Received May 11, 2022; | Revised xx. xx., 2022; | Accepted xx. xx., 2022

DOI: xxx/xxxx

CLASS: UST PRML 2022-1

REPORT FOR THE STUDY

Predict whether the cancer is benign or malignant

Wonjun Lee[1] | **Junseong Bang**[2,3]

[1]Intelligence & Manufacturing Convergence Laboratory, ETRI, Daejun, Republic of Korea

[2]Police Science & Public Safety ICT Research Center, ETRI, Daejun, Republic of Korea

[3]Department of Computer Software, UST, Daejun, Republic of Korea

Abstract

Breast cancer refers to all the malignant tumors that occur in the breast. In this paper, we used a dataset that transformed images examined for tissues of these breast cancers into csv, and we raised the existing performance to 99% using existing machine learning techniques and deep learning techniques.

KEYWORDS:
Breast Cancer, Machine Learning, Linear SVC, Deep Learning, Keras

1 | INTRODUCTION

Breast cancer is the most common malignancy among women, and it is the second leading cause of cancer death among women.[1] Breast Cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor.

A tumor does not mean cancer. tumors can be benign, pre-malignant, or malignant. Tests such as MRI, mammogram, ultrasound, and biopsy are commonly used to diagnose breast cancer performed.

The object of paper is to classify whether the breast cancer is benign or malignant.

In Section 2, we examine dataset information and variables, and statistically analyze them through exploratory data analysis. Section 3 we examine the breast caner data in depth through exploratory data analysis and goes through the process of preprocessing the data necessary for learning to solve the problems found through exploratory analysis. We conducted learning, results and tuning are performed with various machine learning algorithms and deep learning in section 4. in section 5, we introduced result and lastly introduces the conclusion of this paper.

Abbreviations: MRI, Magnetic Resonance Imaging;

2 | DATASET

Breast Cancer data set was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin,USA. [2]

To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan.

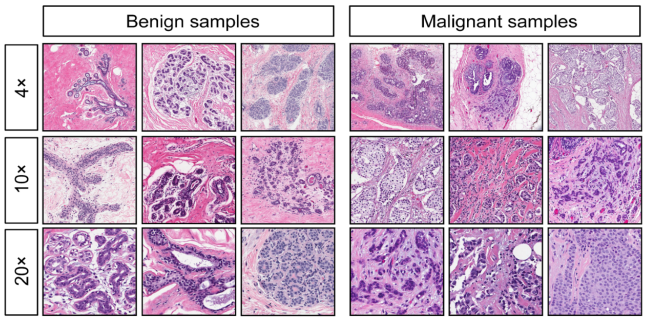

FIGURE 1 Benign and Malignant Cells of Breast Cancer

2.1 | Dataset Attribute

The dataset consists of a total of 569 rows and 33 columns, and is data without missing values. and The dataset Attribute consist of :

1. ID number
2. Diagnosis (M = malignant, B = benign)
3. radius (mean of distances from center to points on the perimeter)
4. texture (standard deviation of gray-scale values)
5. perimeter
6. area
7. smoothness (local variation in radius lengths)
8. compactness
9. concavity (severity of concave portions of the contour)
10. concave points (number of concave portions of the contour)
11. symmetry
12. fractal dimension ("coastline approximation" - 1)

3 | EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis(EDA) is the process of observing and understanding data from various angles when it comes in. [3]

EDA is necessary because by examining the distribution and value of the data, We can better understand what the data represents and discover potential problems with the data.

In addition, through the process of looking at it from various angles,can discover various patterns that would not have occurred in the problem definition stage, and based on them, you can modify existing hypotheses or build new hypotheses.

3.1 | Synthetic Minority Oversampling Technique

When visualizing the target variable through a counter plot to find out the distribution, it was confirmed that the malignant and benig were not balanced.

SMOTE(synthetic minority oversampling technique), an oversampling technique, was used to resolve this imbalance.

3.2 | Histogram

we used a histogram to visualize the distribution of the data.

The reason for using the histogram is that each independent variable except the target variable is continuous data.

After checking the histogram, We can see the variables that can clearly distinguish between positive and malicious.

However, there are other qualities that make it impossible to distinguish between malicious and benign. In addition, although the data generally draw a normal distribution, not all of them are, and there are some areas where skewness is severe, so we will proceed with normalization to increase the effectiveness of the analysis.

