Chapter 3

Classification

In this chapter, we will make a comprehensive survey of our whole diagnosis system first. In the following sections, the techniques of constructing a probabilistic classifier will be brought out such as Bayes' Theorem and Parzen windows. In last section, an accuracy evaluation, leave-one-out cross validation, will be shown and is the basis for the robustness of the results.

3.1 Framework of Computer-Aided Diagnosis System

As mentioned, there are several classification models in our proposed evaluation system. Here, we introduce the process of a single model where Figure 3.1 illustrates the whole procedure constructed by this work. The entire procedure can be mainly divided into two parts: feature selection, accomplished by a ROI selection and principal components selection mentioned in the Chapter 2, and classification to train a classifier of a specific disease, introduced in this chapter. Detailed techniques are summarized in the following steps.

1. **Voxel-based morphometry**

   The main idea of this process is to detect some useful features for post-processing. Once having MR images of normal subjects and patients with the same disease, we apply a VBM analysis on all patients and proper normal subjects who are picked out from all collected normal subjects and have age-matched and gender-matched properties with patients. These selected images are then segmented into GM, WM and CSF partitions, separately normalized to customized GM, WM and CSF templates, modulated to restore volume changes and smoothed using an isotropic Gaussian kernel. Later, a voxel-based statistical analysis is applied to find out significant brain structural differences between selected subjects. A t-test map is then produced and reveals the significance of each voxel. Thus, a mask is built up by setting a threshold to select voxels whose absolute T value is larger than it. In this work, we set

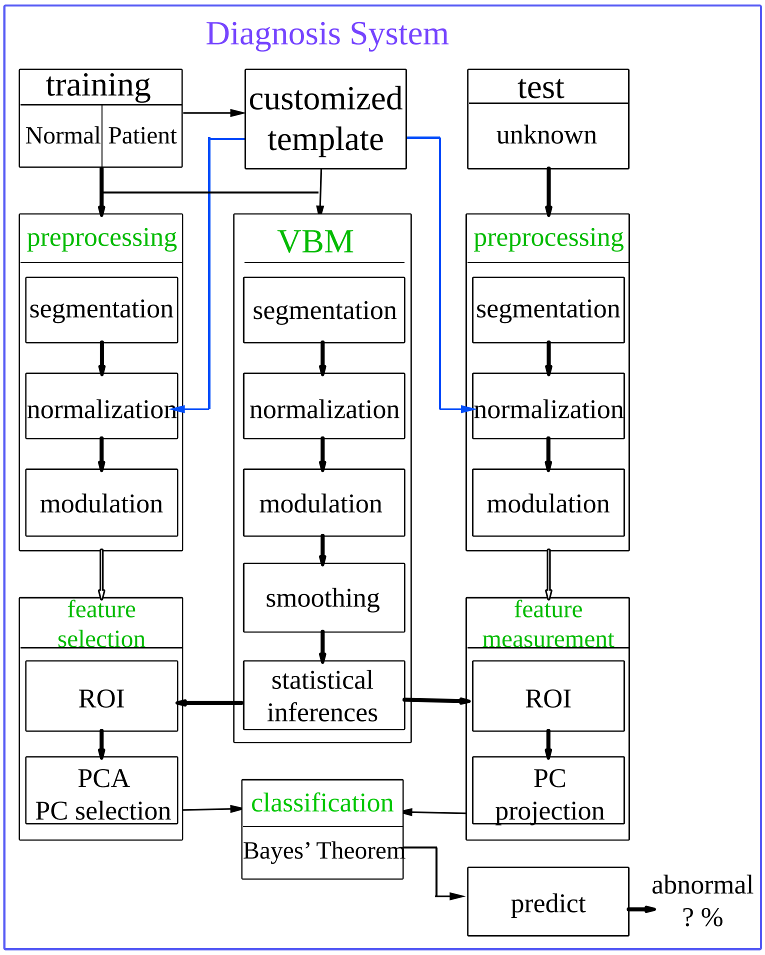

Figure 3.1: **Overview of computer-aided system.** Initially, a voxel-based morphometric analysis is taken between normal and abnormal groups. A registered space, called template, and a t-map are obtained. All training data set are transformed into the registered space and then are applied ROI selection according to the t-map and PCA in order to create a new classification model. Also, an unknown sample is transformed into the registered space and projected into the classification model. Therefore, a prediction can be made by computing the posterior probability of the unknown sample.

