LONTAR KOMPUTER VOL. 9, NO. 3 DECEMBER 2018 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2018.v09.i03.p08 **e-ISSN 2541-5832**
Accredited B by RISTEKDIKTI Decree No. 51/E/KPT/2017

Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction

Ade Jamal[a1], Annisa Handayani[a2], Ali Akbar Septiandri[a3], Endang Ripmiatin[a4], Yunus Effendi[b5]

[a]Informatics Department, Faculty of Science and Technology,
University Al-Azhar Indonesia, Jakarta, Indonesia
[1]adja@uai.ac.id

[b]Biology Department, Faculty of Science and Technology,
University Al-Azhar Indonesia, Jakarta, Indonesia

Abstract

Breast cancer is the most important cause of death among women. A prediction of breast cancer in early stage provides a greater possibility of its cure. It needs a breast cancer prediction tool that can classify a breast tumor whether it was a harmful malignant tumor or un-harmful benign tumor. In this paper, two algorithms of machine learning, namely Support Vector Machine and Extreme Gradient Boosting technique will be compared for classification purpose. Prior to the classification, the number of data attribute will be reduced from the raw data by extracting features using Principal Component Analysis. A clustering method, namely K-Means is also used for dimensionality reduction besides the Principal Component Analysis. This paper will present a comparison among four models based on two dimensionality reduction methods combined with two classifiers which applied on Wisconsin Breast Cancer Dataset. The comparison will be measured by using accuracy, sensitivity and specificity metrics evaluated from the confusion matrices. The experimental results have indicated that the K-Means method, which is not usually used for dimensionality reduction can perform well compared to the popular Principal Component Analysis.

Keywords: *Dimensionality Reduction, Machine Learning, Principal Component Analysis, K-Means Clustering, Breast Cancer*

1. Introduction

The malignant tumor, also known as cancer is one of the prominent death causes globally. As stated by the American Cancer Society, malignant breast tumors or breast cancer is the second leading death cause among women after lung cancer. In poor or developing countries where there is a lack of experienced doctor or physicians to perform a good prognosis of a tumor, the situation is much worse. Many die from this disease although a timely diagnosis of breast cancer can provide a higher possibility of survival. Therefore, a large number of studies are currently ongoing to find methods that can predict breast cancer in its early stages.

In the field of biomedical engineering, principles of engineering, medical science and technology are conjoined for the initiation of prognostic and diagnostic instruments to fill the gaps between medicine and engineering. The need for accurate prognostic tools is strengthened due to most of the clinicians are susceptible to misjudge the disease test results. In the case of breast cancer, the tools should able to classify accurately whether patients' tumor is a harmful malignant tumor or not harmful benign tumor. Many researchers have been carried out for prediction of breast cancer using publicly available data for comparative study.

One of the most frequently used breast cancer data is Wisconsin Breast Cancer available at UCI Machine Learning Repository [1]. This dataset created by Dr. William H. Wolberg of the University of Wisconsin Hospital as the result of accurately diagnosing breast masses based solely on Fine Needle Aspiration (FNA) test. This dataset consisting of 699 instances of clinical data, 458 (65.52%) of them are categorized as benign (benign breast tumor), whereas 241 (34.47%) were categorized as malignant (malignant breast tumor). Each instance consists of 9 attributes with assigned integer value with range 1-10 and one class category with the binary value of either 2 (benign) and 4 (malignant).

A large number of researches on Wisconsin Breast Cancer (WBC) datasets are found in the literature [2]-[10]. The classification performances of four fuzzy rule generation methods on WBC data were examined in [2]. The classification accuracies of five different classifiers namely multilayer perceptron neural network combine neural network, probabilistic neural network, recurrent neural network and support vector machine [3]. The study has shown that the SVM achieved higher diagnostic accuracies than the other four neural network family methods. A study implemented Fuzzy C-Means to classify WBC data into two clusters, benign and malignant [4]. The experimental results show that Fuzzy C-Means has True Positive 100%, True Negative 87%, False Positive 0%, and False Negative 13%. Another study compared Extreme Learning Machine Neural Network (ELM-ANN) and Back Propagation Neural Network (BP-ANN) [5]. The ELM-ANN algorithm excels in accuracy and specificity, but in metric sensitivity, BP-ANN algorithms perform better than ELM-ANN.

