International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Breast Cancer Detection Using Reduced Feature Representation

Dr. T. Seshu Chakravarthy[1], Associate Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh, India.

Tumati Tejasri [2], **Thatha Bharath**[3], **Sadhupati Pranitha**[4],**Thota Oohitha Ramesh**[5]

[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh, India.

[1]tschakravarthy@vvit.net, [2]22bq1a05m3@vvit.net, [3]22bq1a05l3@vvit.net, [4]22bq1a05j1@vvit.net, [5]22bq1a05l8@vvit.net,

Abstract*—Breast cancer remains one of the most common causes of mortality among women worldwide. Accurate and early diagnosis plays an essential role in improving survival rates. Machine learning techniques have increasingly been applied in medical decision-support systems to assist physicians in identifying malignant tumors more reliably. This research investigates the use of Independent Component Analysis (ICA) as a feature reduction technique for breast cancer classification. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is utilized, which initially contains thirty diagnostic attributes extracted from digitized biopsy images. ICA is applied to transform the original feature space into a reduced representation consisting of a single independent component. To evaluate the effectiveness of this dimensionality reduction approach, several machine learning classifiers are employed, including k-Nearest Neighbor (k-NN), Artificial Neural Networks (ANN), Radial Basis Function Neural Networks (RBFNN), and Support Vector Machines (SVM). The classification performance is examined using both the original 30-feature dataset and the reduced feature representation. Different validation strategies such as 5-fold cross-validation, 10-fold cross-validation, and random data partitioning are used to assess performance. The classifiers are evaluated using multiple metrics including accuracy, sensitivity, specificity, F-score, Youden's index, discriminant power, and Receiver Operating Characteristic (ROC) analysis. Experimental results indicate that reducing the feature dimension through ICA significantly decreases computational cost while maintaining competitive diagnostic accuracy. These findings suggest that ICA-based feature reduction can be beneficial for developing efficient computer-aided breast cancer diagnosis systems.

***Index Terms*—Breast cancer, ICA, Machine learning, Classification**

I. Introduction

Breast cancer is among the most frequently diagnosed cancers affecting women and remains a major contributor to cancer-related deaths globally. Early identification and correct classification of tumors significantly improve treatment outcomes and patient survival rates. However, traditional diagnostic procedures often depend heavily on the expertise and subjective judgment of medical specialists.

Although physicians are capable of recognizing visual patterns effectively, assigning accurate probabilistic interpretations to clinical observations can be challenging. Consequently, automated diagnostic systems based on computational methods have been developed to support medical professionals in making reliable decisions.

Machine learning has demonstrated promising results in medical data analysis. Studies show that computer- assisted diagnostic systems can achieve higher accuracy than manual diagnosis in certain conditions. Such systems analyze clinical data to identify patterns associated with benign or malignant tumors.

Breast tumors are generally classified into two categories:

· Benign tumors, which are non-cancerous and typically not life-threatening.

· Malignant tumors, which are cancerous and capable of spreading to other tissues Although benign tumors are less dangerous, they may still increase the risk of developing cancer later. Therefore, accurate classification of tumor type is crucial.

Various machine learning models have been applied to this classification problem. Artificial neural networks have been widely used due to their ability to learn complex nonlinear relationships. Radial basis function neural networks are particularly known for their fast learning capability and strong approximation properties.

Another widely used technique is the Support Vector Machine, which constructs an optimal hyperplane to separate different classes. SVM models are known for their strong generalization ability and effectiveness in high-dimensional datasets.

High-dimensional datasets often contain redundant or irrelevant features that increase computational complexity and reduce model performance. Therefore, dimensionality reduction techniques are frequently applied before classification.

Principal Component Analysis (PCA) is one such technique that reduces dimensionality using second-order sta-

tistical information. However, PCA focuses on uncorrelated components rather than statistically independent ones.

In contrast, Independent Component Analysis (ICA) extracts statistically independent components using higher-order statistics. ICA has been successfully used in pattern recognition, signal processing, and biomedical data analysis. The main objective of this research is to analyze how reducing the dimensionality of breast cancer data using ICA affects classification performance. The WDBC dataset is transformed into a single independent component, and the performance of several machine learning classifiers is evaluated using different validation techniques.

II. Materials and Methods

A. Dataset Description

The experiments conducted in this study utilize the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is widely used for evaluating machine learning algorithms in medical diagnosis tasks. The dataset contains 569 patient samples, each representing measurements derived from breast tissue images. Among these cases, 357 samples correspond to benign tumors, while 212 samples represent malignant tumors.

Each record in the dataset consists of an identification number, a diagnosis label indicating whether the tumor is benign or malignant, and thirty numerical attributes describing characteristics of cell nuclei extracted from digitized biopsy images obtained through Fine Needle Aspiration (FNA).

These attributes represent various morphological properties of the cell nuclei, including measurements related to size, shape, and texture. Examples of these characteristics include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

For each of these ten characteristics, three statistical values are computed:

· the mean value
· the standard error
· the largest (worst) value, calculated as the mean of the three largest observations.

As a result, the dataset contains a total of thirty features for each sample. These features provide detailed information about the structural properties of the cell nuclei and serve as input variables for the classification algorithms used in this study

B. Independent Component Analysis

Independent Component Analysis (ICA) is a statistical technique used to separate a set of observed signals into underlying independent components. The main objective of ICA is to identify hidden source signals that are combined within the observed data. Unlike traditional dimensionality reduction methods such as Principal Component Analysis (PCA), which focuses on removing correlation among variables, ICA aims to extract components that are statistically independent.

In ICA, the observed data are assumed to be generated by a linear combination of several unknown source signals.

