Machine Learning Theory and Application Breast Cancer Classification

Byung Chan Choi

***Abstract*—This paper focuses on building an classifier for breast cancer dataset using machine learning (ML) techniques. The first step for building an effective classifier is to analyze the features of the dataset and determine the multi-colinearity between features using correlation matrix and box plots. The second step is to select appropriate preprocessing methods for each type of ML-based models. Since not all the features in the dataset use the same unit and scale, the dataset needs to be standardized or normalized. This is to prevent the negative effects on model training caused by different feature scales and units. Also, in this paper, Principle Component Analysis (PCA) will be used to extract principle components and check their effects on classification performance. The third step is to build and train various types of ML models for the breast cancer dataset classification. This paper tries both supervised learning-based models and unsupervised learning-based models. Training process will use grid search and 5-fold cross validation to determine the optimal parameters for supervised learning-based model. Lastly, the performance of each model will be evaluated and compared on receiver operating characteristics (ROC) curve.**

***Index Terms*—Machine Learning, Supervised Learning, Unsupervised Learning, Classification**

I. Introduction

ML-based models have been widely used in various classification problems. Each ML model has its advantages and disadvantages stemmed from its unique characteristics. ML-based approaches can be split into two major approaches : supervised learning-based approach and unsupervised learning-based approach.

Supervised learning-based approach uses the dataset with labels and teaches its models to classify the input data into one of pre-determined labels. Linear Regression, Support Vector Machine (SVM), Random Forest, and Neural Network are well-known supervised learning-based models. Recent developments in graphic processing unit and the increased availability of big data have led to the ever-increasing popularity of Deep Neural Network, an advanced version of neural network with deeper and more layers. Although supervised learning is a straightforward methodology in training the model, it must have the labels in the dataset for model training. In this paper, Logistics Regression, SVM, Random Forest, and Multi-Layer Perceptron (MLP) will be used to build the classifier for breast cancer dataset.

Unsupervised learning-based approach does not require pre-determined labels in its dataset. This is because unsupervised learning approach focuses on clustering the data or producing its own optimal answers based on input features. In clustering algorithms, such as K-Means and Gaussian Mixture Model (GMM), each data cluster is later classified into labels by the user. This paper focuses on clustering-based models of unsupervised learning approach for classification task. In this paper, K-Means, GMM, Agglomerative Clustering, and Density-based Spatial Clustering of Applications with Noise (DBSCAN) will be used to build the classifier for breast cancer dataset.

II. Breast Cancer Dataset Analysis and Preprocessing

Breast cancer dataset is a classic binary classification dataset from Diagnostic Winsconsin Breast Cancer Database. It contains 569 samples with 30 features. Among the samples, 212 out of 569 samples are classified as malignant and the rest is classified as benign. A classifier for this dataset has to determine whether a patient has a breast cancer or not based on some or all of 30 features.

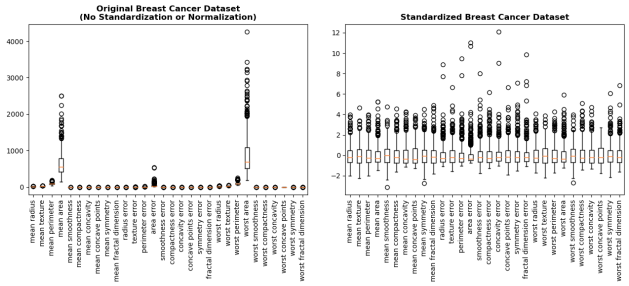

Fig. 1: Distribution of Feature Data. Original Dataset Distribution (Left). Standardized Dataset Distribution (Right)

The first step in dataset analysis is to check the scale difference of each feature through its distribution. If the feature data uses different scales and units from others, its distribution will be different from others. This can directly affect the performance of the classifier through its optimization target. Figure 1 shows the distribution of original feature data and standardized feature data. Original feature data distribution shows that some features' distributions are significantly different from others. These features can cause negative impact on the distance-based classifiers, such as SVM. This is because feature values in different scales can produce different distance values. If those features are correlated to others, they will also affect the training results and parameters of others. Therefore, it is important to apply data rescaling in order to unify the feature data under same scale and unit. Standardized feature data distribution from Figure 1 shows that standardization rescales the dataset with similar mean and variance under z-score. Through standardization, all the features will have same amount of influence upon the classifier's performance and optimization target.