Another study [6] evaluate the value of Area Under Curve (AUC) and scores cost of three different algorithms, namely Extreme Gradient Boosting, Support Vector Machine Kernel RBF and Multi-Layer Perceptron. A hyper-parameter tuning was performed to find the best parameters for each algorithm using detection cost false positive and cost false negative. Cost false positive is the cost incurred for performing FNA test. While the cost false negative is calculated based on how many years of potential life are lost at the time of death caused by breast cancer multiplied by the value of a year of life. The results show that SVM algorithm outperforms other algorithms based on both AUC and cost values. SVM get $2,740.2 for cost score and 99.23 for AUC score with detail as follows: 94.6% accuracy, 92.0% specificity and 100% sensitivity.

Bioinformatics data is usually high dimensionality in terms of attribute number and record numbers. A high attribute or feature dimensionality affects the performance of the machine learning algorithm used for classification [11]. Hence, prior to classification, a so-called dimensionality reduction is frequently employed to diminish the amount of feature. It can be done either by choosing only the most important feature or by extracting new features from raw data. Feature extracting technique based on eigenvector decomposition known as principal component analysis (PCA) is the most popular employed in the breast cancer prediction research. PCA combined with bio-inspired machine learning method, namely artificial immunity was used to predict breast cancer on WBC datasets in [12]. Several measurements calculated from the confusion matrix, namely accuracy, detection rate and false alarm rate were evaluated and yielded satisfactory results except for false alarm rate. PCA was also utilized in a dimensional reduction in conjunction with several models, namely fixed architecture evolutionary neural network, variable architecture neural network, modular neural network and symbolic adaptive neuro evolution (SANE) for breast cancer prediction in [13], which has shown that SANE model yields the highest accuracy.

An article in a just recently published manuscript [14] presented a comprehensive study for dimensionality reduction on WBC datasets. Both feature selection and feature extraction were studied in conjunction with two classification methods, namely fuzzy logic and artificial neural network. Feature selection was done by ranking the feature according to some measurement such as information gain, gain ratio, one R-algorithm and other more. Without any transformation, features which are in lower ranks are ignored in the classification model generation. In feature extraction technique where feature transformation takes place, four algorithms were employed namely PCA, factor analysis, linear discriminant analysis and multi-dimensional scaling. The result of simulation on WBC dataset showed that maximum accuracy is obtained by the use of PCA and backpropagation neural network.

K-Means clustering method is seldom used for dimensionality reduction, though recently published paper in [15] K-Means was used for hashing clustering to reduce feature dimensionality for image classification. This published work has explained the difference between image clustering and feature clustering for image classification purpose. From N images, image features were extracted that yields originally d number of features. Using k-

Means based feature clustering, k new features are obtained that in turn was used to generate similarity-preserving binary codes of the original N images.

2. Proposed Methodology

All used methods involved in the breast cancer prediction tools will be briefly explained here. Basically a breast cancer prediction is a classification technique that doing a prognosis whether breast tumors are malignant or benign. In the presented work, two different methods for dimensionality reduction are utilized and compared. The first method is the most popular dimensionality reduction, namely PCA. The second method is an unusual method for this purpose, namely clustering technique, in this case the K-means method is chosen. K-means clustering method as an unsupervised machine learning can be used to create clusters as new features for the classification models. Fig.1 shows the functional block diagram of the suggested breast cancer prediction model. It consists of two phases namely: a training phase and a testing phase. Each phase performs Principal Component Analysis (PCA) and K-Means clustering method which will reduce the size of the dimensional data. The result of the dimensionality reduction process is a set of new features. In the training phase, the set of new features subsequently is used as features to generate a model. Afterward, the generated model is used to classify the test set in the testing phase.

