Characterization of mammogram using ensemble classification technique for detection of breast cancer

Subhankar Ghosh

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.

Characterization of mammogram using ensemble classification technique for detection of breast cancer

Thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Subhankar Ghosh
(Roll: 111CS0463)

under the guidance of

Prof. Banshidhar Majhi

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.
May‘ 2015

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India.

May 4, 2015

Certificate

This is to certify that the work in the thesis entitled ***Characterization of mammogram using ensemble classification technique for detection of breast cancer*** by ***Subhankar Ghosh*** is a record of an original research work carried out under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Banshidhar Majhi
Professor

Acknowledgment

I owe deep gratitude to the ones who have contributed greatly in completion of this thesis.

Foremost, I would also like to express my gratitude towards my project advisor, Prof. Bansidhar Majhi, whose mentor-ship has been paramount, not only in carrying out the research for this thesis, but also in developing long-term goals for my career. His guidance has been unique and delightful. He provided his able guidance whenever I needed it. Yet he always inspired me to be an independent thinker, and to choose and work with independence.

I am also very thankful to Mr. Shradhananda Beura for listening, offering me advice, and gladly extending his support throughout the process.

I would also like to extend special thanks to my project review panel for their time and attention to detail. The constructive feedback received has been keenly instrumental in improvising my work further.

I would like to thank other researchers in my lab and my friends for their encouragement and understanding.

My parents receive my deepest love for being the strength in me.

Subhankar Ghosh

Abstract

Breast cancer is one of the most common known cancers in women today. Just like any other form of cancer an early detection of cancer provides better chances of cure. However, it is an arduous task for the radiologists to detect cancer accurately. Thus computer aided diagnosis of the mammographic images is the most popular medium to aid the radiologists in accurately classifying benign and malignant mammographic lesions.

In this thesis an efficient approach is presented to classify the mammographic lesion for the detection of breast cancer. In this approach the extracted feature coefficients are balanced using Gaussian distribution. This distribution balances the class unbalanced dataset providing for better classification. This scheme uses Logit Boost classification technique. Logit Boost uses least squared regression cost function on the additive model of Adaboost. The standard MIAS database was used to obtain the mammographic lesions. With a classification accuracy rate of 99.1% and a performance index value of AUC = 0.98 in receiver operating characteristic (ROC) curve the results are pretty much optimal. These results are very promising when compared with existing methods.

Contents

List of Figures

Chapter 1

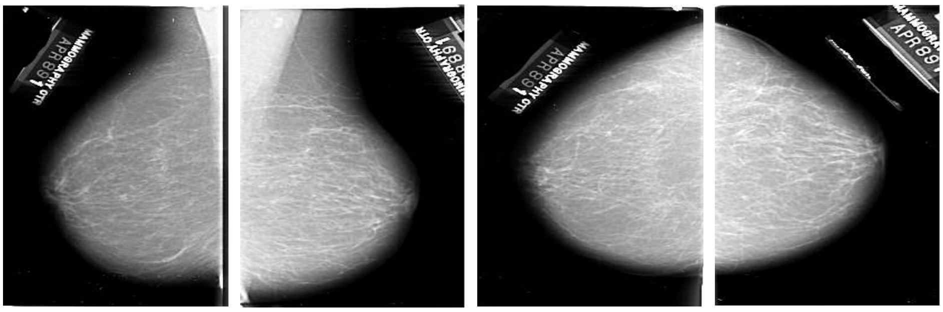
Introduction

One of the main sources of death in ladies is breast cancer. The number of cases of breast cancer reported and deaths due to breast cancer are roughly 232,340 and 39,620 respectively, in US in 2013[1]. Similar situation prevails in India. By 2020, it is expected to see the number of cases of Breast cancer surpass that of cervical cancer among the women in India. According to the Lancet report an imminent threat of a cancer epidemic is lurking over India: it is estimated that by the year 2020 one fifth of the total number of cancer patients of the world will be in India. [2].

Another study by GE Healthcare, the incidents of new cases of breast cancer in India, which amounts to 115,000 per year, would increase to around 200,000 per year, by 2030. [3].

The genesis of breast cancer has been attributed to some of the well recognized risk factors both exogenous and endogenous. Some of the exogenous factors are alcohol abuse, cigarette smoking, lack of physical activity, pesticides, socio-economic status, exposures to pollutants, high fat intake. The endogenous factors include the duration of exposure to steroid hormones. Transitively this is dependent on numerous factors like late pregnancy, obesity and late menopause.

The odds of recovering from breast cancer increases if it is detected at an earlier stage by periodic screening. Mammography has evolved as one of the most reliable techniques for early detection of breast cancer. It is highly recommended to all women above the age of 40 to undergo mammogram on a yearly basis by the American Cancer Society for an early detection of breast cancer. [1].