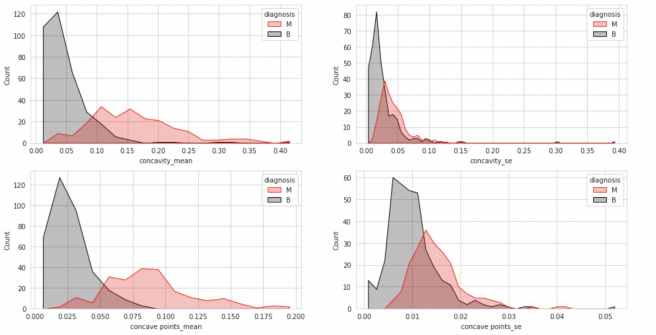

FIGURE 2 Distribution of Data

3.3 | Outlier Detection

Next, we detect the Outlier with Boxplot visualization. When we first used BoxPlot, could find outliers. There are many ways to remove this outlier, but the criteria for determining how far the outlier data is takes precedence.

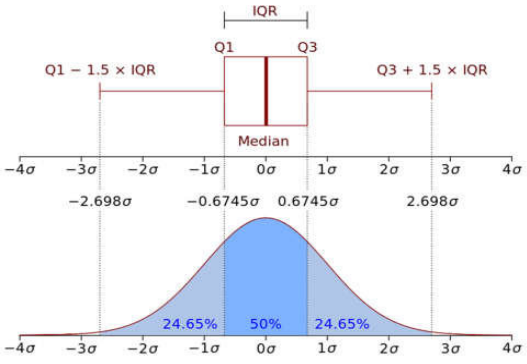

FIGURE 3 Inter Quantile Range

Among the various methods of processing outliers, We used the Inter Quantile Range (IQR) method. The entire data are sorted in ascending order, and divided into four equal parts (25%, 50%, 75%, 100%), or quadrants. The difference between the value at 75% and the value at 25% is called the IQR. Multiply this IQR by 1.5 and add it to the value at 75% points

and subtract it from the value at 25% points to determine the minimum value.[4]

In this case, a value greater than or less than the determined maximum value is regarded as an outlier.

3.4 | Multicollinearity

Multicollinearity is a problem in which there is a strong correlation between independent variables in statistical regression analysis.[5] When such multicollinearity is found, inaccurate regression results are derived, so inaccurate results can be derived.

Therefore, variables with a very high correlation were identified through heatmap, and variables with a correlation of 0.92 or higher were removed to remove multicollinearity.

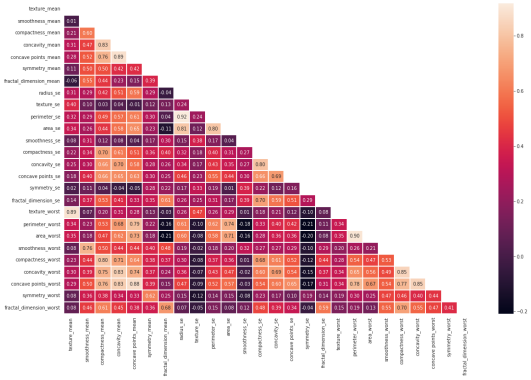

FIGURE 4 Heatmap after multicollinearity removal

4 | EXPERIMENT AND RESULT

Two methods were used as experiments, a machine learning method and a deep learning method.

In common, data preprocessing is performed, normalization and SMOTE are applied, and then learning data and test data are divided into 8:2 ratios.

4.1 | Machine Learning

The algorithms used in this model were the representative Logistic Regression, Linear SVC, and K Neighbors Classifier of the regression model. and we used Random Forest Classifier, Decisition Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, LGBM Classifier, and XGBC Classifier, which are rated as the best algorithm in Kagle.

When all of this model is learned, the highest resulting model is the Linear SVC, which shows accuracy.

