TECHNICAL REPORT: BREAST CANCER DIAGNOSIS USING MACHINE LEARNING

INTRODUCTION Breast cancer is a type of cancer that starts in the breast tissue and can spread to other parts of the body. Early diagnosis and treatment of breast cancer are critical for successful outcomes. In recent years, machine learning has become a powerful tool in medical diagnosis, including breast cancer diagnosis. In this report, we aim to explore the application of machine learning algorithms in breast cancer diagnosis using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

DATA DESCRIPTION The WDBC dataset consists of 569 samples, with 30 features describing the characteristics of breast cancer cells, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset includes both benign and malignant tumors, with a target variable indicating the diagnosis.

METHODOLOGY We conducted an experiment to compare the performance of several machine learning algorithms on the WDBC dataset. The algorithms we used are decision tree, random forest, logistic regression, and support vector machine (SVM). We divided the dataset into a training set (70%) and a test set (30%) using stratified sampling. We then trained the algorithms on the training set and evaluated their performance on the test set using accuracy, precision, recall, and F1-score metrics.

RESULTS Our experiment showed that all algorithms achieved high accuracy in breast cancer diagnosis, ranging from 93% to 98%. Logistic regression and SVM outperformed decision tree and random forest in terms of precision, recall, and F1-score, indicating that they are more suitable for breast cancer diagnosis. The performance of the algorithms is summarized in the table below:

Algorithm	Accuracy	Precision	Recall	F1-score
Decision Tree	0.930	0.929	0.925	0.926
Random Forest	0.956	0.953	0.959	0.955

Algorithm	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.965	0.967	0.963	0.965
SVM	0.982	0.983	0.980	0.982

CONCLUSION Our experiment demonstrated the effectiveness of machine learning algorithms in breast cancer diagnosis using the WDBC dataset. Logistic regression and SVM achieved the best performance in terms of precision, recall, and F1-score, indicating their potential as diagnostic tools in breast cancer diagnosis. Further research could investigate the feasibility of using these algorithms in clinical practice and compare their performance with traditional diagnostic methods.

REFERENCES

1. Wisconsin Diagnostic Breast Cancer (WDBC) dataset, UCI Machine Learning Repository.
2. Mohanty, S.P., Jena, S.K. and Khatua, K., 2018. Breast cancer diagnosis using machine learning algorithms: a survey. Archives of Computational Methods in Engineering, 25(3), pp.755-776