the threshold value T manually according to the height threshold of the statistical analysis from SPM2. In short, three customized templates and three masks are separately created to construct an identical space and to select those useful features for comparisons according to GM, WM and CSF tissue.

2. **ROI selection**

   We can view three customized templates and three masks as three individual models along with their nature. That is, there are a GM model, a WM model and a CSF model. As soon as these models are set up, MR images of all subjects are segmented into GM, WM and CSF images with segmentation technique and are put into corresponding models. For each model, input images are registered to the customized template to be in the same space, modulated with SPM2 software to correct volume changes and masked with the mask to choose proper voxels as features for classification. Thus, there are three feature sets which are based on GM, WM and CSF partitions and used to establish three separate classifier, named as GM classifier, WM classifier and CSF classifier. From now on, all of three feature sets are processed parallelly.

3. **Principal component analysis**

   The primary idea of this step is to reduce the dimensionality of feature set and to pick out more useful features for classification. Principal components of the feature set are found by applying principal component analysis and form a new projection coordinate for comparisons. However, the new space composed of all principal components may not be a good measure to differentiate normal and abnormal groups. So, we select some principal components to create a classification space instead of using all found principal components. Variance-based and significant-based principal component selections are two methods proposed in our work to find proper principal components which form more differentiable space. Once the space is built up, each data point in the features set is transformed into the space and represented with fewer random variables.

4. **Classification**

   Once a feature extraction or a feature selection finds proper representations for a data set, a classifier can be designed using some possible techniques. In our thesis, a classifier was built up based on a probabilistic approach. We applied Bayes' theorem with a nonparametric density estimation technique, Parzen windows, to our construction of computer-aided diagnosis system. A Parzen-window approach is to estimate the probability density function of normal group and abnormal group in the data set. Thus, it is easy to find the class-conditional probability of an observation whose nature is unknown. Moreover, we decide a prior, a parameter in Bayes formula, according to the ratio of the number of patients to the number of total samples in the training data set. Therefore, posterior probability of the unknown sample is computed by Bayes' theorem and then is predicted.

5. **Combination of GM, WM and CSF classifiers**

   Steps from two to four are separately done in three models. In other words, a subject is diagnosed with GM, WM and CSF classifiers. For each classifier, it produces a probability for a subject that presents the possibility of being abnormal according to a particular tissue. In order to make a final prediction, we combine the results from three classifiers by choosing the maximum probability of being abnormal rather than emphasize the result of a specific classifier. Moreover, we found that three classifiers were not identically independent because the loss of tissue A might lead to the increase of tissue B. However, it was not absolute in every classification model because the loss of tissue A might result from the growth of both tissues B and tissue C. In short, we put equal emphasis on all classifier. Figure 3.2 illustrates the combination method with simple drawings. Finally, the system provides a physician and a test subject with only a probability that represents a reference target to fall ill.