	Model	Train Accuracy	Test Accuracy
1	LinearSVC	0.978022	0.982456
0	LogisticRegression	0.971429	0.982456
2	RandomForestClassifier	1.000000	0.964912
7	LGBMClassifier	1.000000	0.964912
8	XGBClassifier	1.000000	0.964912
5	AdaBoostClassifier	1.000000	0.956140
6	GradientBoostingClassifier	1.000000	0.956140
3	KNeighborsClassifier	0.980220	0.956140
4	DecisionTreeClassifier	1.000000	0.929825

FIGURE 5 K-means Clustering result according to cluster number

The Linear Support Vector[6], which was evaluated with the highest performance, was subjected to parameter tuning using grid search. Parameter C regulates the amount of error allowed. The larger the C value, the less error is allowed, and this is called a hard margin. Conversely, the smaller the C value, the more errors are allowed to make a soft margin. In this paper, a soft margin was selected by giving a parameter value of 0.01. Hinge was selected as the loss, and L1, the sum of the absolute values of the differences between the elements, was selected as the penalty to prevent over fitting.

As a result of parameter tuning through grid search, it showed that the accuracy increased by about 2% compared to the result of existing machine learning, and finally, a very high accuracy of 0.99 could be derived.

4.2 | Deep Learning with Keras

The following is data analysis using Keras.[7] Keras is an open source neural network library written in Python designed to enable rapid experimentation of deep learning models and is part of the(ONEIROS)(Open-ended Neuro-Electronic Intelligent Robot Operating System) study.

In this paper, we built a deep learning model using Keras and experimented to obtain more compliant efficiency than the existing results.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_3 (Dense)              (None, 8)                 216
_________________________________________________________________
dense_4 (Dense)              (None, 8)                 72
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 9
=================================================================
Total params: 297
Trainable params: 297
Non-trainable params: 0
```

FIGURE 6 Structure of deep learning model

A deep learning model was constructed using a Sequential function. The reason why deep learning model was not built is

that the amount of dataset is not large and it is built to prevent overfitting.

The deep learning model structure of three stages is used. ReLu is used to build a hidden layer as an activation function, and Sigmoid is used for the last output layer because it is a model to determine whether it is positive or malicious.

The model compiled optimizer was used with sigmoids, and binary classification was used and accuracy was evaluated. The result of learning by setting the total size of Epoch and 30 times in total shows that the total loss value is 0.08 accuracy is 0.97.

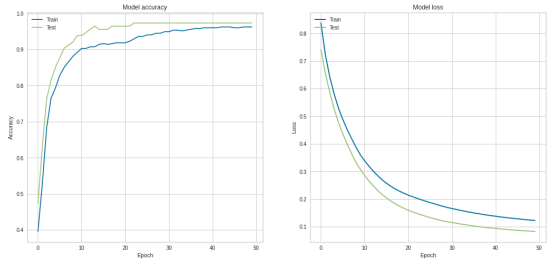

FIGURE 7 Model Acuuracy and Loss graph

5 | RESULT

As a result of the experiment, the accuracy of tuning with Grid Search using the Linear Support Vector algorithm was the highest.

	Linear SVC	Grid Search	Deep Learning
Accuracy	0.98	0.99	0.96

TABLE 1 Result of Linear SVC, Grid Search, Deep Learning

6 | CONCLUSION

In this paper, we applied various machine learning techniques and constructed a deep learning model based on keras to compare the results.

The results of the experiment are the highest after applying the Linear Support Vector algorithm among machine learning techniques, which is because the accuracy of the tuning results is too high, so the problem of overfitting can be considered.

Future work will apply these techniques not only to breast cancer but also to pancreatic cancer and liver cancer, and will apply deep learning model with bigger dataset and improve performance.

References

1. Pamela Cowin, Tracey M Rowlands, and Sarah J Hatsell, *Cadherins and catenins in breast cancer*, Current Opinion in Cell Biology **17** (2005), no. 5, 499–508, Cell-to-cell contact and extracellular matrix.
2. Dheeru Dua and Casey Graff, *UCI machine learning repository*, 2017.
3. Chong Ho Yu, *Exploratory data analysis in the context of data mining and resampling.*, International Journal of Psychological Research **3** (2010), no. 1, 9–22.
4. Husein Perez and Joseph Tah, *Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-sne*, Mathematics **8** (2020), 662.
5. Aylin Alin, *Multicollinearity*, Wiley Interdisciplinary Reviews: Computational Statistics **2** (2010), no. 3, 370–374.
6. Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al., *A practical guide to support vector classification*, 2003.
7. François Chollet et al., *Keras*, https://keras.io, 2015.