Breast Cancer Classification with Principal Component Analysis and Smote using Random Forest Method and Support Vector Machine

Rian Oktafiani
Master Program of Information Technology
Universitas Teknologi Yogyakarta
Sleman, Yogyakarta, Indonesia

Enny Itje Sela
Master Program of Information Technology
Universitas Teknologi Yogyakarta
Sleman, Yogyakarta, Indonesia

ABSTRACT

Patients' lives may be at risk due to low-accuracy and inaccurate breast cancer classification results. The high dimensionality and unequal distribution of classes in breast cancer medical data presents a challenge for the application of machine learning techniques. Subsequently, studies that examine the parameters in the algorithm model are still scarce. Inappropriate parameter selection may lead to low accuracy. To classify breast cancer, this study compares the Random Forest and Support Vector Machine algorithms. The max depth parameter in Random Forest and Linear, Polynomial and RBF kernels in Support Vector Machine are the parameters analyzed in this study. Principal Component Analysis (PCA) is used for feature reduction and Synthetic Minority Oversampling Technique (SMOTE) method is used to overcome class imbalance. The results of this study are, the best accuracy obtained from the SVM method is 99.07% with precision, recall and f1 score 99% by using the RBF kernel and at n component PCA = 6, while Random Forest has the best test accuracy of 98.32%, with precision, recall and f1 score 98% by using max depth = 8 and n component PCA = 6. Therefore, it can be concluded that the method of using SMOTE and PCA can improve accuracy, and the SVM method is better than RF for breast cancer classification. Future studies can test various datasets to examine the impact of additional parameters and classification techniques.

General Terms

Data Mining, Pattern Recognition, Classification, Machine Learning

Keywords

Classification, Breast Cancer, Principal Component Analysis, SMOTE, Random Forest, Support Vector Machine

1. INTRODUCTION

A malignant tumor called breast cancer grows in the breast cells and has the potential to spread to other parts of the body [1]. The diagnosis of breast cancer is ineffective and causes significant harm to the sufferer and the most fatal is that it can cause death [2]. Machine learning and classification-based strategies are among the data mining techniques that can help in the early identification of cancer [3]. However, inaccurate classification results and accuracy can result in a wrong diagnosis and endanger the patient's life [4].

The high dimensionality of medical data presents one of the challenges in applying machine learning techniques, which will affect the process of analysis. PCA, or principal component analysis, is a popular feature reduction method [5]. PCA is a statistical method that can reduce the number of dimensions in complex datasets while maintaining the maximum amount of pertinent information [6]. Research conducted by [7] stated that for breast cancer classification using PCA can increase accuracy, the Adaboost algorithm increased from 95.43% to 96.4%, Decision Tree (DT) from 89.1% to 92.3%, K Nearest Neighbor (KNN) from 92.9% to 97%, Logistic Regression (LR) from 94.5% to 98.8%.

In addition, class imbalance is also an important issue in breast cancer classification. Most of the datasets in breast cancer cases have more samples in the negative class (non-cancer) compared to the positive class (cancer) [8]. Synthetic Minority Oversampling Technique (SMOTE) is an oversampling technique that creates new samples from minority classes by interpolating existing samples [9], [10]. The test results in research [11] handling the imbalance of the Hepatitis C classification dataset, using the SMOTE method in the oversampling technique. Random Forest classification results with SMOTE produced 98% of the test results. This result increased from testing without SMOTE which got 93% results, while using Naïve Bayes (NB) with SMOTE produced 89% results from 88% without SMOTE.

Previous studies have shown that several factors, such as feature selection, class imbalance management, and algorithm selection, can affect classification accuracy. To better understand how handling class imbalance using SMOTE affects classification accuracy using Random Forest (RF) and Support Vector Machine (SVM) on breast cancer datasets, this study will examine the impact of using PCA in feature reduction. When dealing with complex and high-dimensional dataset classification problems, the SVM algorithm is thought to be effective [12]. The advantages of Random Forest (RF) include its capacity to manage dependencies and interactions between complicated and complex variables and its ability to handle overfitting. In addition, Random Forest (RF) has excellent classification performance, can handle very large amounts of training data easily, and is a useful method for guessing missing data [13]. This research will also consider parameters in SVM such as the use of kernel and max depth analysis and other parameters in the Random Forest algorithm.

2. LITERATURE REVIEW

2.1 Previous Study

Problems regarding cancer were also studied by [14] conducted experiments using three datasets Wisconsin Breast Cancer Original (WBCO), Wisconsin Breast Cancer Diagnostic (WBCD) and Wisconsin Breast Cancer prognostic (WBCP) using machine learning algorithms. The results of this study are KNN and RF on the WBCO dataset, producing the highest testing data accuracy of 97.14% and LR 96.56%. In WBCD, the highest accuracy is 96.50% using SVM (linear), SVM (RBF) and Random Forest. Then WBCP with the highest accuracy of 82% with LR, KNN, and RF algorithms. Furthermore, research [15] using the WBCD dataset using the Recursive Feature Elimination method resulted in the highest test accuracy of SVM (linear) 95.61% and SVM (Polynomial) 98.25%. Then [16] applied the Linear Discriminant Analysis (LDA) feature extraction method with Random Forest (RF) and Support Vector Machine (SVM) on the Wisconsin Breast Dataset. Testing accuracy results for SVM with LDA and RF with LDA are 95.6% and 96.4%, respectively. Differences with previous research, this study uses a combination of handling class imbalance using SMOTE and feature reduction methods using Principal Component Analysis (PCA) and using machine learning classification algorithms Support Vector Machine (SVM) and Random Forest (RF). This research also considers SVM parameters such as kernel and Random Forest parameters, namely max depth.

