Breast Cancer Detection using Machine Learning

For Term Project of Data Mining

Choeng Veyseng[a]

[a]*Department of Information and Communication Engineering, Institute of Technology of Cambodia, Phnom Penh, Cambodia*

June, 2022

Abstract

Breast Cancer Detection is one of the most trending studies in medical use case and applying machine learning and deep learning in order to make a better detection and reduce the time consuming of analysing medical images or data. This term project is about detecting breast cancer data with two labels: benign and malignant from the extraction from images and using machine learning to train and predict the data. There are around four main algorithms are using within the project: Random Forest Classifier, K-nearest Neighbour Classifier, Logistic Regression, and Support Vector Machine. As the result, the model using Support Vector Machine (SVM) with Radial Basis Function or rbf kernel tends to perform better than other models and the accuracy is about 98.25% with testing set.

1. Introduction

Tumor detection is one of the most trending study for machine learning and deep learning to support medical field to enable the efficiency and effectiveness of speeding up the medical operation and caring flows. Moreover, being able to understand the dataset before the analysis process would be crucial for choosing the right model to train our data and get the output after processing the data. In this report, breast cancer detection would be the topic to work with machine learning models while try to integrate deep learning models to get more understanding of analysing the data and get the most from our training process. There is a feature selection method is known as Principal Components Analysis (PCA) would be used to support to reduce the time consuming of training process. The expectation from this project is to understand clearly on how to process the data with the right models and have some comparisons to consider on choosing the right models for the dataset.

I would like to formally introduce the objectives of this term project:

- Understand the dataset and visualize the data
- Implement the preprocessing methods such as PCA, standard scaling method and other more
- Apply machine learning algorithms such K nearest neighbour, random forest classification, logistic regression and support vector machines
- Get the results to do the comparison

2. Background

In many developing countries, detecting cancer is the challenging task within the medical which is required technical skills and the support of the recent technology. Breast Cancer is one of the most harmful diseases that are still occurring around the world that need a lot of attentions from all stakeholders to regularly track on the health and do the checking when there is a minor symptom. According to numerous international researches [1–4, 7, 9, 12–15, 17], breast cancer is one of the most concerning diseases that we need to take care in the very first stage of detection with the right treatment and accurate analysis. There are many methods and customization of the existing models to fit on the dataset and study use case in the breast cancer diagnosis in general. This project tends to be a fundamental of study on detecting the breast cancer while knowing how to work with data effectively but an important milestone in order to get into another level of advancement in detecting this kind of disease or other similar disease that machine learning and deep learning would come to have the role of automatic detection in the disease.

3. Literature Review

In order to ensure the project is feasible for the implementation and understand the previous study

of the use case and technology, I would like to formally do literature review on highlight the existing studies and introduce of techniques that are using for improving on the models to maximize the accuracy of detecting breast cancer tumors with images or data extracted from images. There are numerous studies on this topic that discuss on how to detect and effectively apply with the methodologies.

