|  |
| --- |
| **Automated Diagnosis for Breast Cancer Tissue Type**  Gautham Nandhakumar, Saiprasad Sankaranarayanan, Jayanthan Ramesh Vellore  1 School of Biomedical Engineering, Drexel University, USA  Course : Bmes550 Advanced Biocomputational Languages  Instructor: Ahmet Sacan  Date : 2017-12-10 |

[[1]](#footnote-2)\*abstract

The correct diagnosis of breast cancer tissue type is one of the major problems in the medical field. From the literature it has been found that different pattern recognition techniques can help them to improve in this domain. Implementing automated diagnostic systems for breast cancer detection will be very beneficial for the doctors. The major objective of the paper is to create a breast cancer detection system that can be used by everyone. Breast cancer tissue type can be classified into Benign or Malignant based on 9 unique component scores. In our system, we use a PHP to create a web interface that collects these nine scores. The web interface is linked with MATLAB program that implements neural network based diagnostic system which is the backend of the project. Based on the nine scores implemented with the NN, the system calcualtes whether the tissue type is B or M. Existing data can also be viewed by the user when he/she inputs the sample number. This is done by integrating php and sqlite through Matlab.

# introduction

Problem There are many systems that help in diagnosis of a health condition such as electrocardiograms (ECGs), electroencephalograms (EEGs), ultrasound signals/images, X-rays, and computed tomographic images. The Conventional method of diagnosis relies on manual reading of these data and concluding. These individual recording put together give different features that help in diagnosis of various diseases. Automated diagnostic systems are important applications of analysis of database and pattern recognition, aiming at assisting doctors in making diagnostic decisions. Due to large number of patients in intensive care units and the need for continuous observation of such conditions, several techniques for automated diagnostic systems have been developed in the past ten years to attempt to solve this problem. But in order to do this there is need to convert the qualitative data into a format that is numeric and easily read by the computing system. So, some of these visual or observatory readings are converted into quantitative data for the purpose of creation of automated diagnostic system.

ANNs are computational architectures composed of interconnected units (neurons). Its name suggests it is inspired by neural systems, though the functioning of today’s ANNs are very different from the way our brain or our neural system functions. Sometimes the term neural network also refers to the corresponding mathematical model, but properly speaking a network is an architecture.

The breast scan images show any abnormality or other tests indicates a possibility of breast cancer, a biopsy is performed to confirm the presence of breast cancer. The need for biopsy does not indicate breast cancer. In most cases the result of the biopsy tends to be negative. There are different types of biopsies. The technique used in our problem is Fine needle aspiration

In an FNA biopsy, a very thin, hollow needle attached to a syringe is used to withdraw (aspirate) a small amount of tissue from a suspicious area.

In today’s world the number of patient visiting the hospital has gone up tremendously. Also, the cost of treatment for patients have increased. This calls for efficient system to diagnose and treat disease. One crucial factor is the time a doctor can dedicate to each patient. Though, it is important for doctor for diagnosis some obvious results can be done by inexperienced too. Also, the inexperience should not affect the diagnosis of critical disease. So, creating an automated decision system for breast cancer will help solve this problem. The purpose of the project is to create a diagnostic system when the various inputs in a web interface and it decides whether the given sample needs doctor’s attention or not. The system that we have created classifies the given data are benign or malignant and the malignant can be sent to doctor for further studies. This system can be handled by a technician. Hence, reducing the number of report the doctor reads.