2.2 Breast Cancer Classification

Classification is one of the tasks in machine learning where models or algorithms are used to predict or categorize data that has labels into different classes or groups based on the attributes or features possessed by the data [17]. Breast cancer, also known as Carcinoma Mammae, is a malignant tumor that originates in the breast tissue, and symptoms can include lumps, changes in breast shape, skin dimpling, nipple discharge, and red, scaly patches on the skin. These cancers can grow in various parts of the breast tissue, including the mammary glands, glandular ducts, and supporting tissues, and have the potential to metastasize or spread to other parts of the body [18], [19]. breast cancer classification is the process of identifying and separating breast cells or tissues into two main categories: benign (non-cancerous) or malignant (cancerous).

2.3 Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) is one of the methods used to overcome class imbalance. SMOTE is an oversampling technique, which performs random sampling [20]. According to [21] through the generation of synthetic data based on the replication of data from the minority class, the data in the minority class is enhanced in SMOTE. The following is the SMOTE equation:

$$X_s = X_i + (X_{knn} - X_i) \times \delta \qquad (1)$$

where, Xs is the synthetic data to be created, Xi is the data to be replicated, Xknn is the neighboring data that has the closest distance from Xi data, δ is a random number between 0 and 1.

2.4 Principal Component Analysis (PCA)

A popular statistical method for data and dimensionality reduction is principal component analysis (PCA), which entails lowering the number of variables in a dataset while preserving as much of the original information as feasible [22]. The following six steps are used to apply PCA [23]. The first step, find the covariance matrix of the normalized d-dimensional dataset. Second, find the eigenvalues and eigenvectors of the covariance matrix. Third, the eigenvalues are sorted from highest to lowest. fourth, select k eigenvectors, where k is the dimension of the new feature subspace, which map to the k largest eigenvalues. Fifth, create a projection matrix using the k selected eigenvectors. Sixth, the original data is transformed to the new k-dimensional feature space.

2.5 Random Forest (RF)

Random Forest (RF) is a popular choice for classifying text due to its easy algorithm, one of the benefits of using Random Forest (RF) compared to other Machine Learning models is its ability to process high-dimensional data and its robust performance even when working with large amounts of data [24]. The steps in the Random Forest algorithm are as follows[25]; First, take the total number of features (m) and divide it by the number of trees to be used (k), if k is less than m. Next, extract N samples from the dataset. Next, select N samples for every tree in the dataset. Third, choose a subset of m predictors at random from each tree, where m < p, the number of predictor variables. Fourth, until the desired number of trees (k) is reached, repeat steps two and three. Fifth, most votes from the classification outcomes of each formed tree are used to determine the prediction result.

2.6 Support Vector Machine (SVM)

The main idea behind the Support Vector Machine (SVM) algorithm, which is based on linear machines and has excellent characteristics, is to build a hyperplane as a decision maker so that the margin of separation between positive and negative classes is as large as possible [26]. The method in SVM used to classify non-linear data is to apply a kernel function to transform data from the initial input space into a new vector space with higher dimensions [27].

3. RESEARCH METHODS

The research method relates to the steps taken in the research, breast cancer classification. In the following figure are the stages of research.

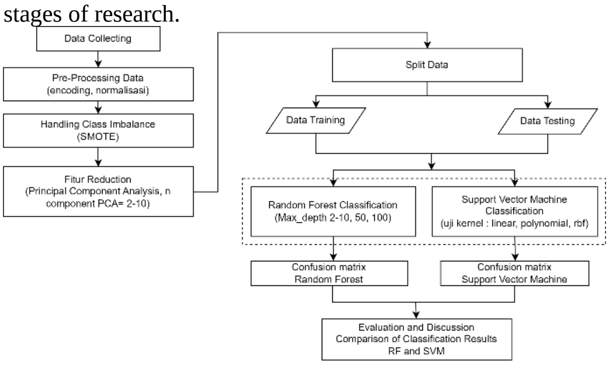
Data Collecting
Pre-Processing Data (encoding, normalisasi)
Handling Class Imbalance (SMOTE)
Fitur Reduction (Principal Component Analysis, n component PCA= 2-10)
Split Data
Data Training
Data Testing
Random Forest Classification (Max_depth 2-10, 50, 100)
Support Vector Machine Classification (uji kernel : linear, polynomial, rbf)
Confusion matrix Random Forest
Confusion matrix Support Vector Machine
Evaluation and Discussion Comparison of Classification Results RF and SVM

Fig 1: Research Stages

The "Breast Cancer Wisconsin" UCI Machine Learning Repository, which is accessible at the following link, provided the data used in this study : https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic [28]. Based on Figure 1, the first research stage is data collection. The data is obtained through digital image analysis of breast masses and consists of 569 data with 32 attributes. The diagnosis attribute is an attribute that is labeled in this study, namely malignant cancer (malignant) and benign cancer (benign). The research data generated 10 main attributes, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, fractal dimension. Each of the 10 main attributes has 3 indicators, namely mean, standard error/se, and worst. Then for 2 other attributes, namely ID number and diagnosis. After the data is collected, proceed with the preprocessing stage. The preprocessing stage in this study is checking for missing values, encoding and normalization. When checking for missing values, there was no empty data, but there was an empty column, namely unnamed 32, and the column was deleted. In addition, the ID column cannot be used for classification because it is unique, therefore the ID column is also deleted. Then the encoding process is carried out to change the categorical data, namely the diagnosis attribute which is categorical data containing Malignant (M) and Benign (B), the B attribute will be encoded into 0, and M into 1. Next, data normalization will be carried out using Standard Scaler or Z Score. This Standard Scaler or Z Score normalization method changes the feature values so that they have a mean of 0 and a standard deviation of 1. After preprocessing is done and the data becomes clean, then class imbalance is handled using the SMOTE method, because the diagnosis data in this data is unbalanced, namely the number of malignant cancer data is 212, while benign cancer is 357. This research uses the PCA method for feature reduction. The features in the breast cancer dataset will be reduced using PCA. The n-component value of PCA used in this study is n 2 to 10. The next step is to split the dataset into training data and test data. The split data method used in this study uses holdout-validation which divides the dataset into two groups, namely the training data dataset and the test data dataset. This research uses 7 holdout validation schemes to divide training data and test data and can be seen in table 1.