In a study of Anisha et al. [2] have proposed the algorithm using the random forest classification algorithm in order to detect the breast cancer tumor with around 98% of accuracy rate compared with other proposed models testing the dataset consisted of benign and malignant labels of the diagnosis. There are 13 selected features that are good for passing to the models within the study such as Body Mass Index (BMI), gender, age, leptin and many more. In another study of Chudhey et al. [4] have recently proposed a method of detecting the data of breast cancer diagnosis using random forest classifier as the main focus with the support of feature selection method of Principal Component Analysis (PCA) and the result turned out to be 97.37% of accuracy rate along with 97.62% of precision, 95.35% of recall, and 96.47% of F1-score. There are total of 32 features as the attributes while the PCA filters to have only 12 features for this study of detection. Furthermore, a proposed method of Chhatwal et al. [3] focusing on implementation of logistic regression for estimate the threats and risk with National Mammography Database through in order to help to detect the breast cancer diagnosis in the very early stage of the disease. The deep learning and other algorithms would play important roles as well to contribute on this study use case like in the study of Ragab et al.[12] has proposed the methodology using Convolutional Neural Network model and Support Vector Machines (SVM) to classify two main labels of diagnosis: benign and malignant. Deep Convolutional Neural Network named as AlexNet is one of the proposed methods in the case study by reaching to 71.01% of accuracy rate with the new-trained model and this method is using for feature extraction to pass to another algorithm. As the result is shown in the paper through using Support Vector Machine for prediction, the accuracy is 87.2%. Another study of trending technology like deep learning and machine learning are mentioned in Alanazi et al. [1]. In this study, there are the discussion of using algorithms of machine learning such as logisitc regression, K-nearest Neighbour and Support Vector Machines to detect whether the data is positive or negative on the breast cancer diagnosis for the labelling procedure. There are 3 models in total of the proposed Convolutional Neural Network model and get around 87% of accuracy from the model 3. With the support of machine learning models, the result is improved around 8% mentioned by the paper.To add on the study of breast cancer detection, Shen et al. [14] shows how deep learning could apply on this study field effectively through the screening process of mammography. With the models using customized convolutional neural network, it proves by the result of 98% of the Area Under the Curve scoring in average of all models. Moreover, in the study of Wang et al. [15], proposed a systematic flow in feature extraction, training and prediction for the breast cancer disease use case. Convolutional Neural Network is one of the core components to extract and measure the prediction of the image data. There are some processes of calculating the features to pass to the models and there are some detection metrics to measure the models as well. Extreme Learning Machine (ELM) tends to perform better for using 400 samples of female mammograms from the northeastern of China. According to the study of Zheng et al. [17], deep learning is effectively applied to the breast cancer detection study. The methodology within the paper is focusing on the usability of Deep Learning assisted Efficient AdaBoost (DLA-EABA) algorithm specified from CNN model. Long and Short Term Memory algorithm is also applied and other methods to build the model and the prediction tends to work well with the model. Among all models, the highest accuracy reaches to 97.2% which is considered to be a very high percentage. According to Gupta et al. [9], working with the medical images are the prominent work that should get the support from deep learning methods while design a model to fit into the use case. In the study, breast cancer images are the histopathology images using the residual networks and Convolutional Neural Networks to do the feature selection and do the prediction with the right customization. The best result from the proposed methods is 99.75% of accuracy rate. A recent study of Das and Mohanty [7] about breast cancer detection also prove other new methodology of detecting the cancer cell through the utilization of deep learning methods. The method is about ensemble recurrent of the combination of the algorithms like Long and Short Term Memory, Recurrent Neural Network, Gated Recurrent Unit and more. There are some preprocessing procedures as well to denoise the images and do the feature selection while also resize the images in the same size of input. In the dataset, there are around 198738 of normal images and 787860 of cancer images to pass to the model by having 80 percents of training set and 20 percents of testing set. At the end

of the study, the accuracy turns out to be around 98.72% with testing set.

4. Data Introduction

The Wisconsin breast cancer dataset [16] which is a trending dataset which could be accessible via Kaggle or other authorised repositories. The dataset is extracted from digital images and there are 30 important features that are the decision boudary for predicting whether the input data is benign or malignant. For the better performance, the labels of each row of dataset are identified in two classes with integer number for benign, and malignant to 0, and 1 accordingly. For the total number of rows, the dataset is consisted of 569 rows in total with 357 rows with benign mark as approximated to 63% and 212 rows with malignant mark as approximateed to 37%.

5. Data Visualization

In order to understand the data for passing to the models more significant, we need to have the visualization to overlook the features if they are good for binary classification or not. There are three types of visualization are using for this use case. First visualization is about heatmap identification to see how each feature correlates to other features that is shown in figure 1. Second visualization is about the strip plot that see 2 classes lie on the graphs with the scatter plot model with the three different pairs of strip plots in figure 2, figure 3, and figure 4. The last visualization is about the using PCA to convert from 30 features to 2 features to represent on how the values of all data lies on the graph and the representation of graph is obviously showen in figure 5.