The ANN system developed by Marceno et al. Has proven that these systems can be used to create decision system that is 95% accurate. They have modified these ANN in lot of way creating new forms of ANN models specifically designed to focus on certain parameters. They have also created SVM based diagnostic system. The latest parameter they are working on is metaplasticity neural network

# Dataset

\* If Breast cancer is a malignant tumor that has developed from cells of the breast. Although scientists know some of the risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity) that increase a woman’s chance of developing breast cancer, they do not yet know what causes most breast cancers or exactly how some of these risk factors cause cells to become cancerous. Research is under way to learn more and scientists are making great progress in understanding how certain changes in DNA can cause normal breast cells to become cancerous

In this study, the Wisconsin breast cancer database taken from fine needle aspirates from human breast tissue was analyzed. They have been collected by Wolberg and Mangasarian (1990) at the University of Wisconsin-Madison Hospitals. The data consists of 683 records of nuclear features. The cells obtained through fine needle aspiration was studied and various parameters are given. Each record in the database has nine attributes. The nine attributes are Clump thickness, uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single epithelial cell size, Bare nuclei, Bland chromatin, Normal nucleoli and Mitoses. The values range 1-10, where 1 is normal state and 10 is the most abnormal state. Of these data 444 are benign tissues and 239 are malignant tissues. Data represents benign as 2 and malignant as 4.

**3 methods and IMplementation**

The Dataset used is imported into the database from the UCI Machine Learning Repository.

***SQLite MANAGER***

The SQLite Manager was used to build a database to create a table with 683 instances and 11 columns. Sample Number, Clump thickness, uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single epithelial cell size, Bare nuclei, Bland chromatin, Normal nucleoli and Mitoses. Pre-processing the data (like removing the incomplete samples), establishing the connection with the database and importing data from database was performed using MATLAB.

***MATLAB***

Visualization of the loaded data was performed using PCA. PCA calculates the eigen values and its corresponding vectors. The vectors of the highest 3 eigen values represent 3 principle component axis that displays the most variance among the data. Next, the classification of the data into benign and malignant is performed using the Neural Networks. All the instances with its 9 attributes were extracted as the input and the last column was extracted as the target data. The exported data was distributed into training data, testing data and validation data at a ratio of 7:3:3. The Network constructed contains 4 hidden layers. The results were published as a GUI created using PHP.

***PHP***

The proposed GUI contains 2 functions. One to act as library search from the database with input being the Sample number. Next, is to test the trained and validated classifier. For this, the GUI asks the user the required 9 features and returns the class (malignant or benign). The complete Data flow is shown in Figure 1.

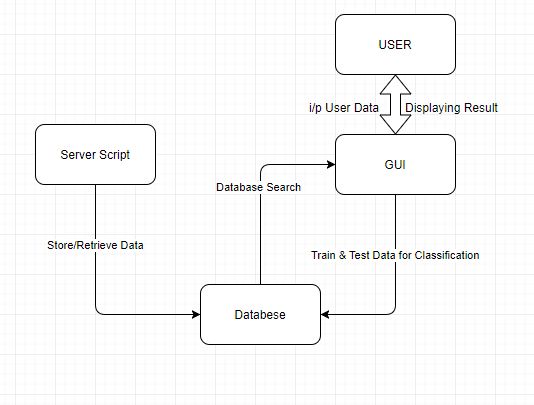


Figure : Flow Chart. The figure describes the exchange of data between the user, the web server, the created database and the constructed GUI.

# experiments and REsults

Interfacing the SQLite manager enables easier exporting of Data for processing. PCA produced excellent visualization of the data as shown in Figure 2.

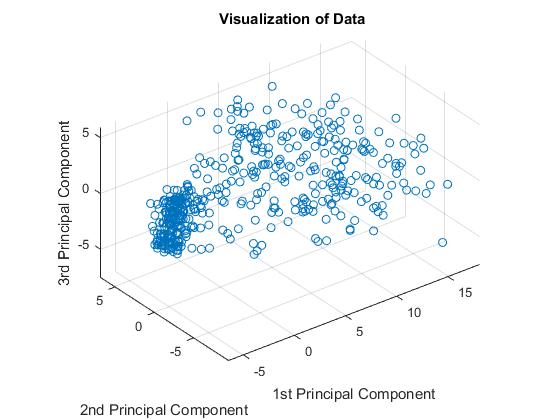
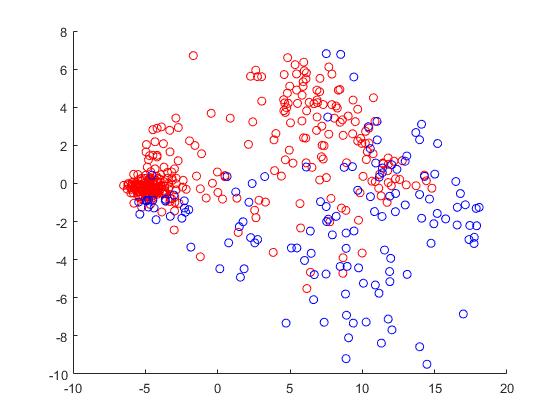