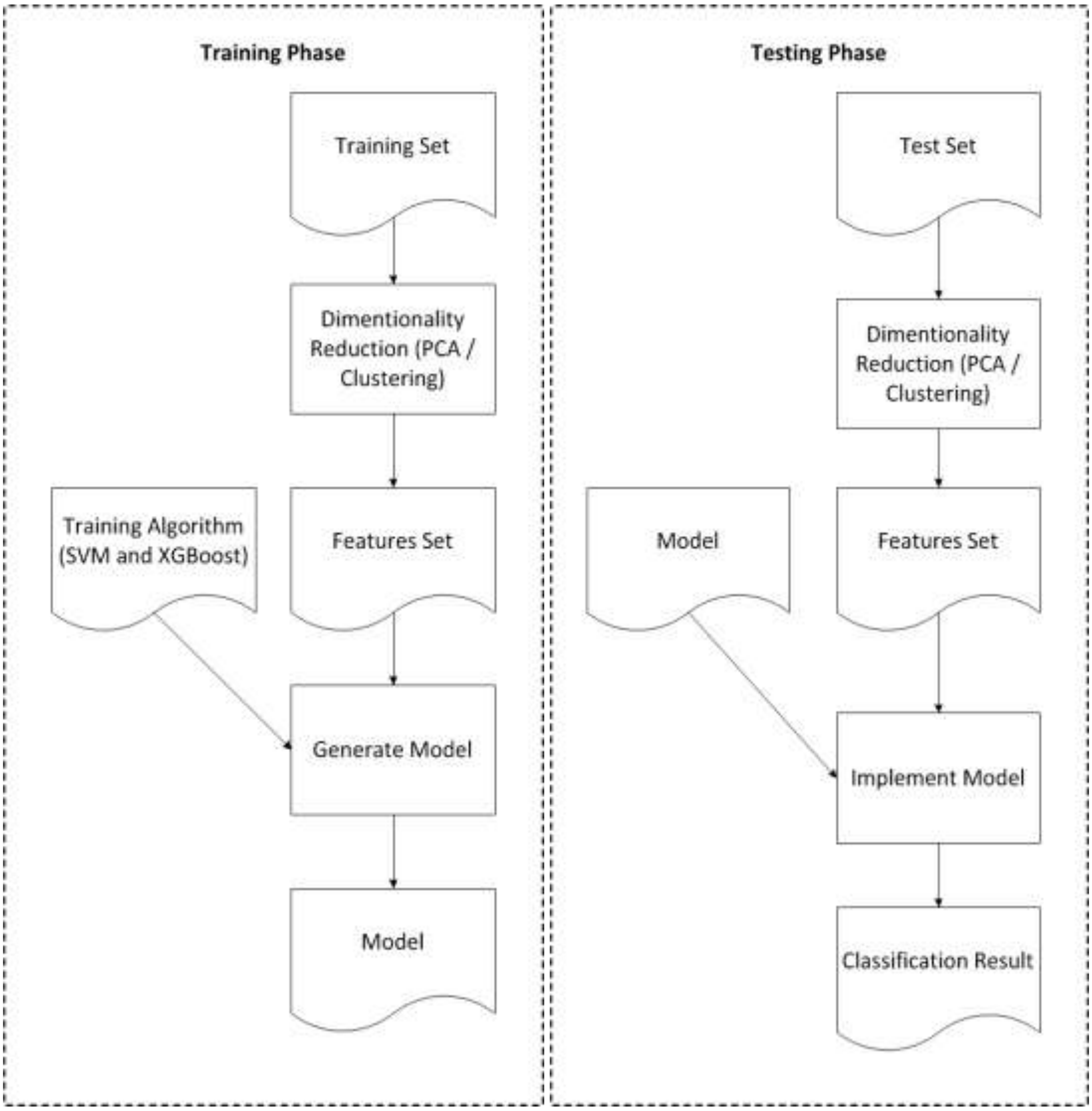

Figure 1. Proposed breast cancer prognosis model

2.1. Classification Methods

A classification method is a systematic methodology to build classifier from an input data set. A classification model is built based on a learning target function $\boldsymbol{f}$that maps each feature set$\boldsymbol{x}$ to one of the predetermined class label $\boldsymbol{y}$s. Classification techniques are most suited for predicting or describing datasets with binary or nominal classes. Classification consists of two-step processes. In the first step, a classification algorithm builds the classifier by examining a training set consisted of database tuples and their related class labels. This phase is also known as supervised learning since the class label of each training tuple is provided. In the second phase, the classifier will be used for classification.

