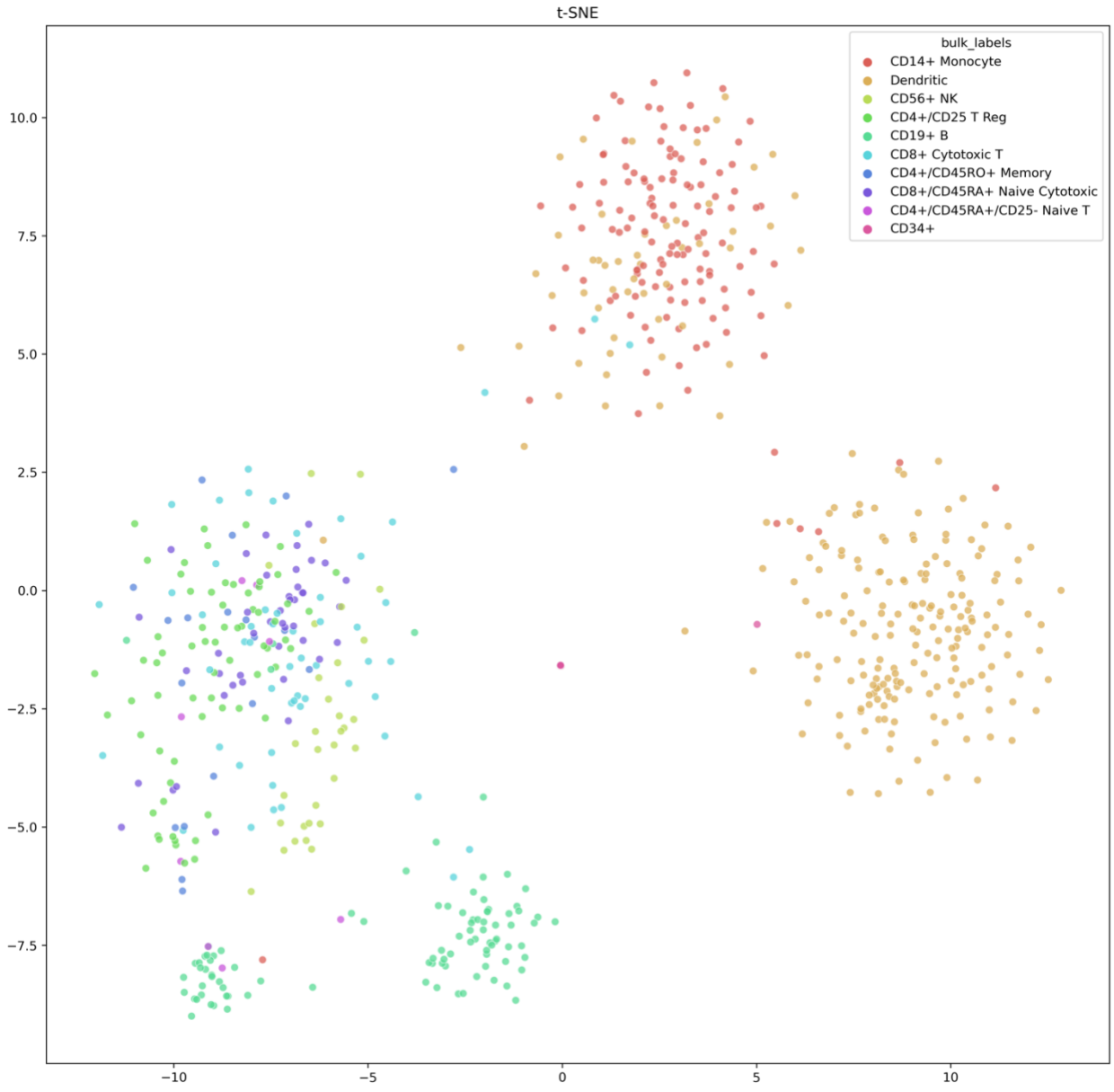
BioE 245

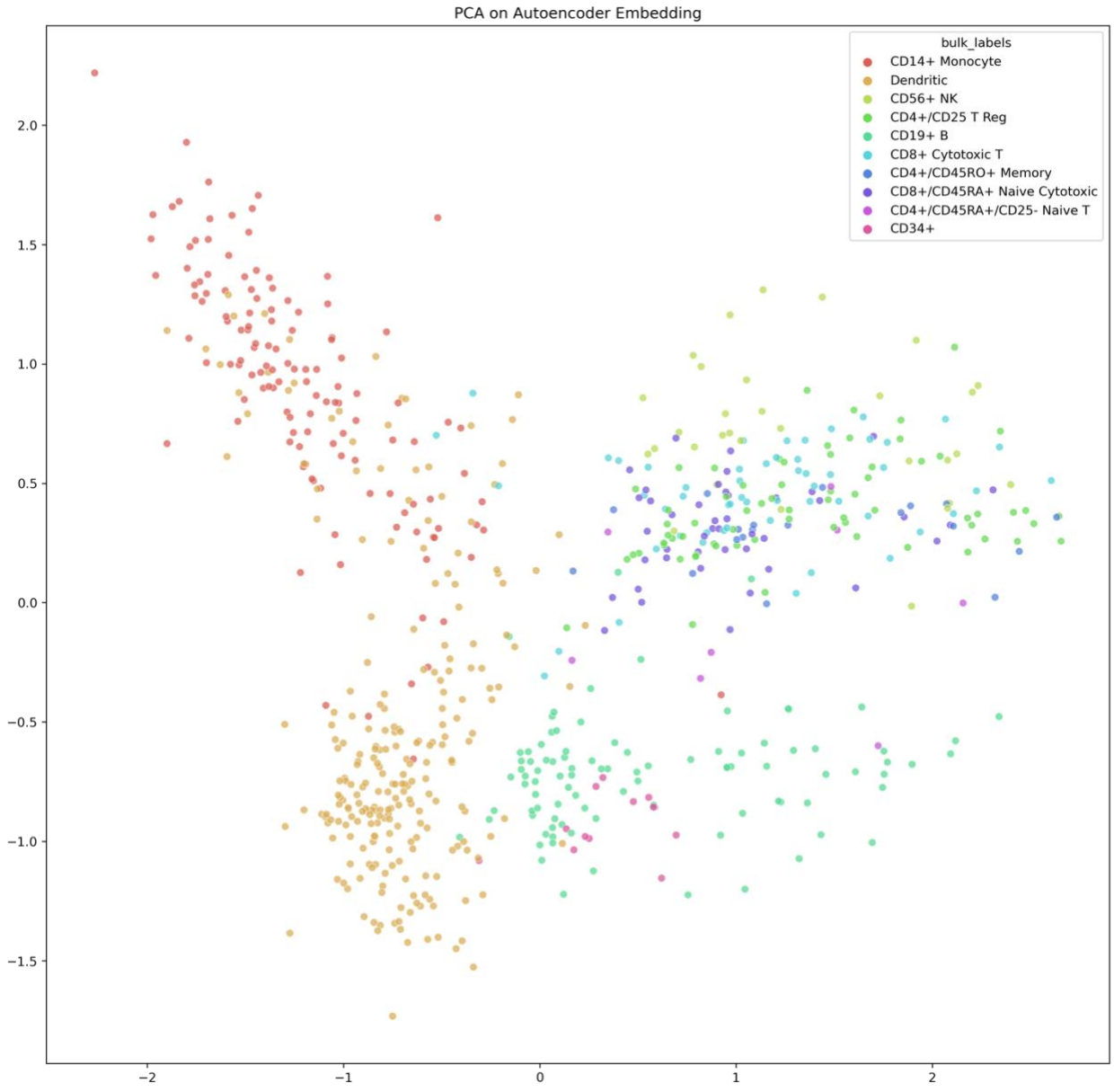
04/29/22

Part 1: Lower-dimension representations of the cells

Overall, t-SNE performs the best dimensional reduction of our PBMC data. After applying t-SNE, the monocyte and dendritic cell subsets are clearly separated from natural killer cells, T, and B cells. Moreover, the different subgroups of T cells (i.e., regulatory T cells, cytotoxic T cells, and memory T cells) are closer in this low-dimensional representation, suggesting a similar adjacency in the high dimensional space, which means the similarity amongst the are preserved after the transformation. Interestingly, B cells are clearly separated too, although there seem to be two separate groups of B cells, hinting at further investigating and updating of the bulk cell labels. PCA also achieves a similar outcome in distinguishing clusters of cell types, but the noise (misplacement of cells) is higher than that of t-SNE, just by eyeballing the difference. PCA on autoencoder's embedding roughly separates the cell types as well, though the spread of the grouping is larger than the previous two methods.


