Cancer has been one of the diseases which causes most death out to people around the world. Predicting cancer can help people to prevent the disease and reduce the number of deaths caused by cancer. With the rapid development of machine learning, it is possible to use machine learning algorithms to predict the risk of having a particular disease. To further make use of medical data and make an effort to improve healthcare service, the risk of having a particular disease is possible to use machine learning algorithms to predict cancer.

For numerical data and medical image data.

The labels are M (malignant) and B (benign) negative and 78,786 IDC positive).

Breast Histopathology Images Dataset

Images of Invasive Ductal Carcinoma (IDC) which is the most common subtype of all breast cancers.

Contains 277,524 patches of size 50 x 50 (198,738 IDC negative and 78,786 IDC positive).

For numerical data’s models, we used Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machines, and K-Nearest Neighbors as classifying models.

The training process is written on Python with models import from Scikit Learn library.

First, we find the correlation between the target column "diagnosis" and other features.

Then we choose features with correlation greater than 0.6 to be training features.

Models will be trained and test by cross-validation with 5 folds.

For the image dataset, we use MobileNetV2 and EfficientNet.

The models are written using PyTorch library.

EfficientNet is a scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient.

In this project, we used the scaled MobileNets.

For the numerical dataset, we use Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machines, and K-Nearest Neighbors as classifying models.

The training process is written on Python with models import from Scikit Learn library.

First, we find the correlation between the target column "diagnosis" and other features.

Then we choose features with correlation greater than 0.6 to be training features.

Models will be trained and test by cross-validation with 5 folds.

For the image dataset, we use MobileNetV2 and EfficientNet.

The models are written using PyTorch library.

EfficientNet is a scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient.

In this project, we used the scaled MobileNets.

The numerical data set will be tested using cross-validation with 5 folds so it is a division of 4 to 1 for train set and test set.

- MobileNetV2
- EfficientNet

For numerical data’s models, we tried two optimizer Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD) and the combination of both with Adam for the first few epochs and SGD for later epochs.

Using only SGD as we did in our method will give the best accuracy (> 90%) compare to only Adam (83%) and combination of Adam and SGD (85-87%).

The image data set will be tested by splitting the original data set to train set and test set by 80% and 20% respectively.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Fold 1</th>
<th>Fold 2</th>
<th>Fold 3</th>
<th>Fold 4</th>
<th>Fold 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>96.231</td>
<td>95.000</td>
<td>95.625</td>
<td>95.000</td>
<td>94.984</td>
</tr>
<tr>
<td>RF</td>
<td>100.000</td>
<td>90.000</td>
<td>92.500</td>
<td>94.167</td>
<td>95.309</td>
</tr>
<tr>
<td>SVM</td>
<td>96.482</td>
<td>93.750</td>
<td>93.750</td>
<td>94.583</td>
<td>94.988</td>
</tr>
<tr>
<td>KNN</td>
<td>92.211</td>
<td>91.250</td>
<td>90.625</td>
<td>90.147</td>
<td>91.230</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNetV2</td>
<td>96.094</td>
</tr>
<tr>
<td>EfficientNet</td>
<td>93.812</td>
</tr>
</tbody>
</table>

Acknowledgements

- I would like to thank Charlie Peck and Igor Minevich for providing detailed feedbacks and helping me finish this project.