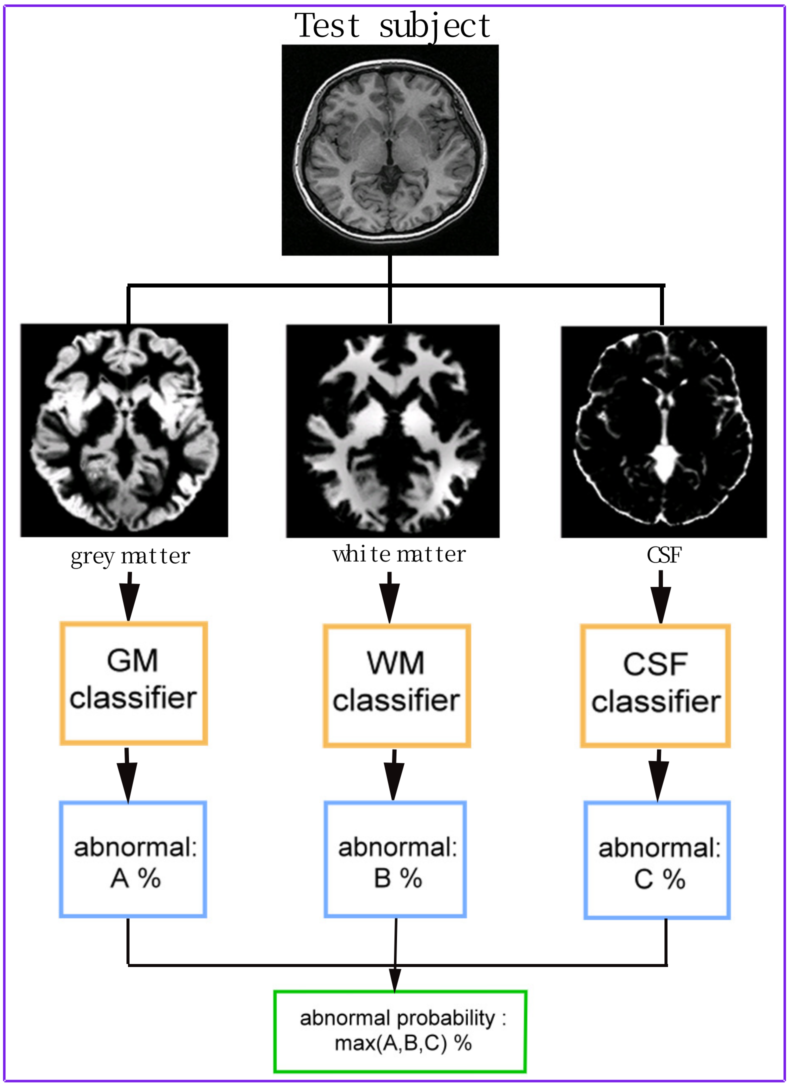

Figure 3.2: **Combination of individual classifiers.** A MR image of test subject is segmented into GM, WM and CSF partitions. These images are then normalized to corresponding customized templates, modulated to correct volume changes, masked with relative masks and inputted into different classifiers according to their natures. Thus, outputs from each classifier are combined to produce a final result.

In our thesis, we only consider two categories classification: normal group and abnormal group, a specific disease. It means that all people in abnormal group suffer from the same illness and all people in normal group are not attacked by this disease. In addition, all these clinical diagnosis of our training data are made by physicians. Thus, our evaluation system is parallelly composed of many classification models to provide probabilities of many disorders instead of making decisions.

3.2 Bayes' Theorem

Bayes' theorem is a fundamental statistical approach to the problem of pattern classification. This approach makes the assumption that the classification problem is posed in probabilistic views. It involves in the class-conditional probability distributions of random variables and a prior of each class. Bayes' theorem is named by Laplace after Thomas Bayes, an English clergyman who set out his theory of probability in 1764 [27]. Unfortunately, Thomas Bayes was not famous for his mathematical works during his lifetime.

The classification problem can be described as follows. Suppose that there are c categories $\{g_1, ..., g_c\}$ with corresponding priors $\{P(g_\mathbf{1}), ..., P(g_\mathbf{c})\}$. A prior of a category represents a probability for a new sample to be a member of the category and is often obtained by statistics. Let x be an observed pattern with d feature values, that is, $\mathrm{x} = (x_1, ..., x_\mathbf{d})$. Thus, the probability density function for x given that the nature is g_i is expressed as $p(\mathrm{x}|g_i)$, so-called the class-conditional probability density or mass function depending on whether the features are continuous or discrete.