Figure : Visualization of Data. The figure shows the distribution of the data at maximum variance across 3 calculated principal axis.

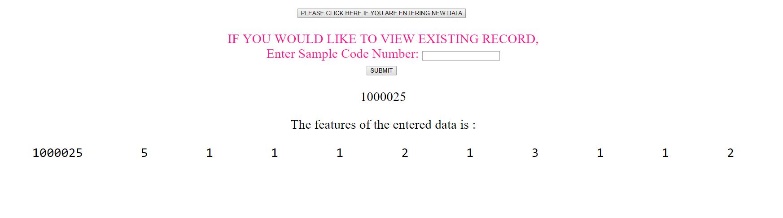
The Neural Network classified the data with an error of 0.2357. The classified data is later visualized again into 2 separate classes (Colored Different) as shown in Figure 3.



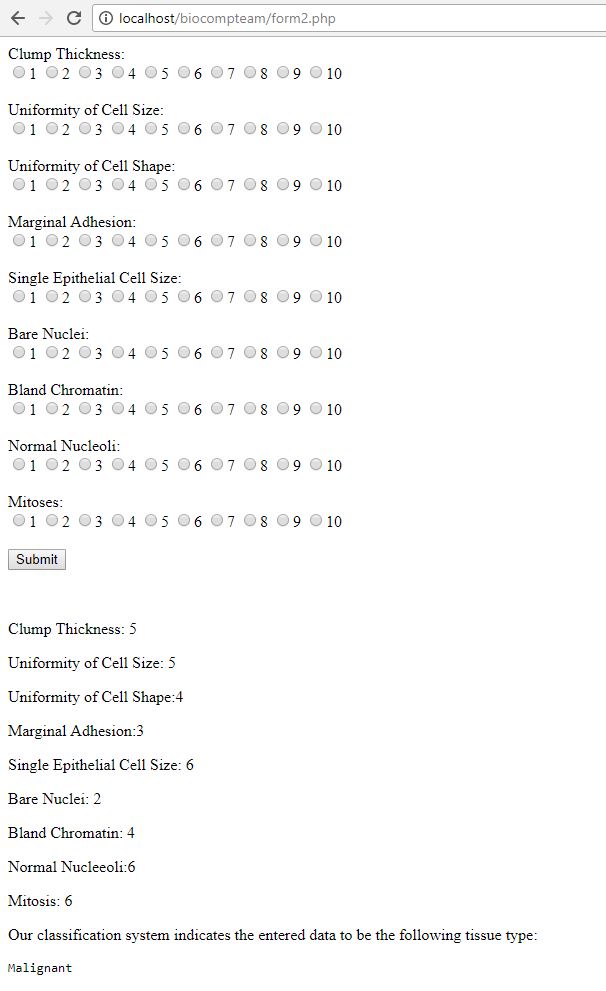
**Figure 3: Classification of Data. The figure shown the 2 separated classes with 2 different markers. The data is represented along PC1 and PC2 ( 2D representation). Red is benign and Blue is malignant.**

The classification function returns the user the tissue type, if malignant or benign. PHP, as discussed is used to display the results. Figure 4 shows the interface to retrieve the instance for the submitted Sample number.