(a) (b) (c) (d)

Figure 1.1: Two types of views of mammogram. (a) MLO view of left breast, (b) MLO view of right breast, (c) CC view of left breast, (d) CC view of right breast.

Radiologists have a very important task of properly interpreting the mammograms since they have to suggest patients for biopsy.

1. However, different radiologists may differ in judging a mammogram as the interpretation of mammograms actually depend on training and experience of the radiologists.

2. Furthermore, factors like different image quality, small and subtle signs of the breast cancer increases the difficulty of correct diagnosis.

3. The probability of human error cannot be left out due factors such as distraction and oversight, fatigue which ultimately leads to inter-observer and intra-observer variations.

4. Computer aided diagnosis of mammographic images not only improves the sensitivity but also the specificity of the diagnosis.

5. Therefore it is on utmost importance that misinterpretation is avoided. It is an observed fact that around $60-90\%$ of the biopsies which are diagnosed to be cancers actually turn out to be benign [4].

Therefore, the current technique that is very popular is computer-aided diagnosis (CAD) for the efficient analysis of digital mammograms. This aids the radiologists in the interpretation of mammograms and double checking their diagnosis.

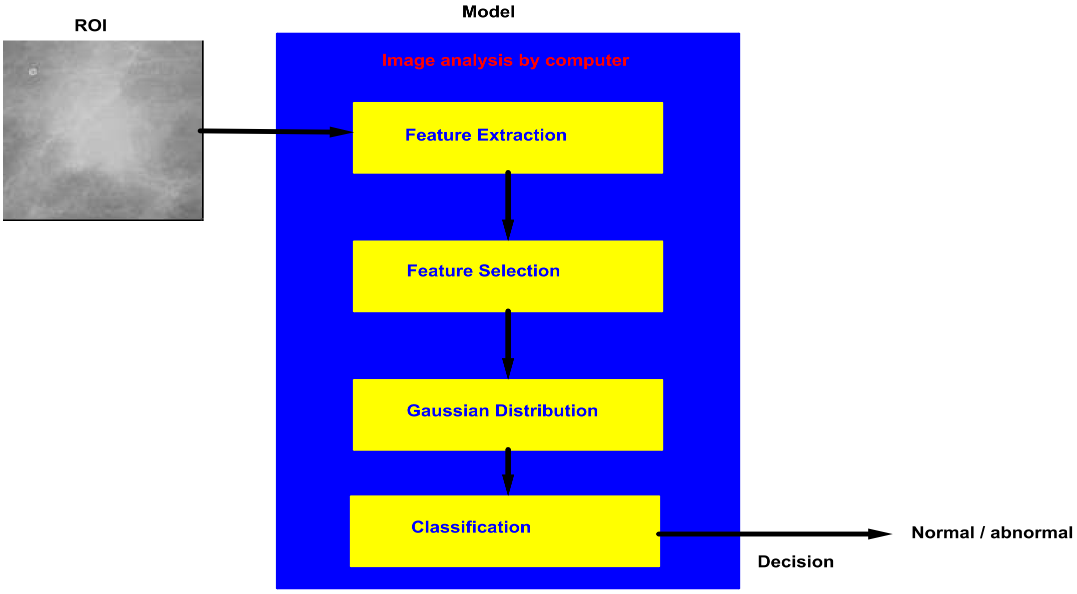


Figure 1.2: CAD for lesion classification.

The organization of the thesis is as follows. In Chapter 2, contains a discussion of the work already done related to CAD of mammographic images. In Chapter 3, we introduce the Gaussian distribution based balancing of dataset. Section. 3.2 contains the description of Gaussian distribution, the properties and the advantages of Gaussian distribution based balancing of dataset. We, then, discuss the ensemble classification methods and then move on to the specific Logit boost classification technique, a variant of Adaboost, in Chapter 4. In Chapter 5, the proposed methodology for classification of mammograms into benign or malignant classes is described. In Chapter 6 contains the final results and simulations. Finally, the scope for further research work and the concluding remarks are presented in Chapter 7.

Chapter 2

Literature Survey

An accuracy rate of 82.3% was obtained by Francisco *et al.* by making use of the possibility of wavelets to analyze different resolutions.Geometrical and cluster classification was used [5].

Zhang *et al.* proposed a unique method using neuro-genetic algorithm for feature selection along with the classification technique of artificial neural network [6]. Statistical features were used for classification purposes. An accuracy rate of 90.5% was obtained for classification.

Moayedi *et al.* combined the human like reasoning of fuzzy techniques along with the classification power of support vector machine and neural networks. This support vector based fuzzy neural network approach gives an accuracy of 97.5% [7].

Talha *et al.* achieved a classification accuracy of more than 90% by reducing wavelet based features using principal component analysis.

Alolfe *et al.* obtained a classification accuracy of 90% for characterization of mammograms [8] using support vector machine classifier in combination with linear discriminant analysis classification.

