Forecast Model of Breast Cancer Diagnosis Based on RF- AdaBoost

Early and accurate diagnosis of breast cancer patients is particularly crucial. This paper considered the Wisconsin female breast cancer tumor dataset as the research object. An integration of random forest and Adaboost algorithms was proposed. This is a classification prediction model that can give a diagnosis result of benign or malignant.

Finally, the result was compared with SVM, logistic regression, k-nearest neighbour, decision tree algorithms. The test results showed that the ensemble model's prediction accuracy increased by 4.3% on average compared to the single algorithm models, with the highest increase up to 9.8%

Confusion matrix was used to evaluate the performance of the classifier. Accuracy, precision, recall and F-measure were the 4 evaluation indicators used to evaluate the performance of models
The experimental results showed that the integrated classification model's accuracy reached 98.6%, with benign classification recall rate reaching 100% and malignant classification recall rate reaching 96%. The F1value of benign and malignant results were 99% and 98%.

GridSearchCV was used to automatically search for hyperparameters, avoiding the occurrence of overfitting and underfitting. Many experiments have shown that this integrated model has performed excently in breast cancer diagnosis and prediction, surpassing the common single algorithm models. There's still room for accuracy improvement. The improved RF model based on Grid Search still has broader research prospects in terms of training time, interpretability, etc. for the doctors to provide more accurate and faster diagnosis results

Research on the Detection Method of Breast Cancer Deep Convolutional Neural Network Based on Computer Aid

Traditional breast cancer image classification methods require manual extraction of features from medical images, which not only require professional medical knowledge, but also have problems such as time-consuming and labor intensive and difficulty in extracting high-quality features. Therefore, the paper proposes a computer-based feature fusion Convolutional neural network breast cancer image classification and detection method. The paper pre-trains two convolutional neural networks with different structures, and then uses the convolutional neural network to automatically extract the characteristics of features, fuse the features extracted from the two structures, and finally use the classifier to classify the fused features. The experimental results show that the accuracy of this method in the classification of breast cancer image datasets is 89%, and the classification accuracy of breast cancer images is significantly improved compared with traditional methods

The study is mainly based on the design of deep learning algorithm for tumor benign and malignant classification based on three-dimensional breast ultrasound data, and focuses on the impact of adjusting the convolutional neural network structure to integrate multiple information on classification performance.In addition to image information, the research also collected the description information of each lesion, and designed a novel network structure that can integrate image information and text information, so that only one network model can be trained to complete the classification and discrimination of all information.
71 benign lesions and 74 malignant lesions were collected to train the neural network.
ReLu is used as the activation function.
Among all the traditional methods, the learning ability of fusion feature of CNN performed best

But if the training dataset is unbalanced, the prediction effect is very poor

ESTROGEN RECEPTOR STATUS PREDICTION FOR BREAST CANCER USING ARTIFICIAL NEURAL NETWORK

The status of estrogen receptor (ER) has been profoundly associated with breast cancer. Numerous studies have been conducted to identify informative genes that are associated with ER status. However, the integrity of the reported genes is still inconclusive as the results are derived from a small cohort of breast cancer patients (< 200 samples). In this paper, the gene signatures from a cohort of 278 breast cancer samples were studied and labelled in ER positive and ER negative classes, using artificial neural network (ANN). The data are divided into 3 sample subsets, i.e. training, test and validation sets. The sample partition ratio is 60:20:20. A 3-layered ANN with backpropagation learning was constructed for predicting the training samples. In the network training process, a stepwise approach is adopted, in which the number of input nodes (i.e. genes) will be increased by one each time the network is trained.
The validation set is then used to further examine the significance of the identified genes. The whole data partition, training and validation processes were repeated for 50 times.

The model showed its efficacy for selecting significant genes compared to the previous study. The result also showed that the highly ranked genes have been previously reported in association to breast cancer development.