Initially, the joint probability density of discovering a pattern in class g_i with feature value x can be computed in two ways: $P(g_i|\mathrm{x})p(\mathrm{x}) = p(\mathrm{x}|g_i)P(g_i)$. After rearranging, we have Bayes formula as follows:

$$P(g_i|\mathrm{x}) = \frac{p(\mathrm{x}|g_i)P(g_i)}{p(\mathrm{x})}, \tag{3.1}$$

where $P(g_i|\mathrm{x})$ is called as a posterior probability and means the probability of a given feature value $\mathbf{x}$ belonging to a class g_i. In other words, Bayes formula helps us convert the prior probability to a posterior probability with respect to an observed sample. Furthermore, $p(\mathbf{x}|g_i)$ is also called the likelihood of g_i with respect to $\mathbf{x}$. $p(\mathbf{x})$ is expressed as

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|g_j)P(g_j), \tag{3.2}$$

and considered as a scale factor for assurance that the posterior probabilities sum to one. Once having the posterior probabilities of a test subject with respect to all categories, we can make a decision that the test subject should be in the class with largest posterior probability. We only had two categories $\{g_1, g_2\}$ in each classification model. Thus, our decision rule could be expressed as

$$\text{Decide } g_1, \text{ if } P(g_1|\mathrm{x}) > P(g_2|\mathrm{x}); \text{ otherwise decide } g_2. \tag{3.3}$$

In this study, all T1-weighted MR images are segmented into GM, WM and CSF images. For each tissue, a corresponding classification model is constructed and determines a posterior probability of a test image. Thus, there will be three probabilities for each test subject to be abnormal. We thought of all probabilities at the same time and selected the maximum of three to have the final probability to be an index and the final decision, a suggestion made in accordance with the index. Lastly, the test subject will be informed how many possibility he or she has to fall ill.

Up to now, it seems that it is easy to work out posterior probabilities. However, class-conditional densities are often unknown in practice and make the problem more difficult. Fortunately, there are some density estimation techniques to estimate an unknown probability density function of an observed value. According to the prior knowledge of training data, these techniques can be separated into two ways: supervised learning and unsupervised learning, also called as clustering. In supervised learning, the number of categories is known and training data are all labeled. In contrast with supervised learning, both of the two properties are unknown in unsupervised learning. A solution of supervised learning problem is brought up in next section and solutions of clustering problem are skipped due to out of scope of this thesis.

In this work, we collected the features of all subjects in a transformed space after applying ROI selection and PCA to the original data set. The data is originally separated into a normal group and an abnormal group by physicians in clinical diagnosis. Also, we have no idea about the density distributions of both groups. It is a typical problem of supervised learning. A nonparametric method was used in our thesis to solve this problem and helped us succeed in calculating posterior probabilities.

3.3 Parzen Windows

As mention in Chapter 1, estimation techniques of probability density function are divided into two categories: parametric estimation and nonparametric estimation in supervised learning problem. Parametric estimation techniques are referred that the shape of a distribution is known so that the problem is simplified to estimate the parameters of distribution instead of estimating an unknown density function. Two common and reasonable methods are often implemented, namely, maximum-likelihood estimation and Bayesian estimation. On the other hand, only labeled data without shape information of probability distributions are known, that is, nonparametric estimation methods. For example, Parzen windows, proposed by Parzen in 1962 [28], is a well-known nonparametric estimation technology.

