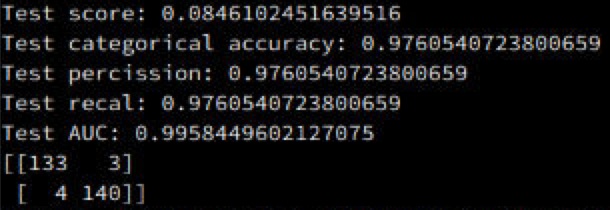
HW4 – Isabelle Hu

Nov 13 2020

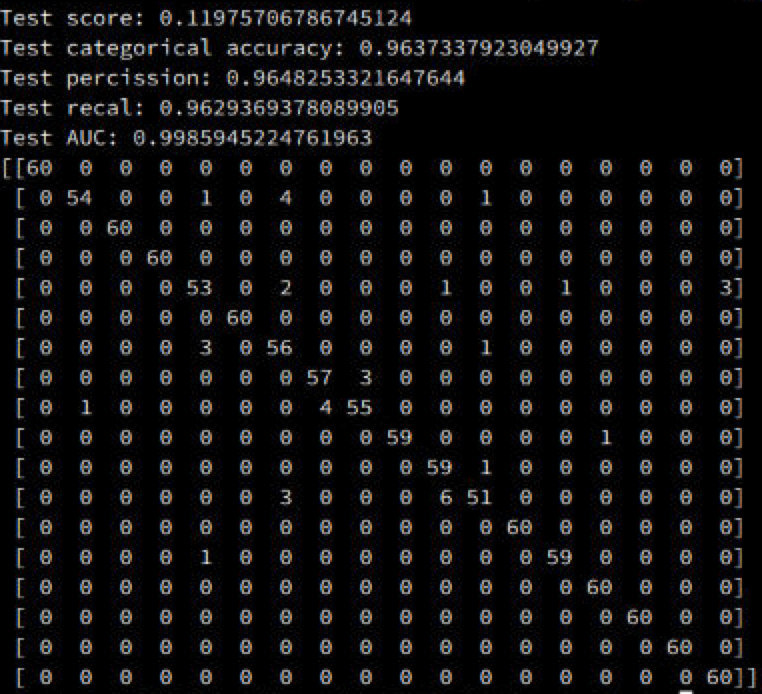
Q0: I trained the cancer type classifier and the normal/tumor classifier using the provided started code with some modification, and the evaluation results are shown below as a reference for the new performances in the other questions. The modifications I made to the starter code include the following: reduced the patience for learning rate reduction, implemented early stopping to save training time when no further progress is made, corrected the loss during evaluation from binary cross-entropy to categorical cross-entropy, changed the accuracy to categorical accuracy since we have multiclass classifiers, and implemented other evaluation metrics including precision, recall, and AUC.

Initial performance evaluated on test set:

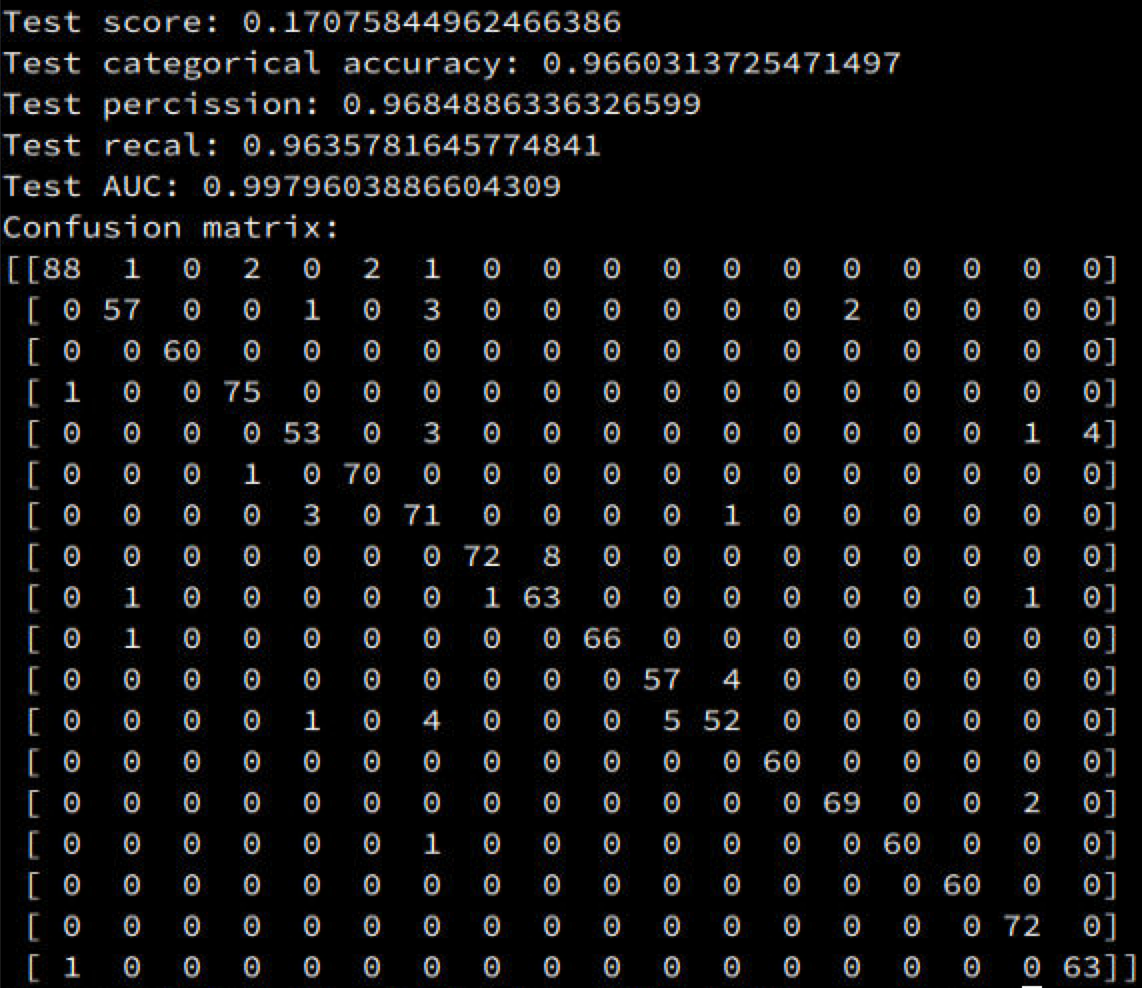
Normal/tumor classifier:



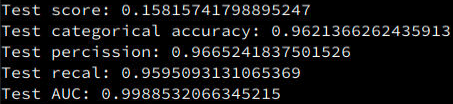
Cancer type classifier:

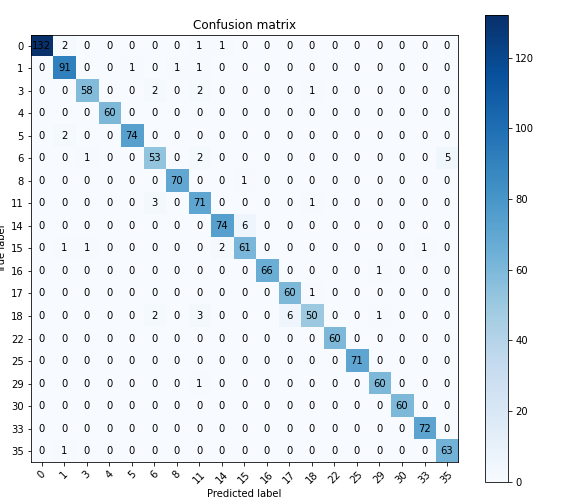


Q1: I used the type classifier from Q0 to classify the tumor samples from the NT data set. The type IDs predicted by the type classifier was used to label the tumor samples. The newly labeled tumor samples’ features were first log transformed to match the scale of the TC dataset, and then normalized in the same fashion as the TC data when training the type classifier in Q0, i.e. scaling each feature by its maximum absolute value. The new tumor data was split into training/testing data according to the training/testing split originally used for the normal/tumor classifier in Q0, which were then added to the TC training and testing data, respectively. Then the type classifier was retrained and evaluated on the expanded set of data, and the performance results are reported below. We can see that the performance generally remained the same as before. This is probably due to the fact that the performance was already very high in Q0 and the dataset size was not a limiting factor, so little changed was observed when the dataset was expanded. Note that the labels for the addition tumor sample were not ground truth, and incorrectly labeled samples, if there were any, would have negatively affected the performance.