Liu *et al.* used level set segmentation and multiple kernel learning and obtained an accuracy of 76% on the morphological features extracted from the segmented regions

[9]. Digital Database for Screening Mammography was used for experimentation.

Javadi *et al.* used particle swarm algorithm along with wavelet transform to pin point the important features. [10]. Fuzzy classification techniques were used for classification purposes obtaining an accuracy of 93.41%.

Dong *et al.* used Gabor filter to classify normal and abnormal and achieved an average of 80% precision in the year 2009 [11].

Li *et al.* modeled each of the region of interests into the texton distributions and in the second stage Fisher classifier was used for classification obtaining an accuracy of 87% [12].

De *et al.* in the year 2014 used zernike moments and applied the results to ELM and SVM neural networks obtaining a best result of 80% accuracy using SVM with RBF kernel [13].

From the literature survey it has been observed that different classification techniques are used in combination with feature extraction and selection techniques for classifying the lesion. Still classification accuracy can be increased.

Thus to increase the accuracy and reduce complexity there is a need to develop some new classifiers as well as feature extraction and selection techniques.

In this paper, Gaussian distribution is used to preprocess the features and Logitboost classifier with Random forest classifier as base classifier is used as classifier to characterize the mammograms into benign and malignant.

Chapter 3

Gaussian Distribution based Balancing of dataset

3.1 Class Imbalance problem

It is a general assumption that most machine learning and data mining algorithms make that the probabilities of the target classes to appear are same. On the contrary in most real world applications, such as breast cancer detection, oil-spill detection, fraud detection, such assumptions are violated. We noticed that the majority of the examples are that of a single class leaving only a small minority of the examples belonging to the other classes, which sometimes turn out to be the more important class of them all. This is known as the class imbalance problem. Many multi-resolution techniques exist to resolve the class imbalance problem. Some of them are :

1. SMOTE
2. Gaussian Distribution
3. Under-sampling

Gaussian-distribution has a lot of advantages over the other two in the problem of breast cancer detection.

3.2 Gaussian Distribution

The Normal (or Gaussian) distribution is a quite common continuous probability distribution, according to the probability theory. The normal distribution gives the information about the probability of any real observation to fall in between any two real limits or real numbers, as the distribution curve approaches zero on either side.

A normal distribution in a variable X with mean μ and variance of σ^2 is a statistical distribution with a probability density function given by:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{3.1}$$

on the domain x in $(-\infty, \infty)$. The term "normal distribution" is used for this distribution by Statisticians and mathematicians generally, and Physicists prefer the name the Gaussian distribution, the name "bell curve" is associated with the distribution because of its bell shape.

3.3 Standard Normal Distribution

If we set $\mu = 0$ and $\sigma^2 = 1$ in general Gaussian distribution then the distribution obtained is called "standard normal distribution". By setting $Z = (X - \mu)/\sigma$, so $dz = dx/\sigma$ any arbitrary normal distribution can be converted to a standard normal distribution, thus yielding:

$$P(x)dx = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{3.2}$$

3.4 Properties of Gaussian Distribution

1. **Symmetric in nature** : The distribution is symmetric about the point $x = \mu$. It can also easily be observed that the mean, median and mode of the distribution is the point $x = \mu$.

2. **It is Unimodal** : For $x > \mu$ the derivative of the curve is positive and for $x < \mu$ it is negative and for $x = 0$ zero.

3. **Double points of inflection** : These are the points where the double derivative of the function is zero. Two such points exists, one at $x = \mu - \sigma$ and other at

$x = \mu + \sigma$.

4. **The density is logarithmically conclave** : In mathematics a non negative function $f : R^n \longmapsto R_+$ is log-conclave if it has a convex set as its domain and also satisfies the following relation:

$$P(x) = f(\theta x + (1 + \theta)y) \geq f(x)^\theta f(y)^{1-\theta} \tag{3.3}$$

$\forall x \in dom f$ and also $0 < \theta < 1$. If f is strictly positive, it can be said that the logarithm of f is concave.

5. **Tolerance intervals of standard deviation** : Almost all values drawn within one σ from the mean amount to 68% of the values. About 95% values lie in between 2σ and 99.7% within 3 σ. This rule is known as the $3-\sigma$ rule.

6. **Limiting case of discrete binomial distribution** : The normal distribution can be proven to be a limiting case on the discrete binomial distribution. If binomial distribution is denoted by $P_p(n|N)$ then if the sample size N becomes very large, then $P_p(n|N)$ is normal with mean and variance given as :

$$\mu = Np$$

$$\sigma^2 = Npq$$

with $q \equiv 1 - p$.

The distribution is normalized properly since:

$$\int_{-\infty}^{\infty} P(x)dx = 1.$$

7. **Cumulative distribution function** : This is nothing but the probability the a variate will take a value $\leq x$. This function is mathematically given by the integral of the normal distribution :

$$D(x) \equiv \int_{-\infty}^{x} P(x')dx' \tag{3.4}$$

$$D(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x'-\mu)^2/(2\sigma^2)}dx' \tag{3.5}$$

$$D(x) = \frac{1}{2}[1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})], \tag{3.6}$$

where erf is the error function.