Table 1. Split data Holdout Validation Scheme

Percentage (Training Data: Test Data)	**Training Data**	**Testing Data**
50%:50%	285	284
60%:40%	342	227
70%:30%	399	170
75%:25%	427	142
80%:20%	456	113
85%:15%	484	85
90%:10%	512	57

Furthermore, the dataset is classified using Random Forest (RF) and SVM algorithms. In Random Forest, the max-depth parameter in RF will be tested, namely max depth 2 to 10, 50 and 100 for optimal accuracy results in breast cancer classification. This approach allows the research to comprehensively investigate and find parameter configurations that provide optimal classification results in this specific context. This study also considered the use of SVM kernels for classification, such as linear, rbf, and polynomial kernels. In the last step, the Random Forest (RF) and Support Vector Machine (SVM) classification models will be tested, and the test results evaluated using Confusion Matrix to see the accuracy, precision, recall and f1 score values. The results of RF and SVM classification will be discussed in relation to the impact of parameter usage, SMOTE and PCA usage.

4. RESULT AND DISCUSSION

4.1 Preprocessing Result

In this dataset there is an Unnamed 32 column which contains NaN or empty data. In addition, there is an ID column, the ID column will be deleted because it is not used for this research. In this research, the encoding process is carried out to change the categorical data, namely the diagnosis attribute which is categorical data containing Malignant (M) and Benign (B), the B attribute will be encoded into 0, and M into 1. The encoding used is using Ordinal Encoder. In the context of breast cancer classification, normalization is performed on the features (variables) used in the model. This research uses the Standard Scaler or Z-Score normalization method.

4.2 SMOTE Implementation

In this study, the SMOTE method is used to handle class imbalance in the breast cancer dataset. Before using SMOTE the dataset was unbalanced, where the B (Benign) class was 357 while the M (Malignant) class was 212. Then after using SMOTE, the M class is oversampled or the data is added to 357, so that the total M and B classes after SMOTE become 714 data. The following is a comparison table of dataset division before and after using SMOTE. Table 2 Comparison of Data Split Before and After Using SMOTE

Table 2. Comparison of Data Split Before and After Using SMOTE

Split data (%)	Before *SMOTE*= 569				After *SMOTE* = 714,			
	Train		*Test*		*Train*		*Test*	
	B	M	B	M	B	M	B	M
50:50	170	114	187	98	185	172	185	172
60:40	209	132	148	80	219	209	148	138
70:30	249	149	108	63	255	244	113	102
75:25	268	158	89	54	271	264	93	86
80:20	286	169	71	43	288	283	74	69
85:15	303	180	54	32	307	299	58	50
90:10	317	195	40	17	326	316	41	31

4.3 PCA Implementation

After the dataset has been balanced using the SMOTE method, the next step is to use the PCA method to reduce the dimensionality of the complex dataset while retaining most of the information. Its application in the context of breast cancer classification aims to reduce the number of features or variables present in breast cancer data to facilitate processing and analysis by classification models. Once applied, PCA will produce principal components which are linear combinations of the original features. In this study using n components 2 to 10.

4.4 Random Forest Classification Results Without SMOTE and PCA

In the first scenario of testing in this study, Random Forest classification was tested without using SMOTE and PCA. In classification using Random Forest, preprocessing and normalization are carried out. The following table 3 is the result of Random Forest classification without using SMOTE and PCA.

Table 3. The result of Random Forest classification without using SMOTE and PCA.

***Split data* (%)**	***Max depth***	**Training Accuracy (%)**	**Testing Accuracy (%)**	***Precision* (%)**	***Recall* (%)**	***F1* (%)**
50:50	6	100	95.57	96	95	95
60:40	**6**	**99.71**	**96.05**	**96**	**95**	**96**
70:30	3	98.74	94.15	94	93	94
75:25	4	99.53	93.71	94	93	93
80:20	6	99.78	95.61	96	95	95
85:15	2	96.89	94.19	94	93	94
90:10	6	99.61	94.74	95	94	94
Average		**99.18**	**94.86**	**95**	**94**	**94.43**

In table 3 above is a summary of the details of RF classification without using SMOTE and PCA, where the best accuracy results obtained in this scheme are test accuracy 96.05%, precision 96%, recall 95%, f1 score 96%, on split data 60%: 40%, max depth = 6. While the lowest test accuracy obtained is 93.71%. The average accuracy obtained in this scheme is 99.18% training accuracy, 94.86% test accuracy, 95% precision, 94% recall, 94.43% f1 score.

4.5 Random Forest Classification Results with SMOTE and PCA

In Random Forest classification using SMOTE and PCA, testing is carried out using split data according to a predetermined scheme. The value of n component PCA in this study was tested using n component PCA from 2 to 10. The following table 4, below is a summary of the results of breast cancer classification research using Random Forest with SMOTE and PCA.

