International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Breast Cancer Detection Using Machine Learning Algorithms

Nandini S R, Jnanesh Gowda K S[1], Bharath S R[2], Kishor L D[3], Manjunath B S[4]

Nandini S R, Assistant professor, BGS Institute of Technology
[1]Jnanesh Gowda K S, Department of Computer Science and Engineering, BGS Institute of Technology
[2]Bharath S R, Department of Computer Science and Engineering, BGS Institute of Technology
[3]Kishor L D, Department of Computer Science and Engineering, BGS Institute of Technology
[4]Manjunath B S, Department of Computer Science and Engineering, BGS Institute of Technology

---

Abstract – The rapid development of in-depth learning, a family of machine learning techniques, has generated a great deal of interest in its application to the problems of medical thinking. Here, we developed an in-depth study algorithm that can accurately detect breast cancer in mammograms testing using a "end-to-end" training method that effectively uses training data sets with complete clinical annotations or only cancer status (label) for the whole picture. In this method, wound annotations are only required for the first stage of training, and subsequent stages require image-level labels only, which removes reliance on wound annotations that are not readily available. Our entire convolutional network method of separating test mammograms has found much better performance compared to previous methods. The rapid development of in-depth learning, a family of machine learning techniques, has generated a great deal of interest in its application to the problems of medical thinking. Here, we developed an in-depth study algorithm that can accurately detect breast cancer in mammograms testing using a "end-to-end" training method that effectively uses training data sets with complete clinical annotations or only cancer status (label) for the whole picture. In this method, wound annotations are only required for the first stage of training, and subsequent stages require image-level labels only, which removes reliance on wound annotations that are not readily available. Our entire convolutional network method of separating test mammograms has found much better performance compared to previous methods.

***Key Words*:** Machine learning, Mammograms

1. INTRODUCTION

Rapid advances in machine learning, especially deep learning, continue to increase the interest of the medical imaging community in applying these techniques to improve the accuracy of cancer screening. Breast cancer is the second most common cause of cancer death in women in US1, and mammography screening has been shown to reduce mortality. Despite the benefits, mammography screening has an increased risk of false positive and false negative results. The average sensitivity of digital screening mammography in the United States is 86.9% and the average specificity is 88.9%. Computer-aided detection and diagnostic (CAD) software 4 has been developed and used clinically since the 1990s to help radiologists improve the predictive accuracy of mammography screening. Given the remarkable success of deep learning in the recognition and detection of visual objects and many other areas8, there is great interest in developing deep learning tools to assist radiologists and improve the accuracy of mammography screening. Recent studies have shown that deep learning-based CAD systems work in both stand-alone and stand-alone modes, improving radiologist performance in support mode. This study proposes an "end-to-end" approach to pre-train a model to classify local image patches using a fully annotated dataset containing ROI information.

Then use the patch classifier weights parameter to initialize the overall image classifier weights parameter. This can be further fine-tuned using datasets without ROI annotations. To develop a patch and frame classifier, we used a large public digitized film mammography database of thousands of images and then transferred the frame classifier to a small public FFDM database of hundreds of images. Breast cancer, one of the most common malignancies, is the leading cause of death for women in developed countries such as the United Kingdom and the United States and in developing countries such as India. With the growth of developing countries, the risk of developing diseases such as breast cancer is increasing in the population.

2. RELATED WORK

A literature review found that the application of machine learning technology to breast cancer prediction is being thoroughly implemented. Machine learning techniques have proven to be much more precise and faster than the latest prediction techniques.

Breast cancer predictions, comparative reviews of machine learning techniques, and their analysis were proposed by NOREEN FATIMA, LI LIU, and HONGSHA. The proposed method was developed to identify breast cancer. This method used a variety of machine learning, deep learning, and data mining algorithms to predict breast cancer. Analysis shows that the SVM algorithm is more accurate than other machine learning algorithms every time. A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm was proposed by ZEXIAN HUANG, DAQL CHEN. They used Variable Importance Measure (VIM) method to optimize the selected feature number for the breast cancer prediction. Deep learning to improve breast cancer detection in screening mammography was proposed by Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, and Russell McBride. They suggest how to combine two steps into one to train the entire image.

To perform classification or segmentation on a large complex image, a common strategy is to use a sliding window type classifier to detect local blobs on the image and generate a grid of probabilistic outputs. Deep learning for the histological diagnosis of breast cancer was proposed by YASIN YARI, THUY V.NGUYEN, HIEUT.NGUYEN. The proposed method was developed to identify breast cancer. This method used the classifier CNN to predict breast cancer. Based on the deep learning framework and the metastasis learning framework, we proposed various models of automatic breast cancer diagnosis.

