**Auxiliary Diagnosis of Breast Cancer Based on Machine Learning and Hybrid Strategy**

**Alternative Title:**

Breast Cancer Prediction using Supervised Machine Learning Algorithms

**Aim:**

The primary aim of this study is to develop a robust and accurate auxiliary diagnostic system for breast cancer by integrating machine learning techniques with a hybrid strategy.

**Abstract:**

Breast cancer has replaced lung cancer as the number one cancer among women worldwide. In this paper, we take breast cancer as the research object, and pioneer a hybrid strategy to process the data, and combine the machine learning method to build a more accurate and efficient breast cancer auxiliary diagnosis model. First, the combined sampling method SMOTE-ENN is used to solve the problem of sample imbalance, and the data are standardized to make the data have better separability. Then, the features of the dataset are initially screened using the mutual information method, and further secondary feature selection is performed using the recursive feature elimination method based on the Logistic Regression algorithm. Thus, the feature dimensionality of the dataset is reduced and the generalization ability of the model is improved. Finally, four different machine learning models are used for classification prediction, the best combination of parameters for each model is found, and the final results of each model are derived. The experiments are conducted using the Wisconsin Diagnostic Breast Cancer dataset (WDBC), and the results of the study find that after the data are processed by the hybrid strategy, the best prediction results are obtained using the Random Forest model with high accuracy, which is better than the previous research methods.

**Introduction:**

 According to the Cancer Statistics, 2023 statistical estimates, breast cancer, lung cancer, and CRC account for 52% of all new diagnoses, with breast cancer alone accounting for 31% of female cancers. Breast cancer, as one of the common malignant tumors in women, has become a focus of public health attention around the world. Its early diagnosis is important for the success of treatment and patient survival. With the rapid development of machine learning and other technologies, more and more research has been devoted to applying these advanced technologies to the diagnosis and assisted decision making of breast cancer.

Machine learning, as an important artificial intelligence technology, has the ability to extract features, discover patterns and build predictive models from a large amount of medical data. It can not only assist doctors in identifying high-risk groups in early screening, but also be used for accurate diagnosis and personalized treatment plan development.

**Existing System:**

 Breast cancer is the leading cause of death in the developed world and second in the developing world, killing almost 8 million people a year. According to the characteristics of the WDBC dataset, the Z-score standardization method is selected to process the dataset. First, the top 20 features with scores are screened by the mutual information method, and then 13 features are screened by the recursive feature elimination method based on the XGBoost algorithm to obtain the final feature subset. So high accuracy need to predict the breast cancer. In existing system, Random forest algorithm used to detect the breast cancer. It gives little minimum amount of accuracy compare to proposed system. So we move on to the proposed system

**Proposed System:**

The incidence and mortality rate of breast cancer is increasing year by year and has become the number one cancer among women worldwide. In the medical field, the diagnosis and treatment of breast cancer relies heavily on early detection and treatment, and the earlier the treatment, the better the clinical outcome for patients**.** Firstly, in the preprocessing sections, some categorical values are found, factorize is used to encode the categorical values into numerical. A combined SMOTE sampling method is used to solve the problem of sample imbalance. Then, the features of the dataset are screened using the mutual information method, and further the recursive feature elimination method based on the Logistic Regression is used to derive the best feature subset. Finally, four different machine learning models Random Forest, SVM, KNN, and Gradient Boost, are used for classification and prediction. The experimental results find that the best prediction results are obtained using the RF model, with the high accuracy. This is better than the previous research methods

**Module Description:**

* Dataset Collection
* Data Pre-Processing
* Algorithm Implementation
* Detection

**Dataset Collection:**

The data used in this paper are from the Diagnostic Breast Cancer Dataset (WDBC) provided by the Wisconsin Center for Clinical Sciences. In the dataset ID is the patient number, Diagnosis is the sample label, and 30 features are presented. Our Breast cancer project WDBC dataset are collected from kaggle.com .Breast cancer data is pre-processed after collection of various records. The dataset contains a more number of patient records. It has some medical parameters.

**Data Pre-Processing:**

The data used in this system are from the Diagnostic Breast Cancer Dataset (WDBC) provided by the Wisconsin Center for Clinical Sciences, publicly available repositories: Kaggle and Github. There is some categorical values are found, factorize is used to encode the categorical values into numerical. There is uneven distribution of categories, and this problem of sample imbalance will affect the training effect of the model. A combined SMOTE sampling method is used to solve the problem of sample imbalance.

**aLGORITHM IMPLEMENTATION:**

The Machine Learning Algorithms used to produce the best results. We are using four different algorithms, SVM, Random Forest, KNN and Gradient Boost to predict the Breast cancer. Apply all four algorithmic modules on this data and find the appropriate one based on their accuracy. The best performing models are identified from the above results based on their low rate of error.

* K Nearest Neighbor
* Support Vector Machine
* Random Forest
* Gradient Boost

**Detection:**

Preprocessed data are trained by ML algorithm and input given by the user from web application, goes to the trained dataset. After prediction the predict value Shown as an output on web application.

**Software Requirements:**

* Operating System : Windows 10 (64 bit)
* Software : Python and Anaconda
* Tools : Jupyter Note Book and Spyder IDE

**Hardware Requirements:**

* Hard Disk : 500GB and Above
* RAM : 4GB and Above
* Processor : I3 and Above

**Architecture Diagram:**

Dataset Collection

Dataset

Breast Cancer Detection

No Breast Cancer

Breast Cancer

Machine Learning

User Input

Preprocessing