Cancer Prediction using Machine Learning Algorithms

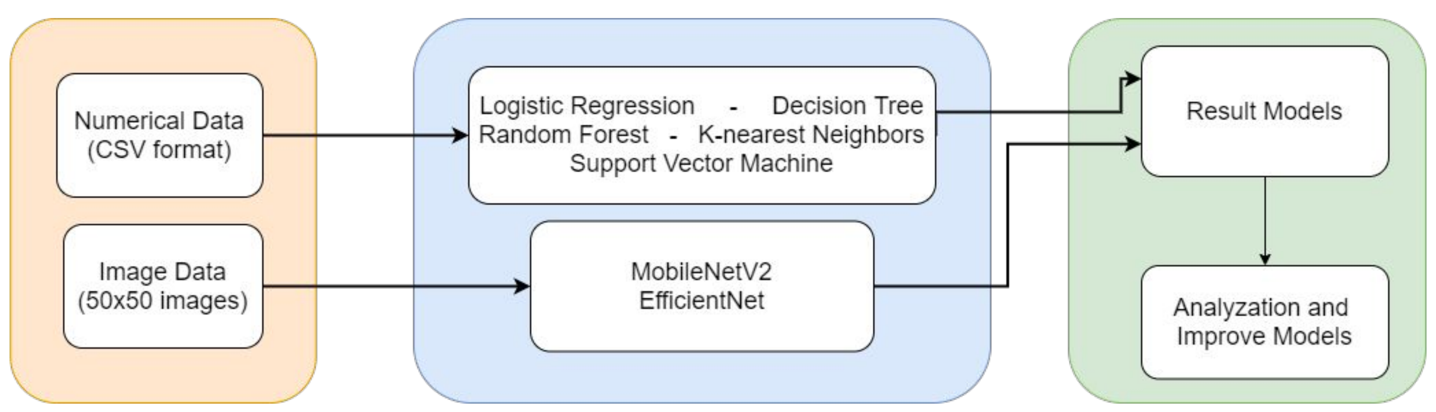
Anh Dang

Department of Computer Science, Earlham College

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,
- I propose a project of analyzing models to predict risk of having cancer base on numerical data and medical image data.

Project Framework

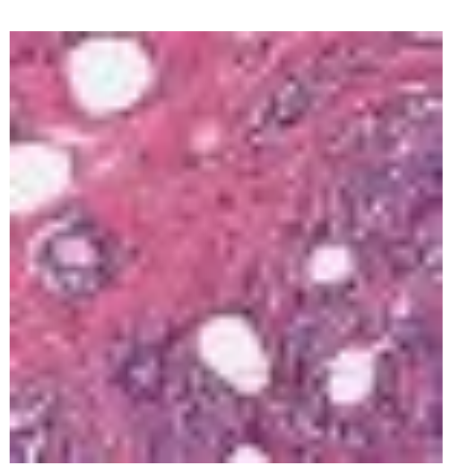


Dataset

- 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)