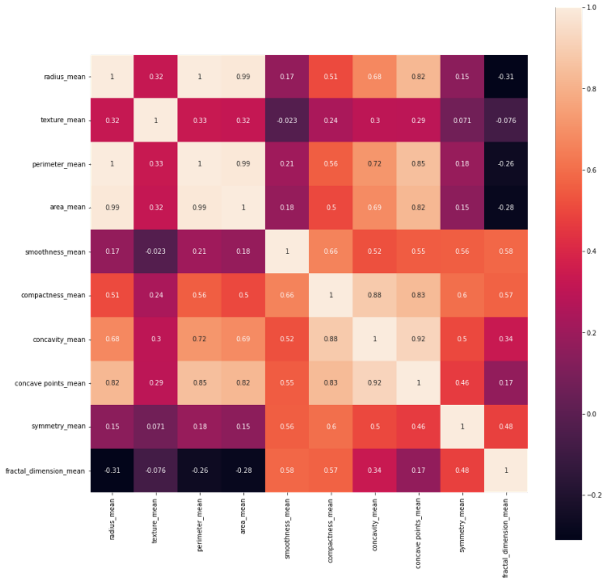

Figure 1: Heatmap with 10 features

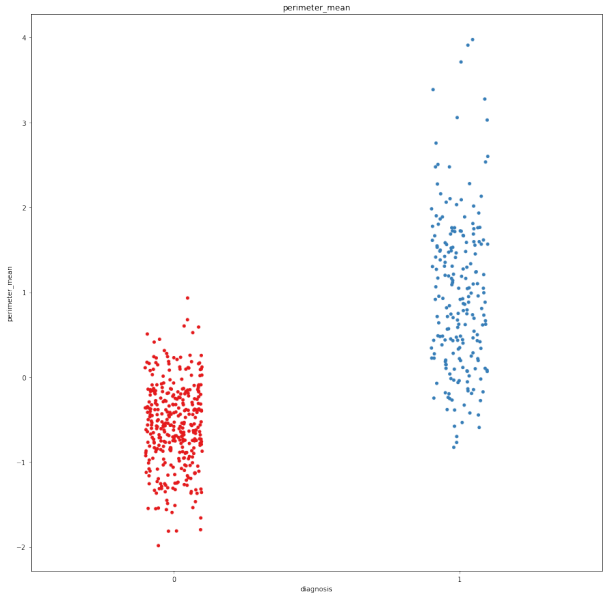

Figure 2: Strip Plot 1 Using Seaborn

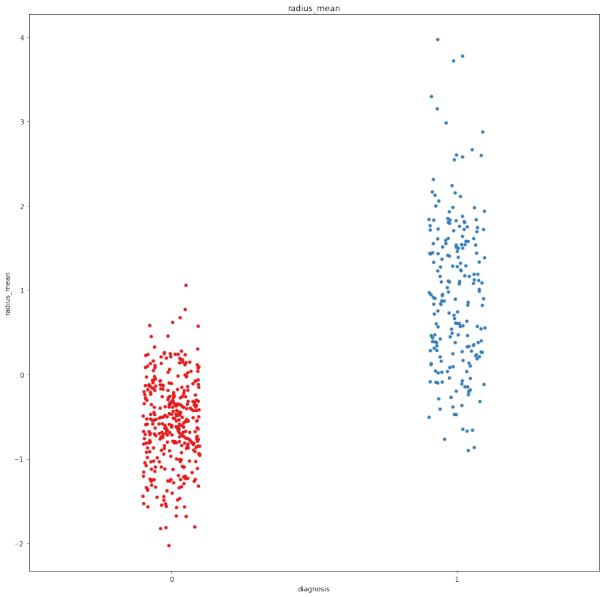

Figure 3: Strip Plot 2 Using Seaborn

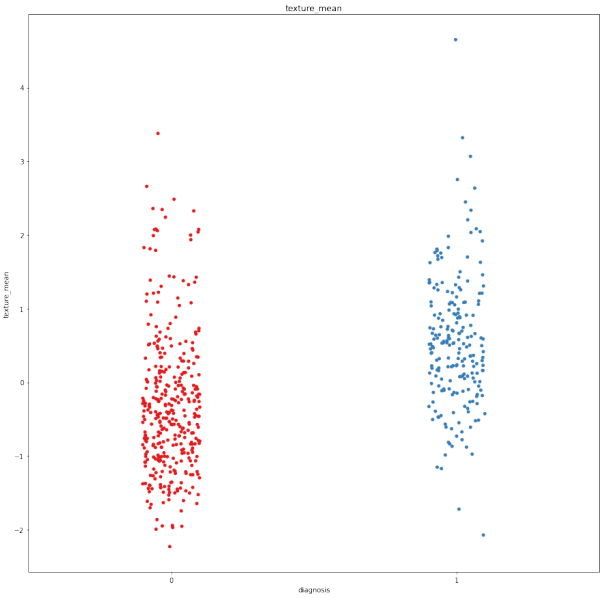

Figure 4: Strip Plot 3 Using Seaborn

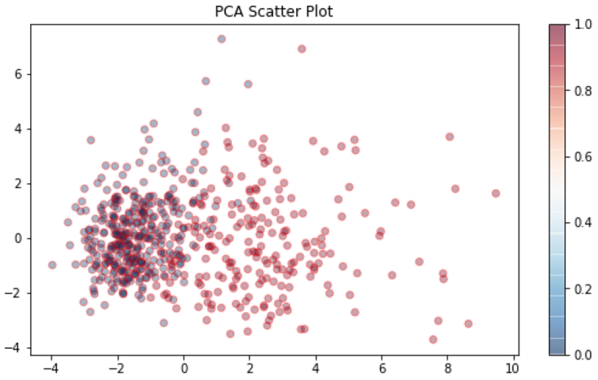

Figure 5: Scatter Plot Using PCA with Matlibplot

6. Technology

There are many algorithms which are considered to be applied on this project effectively to be able to get the most results from them. This section will introduce four algorithms which are using and implementing within this project.

6.1. Random Forest Classifier

Random Forest Classifier [10] can be known as a supervised machine learning that follow the concepts of classifying large amount of data with the decision tree algorithm with the learning called Ensemble Learning.

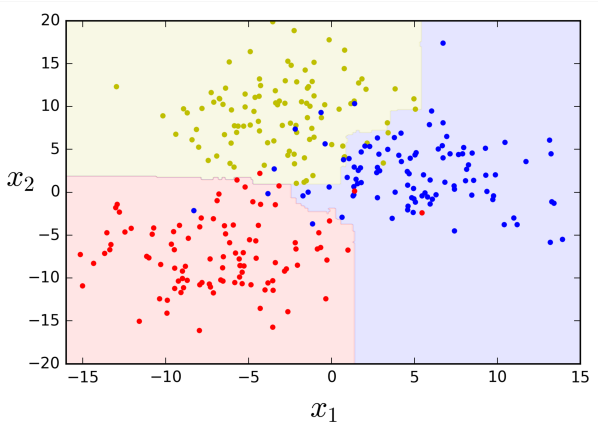

Figure 6: An Example Random Forest Classifier with three boundaries

6.2. K-nearest Neighbour Classifier

K-nearest Neighbour [11] is one of the most commonly used algorithm for supervised and unsupervised learning through the identification of pattern and explore the nearest location or relation to K indicator.