PCA
bulk_labels
CD14+ Monocyte
Dendritic
CD56+ NK
CD4+/CD25 T Reg
CD19+ B
CD8+ Cytotoxic T
CD4+/CD45RO+ Memory
CD8+/CD45RA+ Naive Cytotoxic
CD4+/CD45RA+/CD25- Naive T
CD34+
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


t-SNE
bulk_labels
CD14+ Monocyte
Dendritic
CD56+ NK
CD4+/CD25 T Reg
CD19+ B
CD8+ Cytotoxic T
CD4+/CD45RO+ Memory
CD8+/CD45RA+ Naive Cytotoxic
CD4+/CD45RA+/CD25- Naive T
CD34+
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10
−5
0
5
10


PCA on Autoencoder Embedding
bulk_labels
CD14+ Monocyte
Dendritic
CD56+ NK
CD4+/CD25 T Reg
CD19+ B
CD8+ Cytotoxic T
CD4+/CD45RO+ Memory
CD8+/CD45RA+ Naive Cytotoxic
CD4+/CD45RA+/CD25- Naive T
CD34+
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2

Part 2: Classifiers

After the dimensional reduction step, the cell type classification was compared on three different classifiers: a decision tree, an ensemble method of random forest, and a feedforward neural network. The best result comes from the feedforward neural network, achieving a loss of 0.60 and an accuracy of 0.87 on the test data. The model consists of an input layer, a flatten layer, and three sets of dense layers (first two with corresponding dropout layers). The model was trained on 30 epochs, achieving a loss of 0.02 and an accuracy of 0.99 on the training data.

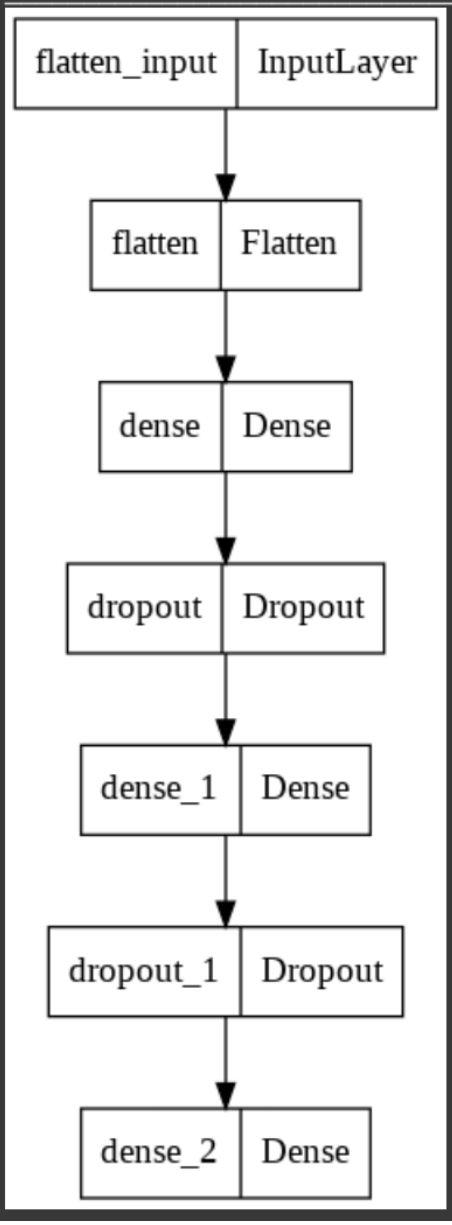
The decision tree is the next better model, averaging a cross validation score of 0.69. The model has a minimum samples per leaf of 10 and does not have a maximum depth.

The random forest model is the least effective model, maybe there are some technical errors involved, averaging 0.61 in cross validation score. The model has 30 feature estimators with max features set to square root and also does not have a maximum depth of the tree.