2.1.1. Support Vector Machine (SVM)

SVM is a learning machine that makes use of a hypothesis linear function space in a high dimensional feature space, trained with a learning technique based on optimization theory that obtained from statistical learning theory. SVM concept can be explained as finding the hyperplane that differentiates the two class, class +1 (positive) and class -1 (negative).

2.1.2. Extreme Gradient Boosting (XGBoost)

Gradient Boosting Machine (GBM) is a combination of boosting method with gradient descent. Gradient boosting is a technique in machine learning for regression problems and generates predictive models in the form of weak predictive model combinations. GBM is built by making a new model to predict errors/residual from the previous model. Iteratively, a new model is added to fix the error from the previous model until no more fixes conducted. Another study [16] proposing additional improvements in the GBM, called XGBoost. XGBoost is a more efficient and scalable GBM version consisting of a collection of multiple classifications and regression trees. XGBoost assigns positive and negative values to every decision made.

2.2. Dimensionality Reduction Techniques

The dimensionality reduction can be divided into two approaches, the first one by just retaining the most relevant features from the initial dataset (feature selection), the second one by examining the inter-dependency of the initial dataset by uncovering a smaller set of new features (feature extraction). The last will be used here.

2.2.1. Principle Component Analysis

The most frequently used algorithm for feature extraction is the Principal Component Analysis (PCA). PCA would find a new set of dimensions (or a set of the basis of views) such that all the dimensions are orthogonal and ranked according to the variance data among them. It converts a set of interrelated variables into a not correlated one so-called principal components. The number of principal components is smaller than the number of initial dataset variables. This principal component is actually the eigenvectors obtained by decomposing the covariance matrix of the data. Before decomposing eigenvalue/eigenvector of the covariance matrix, it is necessary to normalize the features by subtracting the mean from each of the data dimensions. Afterward, the covariance matrix of data points will be calculated and then its eigenvectors and corresponding eigenvalues are solved. Next, the eigenvectors according to their eigenvalues are sorted in decreasing order. Choosing the first k (number of components) eigenvectors will yield the new k dimensions. Finally, PCA would transform the original dimensional data points in the new reduced dimensions.

2.2.2. K-Means Clustering.

In this research we also use K-Means clustering to perform dimensionality reduction. The more common approach is the other way around, namely the dimensionality reduction used for clustering as in [17]. Clustering is a kind of learning by observation rather than learning by examples. Hence, clustering is unsupervised learning which does not need class-labeled training examples. Clustering is also called data segmentation, because clustering divides a large dataset into several segments according to their similarity. K-Means algorithm initially takes a *k* input parameter, each of which becomes a center of *k* clusters. The remaining object in datasets is taken subsequently and allocated to the cluster which yields highly intra-cluster similarity. Cluster similarity is measured with respect to the cluster center, namely the mean value of the objects in a cluster [18]. The squared Euclidean distance is used as the measure of dissimilarity between the data point and a prototype vector. This process is repeated until the criterion function converges. Once the centroid is obtained, the newly extracted features are the distance of any object in the dataset in respect to the k centroids.

K-means clustering was used for dimensionality reduction in [15] for image classification and dubbed as Feature Clustering Hashing method. In this work, we have implemented K-means clustering straightforward as proposed in [19] where the number of clusters is provided as new labels used as the new features. However, in [19] the new features from clustering can be an additional feature to the original feature or as a complete replacement of the original features. In the first case, i.e. an additional feature, the objective is to improve the classification models. The second case is the dimensionality reduction as discussed in this article.

2.3. Classifier performance metrics

A classification model or classifier is a mapping from data instances to predicted classes. In medical cases like the current breast cancer prediction, the predicted classes are discrete and only have two values, namely positive value for a breast cancer class (malignant) or negative value for an un-harmful tumor (benign). There are four possible outcomes. If the instance is actually positive and it is classified as positive, it is called as a true positive (TP); if it is classified as negative, it is counted as a false negative (FN). If the instance is actually negative and it is classified as negative, it is considered as a true negative (TN); if it is classified as positive, it is considered as a false positive (FP). Given a classifier and a set of instances (the test set), a two-by-two confusion matrix can be constructed with the number of instances counted as TP, FP, FN and TN.