This relationship can be expressed mathematically as

$$x = As$$

where xrepresents the observed data vector, sdenotes the vector of independent source signals, and A is an unknown mixing matrix describing how the sources are combined.

The goal of ICA is to estimate a transformation matrix that can recover the independent components from the observed signals. This is achieved by computing a separating matrix W, which approximates the inverse of the mixing matrix A. The independent components can then be obtained as

$$\hat{s} = Wx$$

Before extracting the independent components, the data are typically preprocessed through two important steps. First, the data are centered by subtracting the mean value of each variable. Second, the data are whitened, which removes correlations between variables and scales them to have unit variance.

After preprocessing, ICA algorithms estimate the independent components by maximizing statistical independence using higher-order statistical measures. In this study, the FastICA algorithm is employed because it provides efficient and reliable estimation of independent components.

Once the independent components are obtained, the component with the highest significance is selected as the reduced feature representation. This transformation reduces the dimensionality of the dataset while preserving important information required for accurate classification.

TABLE I
Real-valued features computed for each cell nucleus

No.	Real-valued Feature
1	Radius (mean distance from center to perimeter points)
2	Texture (standard deviation of gray-scale values)
3	Perimeter
4	Area
5	Smoothness (local variation in radius lengths)
6	Compactness ($\text{perimeter}^2/\text{area} - 1.0$)
7	Concavity (severity of concave portions of the contour)
8	Concave points (number of concave portions of the contour)
9	Symmetry
10	Fractal dimension (“coastline approximation” −1)

C. Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of biological neural systems. These models are capable of learning complex relationships between input data and output labels by adjusting internal parameters during training.

A typical neural network consists of multiple layers of interconnected processing units called neurons. The most widely used architecture is the feedforward neural network,

in which information flows sequentially from the input layer to the output layer through one or more hidden layers.

Each neuron receives input signals from neurons in the previous layer. These inputs are multiplied by corresponding weights and combined with a bias term to produce a weighted sum. The neuron then applies a nonlinear activation function to generate its output. This process can be expressed as

$$y_{\text{net}} = \sum_{i=1}^{n} x_i w_i + b$$

where $x_i$represents the input variables, $w_i$denotes the connection weights, and bis the bias term.

The resulting output of the neuron is computed by applying an activation function such as the sigmoid function:

$$y_{out} = \frac{1}{1 + e^{-y_{net}}}$$

During the training phase, the network adjusts its weights using a learning algorithm so that the predicted output becomes closer to the actual target values. In many applications, the backpropagation algorithm is used to update weights by minimizing the error between predicted and actual outputs.

In this study, a feedforward neural network with a single hidden layer is used to classify breast tumor samples. The number of neurons in the hidden layer is varied during experiments to determine the configuration that produces the best classification accuracy.

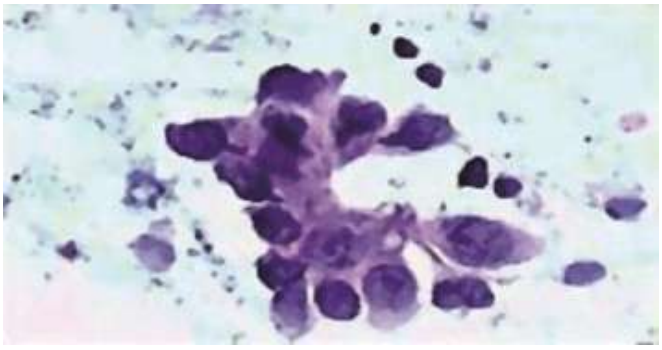

Fig.1(a) Malignant

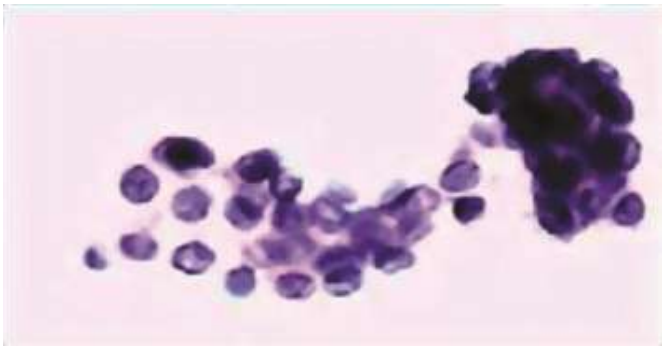

Fig.1(b) Benign

Figure 1. FNA biopsies of breast tumors [24].

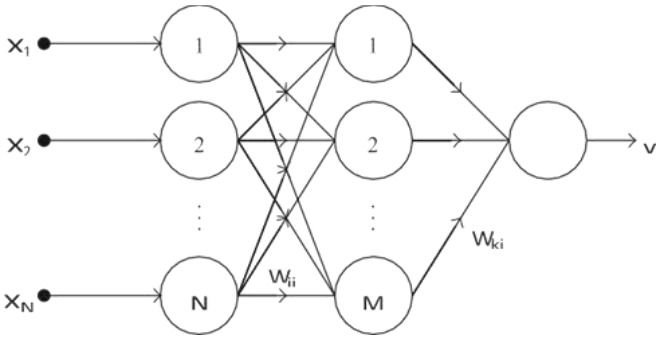

Figure 2. Architecture of feedforward neural network.

A RBFNN also consists of feedforward architecture with three layers, but the hidden layer uses Gaussian function mostly and is called radial basis layer. Each neuron consists of a radial basis function (RBF) centered on a point. The centers and spreads are computed by the training. A hidden neuron computes the Euclidean distance of input vector and the test case from the neuron's center point. Thus, it applies the RBF kernel function to the distance using the spread values.

D. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification tasks. The fundamental idea behind SVM is to determine a decision boundary that separates data samples belonging to different classes.