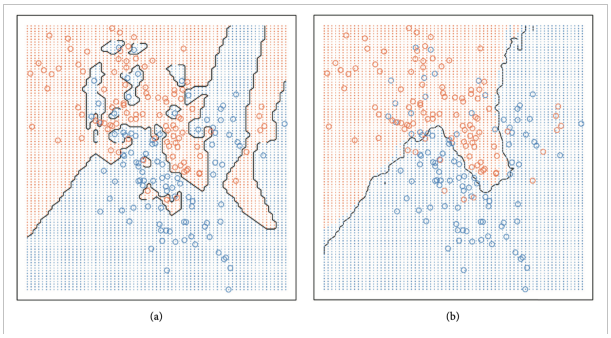

Figure 7: KNN samples

6.3. Logistic Regression

Logistic Regression [6] can be defined as a supervised learning like the following algorithms but it is the improvement of Linear Regression model where it filters better prediction by forecasting binary outcomes via the use of logistic function as the core concept. The binary classification can be labeled as True or False, 0 or 1, Yes or No... In figure 8 indicated the regressions lie on the graphs with clusters of data, it shows the logistic regression would perform better than linear regression with the better decision boundaries of the clusters.

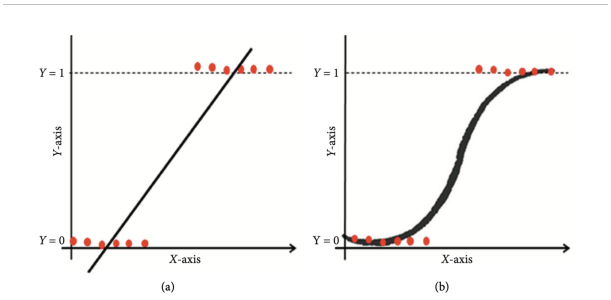

Figure 8: Example of Linear Regression and Logistic Regression

6.4. Support Vector Machine

Support Vector Machine algorithm [5] is applied on the supervised learning in the case of machine learning that focuses on the hyperplane formation and also the improvement of the Linear Regression as well. The purpose of this algorithm is to set a better or fair boundary for distinguishing two classes like showing in figure 9