8. **Bernstein's Theorem** : According to Bernstein's Theorem if we consider two independent variables X_1 and X_2 then $X_1 + X_2$ and $X_1 - X_2$ are also independent then we can conclude that X_1 and X_2 must necessarily have normal distributions.

3.5 Over Sampling

According to the central limit theorem regardless of the actual sampling distribution, the sampling distribution of the mean will always approache normal distribution. Based on this we can manufacture the synthetic examples for minority class even if we do not know the actual real sampling distribution. We expect to create datasets almost complying with the actual dataset. After the new instances are put together with the original minority ones, the original sampling distribution is kept almost intact. The following assumptions are made about independence of the attributes :

1. Every attribute of the dataset is taken to be random.

2. All attributes are considered to be independent of each other.

We are give k attributes b_1, b_2, b_3,...,b_k, thus we have k random variables. In this method the expected value of each variable is calculated using the data of the minority classes of the training set. Let us call the standard deviation and mean of b_i as σ_i^2 and μ_i respectively, for all $i \in 1, 2, ..., k$.

Consider μ_i as the mean and σ_i' as the standard deviation of the unknown distribution controlling the random variable b_i. For minority training data we assume that all the values of the attribute b_i are independent and random variables that are similarly dostributed and the reason for such assumption is that they are results of different experiments, and each following the same distribution function.

So, according to the central limit theorem, as the value of n of samples tends towards infinity the underlying distribution tends towards a standard normal distribution.

$$\frac{\mu_i - \mu_i'}{\sigma_i'/\sqrt{n}} \longrightarrow N(0,1). \tag{3.7}$$

where n denotes the number of minority class examples. We know the following equation if we are given the random variable a_i that obeys standard distribution $N(0,1)$.

$$\mu_i' = \mu_i - a_i \bullet \sigma_i'/\sqrt{n}. \tag{3.8}$$

where μ_i shows the mean of b_i for the minority class of the training set, and assume that it represents the original minority class dataset. μ_i' shows the mean of b_i for the unknown minority class data, and we make the assumption that it represents the unknown minority class data.

So if we are given any example with the value of b_i, it is easy to synthesize value for that attribute using the following equation :

$$b_i' = b_i - a_i \bullet \sigma_i'/\sqrt{n}, i \in 1, 2, ..., k. \tag{3.9}$$

In the above equation σ_i' is not known so its approximation is done using σ_i. Thus leading to the following equation :

$$b_i' = b_i - a_i \bullet \sigma_i/\sqrt{n}, i \in 1, 2, ..., k. \tag{3.10}$$

The above equation forms the basis of the normal distribution model.

Chapter 4

LogitBoost Classifier

4.1 Ensemble Classification Methods

The underlying principle of any ensemble classification technique is to take the aggregate of multiple classifiers. In machine learning, ensemble methods make use of many learning algorithms to come up with better predictive model than that of any of the single learning algorithms. An ensemble classification model creates a set of base classifiers from training data and then does classification taking a vote of each of the base classifiers' predictions.

4.2 Rationale for Ensemble Method

Any classifier is trained such that its training error is minimum. However, a classifier is only said to be useful if it can make an informed prediction about the class labels of the instances it has never seen before. This can be possible if the classifier is designed such that it can generalize its decision boundaries to the regions where no training example is located. This choice is made during the choice of design of the classifier.

These design choices are responsible for introducing a **bias** into the system. If stern assumptions are made by the classifiers about their decision boundaries more will the classifier's bias. For example, a very important design decision in decision tree induction is the amount of pruning is required to get low expected error for the tree. If one of the trees performs very high pruning then it is expected to have a larger bias than the tree which performs very little pruning.

Another important factor affecting the expected error of a classifier is the composition of the training data. Since a different composition of the training data can lead to variability in decision boundaries. This factor is commonly known as **variance**.

There may be cases where the class labels are non deterministic. Which means that examples with same attribute values can have different class labels. Such cases are known as **noise** and are unavoidable.

So, the motivation behind using ensemble techniques are :

1. **Reduction of variance** : The dependency of the results on the peculiarities of the training dataset.

2. **Reduction of bias** : A combination of multiple classifiers may have an even more expressive class than the single classifier.

4.3 Construction of Ensemble Classifier

The basic idea is to create many classifiers using the same training set and then aggregate their results for classifying unknown examples :

1. **Manipulation of training set** : In this approach, many training sets are generated by resampling of the original training data using some specific sampling distribution. Each such training set is used to train the base classifiers. **Bagging** and **Boosting** are examples in this category.

2. **Manipulation of input features** : In this approach a subset of the input attributes are chosen to produce the training dataset. The choosing of the subset can be random or according to some specific statistical method. **Random Forest** is one such example which manipulates its input features.