For binary classification problems, SVM attempts to construct an optimal hyperplane that divides the dataset into two categories while maximizing the distance between the nearest data points of each class. These nearest points are known as support vectors, and they play a crucial role in defining the position of the decision boundary.

The decision function for a linear SVM model can be expressed as

$$g(x) = w^T x + b$$

where x represents the input vector, w denotes the weight vector that determines the orientation of the hyperplane, and bis a bias term that shifts the hyperplane from the origin.

In situations where the data cannot be separated using a linear boundary, SVM employs kernel functions to transform the data into a higher-dimensional feature space. This transformation allows the algorithm to identify nonlinear decision boundaries that effectively separate the classes.

Commonly used kernel functions include:

· Linear kernel
· Polynomial kernel
· Radial Basis Function (RBF) kernel

By applying these kernels, SVM can handle complex classification problems while maintaining strong generalization performance.

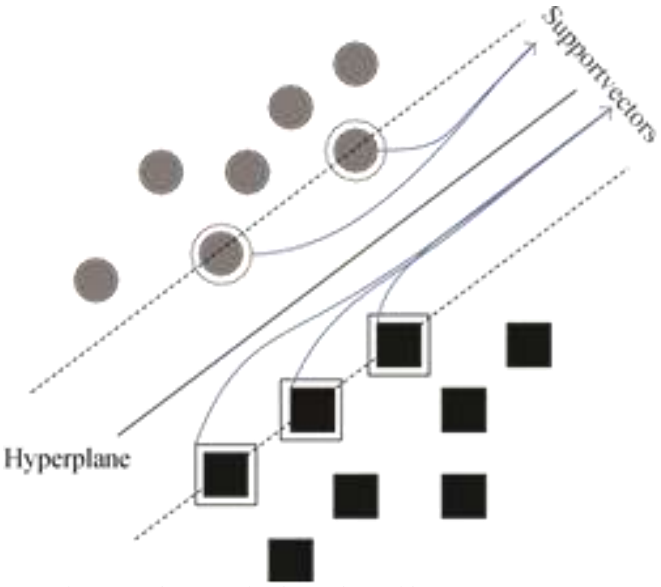

Figure 3. The separating hyperplane with support vectors.

In this study, different kernel functions are evaluated to determine which configuration produces the most accurate classification results for breast cancer detection.

This is a quadratic optimization task with respect to a set of linear inequality constraints. From Karush-Kuhn- Tucker (KKT) conditions the Lagrange function is found by

$$L_p(w, b, \alpha) = -\frac{1}{2}|w|^2 - \sum_{i=1}^{n} \alpha_i [y_i(wx_i + b) - 1]$$

where α_i are Lagrange multipliers and L_i must be minimized to find out optimal w and b. The optimization equation can be written as

The other usage of SVM is that it can solve nonlinear classification problems through the trick of a kernel function. The kernel function maps data points onto a higher-dimensional space in order to construct a hyperplane separating the classes. The new discriminant function is found by

$$g(x) = W^T \Phi(X) + b$$

TABLE II
A CONFUSION MATRIX FOR BINARY CLASSIFICATION

	Recognized value	
Actual value	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

where (X) represents the mapping of input vectors, onto the kernel space X. Therefore, the optimization equation can be written as:

where $K(x_i,x_j)$ is the kernel function equals to $\{(x_i), (x_j)\}$. The kernel functions can be radial basis function (RBF), polynomial or any symmetric functions which satisfy the Mercel conditions

E. Radial Basis Function Neural Network

The RBF Neural Network also follows a feedforward architecture but uses radial basis functions in the hidden layer.

Each neuron computes the distance between the input vector and a center point. The output is obtained using a Gaussian function:

$$\phi(x) = e^{\frac{\|x-c\|^2}{2\sigma^2}}$$

Where:
c is the center
σ represents spread.

RBF networks are effective in nonlinear classification tasks and typically converge faster than traditional neural networks.

F. Performance Metrics

The effectiveness of the classification models is assessed using several statistical indicators derived from the confusion matrix. A confusion matrix summarizes the number of correct and incorrect predictions produced by a classifier for each class.

In binary classification problems such as breast cancer detection, four possible outcomes are considered:

- True Positive (TP): malignant cases correctly identified as malignant
- True Negative (TN): benign cases correctly identified as benign
- False Positive (FP): benign cases incorrectly predicted as malignant
- False Negative (FN): malignant cases incorrectly predicted as benign

Using these quantities, several performance measures are computed to evaluate the reliability of the classifiers. Accuracy

Accuracy indicates the overall proportion of correctly classified samples among all observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity, also known as the true positive rate, measures how effectively the classifier detects malignant tumors

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity evaluates the model's ability to correctly identify benign cases

$$Specificity = \frac{TN}{TN + FP}$$

The F-score combines precision and recall into a single performance indicator and provides a balanced measure of classification accuracy.

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where precision and recall are defined as

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Additional Evaluation Measures

Two additional metrics are considered to further analyze classifier performance: Youden's Index

Youden's index measures the effectiveness of a diagnostic test in distinguishing between positive and negative cases.

$$Y = Sensitivity + Specificity - 1$$

Discriminant power evaluates the ability of a classifier to separate the two classes effectively. Higher values indicate better separation between benign and malignant samples.

ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve provides a graphical representation of classifier performance by plotting the true positive rate against the false positive rate at different threshold values. The Area Under the Curve (AUC) summarizes the ROC curve into a single numerical value, where higher values correspond to better classification capability.

To ensure reliable evaluation, the classifiers are tested using 5-fold cross-validation, 10-fold cross-validation, and random data partitioning. In cross-validation, the dataset is divided into several subsets, and each subset is used as a test set while the remaining subsets are used for training. This iterative process allows a more robust estimation of model performance.

