Human Sperm Health Diagnosis With Principal Component Analysis and K-nearest Neighbor Algorithm

Jiaqian Li[1], Kuo-Kun Tseng[1*], Haiting Dong[1], Yifan Li[1], Ming Zhao[1] and Mingyue Ding[2]
[1]Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen Graduate School, China
[2]Huazhong University of Science and Technology, China

***Abstract*—Sperm morphology is an important diagnostic basis to identify if a sperm cell is healthy or not. This paper presents a method that using principal component analysis (PCA) to extract image features and k-nearest neighbor (KNN) algorithm to diagnose sperm health. We first accurately locate the position of sperm in the microscope images, and segment some small sperm division with a fixed size. Then some of divisions are selected as the training set to classify the remaining small sperm divisions. In this experiment, while the diagnosis accuracy depends on the training set, we have already selected a better training set and obtained a good performance with 87.53% compared with other feature extraction methods such as scale-invariant feature transform (SIFT) and other classifier such as backpropagation neural network (BPNN).**

Keywords—sperm morphology; health diagnosis; principal component analysis; linear discriminant analysis

I. INTRODUCTION

With the development of modern computer technology, medical imaging has played an important role in clinical diagnosis and treatment. More and more medical images, such as electrocardiogram (ECG), antinuclear antibody image and brain waves, have been applied to diagnose diseases. Certainly, the sperm image, which is got under a microscope, is a kind of medical image and is also a valuable diagnosis criterion for some male diseases. Moreover, since people have become more and more concerned about the health of the next generation, it is essential to inspect the health of the sperm. For experts and doctors, the sperm morphology is the most intuitive evidence to assess whether the sperms are healthy or not.

Medical image automatic diagnosis is facing the challenges of precisely extracting information from the medical image and how to accurately diagnose diseases. Low diagnostic accuracy is the important reason for the medical image automatic diagnosis is difficult to promote in the hospital. Sperm image diagnosis is no exception due to the numerous types of sperm shape. Moreover, few studies automatic image diagnostic based on the microscope sperm image, which also is a difficult for us since there were no precedents. Currently, sperm quality assessment is mostly judged by experts and doctors. Because of the numerous types of sperm shape, the efficiency and accuracy relying on human assessment is not ideal. As computer morphology technology develops, quantitative analysis of sperm morphology is demanded to assist doctors in some diseases diagnoses. Thus, this research is intended to find a helpful sperm classification mechanism with better performance.

The aim of this paper is to investigate the probability that automatic diagnosing sperm health or not based on specified feature extraction method and classification method in the microscope sperm image database. In particular the following points will be investigated in the present study:

1) Precisely segment the small sperm divisions. We proposed an available automatic system to accurately position all the sperms in the microscope image and segment the sperm divisions with a fixed size. In this system, we mainly use the knowledge of image morphology, such as erosion, expansion, morphological orientation.

2) Extract PCA features of sperm image for identification. In this study, we extract the PCA features from the small sperm divisions, and then use k-nearest neighbor algorithm to implement the auto-classification of sperm health or not.

3) Compared feature extraction and classification methods. Except for PCA features, we also extract such features as SIFT as the control groups in our experiments. And we also use BPNN for classification.

The rest of this paper comprises four parts. In section II, we introduce some related studies about sperm morphology and image feature extraction methods. Section III presents the used feature extraction methods, that is, PCA feature and SIFT feature. Section IV provides the steps of the sperm health diagnosis, which consists of sperm image segmentation, feature extraction using PCA and classification using k-nearest neighbor algorithm. And the experimental results and comparison will be introduced in Section V. Finally the last part is the acknowledgment and references.