3. METHODOLOGY

To perform this entire operation, we first cleaned up the data using data mining techniques and then applied the Naive Bayes algorithm to classify breast cancer types as benign or malignant. The dataset used in this study is from the UCIML repository and consists of 699 instances and 10 attributes. There are positive and negative samples, and each sample has 10 attributes defined. The Naive Bayesian method is based on the famous Bayesian approach, which follows a simple and clear fast classifier. The naive Bayes classifier is a simple stochastic classifier based on the application of Bayes' theorem with strong (naive) independence assumptions. A more meaningful term for the underlying probabilistic model is "independent feature model". The naive Bayes classifier assumes that given a class variable, the presence (or absence) of a particular feature of a class is independent of the presence (or absence) of another feature. From the confusion matrix, the classifier's performance criteria are analyzed. Breast cancer detection calculated accuracy, accuracy (for multiclass datasets), sensitivity, and specificity to provide deeper insight into automated diagnosis. Accuracy is a percentage of correct predictions. Accuracy is a measure of accuracy when a particular class is predicted. Sensitivity is a measure of the ability of a predictive model to select an instance of a particular class from a dataset. Specificity corresponds to the true negative rate commonly used in two-class problems. Accuracy, accuracy, sensitivity, and specificity are calculated using the above equations. Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

A. Convolution Layer

In this layer, the entire image is scanned for patterns and formulated in the form of a 3x3 matrix. This convolutional feature matrix of the image is called the kernel. Each value in the kernel is called a weight vector.

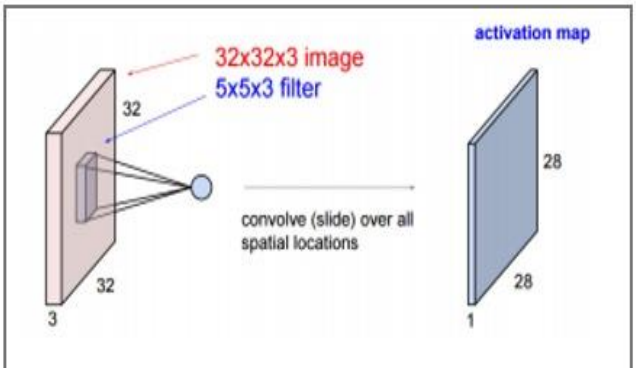

Fig -1: Convolution Layer

B. Pooling Layer

When the convolution here becomes pooling, the image matrix is decomposed into a set of four non-overlapping rectangular segments. There are two types of pooling: maximum pooling and average pooling. Maximum pooling specifies the maximum value of the relative matrix range to be retrieved. Average pooling shows the average of the relative matrix regions. The main advantage of the pooling layer is that it improves computer performance and reduces the possibility of overfitting.

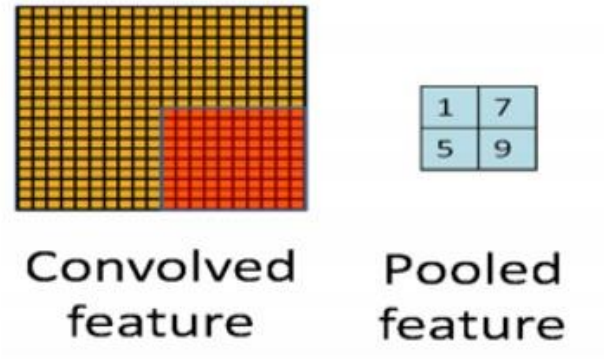

Fig-2: Pooling Layer

C. Activation Layer

This is part of a convolutional neural network whose values are normalized. That is, the value fits into a particular range. The convolution function used is ReLU, which allows only positive values and then rejects negative values. This is characterized

by a small amount of calculation.

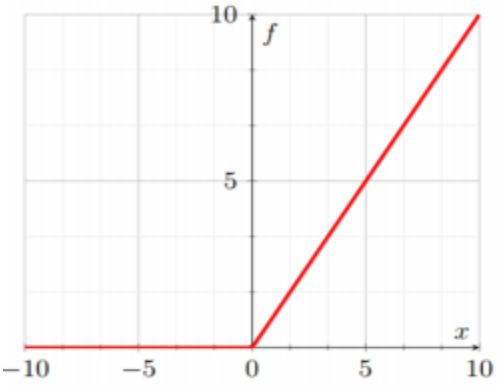

Fig-3: Activation Layer

C. Fully Connected Layer

Features are compared to the features in the test image and similar features are assigned to the specified label. Labels are generally encoded in a numeric format for ease of calculation and later converted to the corresponding string.

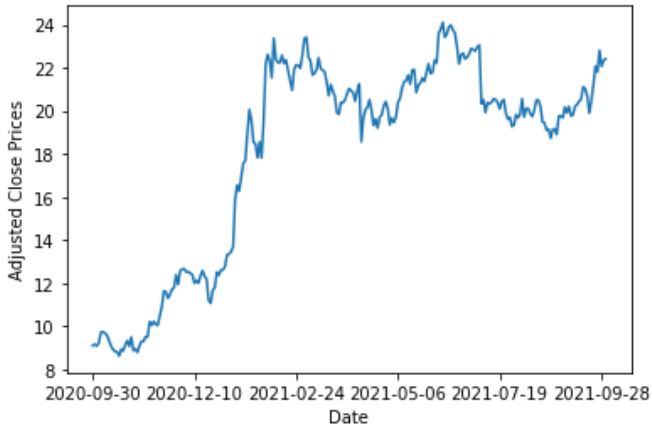

Fig -2: Visualizing the fetched data

Stock data is loaded into a data frame and converted into a CSV file (comma separate value). We plot the line chart of the adjusted close prices over time. The graph shows the data fetched from 30th September 2020 to 28th September 2021.