III. Methodology

The proposed framework evaluates the effect of feature reduction on breast cancer classification.The process consists of the following steps:

1) Load the WDBC dataset containing 569 samples.
2) Apply ICA to reduce the original 30 features to a single independent component.
3) Perform cross-validation.
   - 5-fold cross-validation
   - 10-fold cross-validation
   - 20% test partition
4) Train classifiers.
   - k-Nearest Neighbor
   - Artificial Neural Network
   - Radial Basis Function Neural Network
   - Support Vector Machine.
5) Evaluate classification performance using multiple statistical metrics.

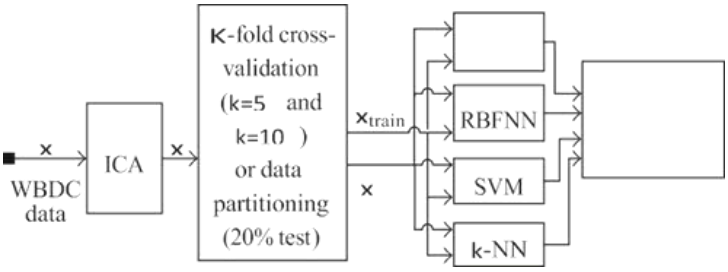

Figure 4. The basic model of the study.

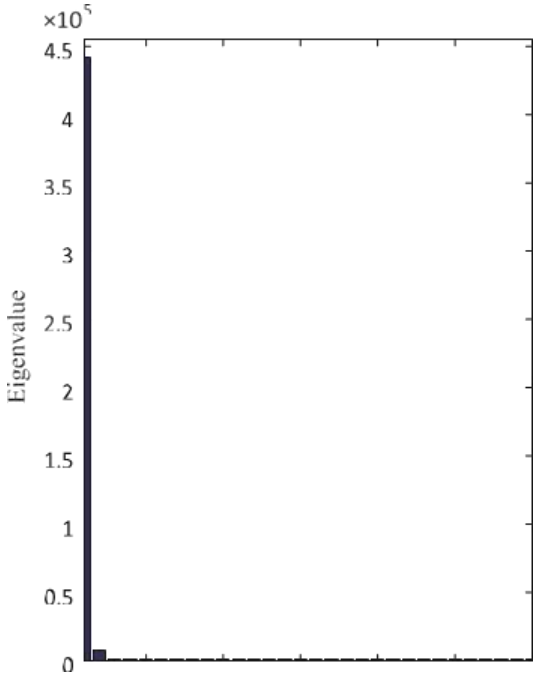

Figure 5. Corresponding eigenvalues of the WDBC data.

The goal is to determine whether reducing the feature dimension can maintain acceptable accuracy while lowering computational complexity.

For training processes, k-NN classifier, one-dimensional. Euclidean distance, $d = \sqrt{(x_{test} - x_{training})^2}$ between test and training samples . The results of k-NN classifier are obtained for the k values from 1 to 25, and then the performance measures at the best k value are stored. The model of ANN is selected as feedforward neural network with one hidden layer. The total number of neurons in the hidden layer is sequentially increased to find the maximum accuracy. Moreover, the activation function of the hidden layer of the network has been chosen as log-sigmoid transfer function. In order to train the network, gradient descent with momentum and adaptive learning rate backpropagation algorithm is used. RBFNN is also evaluated varying the spread value (σ). For SVM, linear, quadratic, and RBF kernels are used to explore which type of separating hyperplane is more suitable for breast cancer classification.

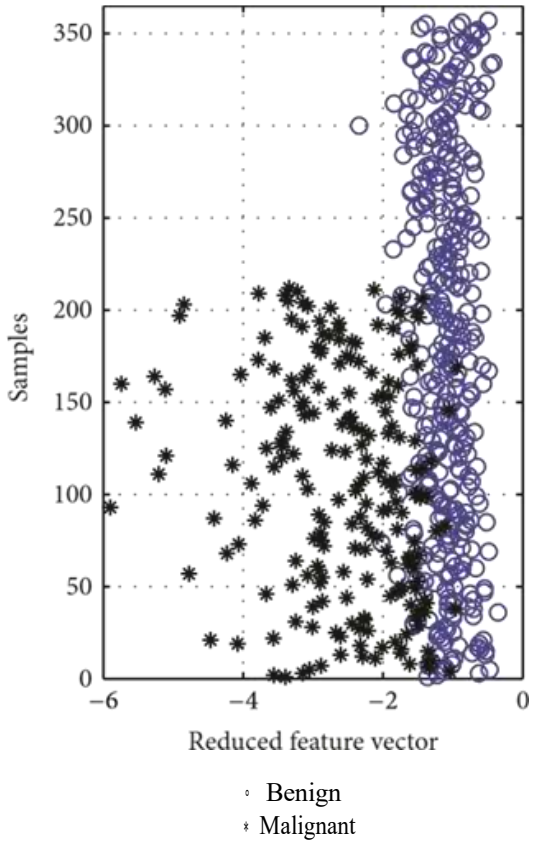

Figure 6. The distribution of computed IC (reduced feature vector).

IV. Results

One-dimensional feature vector of WDBC data reduced using ICA is used for training and testing the classifiers. The accuracy, sensitivity, and specificity of one dimensionality have been performed using 5/10 CV technique and 20% of data as test data. Also, the success of the breast cancer classification is generally evaluated on the basis of sensitivity value because the classifying of the malignant mass is more important than the benign mass.

The accuracy of the k-NN classifier has been computed for varying k values between 1 and 25. The comparison graph of the effect of ICA on accuracy of k-NN classifier is shown in Figure 7.