II. RELATED STUDIES

A. Sperm Morphology

Sperm morphology has long been used to evaluate male fertility potential [1]. The percentage of normal forms has been shown to correlate with some diseases [2, 3]. However, the definition of "normal" sperm morphology has never been empirically derived with respect to fertility endpoints. Instead, abnormal forms have been cytologically defined, a priori, as cells that exhibit gross defects, such as: two heads, no head, incorrect insertion of the midpiece, or a broken flagellum, and by sperm head sizes and shapes that deviate from an aesthetically pleasing, ideal oval form[4, 5]. But, here we focus on studies that identifying the sperm health or not, instead of the sperm morphology is normal or not. Because the sperm

978-1-4799-4013-4/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMB.2014.26

health is automatically classified by technology of classification and identification, we do not need to make specific and accurate categories of sperm morphology. But sperm morphology types and classification methods are still worthy of our study in order to better determine sperm health status.

Moruzzi JF et al. [6] proposed a quantitative, semi-automated method for classifying human sperm based on objective measurements of head shapes and sizes has been developed. Measurements included stain content, length, width, perimeter, area, and arithmetically derived combinations. The classification procedure distinguished normal from abnormal sperm with 95% accuracy and correctly assigned 86% of the sperm to one of 10 shape classes. The classification results demonstrate the ability of automated image analysis to classify individual sperm into clinically familiar shape categories.

Davis RO et al. [7] try to evaluate the accuracy and precision of the CellForm-Human (CFH) automated sperm morphometry instrument. At its present stage of development, the CFH instrument exceeds the accuracy and precision of most manual approaches. With improvements in sperm recognition and type classification algorithms, it could significantly improve the reliability of morphology assays in clinical and research laboratories. Yániz, JL et al. [8] develop a new method based on fluorescence microscopy and image analysis for the automatic assessment of sperm morphometry. In this paper, the sperm morphology study demonstrated that of the total number of spermatozoa analyzed (19,000) 96.58% of the cells were normal. Moreover, 2.02% of the abnormalities were of the head, 0.24% of the midpiece, and 1.09% of the tail. These studies demonstrate that it is feasible to analyze the sperm health status based on the sperm morphology features.

B. Image Feature Extraction

Feature extraction[9] is a concept in computer vision and image processing. It is based on the subject of computer vision and image processing, using computer technology to determine the invariant information in the image and extracting them to handle the actual problem, such as clustering, classification and identification. Based on current research, there are four common kinds of features, that is, color feature, texture feature, shape feature and spatial feature.

1) Color Feature: Color feature is a global characteristic,describing the surface properties of image or image area. General color feature is based on the characteristics of the pixel, that is, all of the pixels in the image or image area has its own contribution. As color is not sensitive to orientation, size etc. of the image or image area, so the local features of the image are not well captured by color feature extraction methods. But it is not affected by changes in image rotation and translation, even sometimes image scale changes. Color feature is generally consisted of color histogram, color set, color moment and color ploymerization vector etc.

2) Texture Feature: Texture feature is an intrinsic and global characteristic, which describes the surface nature of the scene corresponding to a specified image or image area. Compared with color feature, texture feature is not a kind of feature based on pixels, which need statistical calculations on the more than one pixels belonging to the region. As a statistical feature, texture feature, often with rotation invariant characteristics, has a strong resistance capability for noise. However, it also has its drawbacks, and one obvious drawback is that when changing the image resolution, the calculated texture feature may has larger deviations. Moreover, sometimes it may be affected by light and reflection. Texture feature extraction methods can be divided into statistical method, structure method and spectrum method, etc. Sometimes, the valuable feature information may be in the frequency domain. So the spectrum method, such as fourier transform and wavelet transform, can help us capture the feature information.

3) Shape Feature: Since the same object may have a variety of different color but similar shape, a lot of queries may be aimed at the shape of the image instead of the color of the image. There are two kinds of methods to present shape features, one is contour feature, and another is regional characteristics. Therefore, shape feature extraction methods should also be divided into feature extraction based on contour (boundary) and feature extraction based on region. But, shape feature is lack of model, and has high requirement about computation and storage.