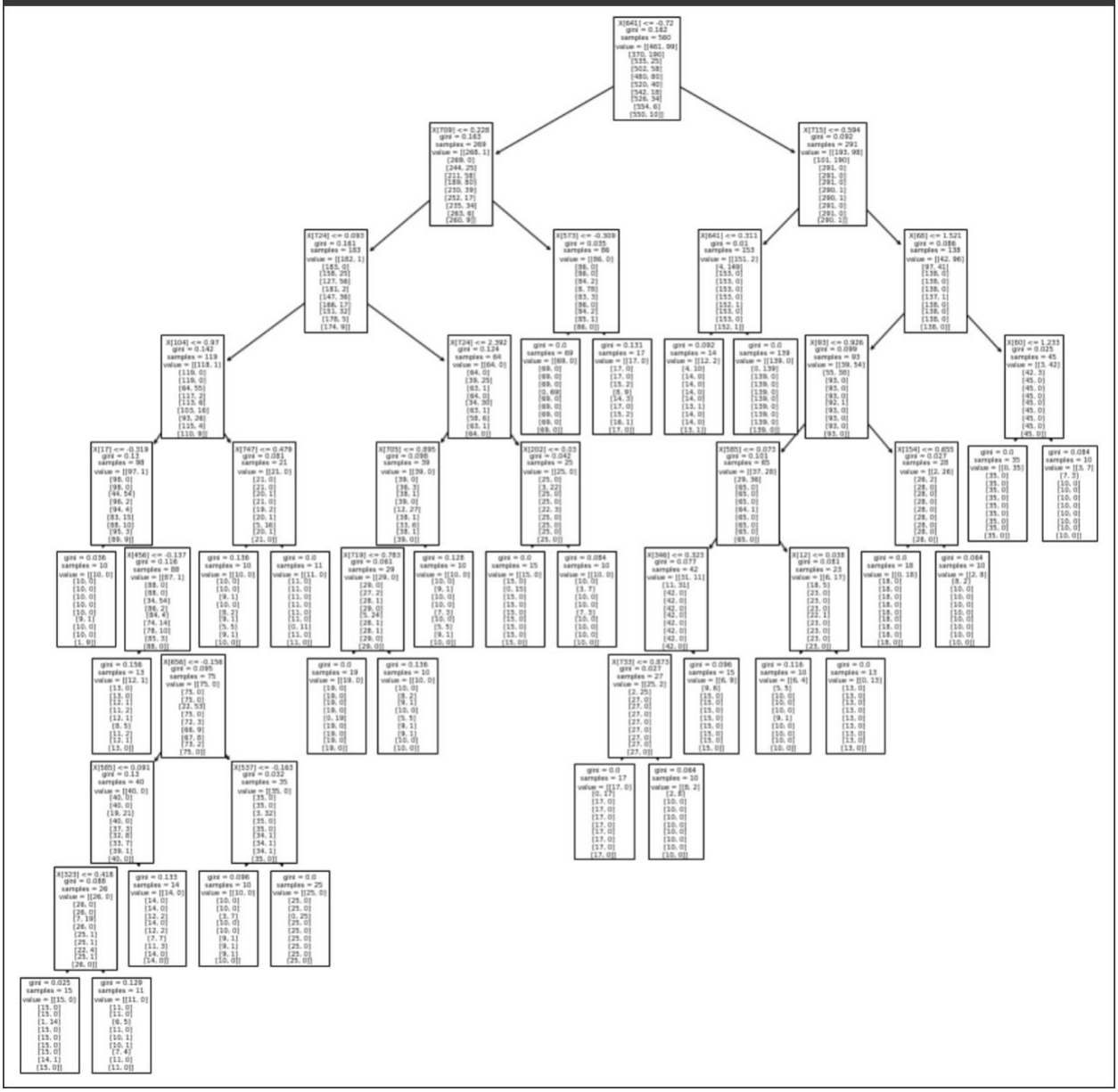
```
5/5 - 0s - loss: 0.6072 - accuracy: 0.8714 - 45ms/epoch - 9ms/step
These are the results from the Feedforward Neural Network classifier

Test loss: 0.6071997880935669

Test accuracy: 0.8714285492897034
 =================================================================
 Total params: 448,418
 Trainable params: 448,418
 Non-trainable params: 0
 _________________________________________________________________
```



```
These are the results from Decision Tree Classifier
Here are the scores:  [0.65714286 0.64285714 0.73571429 0.73571429 0.67857143]
Here is the mean score:  0.6900000000000001
```



```
These are the results from the Random Forest Classifier
Here are the scores:  [0.57142857 0.62142857 0.60714286 0.63571429 0.61428571]
Here is the mean score:  0.61
```