Received March 25, 2022; | Revised xx. xx., 2022; | Accepted xx. xx., 2022

DOI: xxx/xxxx

CLASS: UST PRML 2022-1

REPORT FOR THE STUDY

# Breast Cancer with Dimension Reduction

**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 is the most common malignancy among women, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a results of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. Using this dataset, after apply dimension reduction method(PCA, Kernel PCA), by Logistc Regression. It compares and analyzes to apply Naive, PCA, and Kernel PCA.

**KEYWORDS:**

Breast Cacner, PDA, Kernel PDA

## 1 | INTRODUCTION

Breast cancer is the most common malignancy among women, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a results 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 (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed.

Using this dataset, after apply dimension reduction method(PCA, Kernel PCA), by Logistc Regression it aims to find key factors that cause breast cancer.

## 2 | DATASET

This is an analysis of the Breast Cancer Wisconsin (Diagnostic) DataSet, obtained from Kaggle. This data set was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin,USA. 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.

## 3 | PCA

### 3.1 | Principal Component Analysis(PCA)

Principal Component Analysis is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

### 3.2 | Kernel PCA

Kernel PCA is an extension of PCA that allows for the separability of nonlinear data by making use of kernels. The basic idea behind it is to project the linearly inseparable data onto a higher dimensional space where it becomes linearly separable.

## 4 | CODE

### Import Library

```
#General Libraries
import numpy as np
import pandas as pd

# Statistic & Machine Learning Libraries
```

```
from sklearn.model_selection import
    train_test_split,
from sklearn.metrics import confusion_matrix,
    accuracy_score
from sklearn.naive_bayes import GaussianNB

# Data Visualiztion Libraries
import seaborn as sns
import matplotlib.pyplot as plt
```

### Load Dataset, Split Data

```
from sklearn.datasets import load_breast_cancer
data = load_breast## Importing Library_cancer()
malignant = data.data[data.target == 0]
benign = data.data[data.target == 1]
X_train, X_test, y_train, y_test =
    train_test_split(data.data, data.target,
    random_state =42)
```

### Standard Scale

```
scaler = StandardScaler()
scaler.fit(X_## Importing Librarytrain)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

### PCA

```
pca = PCA(n_components=2) #
X_train_pca  = pca.fit_transform(X_train_scaled)
X_test_pca = pca.fit_transform(X_test_scaled)
```

### Visualization

```
df = X_tn_pca_df
markers=[’o’,’x’,’^’]

for i, mark in enumerate(markers):
    df_i = df[df[’target’]== i]
    target_i = data.target[i]
    X1 = df_i[’pca_comp1’]
    X2 = df_i[’pca_comp2’]
    plt.scatter(X1, X2, marker=mark, label=
    target_i)

plt.xlabel(’pca_component1’)
plt.ylabel(’pca_component2’)
plt.legend()
plt.show()
print(’# Data with labels is visualized.’)
```

### Kernel PCA

```
from sklearn.decomposition import KernelPCA
## Importing Library
k_pca = KernelPCA(n_components=2, kernel=’poly’)
k_pca.fit(X_train_scaled)
X_train_kpca = k_pca.transform(X_train_scaled)
X_test_kpca = k_pca.transform(X_test_scaled)
```

## 5 | RESULT

This is the result of applying logistic regression to data applied with Naive which is version that have not been processed, PCA, and Kernel PCA.

| | Naive | PCA | Kernel PCA |
|---|---|---|---|
| Testing Score | 0.97 | 0.85 | 0.90 |

**TABLE 1** Result of Naive, PCA, Kernel PCA

## 6 | CONCLUSION

In order to analyze breast cancer data, the dimension was reduced to PCA and Kernel PCA, and then a model was created by applying a logistic regression.

However, on this dataset, it can be seen that the non-application method performs better than the application of dimensional reduction. Future research plans aim to increase accuracy and performance by using more diverse dimensional reduction techniques.