The basic concepts of density estimation is that a density in a region is proportional to the amount of data in this region, shown in Figure 3.3. It can be expressed as

$$\mathbf{p}(\mathbf{x}) \approx \frac{\mathbf{k}/\mathbf{n}}{\mathbf{V}}, \tag{3.4}$$

where $\mathbf{x}$ is a sample inside a region R, $\mathbf{n}$ is the total number of samples, $\mathbf{k}$ is the number of

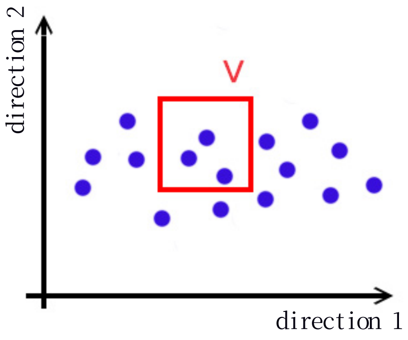

Figure 3.3: **A diagram of density estimation.** This graph shows the basic concepts of density estimation. The number of samples inside a specific region, V, is computed to estimate the density in this region. In this method, the density of whole group could be calculated.

samples inside the region R, and $\mathbf{V}$ represents volume of region R. There are two ways to improve the accuracy of this density approximation. One is to increase the size of $\mathbf{n}$, and the other is to decrease the size of region R. Nevertheless, the number of samples is always fixed in reality. Thus, the only available method to enhance accuracy is by shrinking the size of R. Notice that the size of region R should be chosen carefully because there may be no samples when the size of R is too small or may be a lot of samples so that $\mathbf{p}(\mathbf{x})$ is approximately constant inside the region R if the size of R is too large. Parzen-window approach is based on this assumption and chooses a fixed value for the size of region R to determine a probability density function of a data set.

In Parzen-window approach, the size and shape of a region R is fixed and then the number of samples inside region R is determined. Assume that the region R is a d-dimensional hypercube with side length $\mathbf{h}$ so that the volume of R is $\mathbf{h}^d$. To estimate the density at a point $\mathbf{x}$, the region R is centered at $\mathbf{x}$ and the number of samples inside R can be counted. Let us define a *window function* to represent whether a point is inside the region R or not with an analytic expression. Suppose that $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ are the known data set and the window function is defined as

$$\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{\mathbf{h}}) = \begin{cases} 1, & \forall\, |\mathbf{x}_j - \mathbf{x}_{ij}| \le \frac{\mathbf{h}}{2}, \mathbf{j} = 1, ..., d, \\ 0, & \text{otherwise} \end{cases} . \tag{3.5}$$

The window function $\varphi(y)$ satisfies that $\varphi(y) \ge 0$ and $\int \mathbf{p}_{\varphi}(\mathbf{u})\mathbf{du} = 1$. Thus, the number of samples inside the region *R* is

$$\mathbf{k} = \sum_{i=1}^{n} \varphi(\frac{\mathbf{x}-\mathbf{x}_i}{\mathbf{h}}). \tag{3.6}$$

By substituting Eq. 3.4 with Eq. 3.6, the analytical expression for the estimate of density is written as

$$\mathbf{p}_{\varphi}(\mathbf{x}) = \frac{1}{\mathbf{n}} \sum_{i=1}^{n} \frac{1}{\mathbf{h}^d} \varphi(\frac{\mathbf{x}-\mathbf{x}_i}{\mathbf{h}}). \tag{3.7}$$

Also,$\mathbf{p}_{\varphi}(\mathbf{x})$ satisfies that $\mathbf{p}_{\varphi}(\mathbf{x}) \ge 0$ and $\int \mathbf{p}_{\varphi}(\mathbf{x})\mathbf{dx} = 1$ for all $\mathbf{x}$.

However, the resulting density is not smooth because all $\mathbf{x}_i$ inside the region *R* centered at $\mathbf{x}$ have the same contributions to the density at $\mathbf{x}$ no matter how close $\mathbf{x}_i$ is to $\mathbf{x}$. Therefore, using a general window function such as Gaussian distribution can solve drawbacks of hypercube window function. A general form of *N*-dimensional Gaussian distribution is

$$f(\mathbf{x}) = (2\pi)^{-\frac{N}{2}} \det \Sigma^{-\frac{1}{2}} \exp\,(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)), \tag{3.8}$$

where μ and Σ are the mean and covariance of the distribution. As a result, the estimative density is smooth. Notice that there is a limitation in computing $f(\mathbf{x})$ due to mathematical considerations. Once the dimensionality *N* exceeds in 810, the value of probability density function will be zero and fail to be estimated.