Many common metrics are deducted from these four values in confusion matrix, including accuracy= (TP+TN)/(TP+FP+FN+TN), in other words accuracy is the proportion of correct classifier with respect to all data sets. Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced. Hence two other metrics frequently used in the medical area [20] will be considered in this work, namely specificity= TN/(TN+FP) and sensitivity= TP/(TP+FN). Specificity is the proportion of not breast cancer patients that are correctly identified by the model. Sensitivity is the proportion of breast cancer patients that are correctly identified by the model. Hence, the sensitivity metric is very important for early detection of breast cancer to avoid death casualty.

3. Results

To evaluate the proposed model three measurements, namely accuracy, sensitivity and specificity were used. Prior to executing classification, data visualization will be presented for granting us an insight of dimensionality reduction results.

3.1. Data Visualization

3.1.1. Principal Component Analysis

PCA is also benefited to simplify data, by altering data linearly so that a new coordinate system with the greatest variance is obtained. Fig. 2 depicts an illustration of PCA with the number of principal component or eigenvector n=2. Different colors are used to differentiate benign and malignant breast tumor data, respectively red and blue. In two principal components these two color are found not separated. Using 3 principal components, these two classes of tumors are well separated as shown in Fig. 3.

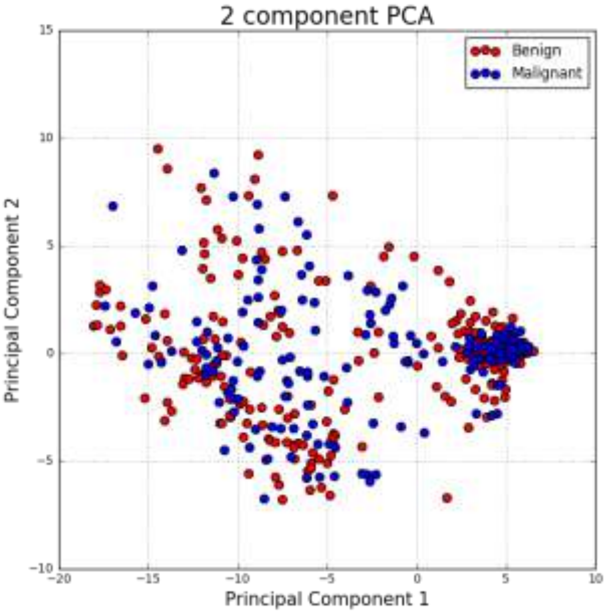

Figure 2. PCA with two components

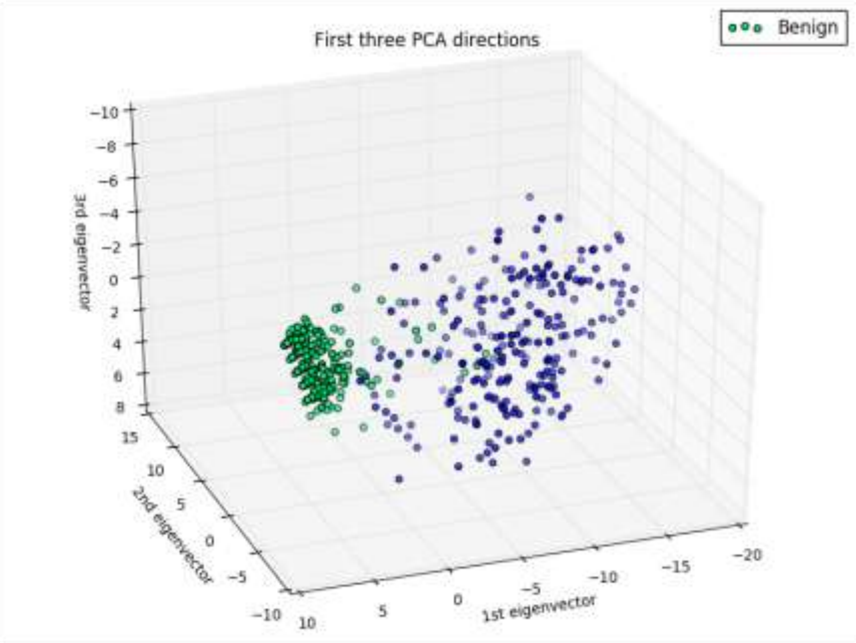

Figure 3. PCA with three components

3.1.2. K-Means clustering

In this research, K-Means clustering also employed to perform dimensionality reduction. Numbers of cluster used in K-means are determined in the range between 1 to 4. The number of cluster incorporation with its new label will replace the original feature, hence the dimension number of the feature is the same as the number of clusters. For the visualization purpose, only result with the number of clusters k=3 is presented in Fig. 4. Stars symbol indicates the centroid of clusters.