3. **Manipulation of the class labels** : This method is generally in use when the number of class labels is very large. The training data is transformed into binary classes. Each set is again recursively transformed into binary class problem to ultimately reach the required number of class problems. **Error-correcting output coding** method is an example in this category.

4. **Manipulation of learning algorithm** : Here learning algorithms are manipulated such that applying them on the same training data may come up with a different model.

4.4 LogitBoost

The very inception of this algorithm is a very interesting procedure called **Boosting**. Boosting focuses on the training examples which are hard to classify, it achieves this by iteratively change the distribution of the training instances.

In this method a statistical framework is used on the basic **Adaboost** algorithm. If we consider Adaboost to be the basic additive model and apply least squared regression cost function then LogitBoost is derived. Adaboost algorithm has the following features. Let $(x_i, y_i)|i = 1, 2, ..., N$ be the set of N training instances.

1. The importance of each base classifier C_j is dependent on its error rate :

$$e_j = \frac{1}{N}\left[\sum_{i=1}^{N} w_i I\left(C_j(x_i) \neq y_i\right)\right], \tag{4.1}$$

   $I(p) = 1$ if p is true else 0.

2. The importance of the classifier C_j is given by

$$\alpha_j = \frac{1}{2}\ln\left(\frac{1-e_j}{e_j}\right) \tag{4.2}$$

   Thus, α_j takes a high value if error rate is close to 0 and negative value if error rate is nearing 1.

3. Same α_j value is used for updating the weights of the training examples.

$$w_j^{(i+1)} = \frac{w_j^{(i)}}{Z_i} * e^{-\alpha_i}, ifC_i(x_j) = y_j \tag{4.3}$$

$$w_j^{(i+1)} = \frac{w_j^{(i)}}{Z_i} * e^{\alpha_i}, ifC_i(x_j) \neq y_j \tag{4.4}$$

   Z_i is the normalization factor such that $\sum_j w_j^{(i+1)} = 1$.

In LogitBoost same procedure is followed with only an additional application of Least Squared cost function. If we are given a J class, N instance dataset LogitBoost algorithm would take the following steps.

Let us define $p_i(x) = P(y_i = 1|x)$ where i=1,2,...,J is the probability of given a feature set x to belong to i^{th} class.

1. We initialize all weights $w_j = \frac{1}{N}, j = 1, 2, ..., N$ and $p_i(x) = \frac{1}{J}, \forall i$
2. For each of the M base classifiers
   (a) Compute the weights and working response of the ith class,
   $$z_j i = \frac{y^*_{ji} - p_i(x_j)}{p_i(x_j)(1 - p_i(x_j))}, \tag{4.5}$$
   $$w_j i = p_i(x_j)(1 - p_i(x_j)), \tag{4.6}$$
   (b) Fit the function $f_{mi}(x)$ by weighted least squared regression.
   (c) update $F_j(x) \longleftarrow F_j(x) + f_{mi}(x)$ also update $p_i(x) = \frac{e^{F_i(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}$, where $\sum_{k=1}^{J} e^{F_k(x)} = 0$.
3. Thus the output class of the input feature set x is obtained by $max_i F_i(x)$.

Chapter 5

Proposed Method

5.1 Materials and methods

The overall block diagram of the proposed method is shown in Fig. 5.1.

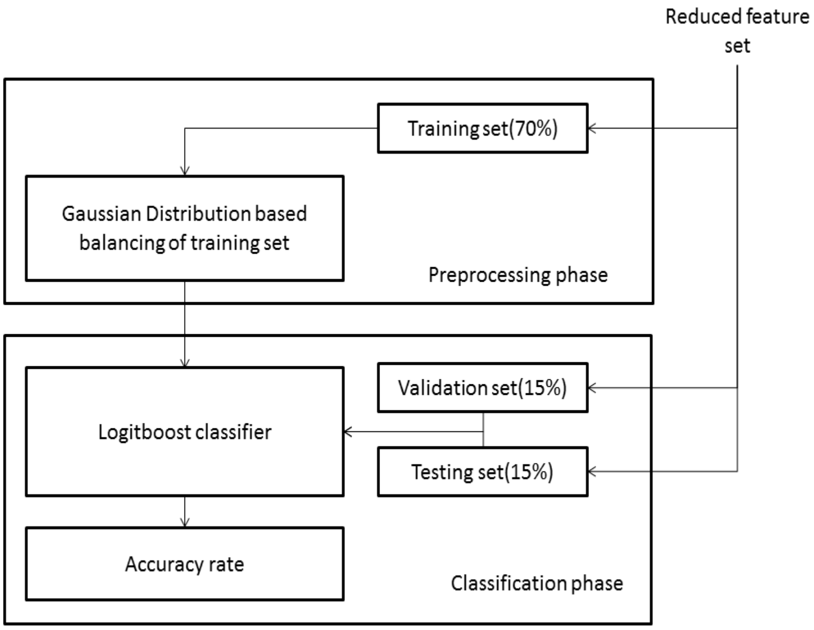


Figure 5.1: Block diagram of proposed scheme for classification of mammographic images using Gaussian distribution and LogitBoost classifier.