Next step in dataset analysis is to check the multicolinearity between features by building the correlation matrix of the dataset. This is to determine whether linear classifiers can work in this dataset and to pick the core features that are independent from each other.

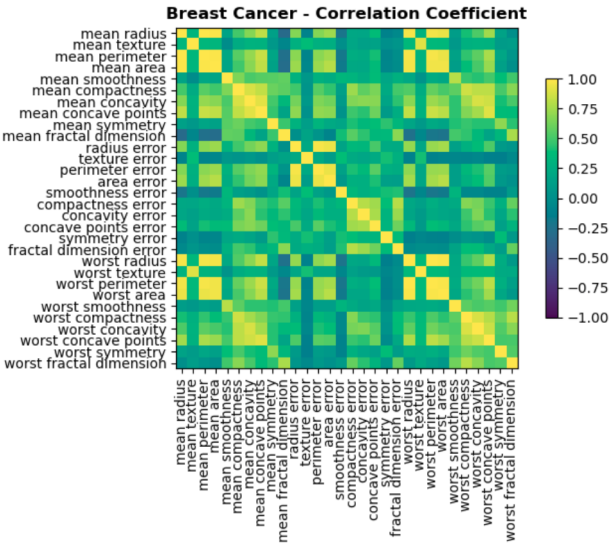

Fig. 2: Correlation Matrix of Breast Cancer Dataset

Figure 2 shows the correlation matrix of the breast cancer dataset. According to this matrix, many of features in the dataset are correlated to each other. This means that many of the features are not truly independent from the influence or changes of other features. During the training process, it would be very difficult and time-consuming to find proper independent parameters for each feature. This is because each feature will be affected by other features' parameters. As a result, it would be ineffective to apply linear classification models, such as

Linear Regression for this dataset. It is recommended to apply more complex classification models that are capable of non-linear classification. Also, since there are too much multi-colinearity among the features, it would be difficult to handpick the core features.

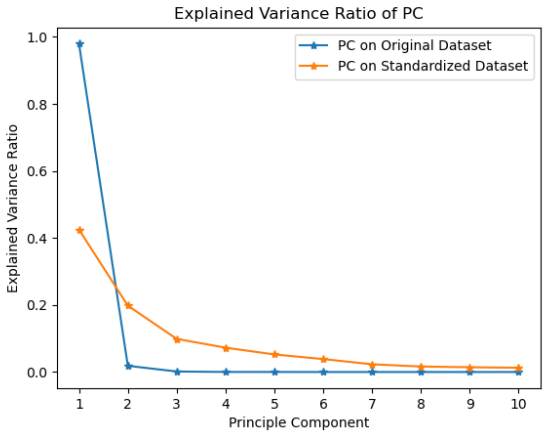

Fig. 3: PCA Results of Breast Cancer Dataset

Correlation matrix has shown that complex classification models are recommended as an effective classifier for breast cancer dataset. It also shows that there are very few independent features for classification task. This raises the question whether all those 30 features in current input space are useful for the classification task. Dimensionaltiy reduction techniques, such as PCA, can reform the entire dataset in the feature space with new features of high explained variance. Figure 3 shows the results of PCA on original dataset and standardized dataset. It shows that original unscaled dataset can be reformed with top 2 Principle Components (PC) under high explained variance of 95%. Standardized dataset requires 5 PCs to be reformed into the dataset with similar explained variance as the original dataset. Dataset reformation through PCA is also highly recommended. It can re-define the entire dataset with fewer features, which will lead to the significant reduction in training time. With new core features that are truly independent from others, the classification performance can greatly increase.