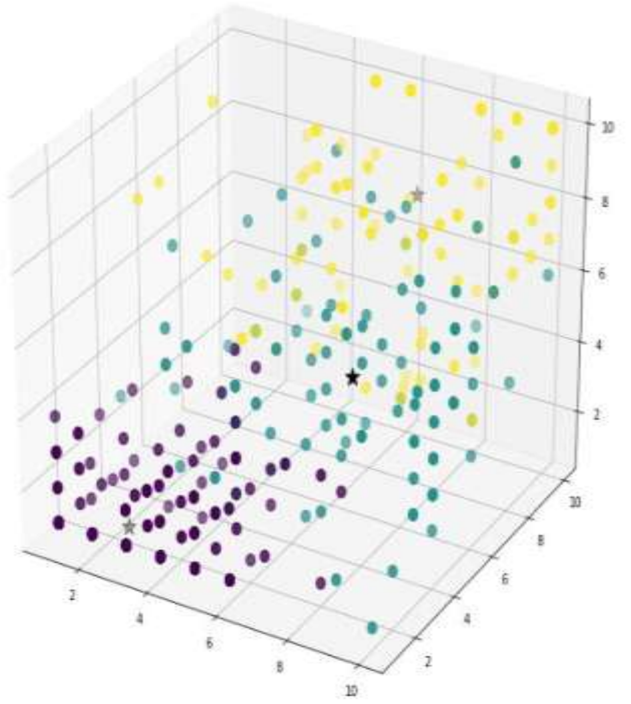

Figure 4. K-Means with 3 clusters

3.2. Metric measurement for classification

Metric measurements employed in the presented work are accuracy that indicates the proportion of correct predictions of a benign and malignant tumor with related of all data sets; specificity, namely the proportion of not harmful benign patients that are correctly identified and sensitivity which describes the percentage of correctly identified malignant tumor among the actual breast cancer patients.

3.2.1. Clustering used for dimensionality reduction

The classifier performance using K-Means clustering for dimensionality reduction combined with SVM and XGBoost are presented in Table 1 thru 4. Up to four clusters using WBC dataset from which 67% is used as a training set and 33% as a testing set are presented in Table 1 and Table 2. Noted that the metric measurement for the number of clusters is one, namely one-dimensional feature is taken into account in the classification for both method SVM or XGBoost is exceptional. The accuracy is very low as also indicated in [19] when K-means clustering used for dimensionality reduction. The specificity, also known as the True Negative Rate which indicates the percentage of healthy people who are correctly identified as not having the condition, scores maximum. The most important measurement to cure breast cancer timely, namely sensitivity is also known as True Positive Rate which indicates the percentage of sick people who are correctly identified as having the condition scores the lowest zero rate. However, when the number of clusters is two or more, all metric measurements are very good, even for accuracy. This suggests breast cancer feature from WBC dataset are highly correlated at least into two clusters.

The portion of WBC dataset used as a training and testing sets are varied and the classifier performance results are presented in Table 3 and Table 4 for three clusters used as feature extractions because from the results three clusters yield the highest sensitivity.