This model showed the ability to identify the most significant gene subset from the breast gene microarray data. The model was able to select 3 most important predictor genes that were previously associated with breast cancer

But also, low sensitivity value of 58.78% in the ER negative class was obtained. This was due to the lack of standard threshold used by immunohistochemistry to label ER status on the samples.

Predicting breast cancer risk using personal health data and machine learning models

This paper talks about different machine learning models which used highly accessible personal health data to predict five-year breast cancer risk better than the Gail model (BCRAT) to improve early detection and prevention of breast cancer. The six models used were logistic regression, Gaussian naïve Bayes, decision tree, linear discriminant analysis, support vector machine, and feed-forward artificial neural network. In order to make a comparison with BCRAT, these models were initially trained only with five of the seven traditional BCRAT inputs (age, age at menarche, age at first live birth, number of first-degree relatives who have had breast cancer, and race / ethnicity) available from the PLCO dataset. When trained with 80% data and tested with the rest of the 20%, it was found that none of the machine learning models were significantly stronger than the BCRAT. Later, the models were trained with broader set of predictors which included age at menopause, an indicator of current hormone usage, number of years of hormone usage, BMI, pack years of cigarettes smoked, years of birth control usage, number of live births, and an indicator of personal prior history of cancer. The logistic regression, linear discriminant analysis, and neural network models with the broader set of inputs effectively predicted five-year breast cancer risk than the BCRAT model. It was observed that the linear discriminant analysis had by far the highest sensitivity which undercut the potential of other models. Conclusion is that these models could serve as the bases of new cost-effective and non-invasive tools to inform and prompt screening and immediate and long-term preventative actions with the potential to increase early detection and reduce the incidence of breast cancer.

Breast cancer risk prediction model based on C5.0 algorithm for postmenopausal women

Breast cancer is one of the most common in the elderly women. So 1031 postmenopausal women (≥43 years old) with breast cancer were considered. Based on characteristics of their breast cancer data, a breast cancer risk prediction model based on C5.0 algorithm was constructed and the model was optimized. The result showed that C5.0 based model provided better performance in constructing breast cancer risk prediction model than machine learning methods such as neural network and support vector machine. Also, the experiment showed that C5.0 model was less superior to C5.0 model with adaptive boosting (uses adaptive enhancement algorithm) which in turn was less superior to C5.0 model with cost matrix. This was because the cost matrix model predicted the risk of breast cancer by strongly correlating with post-menopausal hormones, age, age of menopause, history of benign breast disease and age of the first childbearing. By doing so, it was found that the sensitivity of the model was around 85.71% which was NILL in some of the other models. Since the sensitivity was high, the model was able to predict the chance of breast cancer in women better which were not predicted at all in other models. The accuracy of the C5.0 model with cost matrix in predicting women with breast cancer was pretty high but slightly less when compared to other models. Thus, this research is a practical application of data mining in the medical field and has certain reference value for the clinical diagnosis of breast cancer.

Predicting Breast Cancer Risk Using Subset of Genes

Accurate prediction of breast cancer risk is challenging. The accuracy of predicting breast cancer using microarray technology is better when compared to other models as the classification happens at gene level. So, a prediction model based on ensemble classifiers technique is used here. Formerly models based on ensemble classifier techniques Deep Neural Network and Support Vector Machine (DNN+SVM), Deep Neural Network and Recursive Feature Elimination (DNN+RF), Deep Neural Network and AdaBoost (DNN+AdaBoost) and Support Vector Machine and Recursive Feature Elimination (SVM+RF) are taken. On these models, predication is done using feature selection algorithms Correlation-based filter method (FCBF), Regularized Random Forest algorithm (RRF), Decision Tree algorithm(DT-FWD) and Symmetrical Uncertainty Criteria(SUC) function. Results are compared against a combination of models and feature selections. Results show that prediction based on genes with feature selection FCBF gave best results for their DNN+SVM model. It also predicted which types of genes are more prone to breast cancer.