D. Data Pre-Processing

This step is the most important part of this project. Data preprocessing is a procedure performed to prepare data for a machine learning model. Preprocessing involves transforming raw data into a format that the model can accept and process. This project aims to have a dataset that the model can accept and the algorithm can understand. The value may be missing from the dataset and the information may be verbose, irrelevant, or noisy. Data cleaning is a form of preprocessing that involves removing missing or inconsistent values and changing the index. The same applies to feature selection, hyperparameter adjustment, and data standardization.

I. CNN based feature extraction

ROI images were used as inputs to pre-trained CNNs to extract CNN-based features. This CNN, me. H. AlexNet was trained on an ImageNet dataset of 1.2 million high resolution images and was used to classify common objects into 1000 classes. This pre-trained CNN architecture included five layers of convolution, three layers of pooling, and three layers of connected layers. Because the CNN was pre-trained, its use was limited to the original architecture and input image size of 227 x 227 pixels, so 227 x 227 patches were extracted from the center of each 256 x 256 ROI. The 4096-length vector, which is the output of the first fully connected layer, served as a CNN-based feature, with dimensions reduced by removing these zero-dispersion features in the dataset. Second, CNN-based features were standardized with zero mean and unit variances before entering the classifier. The feature extraction was performed on a computer running the openSUSE Linux operating system with a 6-core / 12-thread Intel Xeon CPU E5-2620 2.10GHz and 24GB of memory.

II. Divide Into Training and Test Datasets

The dataset should be split into a training dataset and a test dataset before modeling.

Training set: A subset of the dataset is used to build and fit the predictive model. Training datasets are generated by creating training dataset scripts that generate training dataset functions from input options. The data is sent to the model for training. The model learns from this data and drives the train set.

Test set: A subset of the dataset used to evaluate the future performance of the model. This is a good benchmark for evaluating your model. The test set is used for the predicted dataset to test the trained model. The model has not seen this part of the set. Used for evaluation purposes.

III. Feature Scaling

This is called data standardization. Sklearn has a feature called the standard scaler that is used to standardize datasets. Standardization is known to improve the numerical stability of the model and improve training speed.

IV. Tuning of Parameters

The parameters are model settings. It is important to adjust the parameters to optimize performance.

Sample code number Id-number

Clump thickness 1-10
Uniformity of cell size 1-10
Uniformity of cell shape 1-10
Marginal Adhesion 1-10
Single Epithelial cell size 1-10
Bare Nuclei 1-10
Bland Chromatin 1-10
Normal Nucleoli 1-10
Mitoses 1-10
Class (2 for benign, 4 for malignant)

V. Model Application and Prediction

This model helps in detecting breast cancer detection using machine learning algorithms and it also helps in finding the stages of the cancer with treatment mentioned for the particular stage.

4. CONCLUSIONS

Although this project is far from complete but it is remarkable to see the success of deep learning in such varied real world problems. In this blog, We have demonstrated how to classify benign and malignant breast cancer from a collection of microscopic images using convolutional neural networks and transfer learning.

REFERENCES

[1] American cancer society. Breast cancer facts and figures 2005-06 (http://www.cancer. org)
[2] Overview: Breast Cancer (http:// www.cancer.org /docroot/CRI/ CRI_2_1x.asp?dt=5)
[3] Breast Cancer (http://www.cancer.gov /cancertopics/types/breast)
[4] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco: Morgan Kaufmann; 2005..
[5] Breast Cancer dataset. http://archive.ics.uci edu/ml /datasets /Breast+ Cancer +Wisconsin +%28Original.
[6] Weka(http://weka. sourceforge.net /doc/ weka/ classifiers)
[7] Naïve Bayes Classifier.www.statsoft.com /textbook/naïve bayes –classifier
[8] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.
[9] Tang, Z., and MacLennan, J., Data Mining with Sql Server 2005. Wiley, 2005
[10] Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. 34:113–127, 2005.
[11] Witten, I. H., and Frank, E., Data mining: practical machine learning tools and techniques. Morgan Kaufmann – Academic Press, America, p. 525, 2005.
[12] Alireza Osareh, BitaShadgar, Machine Learning Techniques to diagnose Breast Cancer. IEEE, 2009
[13] Subbalakshmi G. , Ramesh K., Chinna Rao M., Decision support in Heart Prediction System using Naïve Bayes, IJCSE ,2010.