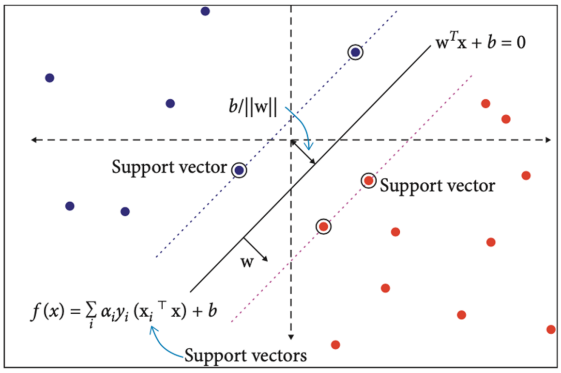

Figure 9: SVM Formation

6.5. Principal Component Analysis

Principal Component Analysis (PCA) [8] is one of the considerable methods for feature selection through reduce the dimensions follow the linear algebra for the computation. It is very useful for graph representation through filtering to get two parameters or get the fair amount of parameters for model training.

7. Experimental Setup

In order to implement this project, there are some important libraries are need to import for different purposes. Google CoLab is using for the experiment process so the connection is a must in order to successfully run the project. I would like formally introduce some libraries and some explanation on how to use these libraries.

- Importing drive from google.colab: It is the function that give us the priviledge to access the data in Google Drive platform where we could store our model or data with CRUD operation from Google CoLab.

```
from google.colab import drive
drive.mount('/content/drive')
```

- Importing numpy library as np in order to work and operate with numpy array form which is better than simple array form.

```
import numpy as np
```

- Importing sklearn library and it is considered to be one of the most important libraries that supporting the running the whole like get the dataset of Wincinson Breast Cancer, some algorithms like support vector machines and many mores, splitting data, evaluation metrics and many more. This is the some captured codes inside the project

```
from sklearn import datasets
from sklearn.preprocessing import
    StandardScaler
from sklearn.model_selection import
    train_test_split
from sklearn.metrics import
    classification_report
from sklearn.decomposition import PCA
```

8. Evaluation Metrics

Classification Report is the evaluation metric for the whole project that it includes accuracy rate, recall, precision and F1-score. Among all rates, accuracy rate is the main focus for evaluating the models with testing set. In order to understand all these calculations clearly, we need to understand the formula of each metric and briefly explain them.

1. Accuracy: it is the sum of all correct predictions divides by total samples that are predicted.
2. Recall: it is focusing on ratio of true positive where all correct predictions of true positive divides by the total number of correct positive samples includes true positive and false negative.
3. Precision: it is focusing on ratio of true positive where all correct predictions of true positive divides by the total number of predicted positive sample includes true positive and false positive.
4. F1-Score: it is the combination of recall and precision to balance these two rates.

9. Methodology

Before introducing the models, preprocessing would be the main methods which would give the best performance for training process. Firstly, shuffling the data is considered to apply for build a randomness for the dataset. After that, I would like to choose 80% splitting for training set and 20% splitting for testing set. For input parameters, standard scaling is applied to normalize the input data for the better performance.
There are 8 models using inside this project to predict the breast cancer data whether benign or malignant:

1. Model 1: Using Random Forest Classifier
2. Model 2: Using Random Forest Classifier with PCA by having only 2 features
3. Model 3: Using K-Nearest Neighbour Classifier with K = 2
4. Model 4: Using K-Nearest Neighbour Classifier with K = 3
5. Model 5: Using Logistic Regression
6. Model 6: Using Logistic Regression with PCA by having only 2 features
7. Model 7: Using Support Vector Machine with rbf kernel and C = 3
8. Model 8: Using Support Vector Machine with linear kernel and C = 3

10. Results

After the implementation of the proposed models, there are the classification reports to represent the accuracy rate, precision, recall and F1-score with testing set as shows in figure 10 to figure 17.