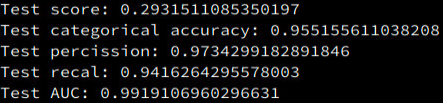


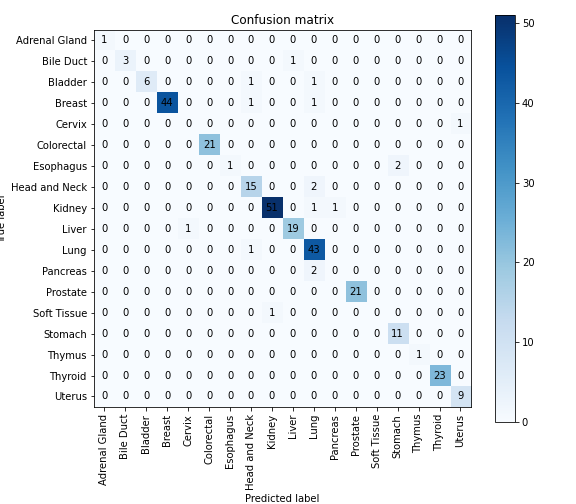
Q2 & Q3: Instead of only selecting the tumor samples as in Q1, in this question all samples in the NT dataset were used. The features were log transformed and normalized in the same fashion as before, and the samples were again split into training/testing data according to the split originally used for the normal/tumor classifier. The tumor samples were labeled using the predictions from the type classifier, and an additional class was created for the normal samples (labeled as type 0). The new data were added to the TC training and testing data. The type classifier was modified to predict 37 classes rather than 36, and was trained and evaluated on the expanded data. The performance results are reported below. The performance still remained about the same as before. Most of the normal samples were correctly predicted as normal – only four of them were misclassified.



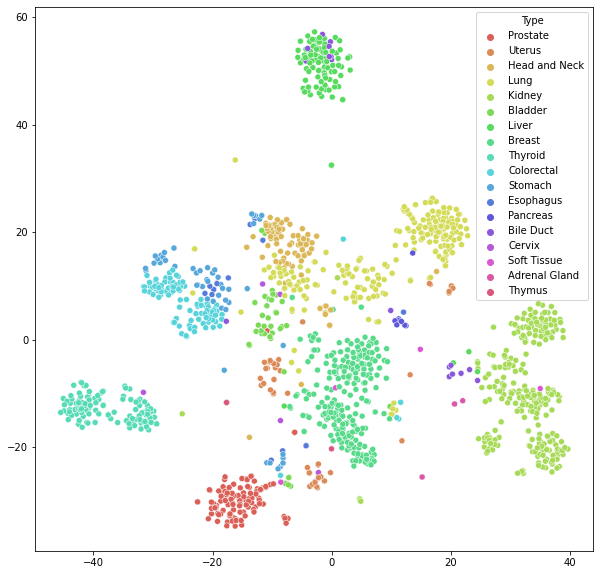


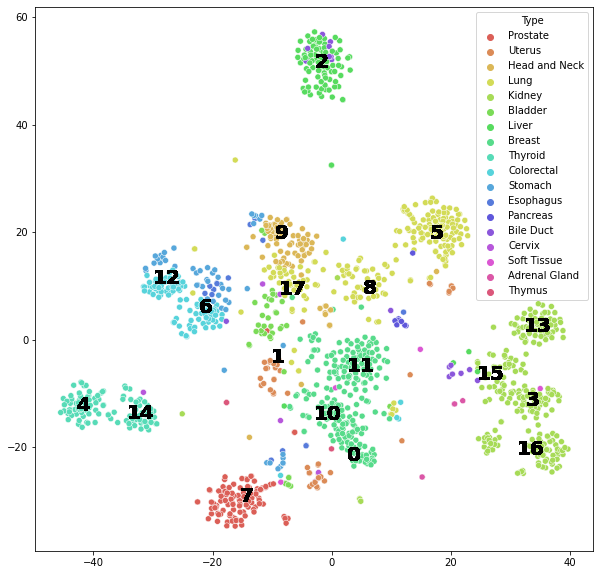
Q4: In this question a new type classifier was trained to predict tissue types for all samples, regardless of their normal/tumor status. The raw data file names in the metadata file was used to read in the raw gene expression data corresponding to each sample, which yielded a dataset consisted of 702 normal samples (after removing duplicates) and 726 tumor samples (no duplicates) whose tissue types are known. The features were log transformed and normalized in the same manner as before. The dataset was split into 80% for training and 20% for testing, stratified using the tissue type labels. The type classifier was modified to predict 18 classes, and was trained and evaluated on this dataset. The performance results are reported below. The performance slightly decreased from the previous questions, likely due to the reduction in dataset size (the TC dataset alone has 5400 samples). The performance was still very high nonetheless, even for some very small classes. Two of the classes that had only one sample in each in the test set were incorrectly predicted, which affected the overall results of the performance metrics. But overall, there were only a handful of misclassifications.