4) Spatial Feature: Spatial relation refers to the relationship of space position or relative direction between multiple target that are formed by image segmentation. It can be divided into adjacency relationship, overlapping relationship and inclusion relationship, etc. Generally, space information has two categories: relative space position information and absolute space location information. The former relation stresses the relative situation between targets. The latter stress the distance and orientation between the targets. Spatial feature can improve the distinguish ability to image content, but it is often sensitive to rotation, inversion and scale change of the image or targets.

III. Feature Extraction Algorithms

Principal component analysis (PCA) [10, 11] and linear discrimination analysis (LDA) [12, 13] are two powerful tools used for dimensionality reduction and feature extraction in most of pattern recognition applications. Here we consider the image after dimensionality reduction as the PCA feature of the original image. And the SIFT [14-16] feature is used as the compared features. Next, a brief introduction will be given.

A. Principal Component Analysis

Let $\{x_1, x_2, \dots, x_N\}$, $x \in R^n$ be N samples from L classes $\{\omega_1, \omega_2, \dots, \omega_L\}$, and $p(x)$ is their mixture distribution. In a sequel, it is assumed that a priori probabilities $p(\omega_i), i = 1,2, \dots, L$, are known. Consider m and Σ denote mean vector and covariance matrix of samples respectively. PCA algorithm can be used to find a subspace whose basis vectors correspond to the maximum variance directions in the original n dimensional space. PCA subspace can be used for presentation of data with minimum error in reconstruction of original data.

Let Φ^p denote a linear $n \times p$ transformation matrix that maps the original n dimensional space onto a p dimensional feature subspace where $p < n$. The new feature vectors $y_i \in R^p$ are defined by:

$$y_i = (\ \Phi^p\ )^t x_i\ \ , i = 1, 2, \dots, N \tag{1}$$

It is easily proved that if the columns of Φ^p are the eigenvectors of the covariance matrix corresponding to its p largest eigenvalues in decreasing order, the optimum feature space for the representation of data is achieved. The covariance matrix can be estimated by:

$$\widehat{\Sigma} = \frac{\sum_{i=1}^{N}(x_i - \widehat{m})\,(x_i - \widehat{m})^t}{N-1} \tag{2}$$

where m in (2) can be estimated by:

$$m_k = m_{k-1} + \partial(x_k - m_{k-1}) \tag{3}$$

where m_k is the estimation of mean value at *k-th* iteration and x_k is the *k-th* input image. PCA is a technique to extract features effective for representing data such that the average reconstruction error is minimized. In the other word, PCA algorithm can be used to find a subspace whose basis vectors correspond to the maximum variance directions in the original n dimensional space. PCA transfer function is composed of significant eigenvectors of covariance matrix. The following equation can be used for incremental estimation of covariance matrix:

$$\Sigma_k = \Sigma_{k-1} + \partial_k(x_k x_k^t - \Sigma_{k-1}) \tag{4}$$

where Σ_k is the estimation of the covariance matrix at *k-th* iteration, x_k is the incoming input vector and ∂_k is the learning rate.

In addition, we use the singular value decomposition (SVD) to implement the principal component analysis. SVD is that a matrix $G \quad R^{N\times M}$ may be decomposed as:

$$G = U\Sigma V^* \tag{5}$$

where $U \quad R^{N\times N}$ is the matrix of left singular vectors.

$$U = [U_1, U_2, \dots, U_N] \tag{6}$$

$V \quad R^{M\times M}$ is the matrix of right singular vectors.

$$V = [V_1, V_2, \dots, V_N] \tag{7}$$

$\Sigma \quad R^{N\times M}$ is the matrix of singular values.

$$\Sigma = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}, \mathrm{S} = \mathrm{diag}(\delta_1, \delta_2, \dots, \delta_r) \tag{8}$$

$$r = min(N, M)\,, \delta_1 > \delta_2 > \cdots > \delta_r > 0 \tag{9}$$

Note that U and V are orthonormal unitary matrices containing orthonormal vectors.