The maximum accuracy results when 20% test data with 30 features is 96.49% where k = 5. However, reduced one feature vector using ICA provides the accuracy of 92.98% where k = 5 and 20% test data is selected. Moreover, the accuracy of k-NN classifier is decreased from 93.15% (30 features) to 91.04% (1 feature by ICA) when 10-CV is used to test and train.

Accuracy graph of ANN has been plotted varying neuron numbers in the hidden layer for 10/5-CV and 20% test data. The accuracy graph of ANN classifier is given in Figure 8. ANN classifier has nearly perfect accuracy value of 99.12% (the number of neurons is four) when original 30 features and 20% test data are selected. The effect of ICA on reducing into one feature is changed accuracy value to 91.23% where neuron number is nine. In addition, the accuracy value is changed from 97.54% to 90.51% when 10-CV is used.

Spread value of RBFNN is adjusted between 0 and 60 to get maximum accuracy for 20% test data ratio and 10/5-CV. The accuracy graph of RBFN is shown in Figure 9.

Referring to the accuracy graph of RBFNN, maximum accuracy, 95.12%, is obtained where spread value is 48 for 20% test data. This value is decreased to 90.35% when reduced one-dimensional feature vector by ICA is used. However, when 10-CV is used, the effect of ICA increases the accuracy from 87.18% (with 30 features) to 90.49% (with 1 feature reduced by ICA).

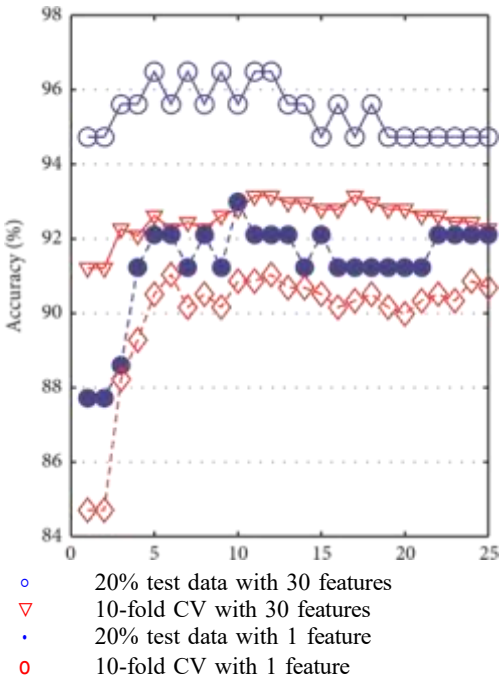

Figure 7. The graph of accuracy of k-NN classifier.

Accuracy evaluation of SVM has been computed for kernel functions including linear, polynomial, and RBF with kernel function parameters such as RBF sigma value for RBF kernel and polynomial degree for polynomial kernel. The accuracy graph of SVM classifier is presented in Figure 10.

where the axes of polynomial degree indicate linear kernel when its value equals one.

Generally, SVM classifier with linear kernel provides more accurate result than polynomial and RBF kernel. Its accuracy is 98.25% for 30 features and 90.35% for reduced 1 feature when 20% of data is used as test data. In contrast to polynomial kernel, effect of ICA increases the accuracy of SVM with RBF kernel from 89.47% (30 features) to 91.23% (1 feature). When 10-CV is used, the accuracy is decreased from 97.54% (30 features, linear kernel) and 95.25% (30 features, RBF kernel) to 90.33% and 90.86% (reduced1 feature by ICA).

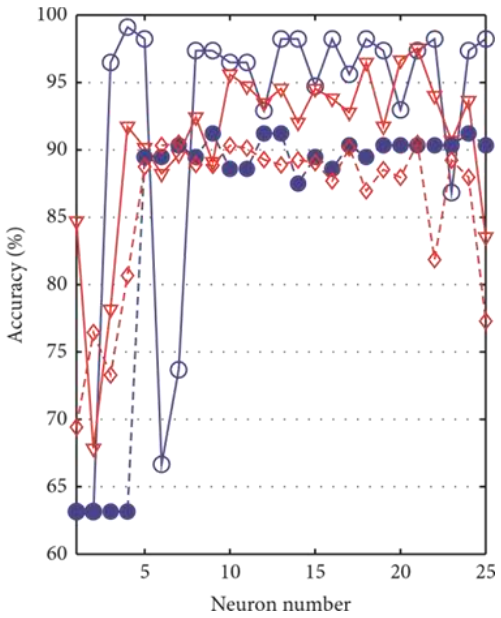

Figure 8. The accuracy graph of ANN.

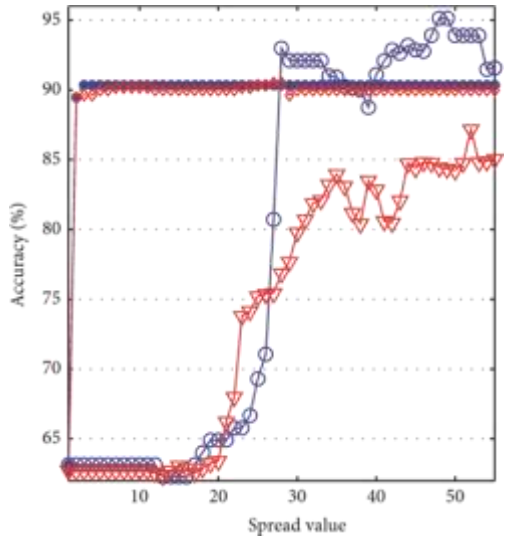

Figure 9. The accuracy graph of RBFNN.

k-NN, ANN, RBFNN, and SVM have been tested and trained to find out maximum accuracy adjusting their parameter. The performance measures such as accuracy, specificity, sensitivity, F-score, Youden's index, and discriminant power of the classifiers are compared to each other. The parameters of the classifiers which provide maximum accuracy are selected to be compared to the other classifiers. In addition to these performance measures, the ROC curve of three classifiers is plotted to enhance visuality of the comparison. 10-CV and one-dimensional feature vector reduced by ICA are used to compare the performances of classifiers. In input data of classifiers, the test data are compared to the original class label to find out TP, TN, FP, and FN values. These values for classifiers are given in the form of confusion matrix in Table 3.