5.1.1 Mammogram dataset

Mammographic Image Analysis Society (MIAS) database has been used for taking the mammographic images [14]. The MIAS database consists of 322 images, which are under seven categories listed in the table below. The 322 images are divided in the following way, 207 images normal, 115 images are abnormal; and among the abnormal images 64 and 51 are the benign and malignant types respectively. Each image has the size of 1024 × 1024 pixels.

Table 5.1: Distribution of MIAS data set

Type	**Benign**	**Malignant**	**Total**
Circumscribed masses	19	4	23
Microcalcification	12	13	25
Asymmetry lesion	6	9	15
Ill-defined masses	7	7	14
Architectural distortion	9	10	19
Spiculated masses	11	8	19
Normal tissue	-	-	207
Total	64	51	322

5.1.2 Feature Preprocessing

In this thesis Gaussian Distribution has been used to balance the feature set obtained from feature selection phase, which is then fed into the classifier as training set.

The training set is manufactured from the Gaussian distribution of the original training feature set. This is obtained by finding μ_{ij} and σ_{ij} of each attribute of the instances belonging to a particular class where i is th *ith* attribute and j is the *jth* class. From this learned Gaussian distribution new instances are sampled for each class, thus removing class imbalance problem of the datasets. The following algorithm illustrates the feature preprocessing process.

Algorithm 1: Feature Preprocessing

```
Require: $feature[1:N,1:K]$, $target[1:N]$
    $K$: Total number of coefficients obtained from an
    image
    $N$: Total number of images in dataset
    $J$: Total number of classes
Ensure: $feature\_preprocess[1:R,1:K]$
    $R$: Total number of sampled instances in training set
Create two empty matrices $m[1:J,1:K]$ and $s[1:J,1:K]$
for $i \leftarrow 1$ to $J$ do
    for $j \leftarrow 1$ to $K$ do
        calculate $\mu_{ij} \leftarrow$ mean of attribute j in class i
        calculate $\sigma_{ij} \leftarrow$ standard deviation of attribute j in class i
        Set m[i,j] $\leftarrow \mu_{ij}$
        Set s[i,j] $\leftarrow \sigma_{ij}$
    end for
end for
for $i \leftarrow 1$ to $J$ do
    for $j \leftarrow 1$ to $NumOfRequiredInstances$ do
        Append new sampled instance from $N(\mu_i, \sigma_i)$ to $feature\_preprocess$
        Set $target[j] \leftarrow i$
    end for
end for
```

5.1.3 Feature Classification

The LogitBoost classifier is used to classify the reduced feature set into different classes. Since it is an ensemble classifier, The base classifier used in this case is the Random Forest classifier. During training, the training set (70% of the total dataset) is preprocessed using Gaussian distribution based balancing. The testing set (15% of the total dataset) provide with an independent evaluation of the classifier performance. During validation, the validation set (15% of the total dataset) is used to evaluate the performance of the classifier. For maximum classification accuracy rate the process is repeated with the new feature set that is with the new number of features and stops when optimum classification accuracy rate is obtained with an optimized feature set. The scheme is described in Fig. 5.1. The LogitBoost algorithm to train the classifier is given as follows :

Algorithm 2: Feature Classification

```
Require: $feature[1:N,1:K]$, $target[1:N]$
    $K$: Total number of coefficients obtained from an
    image
    $N$: Total number of images in dataset
    $J$: Total number of classes
Ensure: $classification\_functionF(x)$
    $R$: Total number of sampled instances in training set
Create two empty matrices $z[1:J,1:K]$ and $w[1:J,1:K]$
for $i \leftarrow 1$ to $M$ do
    for $j \leftarrow 1$ to $J$ do
        Calculate working response $z[m,j]$ for all instances m=1,2,...,n with class j.
        Calculate weights $w[m,j]$ for all instances m=1,2,...,n with class j.
        Fit the function $f_{ij}(x)$ by weighted least squared regression technique of
        z[m,j] to $x_m$, using the weights $w[m,j]$.
        Update function $f_{ij}(x) \leftarrow \frac{J-1}{J}\left(f_{ij}(x) - \frac{1}{J}\sum_{k=1}^{J} f_{ik}(x)\right)$
        Update $F_j(x) \leftarrow F_j(x) + f_{ij}(x)$.
        Update $p_j(x)$.
    end for
end for
The output of the classifier is given by the expression arg $max_j F_j(x)$.
```

The confusion matrix helps in evaluating the performance of the LogitBoost classifier [15]. A confusion matrix is a tabular representation showing the comparison between actual and predicted classification. The confusion matrix for two classes (benign and malignant) and corresponding measures of performance are represented in TABLES 5.2 and 5.3 respectively. Sensitivity and specificity are measures for performance evaluation which calculate the percentage of true positive rate and true negative rate respectively. An ideal performance would show both specificity and sensitivity to be high. The evaluation of a classifier performance can also be accomplished by means of receiver operating characteristics (ROC) curves [4]. It is a two dimensional graph which plots sensitivity versus false positive rate (1-specificity). The area under the ROC curve is an important factor for evaluating the classifier performance. AUC with value 1.0 shows ideal performance of the classifier.