Table 4. Random Forest Classification Results Using SMOTE and PCA

***Split data* (%)**	***Max depth***	**Training Accuracy (%)**	**Testing Accuracy (%)**	***Precision* (%)**	***Recall* (%)**	***F1* (%)**
***n component* PCA = 2**						
50:50	5	96.64	94.12	94	94	94
60:40	5	98.13	93.36	93	93	93
70:30	4	96.39	93.02	92	92	92
75:25	6	97.20	95.53	96	96	96
80:20	5	96.85	93.71	94	94	94
85:15	4	95.21	94.44	95	94	94
90:10	**7**	**95.60**	**95.83**	**97**	**95**	**96**
***n component* PCA = 3**						
50:50	10	100	94.68	95	95	95
60:40	9	99.77	95.10	95	95	95
70:30	8	100	95.35	95	95	95
75:25	**4**	**96.45**	**96.09**	**96**	**96**	**96**
80:20	7	99.30	95.80	96	96	96
85:15	5	97.85	94.44	95	94	94
90:10	7	99.38	95.83	97	95	96
***n component* PCA = 4**						
50:50	7	100	95.52	96	96	96
60:40	7	99.77	96.15	96	96	96
70:30	8	100	96.74	97	97	97
75:25	8	100	97.21	97	97	97
80:20	7	99.65	97.20	97	97	97
85:15	**6**	**99.01**	**97.22**	**98**	**97**	97
90:10	8	100	97.22	98	97	97
***n component* PCA = 5**						
50%:50%	9	100	96.36	96	96	96
60%:40%	8	100	96.85	97	97	97
70%:30%	8	100	96.28	97	97	97
75%:25%	**8**	**100**	**98.32**	**98**	**98**	**98**
80%:20%	6	99.30	97.90	98	98	98
85%:15%	7	99.50	98.15	98	98	98
90%:10%	6	99.22	92.98	98	97	97
***n component* PCA = 6**						
50%:50%	9	100	96.36	96	96	96
60%:40%	8	100	96.85	97	97	97
70%:30%	8	100	96.28	96	96	96
75%:25%	**8**	**100**	**98.32**	**98**	**98**	**98**
80%:20%	6	99.30	97.90	98	98	98
85%:15%	7	99.50	98.15	98	98	98
90%:10%	6	99.22	97.22	98	97	97
***n component* PCA = 7**						
50%:50%	9	100	96.36	96	96	96
60%:40%	6	99.30	96.15	96	96	96
70%:30%	8	100	95.81	96	96	96
75%:25%	**50**	**100**	**98.32**	**98**	**98**	**98**
80%:20%	9	100	97.92	97	97	97
85%:15%	8	100	98.15	98	98	98
90%:10%	8	100	97.22	98	97	97
***n component* PCA = 8**						
50%:50%	50	100	95.24	95	95	95
60%:40%	50	100	96.50	96	97	97
70%:30%	8	100	96.74	97	97	97
75%:25%	6	99.25	97.77	98	98	98
80%:20%	8	100	97.90	98	98	98
85%:15%	**6**	**99.50**	**98.15**	**98**	**98**	**98**
90%:10%	6	99.50	98.15	98	98	98
***n component* PCA = 9**						
50%:50%	8	100	96.08	96	96	96
60%:40%	6	100	95.80	96	96	96
70%:30%	7	100	96.28	96	96	96
75%:25%	100	100	97.26	97	97	97
80%:20%	**4**	**97.20**	**97.90**	**98**	**98**	**98**
85%:15%	6	99.17	97.22	97	97	97
90%:10%	6	99.69	97.22	98	97	97
***n component* PCA = 10**						
50%:50%	7	100	95.80	96	96	96
60%:40%	7	100	95.45	95	95	96
70%:30%	8	100	95.81	96	96	96
75%:25%	**7**	**99.81**	**97.77**	**98**	**98**	**98**
80%:20%	7	99.82	97.20	97	97	97
85%:15%	6	99.50	97.22	98	98	97
90%:10%	7	99.69	97.22	98	98	97

Then at n component PCA = 3, the best test accuracy obtained is 96.09%, with precision, recall and f1 score 96% in the split data scheme 75%: 25% and at max depth = 4. At n component PCA = 4, the best test accuracy obtained is 97.22%, with 98% precision, recall and f1 score 97% in the split data scheme 85%: 15% and at max depth = 6. At n component PCA = 5, 6 and 7, the best test accuracy obtained is 98.32% with precision, recall and f1 score 98%. This accuracy is the best accuracy in the Random Forest classification scenario using SMOTE and PCA, but at n component PCA = 7, the best accuracy is obtained at max depth = 50, which shows the complexity of the model built. In this study, n component PCA 5 and 6 produced the best test accuracy. Then the next n component PCA decreased testing accuracy, where at n component PCA = 8 the highest test accuracy was 98.15%, n PCA = 9 was 97.90% and n PCA 10 = was 97.77%. Based on table 4.5, RF classification results using SMOTE and PCA are better, compared to RF without using SMOTE and PCA. There is an increase in accuracy from 94.86% to 98.32% or an increase of 3.46%, when using SMOTE and PCA. Likewise in precision, recall and f1 score, when using SMOTE and PCA, the value of precision, recall and f1 score also increased and stabilized as in precision from 96% to 98%, recall from 95% to 98% and f1 score from 96% to 98%. The following figure 2 is a diagram that illustrates the difference before and after using SMOTE and PCA in Random Forest.

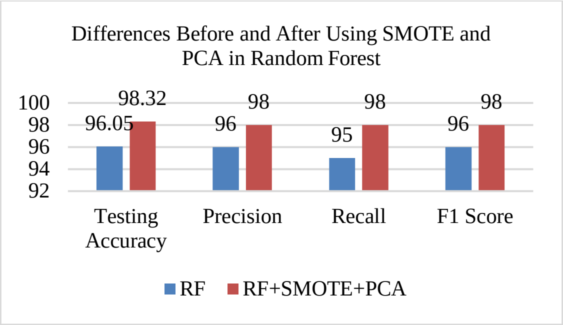

Fig 2. Differences Before and After Using SMOTE and PCA in Random Forest

4.6 SVM Classification Results without SMOTE and PCA

In the third scenario in this study, namely SVM classification without using SMOTE and PCA. In this scenario, SVM will be tested using three kernels, namely linear, rbf and polynomial kernels. The following is table 5 of the third scenario test results.