In this analysis phase, original breast cancer dataset needs to be standardized in order to make all the features have same amount of influence upon the performance of classifier. Correlation matrix has shown that it is recommended to utilize the complex models capable of non-linear classification for current classification task. PCA results from Figure 3 shows that it is also recommended to re-organize the dataset under PCs in order to improve training and prediction performance. In this paper, each model from supervised learning and unsupervised learning will be applied on four types of dataset : original unscaled dataset, standardized datatset, PCA-transformed original dataset, and PCA-transformed standardized dataset.

III. Supervised Learning-based Approaches

Through Section II, breast cancer classification task requires complex models that are capable of handling non-linear data composition. This paper will apply 6 types of supervised learning-based models : Logistics Regression, SVM, K-Nearst Neighbors (KNN), Gaussian Naive-Bayes Classifier, Random Forest, and MLP for breast cancer classification task.

Logistics Regression is an effective classifier for binary output. It uses curve-shape decision boundary to make classification decisions on the data. Its curve shape can fit to the distribution of binary-labeled dataset. Since breast cancer classification task is binary classification task, logistics regression is a strong candidate for becoming an effective classifier.

SVM is a highly well-known and widely-used classification method for its adaptability. SVM focuses on building the decision boundary that maximizes the margin between labels. Through non-linear kernel tricks and dimensional reduction, SVM can be modified to handle

non-linear dataset. However, since SVM relies on the distance values between data, it is easily influenced by scale difference between features.

KNN is a simple classification algorithm that selects K number of nearest neighbors to input data and classifies the data into the label with majority votes. It is based on the assumption that data with same label are clustered close together. Although KNN is very simple to implement, it requires a lot of data in the pool to reach high performance. Also, since it is based on distance measurements between data, like SVM, its performance is heavily influenced by scale and distribution of data. It requires appropriate dataset preprocessing.

Gaussian Naive-Bayes is a classifier based on the assumption that there is a probability that target output possesses its input features. Given features, the classifier is trained to find the label with highest probability. It uses Bayes theorem to indirectly infer the target output of input features.

Random Forest is a classifier composed of various versions of decision trees. Each decision tree in the forest is made of different features and standards. Each decision tree in the forest produces its own inference result from input features. Random Forest conducts majority voting on the inference results from the decision trees. The label with the highest votes are eventually classified as final classification output for input features. Random Forest is effective in passing classification inference on non-linear dataset distribution. However, since it consists of many different version of decision trees, it requires a large memory space and cannot produce the inference results in real-time.

MLP is a classifier with mulitple layers of perceptrons. Perceptrons in each layer is densely connected to the ones in the next layer. Single perceptron is a simple linear classifier. By combining many perceptrons in layers, the network of perceptrons can produce the decision boundary that is much more complex than the one made by single perceptron. The decision boundary made by MLP can handle non-linear data classification. As the layers grow and nodes increase, it also becomes more complicated and capable for non-linear classification task. MLP uses backpropagation to update its weight parameters in order to optimize them to training dataset.

IV. Unsupervised Learning-based Approaches

Clustering algorithms of unsupervised learning can be considered as the complex classifier for breast cancer classification task. However, unlike supervised learning-based approach, clustering algorithms do not explicitly produce the labels that match the ones in the dataset. It is capable of forming the clusters. However, the index of each cluster does not always align with label distribution of the dataset. In unsupervised learning, it is imperative to apply label matching after forming the clusters from the dataset. In binary classification task, as there are only two types of output labels, malignant (1, Positive) and benign (0, Negative), the label matching only involves either keeping the output cluster labels or inverting them between malignant and benign. In this paper, the label matching results that produce the highest accuracy will be used. There will be 4 types of clustering algorithms used in this paper : K-Means, GMM, Agglomerative Clustering, and DBSCAN.

K-Means is a hard clustering algorithm that positions each cluster to the centroid of it nearest data cluster. It assigns each data in the dataset to the nearest cluster index and keeps the record in the partition matrix. Based on the data assignments from the partition matrix, each cluster updates its position to the centroid of the assigned data and saves the results in the codebook. By continuously updating the codebook and partition matrix until the overall distance error is converged to

the minimum, each cluster will be moved to the centroid of the data clusters. However, as K-Means relies on the distance measurements between cluster centers and data, it cannot effectively identify non-circular shaped cluster and eliptical shaped clusters. Also, clustering results greatly change based on initial positions of each cluster. It is easily influenced by outliers.