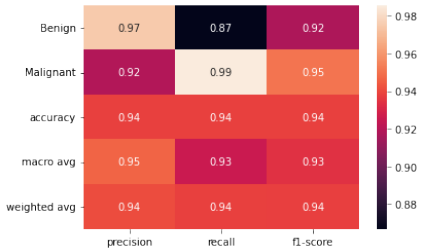

Figure 10: Classification Report Model 1

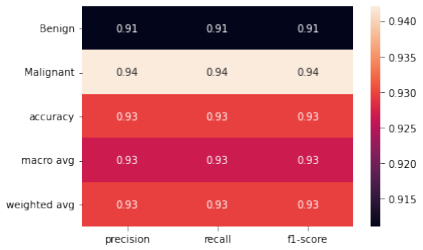

Figure 11: Classification Report Model 2

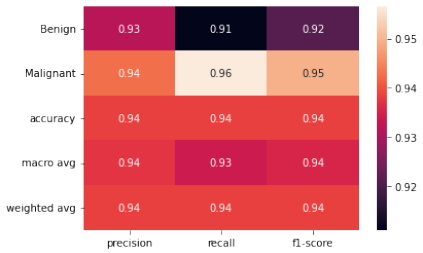

Figure 12: Classification Report Model 3

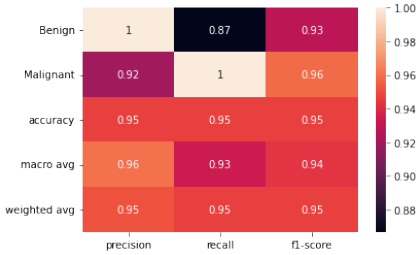

Figure 13: Classification Report Model 4

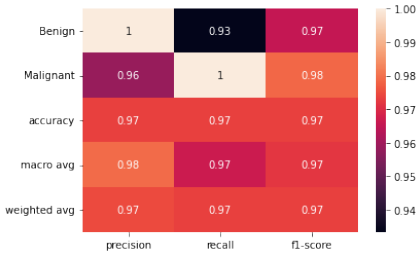

Figure 14: Classification Report Model 5

Figure 15: Classification Report Model 6

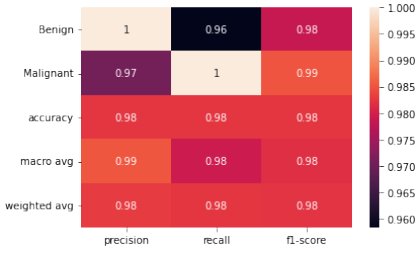

Figure 16: Classification Report Model 7

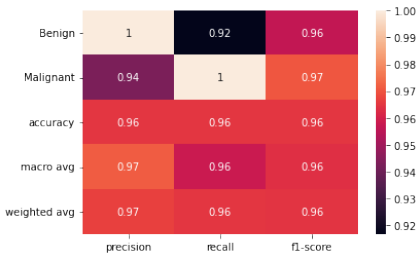

Figure 17: Classification Report Model 8

To be easier to evaluation the eight models, I would like to focus on the accuracy rate of all models with testing set. The results in figure 18 have notified that Support Vector Machines using rbf kernel tends to be the highest rate of accuracy with 98.25% and following by logistic regression without using PCA with 97.37% of accuracy rate.

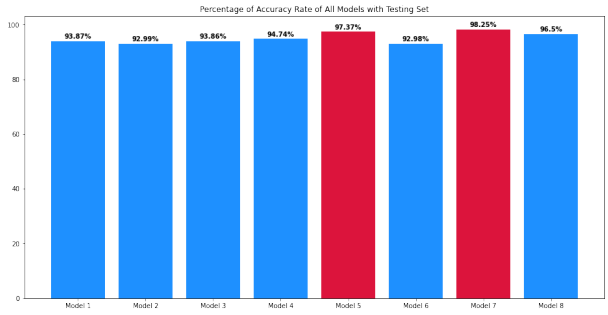

Figure 18: Results of all models with testing set

	Accuracy with Test Set
Model 1	93.867%
Model 2	92.988%
Model 3	93.863%
Model 4	94.744%
Model 5	**97.37%**
Model 6	92.982%
Model 7	**98.25%**
Model 8	96.946%

11. Conclusion

After simulating all models, it is crucial to understand data at the first glance of observation through visualization and read the description of how the data extracted. After working in this project, it is remarkable that we need to consider on the cleaning data in order to reach to maximum of outcomes from the data. After the implementation of preprocessing methods, the training process using variety of algorithms and evaluate them. The results turns out to be good for all methods by using random forest classifier, k-nearest neighbour algorithm, support vector machines, and logistic regression with more than **90%** of accuracy rate with testing and training set. In the computation concerning, PCA would come to play an important role to reduce the time consuming of the project spending and we can customize the models based on the real data and specific scenarios. Moreover, By doing this project, it gives the clear understanding on data mining and machine learning fields while boosting the creativity to adjust the existing technology to align with the specific uses of dataset.