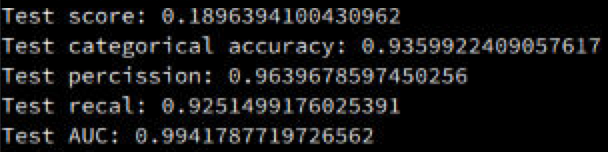


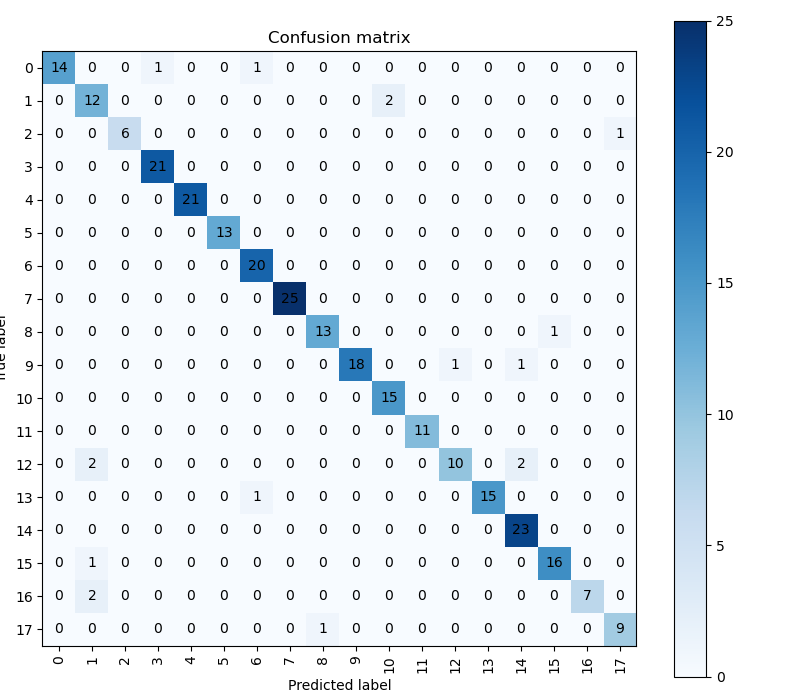
Q5: The dataset obtained in Q4 from the metadata and the raw data files was clustered using t-SNE, which is visualized in the first figure below, where the ground truth tissue types were labeled with color and shown in the legend. K-means clustering was then applied on the 2D embedding obtained with t-SNE in order to assign the clustered samples into 18 types. The cluster assignment is visualized in the second figure below.





The dataset was split into training and testing data in the same manner as in Q4, and the type classifier (same architecture as in Q4) was then trained and evaluated on the re-labeled data. The performance results are reported below. The performance slightly decreased compared to Q4, but still remained high. The decreased can be attributed to the imperfection in the clustering and label assignment process. If more sophisticated unsupervised learning methods were applied, the samples could have been separated into more distinct clusters corresponding to their actual tissue types, which would have improved the classification performance.





Q6: I used a variational autoencoder (VAE) to generate new data samples labeled as normal and tumor. The VAE was trained using the NT dataset, which was normalized the same way as when training the NT classifier in Q0 (i.e. scaling each feature by its maximum absolute value). The reconstruction of samples in the NT dataset predicted by the VAE formed the set of newly generated samples, which inherited labels from the original source NT dataset. Then I ran the generated data through the normal/tumor classifier that I trained in Q0, and the evaluation results are reported below. The performance was slightly lower than the classifier’s performance on the original NT dataset, but still remained very high. The slight performance drop was expected because the data generated by VAE, although achieved high similarity compared with the original data, still had discrepancies from the original. Improvements could be made in the future by using GAN or more variations of VAE if computational resources and time allow.