GMM is a soft clustering algorithm that applies Gaussian distribution model on cluster formations. It is based on the assumption that if the data is densely populated, it is highly likely that most of the data is centered around the mean of the cluster. Each cluster in GMM is a multivariate Gaussian distribution model. GMM tries to fit each Gaussian-shape cluster as close as possible to the nearest data cluster centroid. By fitting the center of Gaussian model of each cluster to the centroid of the data cluster, the occurrence probability of data will be maximized. GMM uses multiple clusters with Gaussian distribution model to handle multiple data cluster and applies covariance matrix to produce eliptial shapes. It uses EM algorithms to make each cluster to form its cluster. In E-Step, all the data is labelled based on the level of partial affiliation to each cluster. Each data is defined as the combination of affiliation ratios. In M-Step, mean, variance, and mixing coeffients are updated so that each cluster's responsiblity is maximized. GMM is capable of passing flexible decisions on data clustering. Although GMM is capable of creating both circular and eliptial cluster shapes, it cannot handle non-ciruclar shaped data clusters. Like K-Means, its clustering performance and results greatly vary according to initial position of each cluster.

Agglomerative clustering is a hierarchical and hard clustering algorithm. At the beginning, each data is considered as cluster. The distance between each data is recorded on proximity matrix. Each cluster can expand and grow bigger by merging with nearby clusters. Cluster growth continues until there is no longer cluster to merge or it reaches target number of clusters. It performance and characteristics depends on the methods of distance measurements.

DBSCAN is a clustering algorithm that focuses on local growth of the clusters. Global distance-based clustering algorithms, such as K-Means, are vulnerable to outliers. Since they include outliers in clusteirng process, the clusters can be skewed, which results into the degradation in classification performance. Instead of global distance measurements, DBSCAN only consider local distance with nearby data or cluster. Each cluster merges with nearby local points and clusters based on their density conditions, maximum cluster radius (Epsilon) and minimum cluster point. Local expansion based on density conditions allows DBSCAN to effectively handle non-circular shaped dataset. It can also easily reject outliers in the dataset. However, if the overall dataset is less dense than density conditions, DBSCAN might reject most of the data as outliers.

V. Model Training and Evaluation Results

Breast cancer dataset is prepared in 4 types : . This is to observe the effect of preprocessing on the performance of the classifier. Each dataset will bet split into 70% training set and 30% test set. Datasets will be applied on 6 types of supervised learning-based models : Logistics Regression, SVM, K-Nearst Neighbors (KNN), Gaussian Naive-Bayes Classifier, Random Forest, and MLP for breast cancer classification task. They will be applied to 4 types of clustering algorithms used in this paper : K-Means, GMM, Agglomerative Clustering, and DBSCAN.

As suggested in Table I, grid search with 5-fold cross validation will be peformed on supervised learning-based models for parameter selection and model training. K-Menas and GMM will produce clustering models with

	Model Types	Grid Search Training Paramters	Training Method
Supervised Learning	Logistics Regression	None	5-Fold Cross Validation
	KNN	Neighbor : 1 $\sim$100	5-Fold Cross Validation
	Gaussian Naive Bayes	None	5-Fold Cross Validation
	Random Forest	Number of Trees : 1 $\sim$100	5-Fold Cross Validation
	SVM	Kernel : Linear, RBF, Sigmoid C : 0.001, 0.01, 0.1, 1.0 Gamma : 0.001, 0.01, 0.1, 1.0	5-Fold Cross Validation
	MLP	1st Layer Perceptrons : 1, 10, 100 2nd Layer Perceptrons : 1, 10, 100	5-Fold Cross Validation
	Model Types	**Training Parameters**	**Training Method**
Unsupervised Learing	K-Means	Number of Clusters : 2	2 Split Dataset
	GMM	Number of Clusters : 2 EM Iterations : 100	2 Split Dataset
	Agglomerative Clustering	Number of Clusters : 2	2 Split Dataset
	DBSCAN	EPS : 0.001, 0.01, 0.1, 1, 10, 100, 1000 MinPts : 1 $\sim$50	2 Split Dataset