RBFNN classification using 30 original features provides worse performance than reduced one-dimensional feature vector; refer to Table 3. The other classification used with 30 features has slightly higher true values when compared to classification with one feature reduced by ICA.

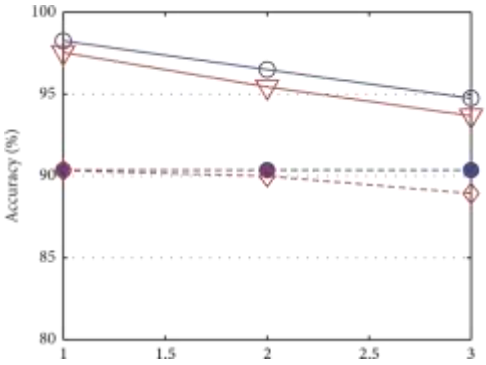

Fig.10(a)

The performance measures of k-NN, ANN, RBFNN, and SVM classifiers such as sensitivity, specificity, accuracy, Fscore, discriminant power (DP), and Youden's index are given in Table 4 to compare the effect of ICA on the classification.

Discriminant power evaluates how well a classifier distinguishes between positive and negative samples. DP of ANN and SVM with 30 original features differs from 3 which means good discriminant. When ICA is used to reduce to one dimensionality, DP falls to 2.769 (SVM) and 2.655 (ANN). In other words, discriminants turn to fair.

TABLE III
THE CONFUSION MATRICES OF THE CLASSIFIERS USING REDUCED ONE DIMENSIONALITY BY ICA (1F DENOTES ONE FEATURE AND 30F DENOTES ORIGINAL FEATURES).

k-NN classifier (*k*									
Actual value	Malignant 1F	30F							
Malignant	338 (TP)	346	19 (FN)	11	Malignant	346 (TP)	357	11 (FN)	0
Benign	32 (FP)	28	180 (TN)	184	Benign	43 (FP)	14	169 (TN)	198
RBFNN classifier (spread = 28)					**SVM classifier (σ = 1.3)**				
Actual value	Malignant 1F	30F	Benign 1F	30F	Actual value	Malignant 1F	30F	Benign 1F	30F
Malignant	345 (TP)	334	12 (FN)	23	Malignant	348 (TP)	343	14 (FN)	9
Benign	43 (FP)	138	169 (TN)	74	Benign	43 (FP)	13	169 (TN)	199

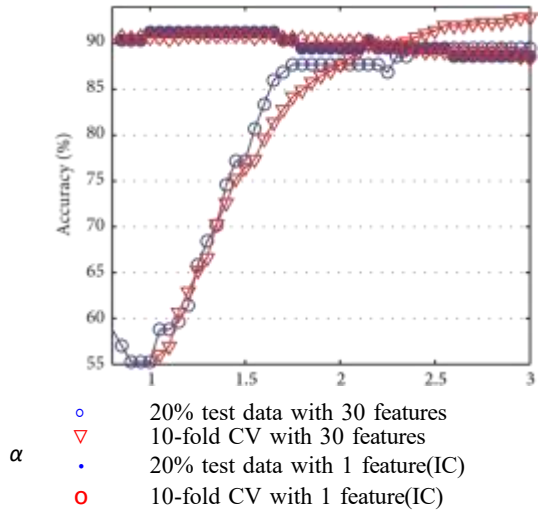

Fig.10(b)

Figure 10. The accuracy graphs of SVM classifiers.

A higher value of Youden's index shows better ability to avoid failure. k-NN results in the highest value of ouden's index; refer to Table 4. Youden's index is used to plot the ROC curve of a classifier. The true positive rate (sensitivity) is plotted in function of the false positive rate (1-Specificity) for cut-off points in a ROC curve. The ROC curve can be used to compute area under the ROC curve (AUC) and 95% confidence interval (CI). AUC equals 1 when all test data is assigned to true class labels. Higher AUC indicates that higher accuracy 95% CI is another indicator of the ROC curve which can be used to test whether a classifier can distinguish the classes. If its value is not 0.5, it means the classifier can distinguish the classes. The ROC curves of the k-NN, ANN, RBFNN, and SVM classifiers using one-dimensional feature vector reduced by ICA and 30 features are presented in Figure 11.