References

[1] S. A. Alanazi, M. M. Kamruzzaman, M. N. Islam Sarker, M. Alruwaili, Y. Alhwaiti, N. Alshammari, and M. H. Siddiqi, «Boosting breast cancer detection using convolutional neural network», Journal of Healthcare Engineering **2021**, e5528622, ISSN: 2040-2295 (2021) 10.1155/2021/5528622.

[2] P. R. Anisha, C. Kishor Kumar Reddy, K. Apoorva, and C. Meghana Mangipudi, «Early diagnosis of breast cancer prediction using random forest classifier», IOP Conference Series: Materials Science and Engineering **1116**, 012187, ISSN: 1757-8981, 1757-899X (2021) 10.1088/1757-899X/1116/1/012187.

[3] J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn, K. A. Shaffer, and E. S. Burnside, «A logistic regression model based on the national mammography database format to aid breast cancer diagnosis», AJR. American journal of roentgenology **192**, 1117–1127, ISSN: 0361-803X (2009) 10.2214/AJR.07.3345.

[4] A. S. Chudhey, M. Goel, and M. Singh, «Breast cancer classification with random forest classifier with feature decomposition using principal component analysis», in Advances in data and information sciences, edited by S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. Mishra, and B. K. Singh, Lecture Notes in Networks and Systems (2022), pp. 111–120, ISBN: 9789811656897, 10.1007/978-981-16-5689-7_10.

[5] C. Cortes and V. Vapnik, «Support-vector networks», Machine learning **20**, 273–297 (1995).

[6] D. R. Cox, «The regression analysis of binary sequences», Journal of the Royal Statistical Society: Series B (Methodological) **20**, 215–232 (1958).

[7] A. Das and M. N. Mohanty, «Design of ensemble recurrent model with stacked fuzzy ARTMAP for breast cancer detection», Applied Computing and Informatics **ahead-of-print**, 10.1108/ACI-03-2022-0075, ISSN: 2210-8327 (2022) 10.1108/ACI-03-2022-0075.

[8] K. P. F.R.S., «Liii. on lines and planes of closest fit to systems of points in space», The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**, 559–572 (1901) 10.1080/14786440109462720.

[9] V. Gupta, M. Vasudev, A. Doegar, and N. Sambyal, «Breast cancer detection from histopathology images using modified residual neural networks», Biocybernetics and Biomedical Engineering **41**, 1272–1287, ISSN: 02085216 (2021) 10.1016/j.bbe.2021.08.011.

[10] T. K. Ho, «Random decision forests», in Proceedings of 3rd international conference on document analysis and recognition, Vol. 1 (IEEE, 1995), pp. 278–282.

[11] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, «K-nearest neighbor classification», in *Data mining in agriculture*, edited by A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, Springer Optimization and Its Applications (Springer, New York, NY, 2009), pp. 83–106, ISBN: 9780387886152, 10.1007/978-0-387-88615-2_4.

[12] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, «Breast cancer detection using deep convolutional neural networks and support vector machines», PeerJ **7**, e6201, ISSN: 2167-8359 (2019) 10.7717/peerj.6201.

[13] A. Rather, «Awareness regarding female breast cancer in Kashmiri males - A study»,

[14] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, «Deep learning to improve breast cancer detection on screening mammography», Scientific Reports **9**, 12495, ISSN: 2045-2322 (2019) 10.1038/s41598-019-48995-4.

[15] Z. Wang, M. Li, H. Wang, H. Jiang, Y. Yao, H. Zhang, and J. Xin, «Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features», IEEE Access **7**, 105146–105158, ISSN: 2169-3536 (2019) 10.1109/ACCESS.2019.2892795.

[16] M. Wolberg, William Street, W., *Breast Cancer Wisconsin (Diagnostic)*, UCI Machine Learning Repository, 1995.

[17] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, «Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis», IEEE Access **8**, 96946–96954, ISSN: 2169-3536 (2020) 10.1109/ACCESS.2020.2993536.