TABLE I: Model Training Parameters and Methods

training set and evaluated on test set. In case of Agglomertiave Clustering and DBSCAN, their clustering capabilities will be evaluated on the entire dataset. In breast cancer classification task, it imperative to identify breast cancer correctly while minimizing the misdiagnosis of cancer. Each model will be plotted on receive operating characteristic (ROC) curve. Through ROC curve, each model will be evaluated on its sensitivity and specificity. The best classifier for breast cancer classification is the one with highest true positive rate (TPR, Sensitivity) and lowest false positive rate (FPR, Specificity). This will be determined base on area under curve (AUC) value.

Table II shows the training and parameter selections results for both supervised learing-based models and unsupervised learning-based models. In supervised learning, overall test set classification accuracy suggests that standardization preprocessing can improve the accuracy of supervised-learning based classifiers. Although PCA does not always increase the accuracy compared with the case of using original dataset, it still maintains high accuracy whether is standardized or not.

In unsuperivsed learning, standardization preprocessing has greater impact on the accuracy of clusteirng-based classifier. This is because preprocessing can dramatically change the distribution of data, which directly influences the clustering results. However, in case DBSCAN, standardization preprocessing caues negative impact on the classifier. This is because standardization increases the overall density of the data by unifying them under similar mean and variance. This can affect local merging of DBSCAN and makes it harder to separate between data clusters. PCA bring possitive effects on Agglomerative Clustering, because transforming the dataset in the feature space can make difference in distance measurements used in Agglomerative Clustering algorithm. However, when PCA is used in tandem with standardization, it does not always bring significant improvements in accuracy. Transforming the dataset in the features with high explained variance can improve the performance of distnace-based clustering algorithms.

On the other hand, standardization changes the overall formation of data clusters. Therefore, when PCA and standardization are used together in preprocessing step, the effect on clustering-based classifier is not determinant.

Figure 4 and 5 show ROC curves of the classifiers from supervised learning-based approaches and unsupervised learning-based approaches. These figures suggest that supervised learning-based models are more reliable than clustering-based classifiers in its TPR and FPR. This suggests that supervised-learning based classifiers are more capable in classifying the patients with cancer while maintaining misdiagnosis at minimum level. This is because clustering-based models do not know how to re-label or adjust their parameters based on the labels, which are not given during the clustering process. PCA and standardization tend to improve the performance of supervised learning-based classifiers. Among supervised learning-based classifiers, SVM shows the best classification performance in overall AOC metrics under various types of dataset. Therefore, when it comes to the classification tasks with clear labels, it is recommended to apply supervised learning-based models along with preprocessing and dimensionality reduction.

VI. Conclusion

In this paper, various ML models from supervised learning and unsupervised learning have been utilized in diverse ways for breast cancer classification task. In analysis phase, the correlation matrix is used to determine whether it is appropriate to apply linear classifiers. In preprocessing, dataset is standardized or dimensionally reduced for effective model training. In training phase, each model is put through various types of dataset through cross validation and grid search. In evaluation phase, each type of models is evaluated and compared on ROC curve. Supervised learning-based models have shown higher classification performance results than unsupervised learning-based models. These experiments and evaluations show that it is important to apply appropriate model preprocessing, training strategies, and evaluation methodologies to determine the optimal classifier for given classification task.

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[3] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