Table 5. SVM Classification Results Without SMOTE and PCA

Split data	Kernel	Accuracy (%)		Prec (%)	*Recall* (%)	F1 (%)
		Train	*Test*			
50%:50%	LINEAR	**96.13**	**95.44**	**96**	**94**	**95**
	RBF	90.14	90.88	94	88	90
	POLY	89.79	89.82	93	86	88
60%:40%	LINEAR	**96.77**	**94.74**	**95**	**93**	**94**
	RBF	90.32	91.23	94	88	90
	POLY	90.32	90.79	94	88	90
70%:30%	LINEAR	**97.24**	**93.57**	**94**	**92**	**93**
	RBF	90.95	90.06	93	87	89
	POLY	90.70	90.06	93	87	89
75%:25%	LINEAR	**96.95**	**93.71**	**94**	**92**	**93**
	RBF	91.08	88.81	92	85	87
	POLY	91.08	89.51	93	86	88
80%:20%	LINEAR	**97.36**	**92.11**	**93**	**90**	**91**
	RBF	91.87	90.35	93	87	89
	POLY	91.21	89.47	93	86	88
85%:15%	LINEAR	**96.27**	**95.35**	**97**	**94**	**95**
	RBF	91.51	94.19	96	92	94
	POLY	91.10	90.70	94	88	89
90%:10%	LINEAR	**96.09**	**94.74**	**96**	**93**	**94**
	RBF	91.60	92.98	95	90	92
	POLY	90.62	89.47	93	86	88

In table 5 above, the best test accuracy results are 95.44%, precision 96%, recall 94% and f1 score 95% on split data 50%: 50%, with a linear kernel. Based on the table above, the best results in each training and test data division scheme are using a linear kernel.

4.7 SVM Classification Results with SMOTE and PCA

In the fourth scenario in this study, SVM classification is used using SMOTE and PCA. In this study, SVM parameters were tested, namely the kernel using three kernels, namely linear, RBF and polynomial. Then tested using n-component PCA 2 to 10. on each predetermined split data scheme. After processing the classification results. These results are written into a table for SVM classification results using SMOTE and PCA. The following in table 6, is a summary of SVM classification results using SMOTE and PCA.

Table 6. SVM Classification Results Using SMOTE and PCA

Split data (%)	Kernel	N-Pca	Train Accuracy (%)	Test Accuracy (%)	*Prec (%)*	*Recall (%)*	F1 (%)
50:50	LINEAR	10	97.48	**98.32**	98	98	98
	RBF	10	98.32	97.76	98	98	98
	POLY	2	94.68	95	95	95	95
60:40	LINEAR	10	98.13	**98.25**	98	98	98
	RBF	9	98.13	97.55	98	98	98
	POLY	8	96.96	95.45	95	95	95
70:30	LINEAR	6	97.80	**97.21**	97	97	97
	RBF	9	98.40	96.74	97	97	97
	POLY	6	96.39	93.95	94	94	94
75:25	LINEAR	5	97.38	98.32	98	98	98
	RBF	10	98.13	**98.88**	99	99	99
	POLY	10	96.82	92.74	93	93	93
80:20	LINEAR	9	98.07	**98.60**	99	99	99
	RBF	9	98.07	**98.60**	99	99	99
	POLY	10	96.67	91.61	92	92	92
85:15	LINEAR	9	98.84	95.37	95	95	95
	RBF	**6**	**97.52**	**99.07**	**99**	**99**	**99**
	POLY	9	96.70	90.74	91	91	91
90:10	LINEAR	8	98.13	97.22	97	97	97
	RBF	10	98.29	**98.61**	99	98	99
	POLY	10	96.73	93.07	93	94	93

In this fourth scenario, namely, SVM classification using SMOTE and PCA, the best results were obtained when using the RBF kernel and n PCA components = 6, with a testing accuracy of 99.07%, training accuracy of 97.52%, and 99% on the precision, recall, and f1 score values. The results obtained are better than the third scenario, which is without using SMOTE and PCA, where in the third scenario the best test accuracy results are 95.44%, precision 96%, recall 94% and f1 score 95% on the division of training and testing data 50%: 50%, with a linear kernel. Accuracy has increased by 3.63%, namely test accuracy from 95.44% without using SMOTE and PCA, increasing to 99.07% when using SMOTE and PCA. In addition to test accuracy, the precision, recall and f1 score also increased to 99%. Figure 3 shows the difference in test accuracy, precision, recall and f1 score before and after using SMOTE and PCA in the Support Vector Machine (SVM) classification method.

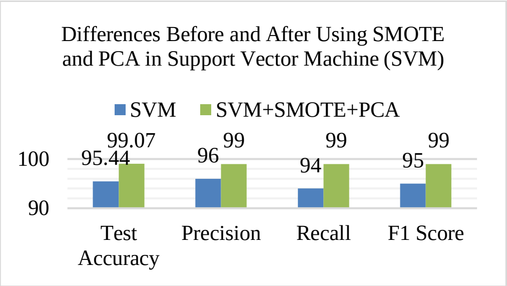

Fig 3. Differences Before and After Using SMOTE and PCA

4.8 Discussion

This breast cancer classification research with Wisconsins Breast Cancer Diagnostic dataset was conducted with four experimental scenarios and involved the use of SMOTE method to overcome class imbalance and PCA method for feature reduction, using Random Forest (RF) and Support Vector Machine (SVM) classification methods. Overall, from the four scenarios, the best results were obtained for Random Forest 98.32% and SVM 99.07%. In previous research, by [14] for breast cancer classification using the Wisconsins Breast Cancer Diagnostic dataset obtained SVM (Linear and RBF) test accuracy results of 96.50% and in Random Forest classification, the resulting test accuracy was also 96.50%. in research by [14], there is no method to overcome class imbalance and no feature reduction or selection, with the method proposed in this study, proving that by applying SMOTE to handle class imbalance and using PCA for feature reduction can improve classification results. The proposed research also tested the max depth parameter in the Random Forest classification method and the kernel parameters (linear, polynomial and RBF) in the Support Vector Machine (SVM) classification method. Testing using parameter settings is needed to get the best results.