Breast Cancer Prediction Based On Backpropagation Algorithm

This paper outlines a system that can classify “Breast Cancer Disease” tumors using a neural network with Feed-forward Backpropagation Algorithm to classify the tumor from a symptom that signifies the breast cancer disease. Breast cancer tumor database used for this purpose is from the University of Wisconsin (UCI) Machine Learning Repository. There are 699 records in this database with nine attributes graded on an interval scale from a normal state of 1–10. 241 (65.5%) records are malignant and 458 (34.5%) records are benign, represented by numbers between -1 and 1. The input layer consists of nine nodes that represent the nine attributes. The hidden layer consisted of 7 nodes. The output layer consisted of a single node representing diagnostic outcome; 0.0 for malignant and 1.0 for benign. Each of the iteration in backpropagation constitutes two sweeps: forward activation to produce a solution, and a backward propagation of the computer error to modify the weights repeatedly until the ANN solution agrees with the desired value within a prespecified tolerance. The training algorithm is divided into three main parts which are feed forward, error calculation and updating the weight. During the execution of the feed forward process which sends in the input signal, an application algorithm for the neural network is used. This algorithm is applied to the real system to give the output and to classify the breast cancer tumor. A total of 7 hidden layers achieves the highest accuracy. The mean square error for this neural network model is small and towards 0.001. The feed-forward backpropagation algorithm is the best classifier to predict breast cancer disease with an accuracy of 96.63%.

Advantages:
Best accuracy(96.63%) as compared to what would’ve been if the SVM classifier was used(96.19%).
Best accuracy(96.63%) as compared to what would’ve been if the Decision Tree classifier was used(92.38%).

Disadvantages:
There still exists a higher possible value of accuracy that can be attained.

Breast Cancer Disease Prediction With Recurrent Neural Networks (RNN)

This study explores deep learning techniques in conjunction with Recurrent Neural Networks (RNN) to predict the occurrence of breast cancer. To assess the efficiency of the proposed method, breast cancer data belonging to the UC Irvine repository were used. The dataset used consisted of 561 instances and 31 attributes, out of which 30 attributes are considered as input attributes and the 1st attribute is considered the target class. As a part of pre-processing, standardization of the data was performed. RNNs were preferred because RNN nodes are more dominant than other models for predicting the outcomes since these models use backpropagation. This method has one input layer consisting of 30 input nodes, three hidden layers consisting of 64, 128, and 256 nodes, and one output layer consisting of one node output of either 0 or 1. ReLU activation function is used in the hidden layer and dropout of about 0.25 is used. The dataset includes data from 569 instances with 31 characteristics. Based on experimental results, the RNN model exhibited the 97% of f1 score and an accuracy of 97.37%.

No adv or dis

Breast Cancer Classification Using Deep Learning ICECOS

The classification based on recurrence and no-recurrence events uses datasets from the University of Medicine Center, Institute Of Oncology, Ljublijana,Yugoslavia. Out of the total 286 datasets, there were 201 No-Recurrences-Events classes, 85 Recurrences-events classes and 10 attributes. The algorithm used for breast cancer classification is the Multilayer Perceptron algorithm with the accuracy level of 96.5%. The dataset still had incomplete or missing values denoted by the "?". Data refinement is performed by filling in missing values on the dataset using the average attribute values of all the samples residing in the same class. General multilayer perceptron (MLP) with back propagation learning rule is used here. Each neuron in the hidden layer and output layer receives output vectors from the previous layer to evaluate the weighted sum and to achieve the output vectors by the activation functions. The dataset is divided into 10 partitions randomly. Then 10 experiments were conducted, each experiment using the 10th partition data for data testing the rest is used as training data. Hence, the number of input neurons is defined by the number of markers and the number of hidden neurons is optimized for each marker combination. The accuracy of classifying recurrent and no-recurrent is 96.5%.