[4] O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

[5] K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

	Model Types	Best Model Parameters	Test Accuracy (Dataset Type)
Supervised Learning	**Logistics Regression**	None	0.9356 (Original)
			0.9766 (Standardized)
			0.9298 (PCA-Original)
			0.9699 (PCA-Standardized)
	KNN	Neighbor : 9	0.9415 (Original)
		Neighbor : 9	0.9707 (Standardized)
		Neighbor : 8	0.9415 (PCA-Original)
		Neighbor : 14	0.9649 (PCA-Standardized)
	Gaussian Naive Bayes	None	0.9649 (Original)
			0.9347 (Standardized)
			0.9298 (PCA-Original)
			0.9298 (PCA-Standardized)
	Random Forest	Number of Trees : 59	0.9824 (Original)
		Number of Trees : 74	0.9707 (Standardized)
		Number of Trees : 53	0.9122 (PCA-Original)
		Number of Trees : 33	0.9766 (PCA-Standardized)
	SVM	Kernel : Linear / C : 0.1 / Gamma : 0.001	0.9532 (Original)
		Kernel : Linear / C : 0.1 / Gamma : 0.001	0.9766 (Standardized)
		Kernel : Linear / C : 0.01 / Gamma : 0.001	0.9298 (PCA-Original)
		Kernel : RBF / C : 1.0 / Gamma : 0.01	0.9766 (PCA-Standardized)
	MLP	1st Layer Perceptrons : 100 / 2nd Layer Perceptrons : 100	0.9356 (Original)
		1st Layer Perceptrons : 1 / 2nd Layer Perceptrons : 100	0.9766 (Standardized)
		1st Layer Perceptrons : 10 / 2nd Layer Perceptrons : 100	0.9356 (PCA-Original)
		1st Layer Perceptrons : 1 / 2nd Layer Perceptrons : 100	0.9766 (PCA-Standardized)
	Model Types	**Best Model Parameter**	**Test Accuracy (Dataset Type)**
Unsupervised Learning	**K-Means**	Number of Clusters : 2	0.8654 (Original)
			0.9122 (Standardized)
			0.8654 (PCA-Original)
			0.9122 (PCA-Standardized)
	GMM	Number of Clusters : 2 EM Iterations : 100	0.9473 (Original)
			0.9473 (Standardized)
			0.9590 (PCA-Original)
			0.8654 (PCA-Standardized)
	Agglomerative Clustering	Number of Clusters : 2	0.7785 (Original)
			0.8804 (Standardized)
			0.8752 (PCA-Original)
			0.8365 (PCA-Standardized)
	DBSCAN	EPS : 100 / MinPts : 50	0.9015 (Original)
		EPS : 10 / MinPts : 50	0.6291 (Standardized)
		EPS : 100 / MinPts : 50	0.9015 (PCA-Original)
		EPS : 0.1 / MinPts : 50	0.9226 (PCA-Standardized)