Based on the research results, in the first scenario, namely Random Forest classification without using SMOTE and PCA, has the lowest test accuracy of 94.15%, precision 94%, recall 93% and f1 score 94% and is in the 70%: 30% data split scheme with max depth = 7, while the best results are test accuracy 96.05%, precision 96%, recall 95% and f1 score 96% in the 60%: 40% data split scheme. Then in the second scheme, namely by applying the SMOTE and PCA methods to Random Forest classification, the best accuracy results are obtained, namely in Random Forest classification using SMOTE and the best PCA is produced at n component PCA = 6 and the composition of the division of training and test data 75%: 25%, namely with 98.32% test accuracy, and with 98% value for precision, recall and f1 score. In Random Forest classification using SMOTE and PCA, there is an increase in accuracy from 94.86% to 98.32% or an increase of 3.46%.

Trials using the max depth parameter have been carried out in the first and second scenarios, which involve classification using the Random Forest method. It is possible to draw the conclusion from this study that the max depth parameter also influences or has an impact on improving classification accuracy. The study's findings indicate that accuracy rises when max depth is used. The depth of Decision Tree and Random Forest models is influenced by max depth, which influences the complexity of these models [29], [30]. Consequently, the more complex the model, the higher the max depth. An excessively high max depth value may cause overfitting, in which the model fails to generalize to new data and instead retains the training set [31]. On the other hand, underfitting, in which the model fails to adequately capture patterns, can result from a value that is too low [32]. Underfitting can lead to general poor performance, while overfitting can produce high accuracy during training but poor accuracy during testing [33]. his research has also been proven in [34] which looks at how max depth parameters are used to classify heart disease. The 90% scheme produces classification results with the best accuracy of 10% and max depth = 7, while the Random Forest algorithm produces the best accuracy results of 99.29% and the Decision Tree (DT) algorithm produces 98.05%. In addition, values for Precision and Recall rise in response to changes in max depth. Most of these maximum depths show an increase in training and testing accuracy, indicating that these depths are the best places to increase accuracy in the proposed research. These are max depths 6, 7, and 8.

Then for the Support Vector Machine (SVM) classification method using SMOTE and PCA in the fourth scenario, the best results were obtained when using the RBF kernel with 99.07% testing accuracy, 97.52% training accuracy, and 99% in precision, recall, and f1 score values. The results obtained are better than the third scenario, which is without using SMOTE and PCA, where in the third scenario the best test accuracy results are 95.44%, precision 96%, recall 94% and f1 score 95% on the division of training and testing data 50%: 50%, with a linear kernel. Accuracy has increased by 3.63%, namely test accuracy from 95.44% without using SMOTE and PCA, increasing to 99.07% when using SMOTE and PCA and using the RBF kernel at n-component PCA = 6. Based on research conducted, the kernel parameter in SVM also has an effect in improving classification accuracy. In this study, the RBF kernel provides the best test accuracy results. The following figure 4 is the result of Random Forest and SVM comparison using SMOTE and PCA.

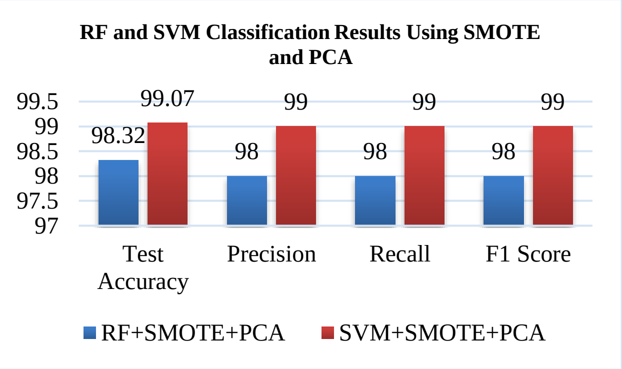

Fig 4. Comparison Diagram of Random Forest and SVM After Using SMOTE and PCA

Based on Figure 4 is the result of the comparison of Support Vetor Machine (SVM) and Random Forest (RF) classifications using SMOTE and PCA. Based on these results, the Support Vetor Machine (SVM) classification method is better than the Random Forest (RF) method for breast cancer classification on the Wisconsin Breast Cancer Diagnostic dataset. The best test accuracy of SVM is 99.07% with precision, recall and f1 Score 99%, while Random Forest best test accuracy is 98.32%, with precision, recall and f1 Score 98%. Comparatively speaking, the test accuracy produced in this study is superior to that of the research by [14], [15], [16] who used the same dataset, namely Wisconsin Breast Cancer Diagnostic. Research by [14], has a test accuracy of 96.50% on SVM and Random Forest methods, in this study did not use feature selection or handling class imbalance. The research [15], used the Recursive Feature Elimination (RFE) method and had a linear SVM test accuracy of 95.61% and polynomial SVM 98.25%. then, research by [16] Based on Figure 4, is the result of the comparison of Support Vetor Machine (SVM) and Random Forest (RF) classifications using SMOTE and PCA. Based on these results, the Support Vetor Machine (SVM) classification method is better than the Random Forest (RF) method for breast cancer classification on the Wisconsin Breast Cancer Diagnostic dataset. The best test accuracy of SVM is 99.07% with precision, recall and f1 Score 99%, while Random Forest best test accuracy is 98.32%, with precision, recall and f1 Score 98%. Comparatively speaking, the test accuracy produced in this study is superior to that of the research by [14], [15], and [16] who used the same dataset, namely Wisconsin Breast Cancer Diagnostic. Research by [14], has a test accuracy of 96.50% on SVM and Random Forest methods, in this study did not use feature selection or handling class imbalance. Rasool et al. [15], used the Recursive Feature Elimination (RFE) method and had a linear SVM test accuracy of 95.61% and polynomial SVM 98.25%. then, research by Adebiyi et al. [16] used Linear Discriminant Analysis (LDA) for feature selection resulting in LDA+RF training accuracy of 95.6% and LDA+SVM 96.4%. The results of the proposed research show that using SMOTE and PCA methods can improve accuracy in breast cancer classification.