Finally, the accuracy of density estimation is involved in window width $\mathbf{h}$. Let us consider the term, $\frac{1}{\mathbf{h}^d}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{\mathbf{h}})$ in Eq. 3.7, and it is named as the function $\delta(\mathbf{x})$. It is clear that $\mathbf{h}$ affects both the amplitude and the width of $\delta(\mathbf{x})$. If $\mathbf{h}$ is very large, the amplitude of $\delta(\mathbf{x})$ is small and the width of $\delta(\mathbf{x})$ is extensive. So, $\mathbf{p}_{\varphi}(\mathbf{x})$ is a slowly changing function and looks oversmoothed. On the other hand, if $\mathbf{h}$ is small, the amplitude of $\delta(\mathbf{x})$ is large and the width of $\delta(\mathbf{x})$ is narrow. Therefore, $\mathbf{p}_{\varphi}(\mathbf{x})$ is a rapidly changing function and looks noisy. Thus, a proper window width $\mathbf{h}$ should be found to have a more accurate density estimation from the training data. In this work, we proposed an efficient method to find a suitable value of window width h shown in Chapter 5.

In our thesis, we implemented multivariate Gaussian distribution as a window function and approximated the estimation to real density distribution with N(0,1), which is a normal distribution with zero mean and identity covariance matrix. Except for the limitation of probability density function, some computation problem also exist and run out of memory on our PC equipped Windows XP with a processor 3.20GHz and 2GB RAM. We have inspected that the dimensionality of data process should be no more than 188600 after ROI selection to avoid lack of memory.

3.4 Accuracy Evaluation

The classification accuracy of our system is evaluated by comparing predictions to the preoperative clinical diagnosis with cross-validation techniques. The theory of cross-validation was posed by Seymour Geisser. It is one of popular approaches to estimate how well the classification model learned from training data is going to perform on future unknown data. The basic idea of cross-validation is not to use entire data set when training a classifier and to partition whole data set into two groups. Thus, one group is considered as the training data for learning and the other group is used as the test data, viewed as new and unknown data set to examine the performance of the learned model. In this work, a leave-one-out cross-validation are implemented and introduced as follows.

In leave-one-out cross-validation, the original data set is divided into K subsets where K equals to the number of samples in the data set. A single sample of all samples is selected to be a test point and the others are retained as the training data to train a classifier. This procedure is repeated N times, the number of data points in the set, such that each sample is used once as the validation data. Finally, classification accuracy is then obtained by averaging the N. Generally speaking, the performance by leave-one-out cross validation is good but it costs heavy computation. Fortunately, there are some efficient methods in some cases to reduce computation [29, 30].

Because the clinical diagnosis of data set is known, we can use the ground truth of these diagnoses to verify the results from our prediction models. In our experiments, the ground truth is made by professional physicians. After making predictions on subjects with a classification model of a specific disease, the ground truth is performed to find the regions of true-positive (TP), false-positive (FP), true-negative (TN) and false-negative (FN). The meaning of each region is listed as below.

- **True-positive (TP)**: the result predicts that a test subject is attacked by a specific disease because of higher possibility of being abnormal and the test subject is a patient with the specific disease in reality.
- **False-positive (FP)**: the result predicts that a test subject is attacked by a specific disease because of higher possibility of being abnormal but the test subject is a healthy person in reality. This phenomenon is so-called as false alarm.
- **True-negative (TN)**: the result predicts that a test subject is healthy because of lower possibility of being abnormal and the test subject is healthy in reality.
- **False-negative (FN)**: the result predicted that a test subject is healthy because of lower possibility of being abnormal but the test subject is attacked by a specific disease in reality. This phenomenon is so-called as misdetection.