Table 1. K-Means and SVM (33% testing set)

Number of clusters	Accuracy	Specificity	Sensitivity
1	0.664	1.000	0.000
2	0.965	0.987	0.921
3	0.978	0.987	0.961
4	0.965	0.980	0.934

Table 2. K-Means and XGBoost (33% testing set)

Number of clusters	Accuracy	Specificity	Sensitivity
1	0.664	1.000	0.000
2	0.965	0.987	0.921
3	0.978	0.987	0.961
4	0.965	0.980	0.934

Table 3. K-Means and SVM (3 clusters)

Ratio	Accuracy	Specificity	Sensitivity
50-50	0.982	0.987	0.975
60-40	0.978	0.983	0.968
67-33	0.978	0.987	0.961
70-30	0.976	0.985	0.958
80-20	0.978	0.989	0.955

Table 4. K-Means and XGBoost (3 clusters)

Ratio	Accuracy	Specificity	Sensitivity
50-50	0.980	0.982	0.975
60-40	0.978	0.983	0.968
67-33	0.978	0.987	0.961
70-30	0.978	0.983	0.968
80-20	0.980	0.982	0.975

3.2.2. PCA used for dimensionality reduction

The classifier performance using PCA for dimensionality reduction combined with SVM and XGBoost are presented in Table 5 thru 8. Up to the first four eigenvectors as new features or principal components are provided using WBC dataset from which 67% is used as training set and 33% as a testing set are presented in Table 5 and Table 6. The portion of WBC dataset used as a training and testing sets are varied and the classifier performance results are presented in Table 7 and Table 8, respectively for three principal components.

Table 5. PCA and SVM (33% testing set)

Number of components	Accuracy	Specificity	Sensitivity
1	0.9707	0.9718	0.9701
2	0.9756	0.9859	0.9701
3	0.9659	0.9859	0.9627

4	0.9659	0.9859	0.9522

Table 6. PCA and XGBoost (33% testing set)

Number of components	Accuracy	Specificity	Sensitivity
1	0.9707	0.9718	0.9701
2	0.9707	0.9718	0.9701
3	0.9659	0.9577	0.9701
4	0.9659	0.9577	0.9701

Table 7. PCA and SVM (3 components)

Ratio	Accuracy	Specificity	Sensitivity
50-50	0.9766	0.9916	0.9686
60-40	0.9745	0.9895	0.9665
67-33	0.9659	0.9859	0.9627
70-30	0.9707	0.9859	0.9627
80-20	0.9781	1.0000	0.9677

Table 8. PCA and XGBoost (3 components)

Ratio	Accuracy	Specificity	Sensitivity
50-50	0.9766	0.9832	0.9731
60-40	0.9745	0.9895	0.9665
67-33	0.9659	0.9577	0.9701
70-30	0.9659	0.9577	0.9701
80-20	0.9708	0.9773	0.9677

4. Conclusions

The presented article has shown that the number of features for classification of breast cancer from the original WBC data set can be reduced by the feature extracting, namely transforming original data using principal component (eigenvector) decomposition and also using K-means clustering technique. The last mentioned technique is quite unusual tools for dimensionality reduction. In that case, the feature extraction is done by transforming data from the original dimensional to new dimensional based on the Euclidian distance from each cluster centroids.

The metric measurement results that the dimensionality reduction using K-means cluster is almost as good as PCA with the reduced feature number at least two clusters. Using only one cluster in K-means clustering yields incorrect classification model regarding True Positive Rate, i.e. sensitivity. Sensitivity as per definition the proportion of breast cancer patients that are correctly identified by the model, is the most important measurement for the sake of early detection of breast cancer.

References

[1] O. L. Mangasarian, “Cancer Diagnosis via Linear Programming" *SIAM News,* vol. 23, no. 5, p. 1-18, 1990.

[2] R. Jain and A. Abraham, “A Comparative Study of Fuzzy Classification Methods on Breast Cancer Data” *Australasian Physics & Engineering Sciences in Medicine,* Vol. 27, no. 4, p. 213-218, 2004.

[3] E. D. Ubeyli, “Implementing Automated Diagnostic Systems for Breast Cancer Detection” *Expert System with Applications,* Vol. 33, no. 4, p. 1054-1062, 2007.