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

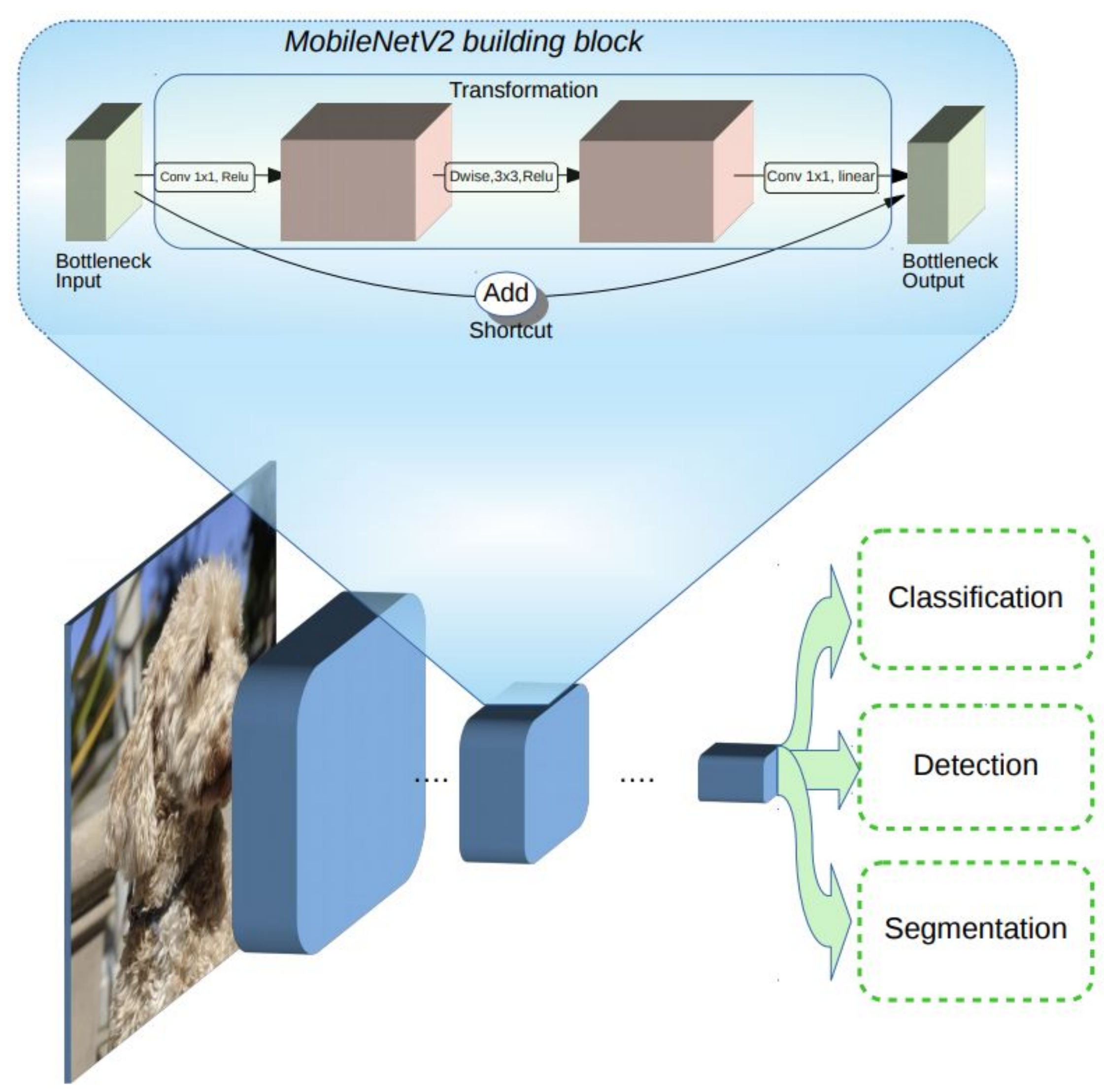
- Breast Histopathology Images Datase
- Images of Invasive Ductal Carcinoma (IDC) which is the most common subtype of all breast cancers.
- Contains 277,524 patches of size 50 x 50 (198,738 IDC negative and 78,786 IDC positive).



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.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224 × 224	32	1
2	MBConv1, k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	28 × 28	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling & FC	7 × 7	1280	1

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.

	Accuracy	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
LR	96.231	95.000	95.625	95.000	94.984	94.975
DT	100.000	90.000	92.500	94.167	95.309	95.234
RF	96.482	93.750	93.750	94.583	94.988	94.978
SVM	92.211	91.250	90.625	90.417	91.230	91.972
KNN	94.975	93.750	94.375	92.917	93.422	93.978

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

	Training Accuracy	Testing Accuracy
MobileNetV2	96.094	92.352
EfficientNet	95.312	91.021

Discussion

- 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.

	Accuracy	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
DT	96.231	93.750	95.000	94.583	94.672	94.472
RF	97.236	92.500	93.125	94.167	94.992	94.728
SVM	96.734	93.750	94.375	95.000	95.934	95.987
KNN	94.975	93.750	94.375	92.917	93.422	93.978

- For image 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 ou method will give the best accuracy (> 90%) compare to only Adam (83%) and combination of Adam and SGD (85-87%).

Acknowledgements

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

References

- This poster is based on "Cancer Predictionusing Machine Learning Algorithms", Anh Dang, available at https://portfolios.cs.earlham.edu/wp-content/uploads/2020/05/aqdang16_paper_final.pdf