Table 3.1 clarifies brief definitions of TP, FP, TN and FN. The rates of TP, FP, TN and FN are defined as

$$TPrate = \frac{TP}{TP + FN}, \tag{3.9}$$

Table 3.1: **Interpretations of TP, FP, TN, and FN.** The items of predicted and actual show the results from a classification model and clinical diagnosis respectively. The item of abnormal represents that a subject is attacked by a specific disease. The term of FN is also called as misdetection and the term of FP is so-called as false alarm.

Predicted Actual	Abnormal	Normal
Abnormal	TP	FN (misdetection)
Normal	FP (false alarm)	TN

$$FPrate = \frac{FP}{TN + FP}, \tag{3.10}$$

$$TNrate = \frac{TN}{TN + FP}, \tag{3.11}$$

$$FNrate = \frac{FN}{TP + FN}. \tag{3.12}$$

The system will fail in two respects FP and FN. In statistics, FN is viewed as type I error and FP is called as type II error. We can assess the performance of our system by using the receiver operating characteristic (ROC), a graphical plot of the relations between sensitivity and (1 - specificity) for a classifier system as its discrimination threshold is varied. In our work, the sensitivity means the ability of a classification model to predict that a test subject is actually attacked by a particular disease and it is the same as the TP rate. The specificity which shows the ability to predict that a test subject does not actually fall ill is the same as the TN rate. Hence, the term, (1 - specificity), in the ROC curve is equal to the FP rate. Figure 3.4 shows a general form of a ROC curve. In this study, the varied parameter is the variance ratio of selected principal components in both of variance-based PC selection method and significant-based PC selection method.

In general, when the curve of a method is close to the top left corner, the performance

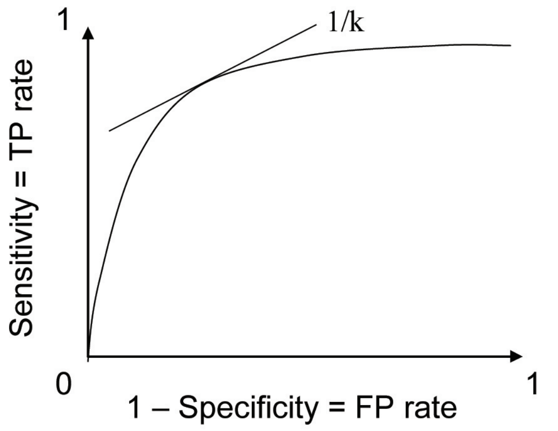

Figure 3.4: **A general form of a ROC curve.** This figure is obtained by varying the value of a parameter in a method and then plotting sensitivity (TP rate) against 1-specificity (FP rate) under this parameter value. In general, methods with ROC curves closer to the top left corner express better performances. Besides, when we define the risks of a misdetection and a false alarm, a line with a slope of $1/k$ can be depicted. The intersection point of the curve and the line represents the best performance of the system with the lowest total cost.

of the method is more accurate with a good parameter. Moreover, the cost of misdetection is quite different from that of false alarm in a CAD system. A misdetection may lead to one's death but a false alarm may just cost one some money to have more and detailed examinations. Thus, the risk of a false alarm is defined as 1 and the risk of a misdetection is defined as k where $k \geq 1$. Therefore, a total cost of a classifier could be defined as

$$Cost = k \times FNrate + 1 \times FPrate \tag{3.13}$$

and is expected as small as possible. Once the best efficiency with the lowest total cost, the intersection point shown in figure 3.4, is found, the optimal parameters of a classifier are hence obtained.

The area under the curve (AUC) which represents the total amount of correct identification is one of the common summary indices to quantify several ROC curves. However, the ROC curves in our work were not informative enough to provide the whole area under the curve. Thus, a partial area under the curve (PAUC) was used to compare different curves in our work. We defined a specific region of the ROC curve and then computed it. Strictly speaking, the computed areas of AUC and PAUC are dissimilar but their concepts are the same that a larger AUC or PAUC index indicates a better performance of the method.