5. CONCLUSION

The results of the tests conducted in this study, which involved classifying breast cancer using the Wisconsin Breast Cancer Diagnostic dataset and using Principal Component Analysis (PCA) for feature reduction and the Synthetic Minority Oversampling Technique (SMOTE) method to overcome class imbalance in Random Forest (RF) and Support Vector Machine (SVM) classification, indicate that using SMOTE and PCA can improve the accuracy of breast cancer classification tests. The best accuracy obtained in this study is from the SVM method, which is 99.07% with precision, recall and f1 Score 99% by using the RBF kernel and at n component PCA = 6, while Random Forest has the best test accuracy of 98.32%, with precision, recall and f1 Score 98% by using max depth = 8 and n component PCA = 6. Therefore, it can be concluded that the SVM method is better than RF for breast cancer classification. Future research can explore more deeply the optimization of the parameters of the Random Forest (RF) and Support Vector Machine (SVM) classification models by testing several datasets. to enhance the model's performance and to better understand how parameters affect it through additional research. Other deep learning or machine learning techniques may also be used in future studies.

6. REFERENCES

[1] A. R. Vaka, B. Soni, and S. R. K., “Breast cancer detection by leveraging Machine Learning,” ICT Express, vol. 6, no. 4, pp. 320–324, Dec. 2020, doi: 10.1016/j.icte.2020.04.009.

[2] K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi, and T. Sato, “Deep-Learning-Empowered Breast Cancer Auxiliary Diagnosis for 5GB Remote E-Health,” IEEE Wirel Commun, vol. 28, no. 3, pp. 54–61, Jun. 2021, doi: 10.1109/MWC.001.2000374.

[3] M. Kirola, M. Memoria, A. Dumka, A. Tripathi, and K. Joshi, “A Comprehensive Review Study on: Optimized Data Mining, Machine Learning and Deep Learning Techniques for Breast Cancer Prediction in Big Data Context,” Biomedical and Pharmacology Journal, vol. 15, no. 1, pp. 13–25, Mar. 2022, doi: 10.13005/bpj/2339.

[4] M. M. Chanu, N. H. Singh, C. Muppala, R. T. Prabu, N. P. Singh, and K. Thongam, “Computer-aided detection and classification of brain tumor using YOLOv3 and deep learning,” Soft comput, vol. 27, no. 14, pp. 9927–9940, Jul. 2023, doi: 10.1007/s00500-023-08343-1.

[5] A. D. Krismawan and E. H. Rachmawanto, “Principal Component Analysis (PCA) dan K-Nearest Neighbor (KNN) dalam Deteksi Masker pada Wajah,” Prosiding Sains Nasional dan Teknologi, vol. 12, no. 1, p. 382, Nov. 2022, doi: 10.36499/psnst.v12i1.7066.

[6] P. F. Eduardo, C. Damián, and M. Fernando, “A comparison of deep learning models applied to Water Gas Shift catalysts for hydrogen purification,” Int J Hydrogen Energy, vol. 48, no. 64, pp. 24742–24755, Jul. 2023, doi: 10.1016/j.ijhydene.2022.09.215.

[7] H. Yilmaz and F. Kuncan, “Analysis of Different Machine Learning Techniques with PCA in the Diagnosis of Breast Cancer,” Journal of Engineering Technology and Applied Sciences, vol. 7, no. 3, pp. 195–205, Dec. 2022, doi: 10.30931/jetas.1166768.

[8] S. Wang, Y. Dai, J. Shen, and J. Xuan, “Research on expansion and classification of imbalanced data based on SMOTE algorithm,” Sci Rep, vol. 11, no. 1, p. 24039, Dec. 2021, doi: 10.1038/s41598-021-03430-5.

[9] A. S. Tarawneh, A. B. A. Hassanat, K. Almohammadi, D. Chetverikov, and C. Bellinger, “SMOTEFUNA: Synthetic Minority Over-Sampling Technique Based on Furthest Neighbour Algorithm,” IEEE Access, vol. 8, pp. 59069–59082, 2020, doi: 10.1109/ACCESS.2020.2983003.

[10] H. Yi, Q. Jiang, X. Yan, and B. Wang, “Imbalanced Classification Based on Minority Clustering Synthetic Minority Oversampling Technique With Wind Turbine Fault Detection Application,” IEEE Trans Industr Inform, vol. 17, no. 9, pp. 5867–5875, Sep. 2021, doi: 10.1109/TII.2020.3046566.

[11] N. Sharfina and N. G. Ramadhan, “Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes,” JOINTECS (Journal of Information Technology and Computer Science), vol. 8, no. 1, p. 33, Jun. 2023, doi: 10.31328/jointecs.v8i1.4456.

[12] S. S. Hameed, W. H. Hassan, L. A. Latiff, and F. F. Muhammadsharif, “A comparative study of nature-inspired metaheuristic algorithms using a three-phase hybrid approach for gene selection and classification in high-dimensional cancer datasets,” Soft comput, vol. 25, no. 13, pp. 8683–8701, Jul. 2021, doi: 10.1007/s00500-021-05726-0.

[13] H. Tantyoko, D. K. Sari, and A. R. Wijaya, “Prediksi Potensial Gempa Bumi Indonesia Menggunakan Metode Random Forest Dan Feature Selection,” IDEALIS : InDonEsiA journaL Information System, vol. 6, no. 2, pp. 83–89, Jul. 2023, doi: 10.36080/idealis.v6i2.3036.

[14] N. Feroz, M. A. Ahad, and F. Doja, “Machine Learning Techniques for Improved Breast Cancer Detection and Prognosis—A Comparative Analysis,” 2021, pp. 441–455. doi: 10.1007/978-981-16-3067-5_33.

