BREAST CANCER DETECTION

ABSTRACT

Breast cancer is a leading cause of mortality among women globally. The World Health Organization estimates that over 600,000 women (about half the population of Hawaii) died from breast cancer in 2020 alone. Early detection significantly increases the chances of successful treatment and survival. However, traditional methods of detection, such as mammography and biopsy, can be invasive, costly, and sometimes inaccurate. This project aims to develop a machine learning model for the early detection of breast cancer using patient feature values, offering a non-invasive, cost-effective, and potentially more accurate alternative.

The proposed model uses patient data features such as age, tumor size, lymph node status, and other clinical measurements as input. These features are often used by medical professionals to assess a patient's risk of breast cancer. The model is trained on a large dataset of anonymized patient records, with labels indicating the presence or absence of breast cancer. This dataset is split into a training set for model development and a test set for model evaluation. The use of a separate test set ensures that the model's performance is evaluated on unseen data, providing a realistic measure of its predictive power.

Data preprocessing is a crucial step in our project. Real-world data is often messy and incomplete. We employ various techniques to handle missing values and outliers. Missing values are imputed using statistical methods, while outliers are identified and handled to ensure they do not skew the model's performance. The data is also normalized to ensure that all features contribute equally to the model. This step is

vital for the performance of the machine learning algorithm, as it ensures that the model is not unduly influenced by features with larger scales.

The preprocessed features are then fed into a machine learning algorithm. We experiment with several algorithms, such as Support Vector Machines, Decision Trees, Logistic Regression, K Nearest Neighbours, Random Forest Classifier, and Naïve bayes Classifier to identify the one that provides the highest accuracy. Each algorithm has its strengths and weaknesses, and the choice of algorithm can significantly impact the model's performance. We also employ techniques such as cross-validation to ensure that our model is robust and not overfitting to the training data.

Preliminary results show promising accuracy rates in detecting breast cancer, outperforming traditional diagnostic methods. This project has the potential to enhance breast cancer screening and diagnosis, making it more accurate, efficient, and accessible. By reducing the need for invasive procedures, our model could also improve the patient experience.

Our future work will focus on improving the model's performance through hyperparameter tuning and feature selection. Hyperparameters are parameters that are not learned from the data but are set before the training process. Tuning these can significantly improve the model's performance. Feature selection involves identifying the most informative features for prediction, which can simplify the model and improve its interpretability.

We also focus on employing advanced image processing techniques to enhance the quality of the images and extract meaningful features. These features are then fed into a deep learning algorithm, which is trained to classify the images as benign or malignant.

We also plan to integrate the model into existing healthcare systems to provide a practical tool for early breast cancer detection. This integration will involve working closely with healthcare professionals to ensure that the model meets their needs and can be seamlessly incorporated into their workflow

This project underscores the potential of machine learning in healthcare, particularly in disease detection and diagnosis. By leveraging machine learning, we can make significant strides in the ongoing fight against breast cancer, ultimately saving lives and improving patient outcomes. We believe that our project will contribute significantly to this important field and look forward to sharing our results with the scientific community.