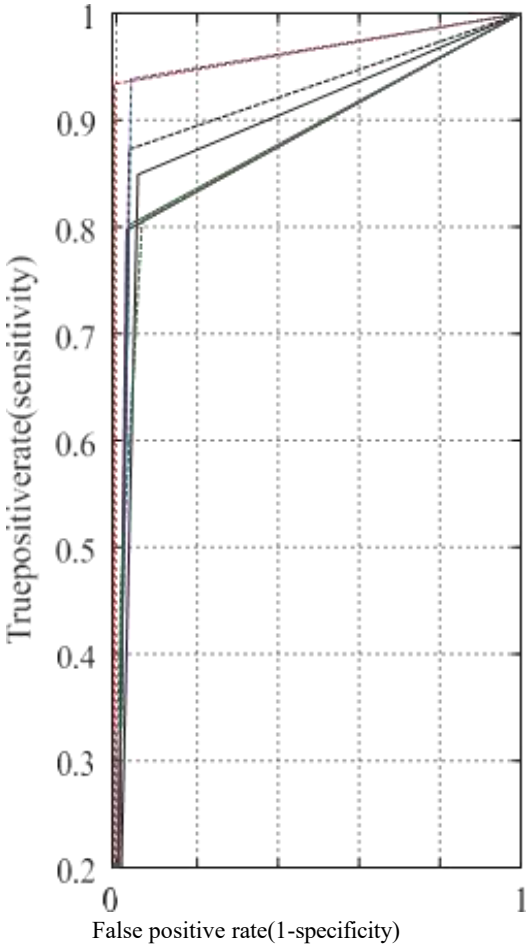

Figure 11. The ROC curves of k-NN, ANN, RBFNN, and SVM classifiers

TABLE IV
THE COMPARISON OF ICA ALGORITHM'S EFFECT ON THE CLASSIFIERS' PERFORMANCE MEASURES (SENSITIVITY, SPECIFICITY, ACCURACY, AND F-SCORE IN %).

	k-NN		ANN		RBFNN		SVM (RBF K.)	
Measures	1F	30F	1F	30F	1F	30F	1F	30F
F -score	92.98	94.65	92.76	98.07	92.61	80.57	93.04	96.21
DP	2.539	2.912	2.655	InF	2.606	1.131	2.769	3.267
Y	0.795	0.839	0.766	0.934	0.763	0.284	0.772	0.899
Accuracy	91.03	93.14	90.5	97.53	90.49	87.17	90.86	95.25
Specificity	84.9	87.26	79.71	93.39	79.71	34.9	79.71	93.86
Sensitivity	94.67	96.63	96.91	100	96.63	93.55	97.47	96.07

TABLE V
CRITERION VALUES OF THE ROC CURVES OF K-NN, ANN, RBFNN, AND SVM.

Criterion	**k-!**		**ANN**		**RBFNN**		**SVM**	
	1F	30F	1F	30F	1F	30F	1F	30F
AUC	0.880	0.911	0.879	0.956	0.881	0.877	0.879	0.945
95% CI	0.86–0.92	0.89–0.94		0.94–0.98	0.85–0.91	0.85–0.91		0.92–0.97

k-NN (1 IC) ANN (1 IC) RBFNN (1 IC) SVM RBF (1 IC)
k-NN (30 features) ANN (30 features) RBFNN (30 features) SVM RBF (30 features)

The criterion values of the ROC curves of classifiers are given in Table 5. AUC of the ANN (0.966) and SVM (0.949) results in higher value when 30 original features are used. However, when classification with 1 feature reduced by ICA is evaluated, k-NN (0.897) and SVM (0.885) result in higher AUC. It means k-NN and SVM classifiers using reduced one feature distinguish samples more correctly.

Table 5 shows that the accuracy of the k-NN (91.03%) is better than the accuracy of ANN, RBFNN, and SVM (90.50%,90.49%, and 90.86%). Generally, one feature reduced by ICA.

TABLE VI
CPU TIME FOR CLASSIFICATION.

Classifying	Partitioning	IC (seconds)	30 features (seconds)
k-NN	20% 10-CV	8.02 13.52	8.31 14.77
ANN	20% 10-CV	11.12 76.72	13.9 118.21
RBFNN	20% 10-CV	14.9 90.49	20.03 129.84
SVM (poly)	20% 10-CV	7.17 7.47	7.28 9.13
SVM (RBF)	20% 10-CV	9.02 10.72	43.30 19.05

decreases the accuracy of k-NN, ANN, and SVM. However, it increases the accuracy of RBFNN.

The afore mentioned classification methods are analyzed in terms of computing time given in Table 6 to compare the computational complexities to the classifications with the original 30 features.

The proposed methods have lower computing time when compared to classification of the original dataset. In case of neural network classifications with 30 features, network constructions consume highly more time than classification with one IC. The measured durations of 13.9 and 20.03 seconds are decreased to 11.12 and 14.9 seconds when ANN and RBFN with 20% partitioning are deployed. Particularly, the effect of using IC as feature on complexity is existed when 10-CV is selected. The consumed time of the ANN and RBFNN is decreased from 118.21 and 129.84 seconds to 76.72 and 90.49 seconds, respectively. In addition, ICA decreases computational times of the SVM and k-NN classifications, but the rates are less than the neural networks.

V. DISCUSSION

Sensitivity/specificity indicates the proportion of actual positives/negatives which are correctly identified. While use of one-dimensional feature vector reduced by ICA decreases accuracy slightly, it increases sensitivity values of SVM and RBFNN classifiers. The maximum sensitivity measure belongs to SVM with RBF kernel when one feature is used. The graph of the effect of ICA on sensitivity measures of classifiers is shown in Figure 12.

Sensitivity refers successfully identified malignant samples in cancer classification. Thus, higher sensitivity means higher diagnostic capability of malignant tumors and it can be used to help physicians to diagnose cancerous mass more correctly. The accuracy and sensitivity measures of previous classification studies and this study on WDBC dataset are given in Table 7 to compare the effect of feature reduction using ICA. It should be noted that the studies on WDBC differ from studies on WBC dataset which consists of 699 instances with 10 attributes.

TABLE VII
COMPARISON OF THE METHODS AND ACCURACY OF PREVIOUS STUDIES AND THIS STUDY.