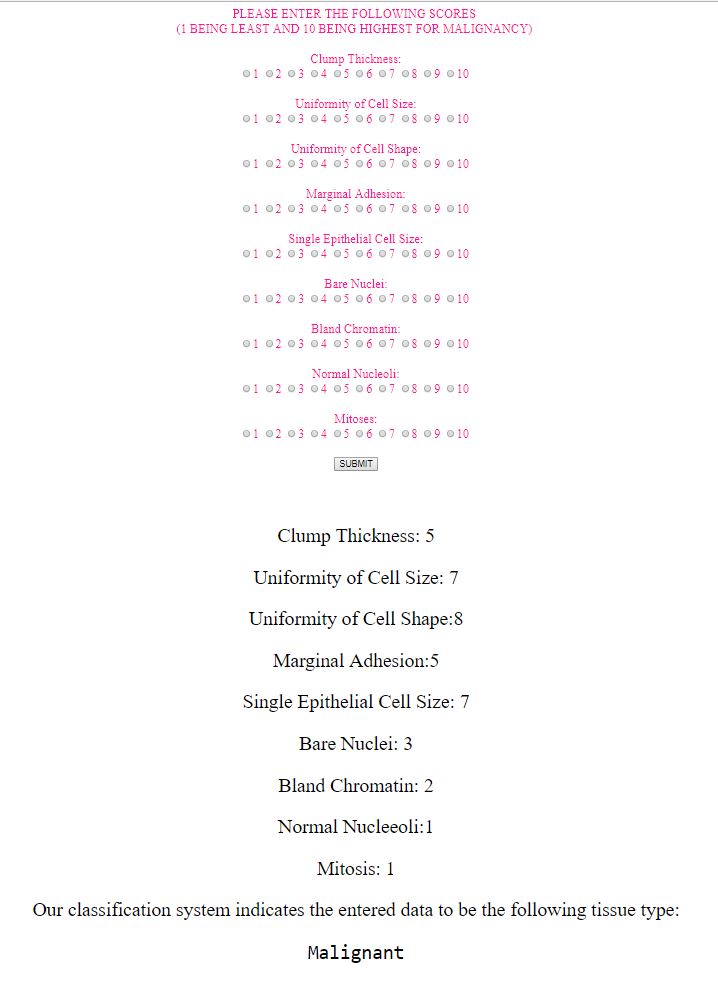
Figure : Interface for Database Search. The figure shows the result of a Database Search. The submitted Sample number displays its attributes.



The result of the classification on the user inputted data is shown in Figure 5.



**Figure 5: Interface to display classification result. The figure displays the label of the classified user input.**



# DISCUSSION

Our results are consistent with the existing dataset. It is important to note that all the nine principal components of diagnosis are essential, and no component is more significant than the other. We tried our testing with values from existing data and it proves to be functional.

The difference between the two types of cancer is very important. It helps is recognizing the subsequent prognosis. It is also important to note that some characteristics are shared between the two types. This makes this study relatively less accurate. With a large verified dataset, it is possible to bring down the error rate. This helps in increasing the accuracy and subsequent prognosis.

Major limitations are scoring system and limited dataset. The scoring pattern is based on selected previous case studies. In order to improve the diagnosis, the user needs to be educated on the scoring system. This can be done with the help of reference tissue images.

In the future, the application can include an option to add input sample data into the dataset upon physician request. The data needs to be verified and validated by professionals before being added to the dataset. This should be done to avoid decreasing accuracy. The application will be made available as a package to all oncologists. This will help increase the dataset.

# References

1. Sacan, A., I.H. Toroslu, and H. Ferhatosmanoglu, *Integrated search and alignment of protein structures.* Bioinformatics, 2008. **24**(24): p. 2872-9.

2. Übeyli. E. D., *Implementing automated diagnostic systems for breast cancer detection*. Expert Systems with Applications 33 (2007) 1054–1062

3. A.Marcana-Cedeno.Quintanilla-Domingues.D.Andina, *Breast cancer classification applying artificial metaplasticity algorithm,*Neurocomputing 74 (2011) 1243–1250.

4. www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy.html

1. [↑](#footnote-ref-2)