TABLE II: Classifier Model Training Results

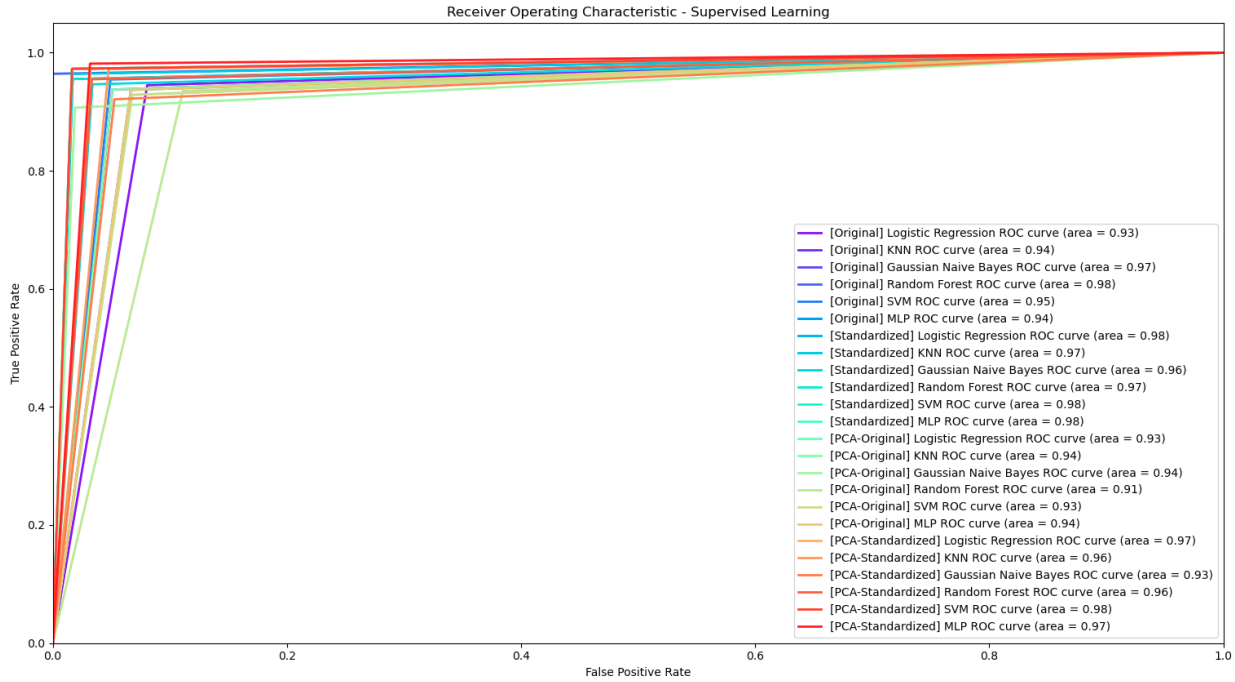

Fig. 4: ROC Curve of Supervised Learning-based Models

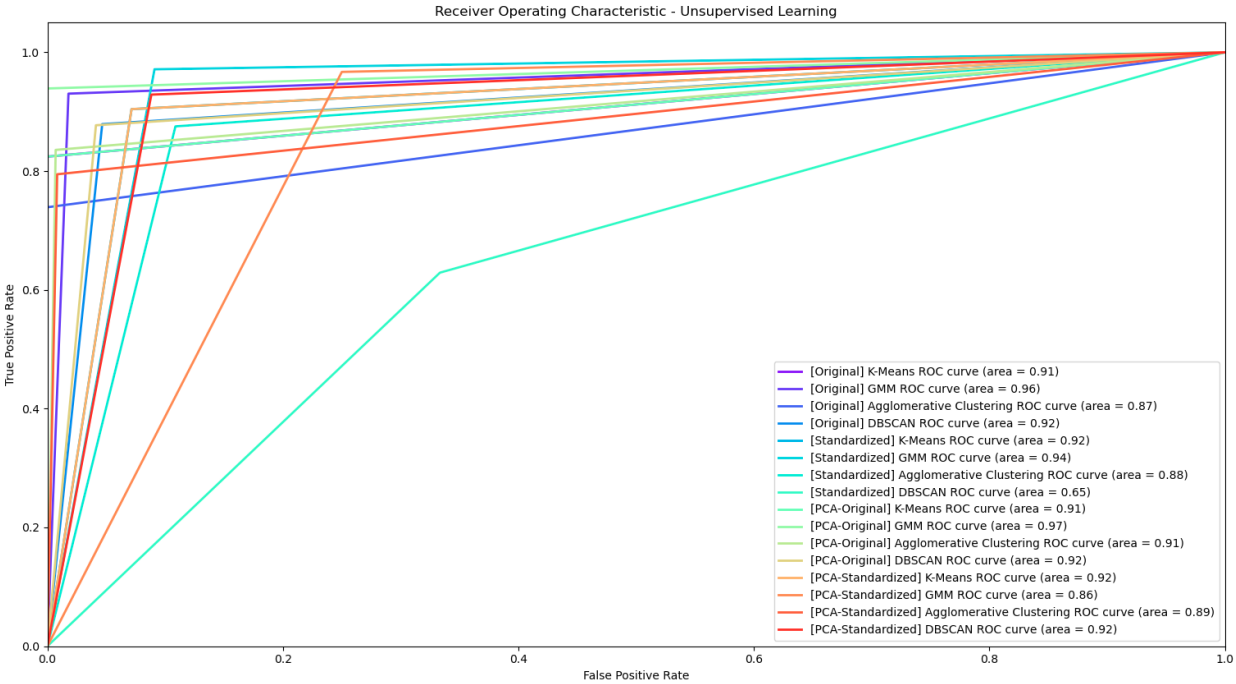

Fig. 5: ROC Curve of Unsupervised Learning-based Models