[4] I. Muhic, “Fuzzy Analysis of Breast Cancer Disease Using Fuzzy C- Means and Pattern Recognition” *Southeast European Journal of Soft Computing,* vol. 2, no. 1, p. 50-55, 2013.

[5] C. P. Utomo, A. Kardiana and R. Yuliwulandari, “Breast Cancer Diagnosis Using Artificial Neural Networks with Extreme Learning Techniques” *International Journal Advanced Research in Artificial Intelligence,* vol. 3, no. 7, p. 10-14, 2014.

[6] A. Handayani, A. Jamal and A. A. Septiandri, “Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara” *Jurnal Nasional Teknik Elektro Teknologi Informasi* vol. 4**,** no. 4, p. 394-403, 2017.

[7] A. Fallahi and S. Jafari, “An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network” *International Journal of Advanced Science and Technology,* vol. 34, p. 65-70, 2011.

[8] A. Aloraini, "Different Machine Learning Algorithms for Breast Cancer Diagnosis," *International Journal of Artificial Intelligence & Applications (IJAIA)*, vol. 3, no.6, p. 21-30, 2012.

[9] K. Sivakami and Nadar Saraswathi, "Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model," *International Journal of Scientific Engineering and Applied Science (IJSEAS),* vol. 1, no. 5, p.418-429, 2015.

[10] K. Menaka and S. Karpagavalli , "Breast Cancer Classification using Support Vector Machine and Genetic Programming," *International Journal of Innovative Research in Computer and Communication Engineering*, vol.1, no. 7, p. 1410-1417, 2013.

[11] M. U. Ali, S. Ahmed, J. Ferzund, A. Mehmood and A. Rehman, “Using PCA and Factor Analysis for Dimensionality Reduction of Bioinformatics Data” *International Journal of Advanced Computer Science and Applications,* vol. 8, no. 5, p. 415-426, 2017.

[12] M. M. Al-Anezi, M. J. Mohammed and D. S. Hammadi, “Artificial Immunity and Feature Reduction for Effective Breast Cancer Diagnosis and Prognosis” *International Journal of Computer Science Issue,* vol. 10, no. 3, p. 136-142, 2013.

[13] R. R. Janghel, R. Tiwari, R. Kala and A. Shukla, “Breast cancer data prediction by dimensionality reduction using PCA and adaptive neuro evolution” *International Journal of Information Systems and Social Change,* vol. 3, no. 1, p. 1-9, 2012.

[14] K. Gupta and R. R. Janghel, “Dimensionality Reduction-Based Breast Cancer Classification using Machine Learning” *Computational Intelligence: Theories, Application and Future Directions* (*Advances in Intelligent System and Computing*), vol. 1, editors N. K. Verma and A. K. Ghosh, Springer Nature Singapore Pte Ltd., p. 133-146, 2019.

[15] T. Yuan, W. Deng, J. Hu, Z. An, and Y. Tang, “Unsupervised Adaptive Hashing based on Feature Clustering” *Neurocomputing*, vol. 323, p. 373-282, 2019.

[16] T. Chen and C. Guestrin, “XGBoost: a Scalable Tree Boosting System” in *KDD'16 Proceedings of the 22nd ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining,* California, 2017, p. 785-794.

[17] D. Napoleon and S. Pavalakodi, “A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Sets”, *International Journal of Computer Applications,* vol. 13, no. 7, p. 41-46, 2011.

[18] D. Rusjayanthi, “Identifikasi Biometrika Telapak Tangan Menggunakan Metode Pola Busur Terlokalisasi, Block Standar Deviasi, dan K-Means Clustering” *Lontar Komputer*, vol. 4, no. 2, p. 265-276, 2013.

[19] M. Khan, “KMeans Clustering for Classification” Towards Data Science, 7 Aug. 2017 [online], Available: https://towardsdatascience.com/kmeans-clustering-for-classification-74b992405d0a [Access 10 Oct. 2018]

[20] Arif Habib, Meshiel Alalyani, I Hussain Musa and M. S. Almutheibi, “Brief review on Sensitivity, Specificity and Predictivities” *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)*, vol. 14, no. 4, p.64-68, 2015.