# Breast Cancer Predication from Mammographic Findings

Yan Zhao, Rui Tian

October 2, 2014

## 1 PROBLEM MOTIVATION

Breast cancer is one of the most common cancers with an estimated 230,000 new cases in 2013, and over 1 million diagnosed cases around the world [1]. Mammography is a widely used imaging technique for breast cancer detection, with a reported specificity of 97% [2]. However, its relatively low sensitivity always results in unnecessary biopsies. In order to reduce the false positive rates of mammography, various computer aided detection (CAD) systems have been proposed recently. Machine learning algorithms such as artificial neural networks (ANN) have been used in the process of clinical decision-making. In this project, we propose to make model based prediction from mammographic findings and medical records and evaluate the performance of different models using receiver operating characteristics (ROC) curve analysis.

## 2 DATASET

We will use the well-known UCI machine learning repository [4]. Mammographic findings are standardized by radiologists in the format of BI-RADS score, which is an integer varying from 1 to 5, mass shape (integer), mass margin (integer) and mass density (integer). Furthermore, medical records including patient age and body mass index (BMI) are also available as input attributes. Severity, namely benign or malignant, which is confirmed by biopsy results, will be treated as the output attribute. The dataset including the input and output attributes of a large number of patients will be used for the model training and performance

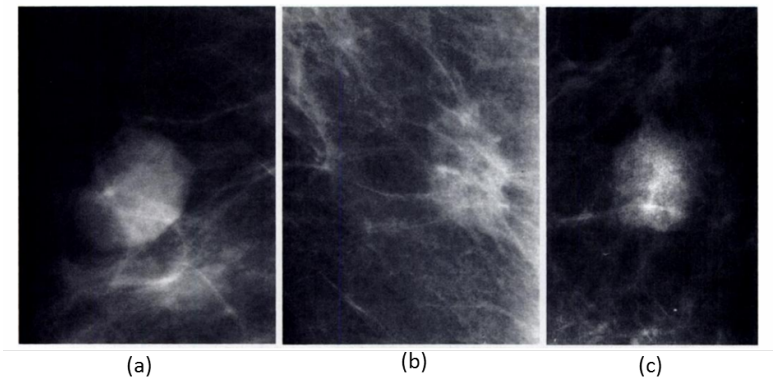


Figure 1.1: Mammographic images of (a) a benign lesion, (b) a malignant lesion, and (c) a mass described as “likely malignant”. The lesions are described by mass, shape, position, et al. Figure courtesy of Jay A. Baker [3]

evaluation. Note that some attributes are missing and the effect of their absence will be investigated as well if time allows.

# 3 METHODS

We intend to implement three different models to make prediction based on the known data. They are Articifical Neural Network, Decision-Tree Learning and Bayesian Networks. These models have been widely applied to medical imaging process and proven to be effective models.

## 3.1 ARTIFICAL NEURAL NETWORK

Artifical Neural Network (ANN) is a computational model first introduced in 1943 but recently it attracted more and more attention. The typical structure of an ANN consists of a seires of interconnected elements which can be called as nodes or neurons. These neurons can be capable of executing the necessary calculation based on activation function and input data in a parallel way [5]. In medical imaging, ANN is widely used in medical image processing and recognition for diasease prediction and diagnose [6].

## 3.2 DECISION-TREE LEARNING

As its name suggests, a Decision-Tree Learning (ID3) uses a decision tree to present a strategy of classifying and acquiring target values. A node in a decision tree represents a test for a certain attribute. The successor nodes of that node represent the possible values of that attribute. A leaf node in a decision tree represents a subclass. Decision-tree learning has been used to classify medical data and make diagnoses [7].

### 3.3 Bayesian Networks

Bayesian network utilizes the probability theory from Bayes' Rule, and it has various applications in the clinical decision-making. The ability to deal with uncertainty is one of the key advantages of Bayesian network. In this project, we will construct a Bayesian network for mammographic decision support.

## 4 Timeline and Milestone

By milestone (Oct. 28, 2014), we intend to finish implementing the three models introduced above and compare the performances based on Operating Characteristic Curve (ROC). In the first week, we plan to explore the database and sort out an available and efficient dataset. During the second and third week, we intend to implement the approaches disscussed above in MATLAB. In the week before the milestone, we will evaluate the performance of the three models using ROC analysis. The detailed timeline is presented as below.

Timeline

| Time | Task |
|---|---|
| Week 1 (Oct. 2 - Oct. 5) | Database exploration and formatting |
| Week 2 & 3 (Oct. 6 - Oct. 20) | Implementation of the models |
| Week 4 (Oct. 21 - Oct. 28) | Traning performance evaluation |

## Reference


[1] American Cancer Society. Cancer facts and figures. 2013. http://www.cancer.org/research/cancerfactsfigures/cancerfactsfigures/cancer-facts-figures-2013.

[2] Per Skaane, Solveig Hofvind, and Arnulf Skjennald. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: Follow-up and final results of oslo ii study 1. *Radiology*, 244(3):708–717, 2007.

[3] Jay A Baker, Phyllis J Kornguth, Joseph Y Lo, Margaret E Williford, and Carey E Floyd Jr. Breast cancer: prediction with artificial neural network based on bi-rads standardized lexicon. *Radiology*, 196(3):817–822, 1995.

[4] Catherine Blake and Christopher J Merz. {UCI} repository of machine learning databases. *University of California, Irvine, Dept. of Information and Computer Sciences*, 1998. http://archive.ics.uci.edu/ml/.

[5] IA Basheer and M Hajmeer. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, 43(1):3–31, 2000.

[6] Jianmin Jiang, P Trundle, and Jinchang Ren. Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 34(8):617–631, 2010.

[7] Chin-Yuan Fan, Pei-Chann Chang, Jyun-Jie Lin, and JC Hsieh. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing*, 11(1):632–644, 2011.