Author	Method	Feature number	Accuracy	Sensitivity
This study	10-CV, k-NN	1 feature reduced by ICA	91.03%	94.67%
	40% test, k-NN		92.56%	94.02%
	10-CV, ANN		90.50%	96.91%
	40% test, ANN		90.89%	97.00%
	10-CV, RBFNN		90.49%	96.63%
	40% test, RBFNN		89.98%	96.01%
	10-CV, SVM (linear)		90.33%	96.35%
	40% test, SVM (linear)		90.01%	95.00%
	10-CV, SVM (quadratic)		89.98%	95.24%
	40% test, SVM (quadratic)		91.01%	96.42%
	10-CV, SVM (RBF)		90.86%	97.47%
	40% test, SVM (RBF)		91.03%	97.56%

The experimental analysis highlights the impact of dimensionality reduction on classification performance.

While using all 30 features generally provides slightly higher accuracy, reducing the dataset to a single independent component significantly decreases computational complexity and training time.

This reduction is especially useful in real-time diagnostic systems where rapid decision making is required. Furthermore, some classifiers such as RBFNN even benefit from the reduced feature representation, suggesting that eliminating redundant information can improve generalization.

Higher number of features used to classify breast cancer as benign and malignant results in slightly higher accuracy. Feature reduction into one using ICA decreases the accuracy of k-NN, ANN, and SVM slightly. However, it increases the accuracy of RBFNN and the sensitivity values of SVM and RBFNN.

Referring to Table 7, the sensitivity measures of the classifiers used with one-dimensional feature vector reduced by ICA in this study perform better than the other studies. However, accuracy rates of the proposed classifications (90.53% ± 0.34) are lower than the previous methods (94.93% ± 2.07). The study of WDBC data creators [39] set has the highest accuracy (97.50%) using multi surface method tree (MSMT) with 3 selected features. Similarly, hybrid methods are

more successful than the others. Breast cancer classifications using probabilistic neural network (PNN) with hybrid feature reduction using discrete wavelet transform (DWT) and ICA [40] or classification using SVM with 6- dimensional feature space obtained by k-means algorithm [41] have accuracy rates of 96.31% and 97.38% for 10- CV. Particularly, SVM based studies [36, 38] with 30 features have near scores to our one dimensional results.

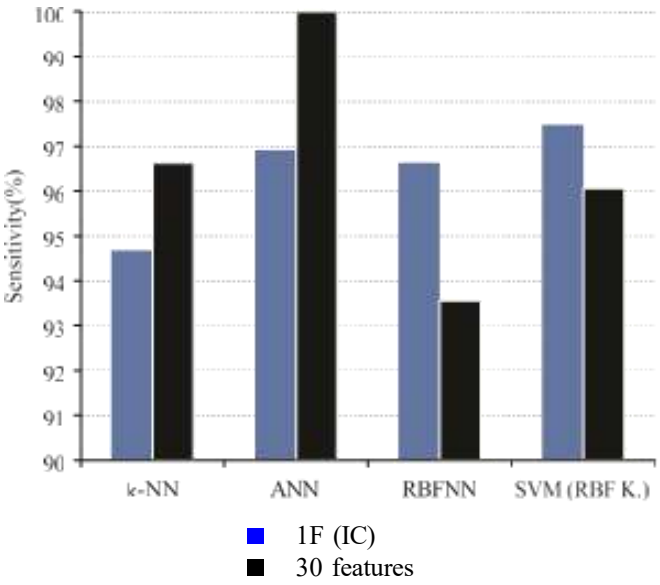

Figure 12. Sensitivity measures of the classifiers

VI. CONCLUSION

This study examined the use of Independent Component Analysis as a feature reduction technique for breast cancer classification. The WDBC dataset was reduced from thirty diagnostic features to a single independent component, which was then used to train multiple machine learning models. The results demonstrate that ICA-based dimensionality reduction can significantly reduce computational cost while maintaining competitive classification accuracy. Among the classifiers tested, Artificial Neural Networks and Support Vector Machines achieved the highest performance. These findings indicate that ICA can be effectively integrated into computer-aided diagnostic systems for breast cancer detection. Future research may explore combining ICA with advanced deep learning models or hybrid feature selection techniques to further improve classification accuracy.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their respected guide Dr. T. seshu chakravarthy for the continuous support, valuable suggestions, and insightful guidance throughout the course of this work. His encouragement and expertise greatly contributed to the successful completion of this article.

We are also thankful to the Project Coordinator, Dr. G. Sanjay Gandhi for providing timely assistance, constructive feedback, and for ensuring smooth progress during all phases of the project.

Our heartfelt thanks go to the Head of the Department, Dr. V. Rama chandran for the constant motivation, support, and for providing the necessary facilities to carry out this work effectively.

We extend our deep appreciation to the Principal, Dr. Y. Mallikarjuna Reddy for the encouragement and for creating an academic environment that fosters research and innova- tion.

Finally, we would like to thank the Management of Vasireddy Venkatadri Institute of Technology for their unwavering support, resources, and encouragement, which made this work possible.

REFERENCES

[1] I. Christoyianni, E. Dermatas, and G. Kokkinakis, "Fast detection of masses in computer-aided mammography," *IEEE Signal Processing Magazine*, vol. 17, no. 1, pp. 54–64, 2000.

[2] N. Salim, *Medical Diagnosis Using Neural Network*, Faculty of Information Technology University, 2013. [Online]. Available: http://www.generation5.org/content/2004/MedicalDiagnosis.asp

[3] A. Tartar, N. Kilic, and A. Akan, "Classification of pulmonary nodules by using hybrid features," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 148363, 11 pages, 2013.

[4] N. Kilic, O. N. Ucan, and O. Osman, "Colonic polyp detection in CT colonography with fuzzy rule based 3D template matching," *Journal of Medical Systems*, vol. 33, no. 1, pp. 9–18, 2009.

[5] A. Mert, N. Kilic, and A. Akan, "Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats," *Neural Computing and Applications*, vol. 24, no. 2, pp. 317–326, 2014.