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 cause 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, possible to use machine learning algorithms to predict cancer. Predicting cancer can help people to prevent the disease and reduce the number of deaths cause by cancer.

Motivation

- 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 cause 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, possible to use machine learning algorithms to predict cancer.

Dataset

- Breast Cancer Wisconsin (Diagnostic) Data Set
  - 33 columns with features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.
  - Describe characteristics of the cell nuclei in the image.
  - The labels are M (malignant) and B (benign).

- Breast Histopathology Images Datase
  - 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).

Methods

- 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.
- MobileNetV2 is a family of general purpose computer vision neural networks designed with mobile devices in mind to support classification, detection and more.
- 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.

Result

- 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.

Discussio

- For numerical data's models, we use GridSearchCV to tune the hyperparameters of each model.
- GridSearchCV will take input of our model and a dictionary of hyperparameters and return the hyperparameter setting for the best possible model.
- The result after we tuned the hyperparameter as follow.

<table>
<thead>
<tr>
<th></th>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNetV2</td>
<td>96.094</td>
<td>92.332</td>
</tr>
<tr>
<td>EfficientNet</td>
<td>93.312</td>
<td>91.021</td>
</tr>
</tbody>
</table>

Acknowledgements

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

References