Disadvantages:
There is still further scope for improvement of the accuracy.
Using DBN-NN gives an accuracy of 99.68%.
Advantages:
Much better accuracy than what was observed when Bayesian Linear Discriminant Analysis was used. The results show an average classification accuracy of about 83.45%.

A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction

Materials and Methods: This retrospective study included 88 994 consecutive screening mammograms in 39 571 women.

Cancer outcomes were obtained through linkage to a regional tumor registry. Comparisons were made to an established breast cancer risk model that included breast density. We excluded 21328 women because they lacked sufficient follow-up or had another form of cancer in their breast.Of the 80 243 mammographic examinations used for training and validation, 3045 were followed by a cancer diagnosis within 5 years. Of the 8751 mammographic examinations used for testing, 269 were followed by a cancer diagnosis within 5 years. In each cell, we reported the fraction of examinations that developed cancer within 5 years, assessing the risk of breast cancer 3–5 years after mammography. This can be especially beneficial to patients who do not know their family history of breast or ovarian cancer. In the States, almost half of all women screened are told that they are at increased risk of breast cancer on the basis of their dense breast tissue. At the same time, this practice can mislead women who do not have dense breast tissue to believe they are not at increased risk for breast cancer.

Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography

The predictive accuracy of the model is highest in annually screened non-Hispanic white women and is lowest in women with different demographic characteristics than the population from which the model was developed. Women with high Gail scores have been encouraged to be screened, to undergo genetic or biomarker evaluation, and to participate in intervention trials. Women with previous breast cancer were excluded. Women with breast augmentation were also excluded because augmentation decreases breast cancer detection by mammography. Screening examinations had to be designated as bilateral screening by the radiology facility and needed to be done at least 9 months after any preceding breast imaging to ensure an accurate designation as a screening mammogram. Patient information was primarily obtained from self-report at the time of the screening mammogram. Age at menarche was often not collected or not reported, but it was tested as a risk factor in women who reported it.
We evaluated the probability of a cancer diagnosis within 1 year of each screening mammogram by use of logistic regression analysis in SAS, Version 9.0, and Stata, Version 8 . The primary goal was to find a model that best predicted a diagnosis of breast cancer separately in premenopausal and postmenopausal women with a minimal

number of predictors. The c statistic ranges from 0.50 to 1.00, with a higher score indicating better prediction for an individual woman.

The particular dataset used in this study was a large cross-classification of risk factors by cancer outcome. This procedure required excluding 7.6% of the mammograms from women aged 45 – 54 years with unknown menopausal status.

Comparing random forest and support vector machines for breast cancer classification

Cancer is one of the deadliest diseases in the world. There are over 100 different types of cancer that affect humans. However, this study, aims to analyse breast cancer, a disease in which cells grow out of control to form a tumour which tends to affect another part of the body. There are three common parts of the breast whose cells have the ability to turn into cancer namely lobules, ducts, and the connective tissue.

The exact causes of breast cancer are still not known, but experts are of the opinion that an interaction between genes with lifestyle, environment, and hormone, tends to provoke abnormal cell growth. There are several factors that increase the risk of getting breast cancer such as age. According to research, in most cases people are diagnosed after the age of 50. Men still have a risk of getting breast cancer even though it is a lot lower than women.

The random forest algorithm follows the steps below:

− Choose a feature to be named as root node and make a branch for all possible features. Boruta feature selection is built around the random forest classification algorithm, which is carried out without tuning of parameters and numerical estimate of the important feature.

− Determine the maximum Z score for the extended feature and assign the feature assuming it has a better score than the extended feature. Furthermore, run a two-sided test of equality using the Z score for the extended feature of each attribute with undetermined importance. − Label the feature with lower Z score for the extended ‘unimportant’ feature and remove it from the system.

− Label the feature with higher Z score for the extended ‘important’ feature.

− Remove all copied features and repeat the procedure till none is removed. Random forest was first introduced by Ho in 1995 to split nodes. It is the ensemble of many decision trees using bootstrapping and random feature selection.