Table 5.2: Confusion Matrix for two classes

Actual class	**Predicted class**	
	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Table 5.3: Measures of classification performance

Measure	**Definition**
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Accuracy	(TP+TN)/Total number of samples

Chapter 6

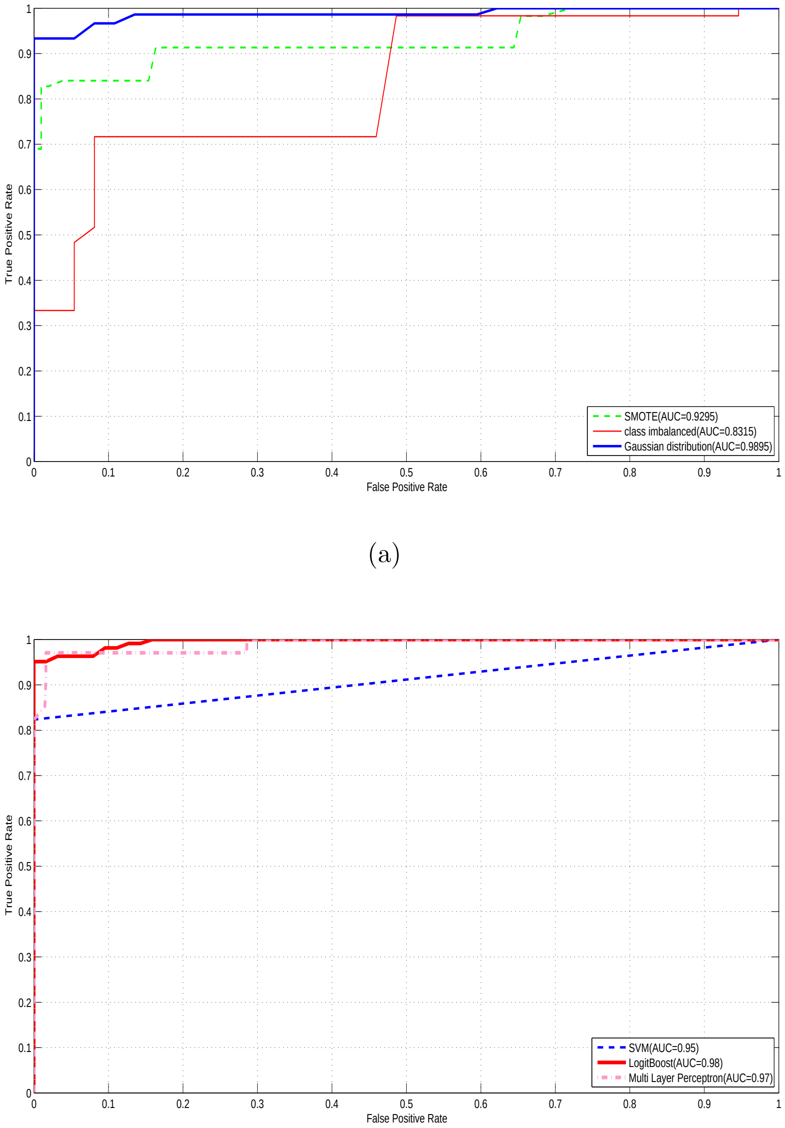
Simulation and Results

Experiments were done in MATLAB environment to validate the proposed scheme. The training set is preprocessed using Gaussian distribution based balancing, and this preprocessed training set is used to train the LogitBoost classifier. In the classifier 70% of the total set was training set and 15% is used for testing and other 15% is used for validation.

During simulation, the feature sets are selected with different dimensions and fed into LogitBoost, Support Vector Machine(SVM) and Multi-Layer Perceptron classifiers. The classification accuracy is observed to be maximum for a dimension of 130 features with LogitBoost classifier.

The maximum classification accuracy rate is found to be 99.1% by using preprocessing the 130 feature training set using Gaussian distribution and using LogitBoost classifier for classification. For the same training set, the SVM gives an accuracy rate of 96.91%.

The ROC curves for benign and malignant classes of lesion using Gaussian distribution and without using Gaussian distribution are presented in Fig. 6.1(a). For the prediction of malignant lesion in the mammogram, the Gaussian distribution based preprocessing provides a more efficient method. As shown in Fig. 6.1(b), the ROC comparison of different classifiers. Different classification performance measures computed during simulations are presented in TABLE 6.1.