[15] A. Rasool, C. Bunterngchit, L. Tiejian, Md. R. Islam, Q. Qu, and Q. Jiang, “Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis,” Int J Environ Res Public Health, vol. 19, no. 6, p. 3211, Mar. 2022, doi: 10.3390/ijerph19063211.

[16] M. O. Adebiyi, M. O. Arowolo, M. D. Mshelia, and O. O. Olugbara, “A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis,” Applied Sciences, vol. 12, no. 22, p. 11455, Nov. 2022, doi: 10.3390/app122211455.

[17] S. Hidayatulloh, M. A. Mustajab, and Y. Ramdhani, “PENGGUNAAN OTIMASI ATRIBUT DALAM PENINGKATAN AKURASI PREDIKSI DEEP LEARNING PADA BIKE SHARING DEMAND,” INFOTECH journal, vol. 9, no. 1, pp. 54–61, Feb. 2023, doi: 10.31949/infotech.v9i1.4530.

[18] M. Lestandy, “Deteksi Dini Kanker Payudara Menggunakan Metode Convolution Neural Network (CNN),” Inspiration: Jurnal Teknologi Informasi dan Komunikasi, vol. 12, no. 1, p. 65, Jun. 2022, doi: 10.35585/inspir.v12i1.2667.

[19] K. Suparna and L. M. K. K. Sari, “Kanker Payudara: Diagnostik, Faktor Risiko, Dan Stadium,” Ganesha Medicina Journal, vol. 2, no. 1, pp. 42–48, Mar. 2022, Accessed: Oct. 26, 2023. [Online]. Available: https://ejournal.undiksha.ac.id/index.php/GM/article/view/47032

[20] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, “Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM,” MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 3, no. 2, pp. 153–160, Oct. 2023, doi: 10.57152/malcom.v3i2.897.

[21] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, “Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier,” Indonesian Journal of Computer Science, vol. 12, no. 5, Nov. 2023, doi: 10.33022/ijcs.v12i5.3413.

[22] K. Younes et al., “Application of Unsupervised Machine Learning for the Evaluation of Aerogels’ Efficiency towards Ion Removal—A Principal Component Analysis (PCA) Approach,” Gels, vol. 9, no. 4, p. 304, Apr. 2023, doi: 10.3390/gels9040304.

[23] M. A. Almaiah et al., “Performance Investigation of Principal Component Analysis for Intrusion Detection System Using Different Support Vector Machine Kernels,” Electronics (Basel), vol. 11, no. 21, p. 3571, Nov. 2022, doi: 10.3390/electronics11213571.

[24] M. Rizky and R. Andarsyah, “Klasifikasi MIT-BIH Arrhythmia Database Metode Random Forest dan CNN dengan Model ResNet-50: A Systematic Literature Review,” Jurnal Teknologi Dan Sistem Informasi Bisnis, vol. 5, no. 3, pp. 190–196, Jul. 2023, doi: 10.47233/jteksis.v5i3.825.

[25] M. I. C. Rachmatullah, A. Wicaksono, and V. Putratama, “Perbandingan Metoda K-NN, Random Forest dan 1D CNN untuk Mengklasifikasi Data EEG Eye State,” Journal of Information System Research (JOSH), vol. 4, no. 2, pp. 669–675, Jan. 2023, doi: 10.47065/josh.v4i2.2998.

[26] S. D. Asri, D. Ramayanti, A. D. Putra, and Y. T. Utami, “DETEKSI RODA KENDARAAN DENGAN CIRCLE HOUGH TRANSFORM (CHT) DAN SUPPORT VECTOR MACHINE (SVM),” Jurnal Teknoinfo, vol. 16, no. 2, p. 427, Jul. 2022, doi: 10.33365/jti.v16i2.1952.

[27] M. A. Saddam, E. Kurniawan, and I. Indra, “Analisis Sentimen Fenomena PHK Massal Menggunakan Naive Bayes dan Support Vector Machine,” Jurnal Pengembangan IT (JPIT), vol. 8, no. 3, Sep. 2023, Accessed: Oct. 25, 2023. [Online]. Available: http://ejournal.poltektegal.ac.id/index.php/informatika/article/view/4884

[28] W. Wolberg, N. Street, and O. Mangasarian, “Breast Cancer Winscoin ,” UCL Machine Learning Repository. Accessed: Apr. 28, 2023. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[29] G. L. Pritalia, “Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum,” KONSTELASI: Konvergensi Teknologi dan Sistem Informasi, vol. 2, no. 1, Apr. 2022, doi: 10.24002/konstelasi.v2i1.5630.

[30] M. Zhou, H. Zhang, W. Zhang, and Y. Yi, “An Improved Random Forest Algorithm-Based Fatigue Recognition With Multiphysical Feature,” IEEE Sens J, vol. 23, no. 21, pp. 26195–26201, Nov. 2023, doi: 10.1109/JSEN.2023.3314316.

[31] B. P. Koya, S. Aneja, R. Gupta, and C. Valeo, “Comparative analysis of different machine learning algorithms to predict mechanical properties of concrete,” Mechanics of Advanced Materials and Structures, vol. 29, no. 25, pp. 4032–4043, Oct. 2022, doi: 10.1080/15376494.2021.1917021.

[32] A. D. Patange, S. S. Pardeshi, R. Jegadeeshwaran, A. Zarkar, and K. Verma, “Augmentation of Decision Tree Model Through Hyper-Parameters Tuning for Monitoring of Cutting Tool Faults Based on Vibration Signatures,” Journal of Vibration Engineering & Technologies, Nov. 2022, doi: 10.1007/s42417-022-00781-9.

[33] H. Zhang, L. Zhang, and Y. Jiang, “Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems,” in 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/WCSP.2019.8927876.

[34] R. Oktafiani, A. Hermawan, and D. Avianto, “Max Depth Impact on Heart Disease Classification: Decision Tree and Random Forest,” Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) , vol. 8, no. 1, pp. 160–168, Feb. 2024, doi: https://doi.org/10.29207/resti.v8i1.5574.