

My thesis short Summary

About Data Set

- An open-source dataset used

Total two classes Malignant and Benin.

In data augmentation file	0 = Benin
	1 = Malignant
Before data augmentation	2 = Benin.
	4 = Malignant

• Before Data Augmentation the two data sets comparison

WBC = Total entries 699 (with 10 feature columns and 1 predictive class column with 16 missing values in the attribute number 7 “bare nuclei”)

WDBC = It contains 32 attributes and 569 instances. This dataset does not have any missing values.

• After Data Augmentation on WBC data Set

Total entries increasing up to, after applying Data Augmentation= 6831

0 = Benin = 4440

1 = Malignant = 2391

Before Data Augmentation (Applying different Machine learning Models)

With Reference results

Performance Evaluation Metrics for (WBC)						
Implemented Technique	Accuracy %	Precision %	Recall %	F1-score %	Training Time (secs)	Prediction Time (secs)
RF+FA	97.8	96.7	100	98.3	0.53	0.06
RF+PCA	95	94.6	97.8	96.2	0.50	0.050
RF+Chi2	95	95.79	96.7	96.24	0.67	0.040
RF+RFE	97.8	97.8	98.9	98.3	0.48	0.044
RF+PCA+RFE	98.6	98.8	98.8	98.8	0.37	0.018
Performance Evaluation Metrics for (WDBC)						
RF+FA	93.86	96.92	92.6	94.7	0.17	0.015
RF+PCA	93.7	97.33	93.6	95.4	0.26	0.02
RF+Chi2	96.49	98.67	96.1	97.37	0.22	0.016
RF+RFE	96.49	98.59	95.89	97.22	0.23	0.021
RF+PCA+RFE	98.2	100	97.3	98.6	0.21	0.011

AFTER Applying Data Augmentation only on (WBC) data set.

Data Augmentation gives astonishing results on (WBC) data sets, when we apply KNN and SVM models. The accuracy increases to 99.94 and precision is 100 percent .

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN	99.94	100	100	100
SVM	99.94	100	100	100
Decision Tree	97.77	98/97	98/97	98/97
Naive Bayes	96.19	98/93	95/97	97/95
Logistic Regression	96.72	97/95	97/95	97/95

Conclusion:

Data augmentation improves the accuracy of various machine learning models in detecting breast cancer, thereby assisting physicians in making more accurate diagnoses.

Before Data Augmentation on WBC data set highest accuracy by using different machine learning models is = 98.6

After Data Augmentation on WBC data set highest accuracy by using different machine learning models is = 99.94