(a)

(b)

Figure 6.1: ROC curves. (a) Classification of mammograms by using Gaussian distribution, SMOTE and without any feature preprocessing, (b) Classification of mammograms by LogitBoost, SVM and Multi-Layer Perceptron classifiers.

Table 6.1: Performance measures for different classifiers at different number of features.

Number of features	Performance measures (Accuracy in %)					
	LogitBoost		SVM		Multi-Layer Perceptron	
	Accuracy	A U C	Accuracy	A U C	Accuracy	A U C
30	96.93	0.95	58.96	0.5	96.9	0.97
50	97.96	0.98	86.97	0.83	96.87	0.97
70	97.96	0.98	86.97	0.83	96.87	0.97
90	98.63	0.98	88.65	0.85	96.93	0.97
110	98.96	0.98	93.81	0.90	96.87	0.97
130	99.11	0.989	96.97	0.95	96.87	0.97
150	98.63	0.98	96.97	0.95	96.93	0.97
170	98.63	0.978	95.78	0.95	96.93	0.97

Chapter 7

Conclusion and Future Work

The characterization of mammographic lesion into benign and malignant to help the decision making of radiologists is presented through a novel scheme in this thesis. The selected features are preprocessed using Gaussian distribution. This helps in getting rid of the class imbalance problem. Finally LogitBoost classifier is used for classifying the mammographic lesions into benign and malignant. The MIAS database was used to get the mammographic images on which the simulation experiments were performed. The proposed scheme achieves the AUC of 0.9895 from the ROC analysis and a 99.1% is the classification accuracy. The simulation results show that the Gaussian distribution based preprocessed features along with a LogitBoost classifier gives better accuracy than its counterparts.

Bibliography

[1] American Cancer Society. Cancer facts & figures 2013. 2013.

[2] Priya Shetty. India faces growing breast cancer epidemic. *The Lancet*, 379(9820):992–993, 2012.

[3] Wilking N. Jonsson B. Prevention, early detection and economic burden of breast cancer. Technical report, GE Healthcare, 2013.

[4] HD Cheng, XJ Shi, Rui Min, LM Hu, XP Cai, and HN Du. Approaches for automated detection and classification of masses in mammograms. *Pattern recognition*, 39(4):646–668, 2006.

[5] FRANCISCO Ballesteros, ALICIA Oropesa, L Martin, and D Andina. Mammography classification using wavelets. In *Automation Congress, 2002 Proceedings of the 5th Biannual World*, volume 13, pages 293–300. IEEE, 2002.

[6] Ping Zhang, Brijesh Verma, and Kuldeep Kumar. A neural-genetic algorithm for feature selection and breast abnormality classification in digital mammography. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 3, pages 2303–2308. IEEE, 2004.

[7] Fatemeh Moayedi, Reza Boostani, Zohreh Azimifar, and Serajedin Katebi. A support vector based fuzzy neural network approach for mass classification in mammography. In *Digital Signal Processing, 2007 15th International Conference on*, pages 240–243. IEEE, 2007.

[8] Mohamed A Alolfe, Wael A Mohamed, A Youssef, Ahmed S Mohamed, and Yasser M Kadah. Computer aided diagnosis in digital mammography using combined support vector machine and linear discriminant analaysis classification. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 2609–2612. IEEE, 2009.

[9] Xiaoming Liu, Jun Liu, and Zhilin Feng. Mass classification in mammography with morphological features and multiple kernel learning. In *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*, pages 1–4. IEEE, 2011.

[10] SMT Javadi and K Faez. Finding suspicious masses of breast cancer in mammography images using particle swarm algorithm and its classification using fuzzy methods. In *Computer Communication and Informatics (ICCCI), 2012 International Conference on*, pages 1–5. IEEE, 2012.

[11] Aijuan Dong and Baoying Wang. Feature selection and analysis on mammogram classification. In *Communications, Computers and Signal Processing, 2009. PacRim 2009. IEEE Pacific Rim Conference on*, pages 731–735. IEEE, 2009.

[12] Xi-Zhao Li, Simon Williams, Gobert Lee, and Min Deng. Computer-aided mammography classification of malignant mass regions and normal regions based on novel texton features. In *Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on*, pages 1431–1436. IEEE, 2012.

[13] Sidney ML de Lima, Abel G da Silva-Filho, and Wellington Pinheiro dos Santos. A methodology for classification of lesions in mammographies using zernike moments, elm and svm neural networks in a multi-kernel approach. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 988–991. IEEE, 2014.

[14] John Suckling, J Parker, DR Dance, S Astley, I Hutt, C Boggis, I Ricketts, E Stamatakis, N Cerneaz, Siew-Li Kok, et al. The mammographic image analysis society digital mammogram database. 1994.

[15] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30(2-3):271–274, 1998.