B. Scale-invariant Feature Transform

David G. Lowe [14] summed up the existing feature detection method based on invariants technology in 2004, and formally proposed an image scaling, rotation, even affine transformation for invariant image with local feature description operator based on scale space SIFT. The SIFT algorithm first undertakes feature detection in scale space and defines the key points' positions and the scale of the key points, and then uses the main direction of the neighborhood gradient of these key points as the direction features of the points in order to achieve the operator independence of scale and the direction. The format which produces the scale space is as follows:

$$L(x, y; t) = \int_{\xi=-\infty}^{\infty}\int_{\eta}^{\infty} \frac{1}{2\pi t} e^{-\frac{\xi^2+\eta^2}{2t}} f(x-\xi, y-\eta)\, d\xi d\eta \tag{10}$$

In format (10), t represents the scaling parameter. By undertaking convolution in the whole domain with a two-dimensional Gaussian kernel and input image, we can achieve scaling corresponding to t.

The SIFT feature vector has the following features:

- It is the local feature of an image which maintains invariance not only on rotation, scale and brightness variation, but also on the viewing angle, the affine transformation and the noise.
- It is distinctive and informative, suitable for fast, accurate matching in a mass signature database.
- It can produce a large number of SIFT feature vector with few objects.
- It is high speed.

The Matlab source code of SIFT is from http://www.vlfeat.org/index.html [15]. We used the SIFT method to extract feature points of all sperm images, and then use its own matching method to classification.

IV. Proposed Sperm Health Diagnosis Method

The proposed sperm health diagnosis method consists of three parts: sperm image segmentation, feature extraction by PCA and classification by k-nearest neighbor algorithm. Finally experimental results will be compared between PCA feature and SIFT feature, KNN classifier and BPNN classifier.

A. Sperm Image Segmentation

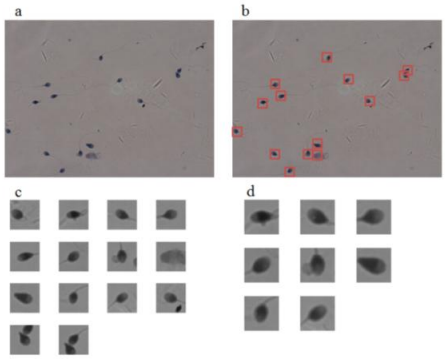

Fig. 1. Process of sperm segmentation. a. the original image (fifth image in the sperm database); b. locating sperms; c. the preliminary segmentation; d. the final segmentation after removing the abnormal.

In our sperm database, there are 80 microscope images containing sperms (Figure 1. a). And each image has different number of sperms. Before diagnosing the health status of the

single sperm, location and segmentation should be implemented in order to obtain the single sperm with a fixed size.

The preprocessing consists of three steps: location, segmentation and removing abnormal. At the first step, we mainly use the image morphology method to locate the position of sperms in the image (Figure 1. b). Then a size small image will be extracted based on each of sperm position (Figure 1. c). But not all the sperm divisions are normal, so we need remove the abnormal sperm divisions which include half sperm or multiple sperms. The final divided small images are shown in Figure 1. d.

In order to diagnose the health status of sperm, we firstly understand what the health status is. So we cooperate with professional doctors to mark the status of sperms. Figure 2 presents the healthy sperms and unhealthy sperms.

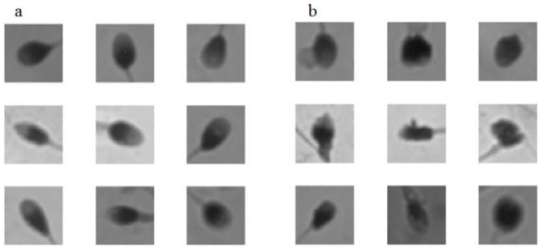

Fig. 2. The health status of sperms. a. healthy sperms; b. unhealthy sperms.

B. PCA Feature Extraction

After sperm segmentation, the database consists of 605 single sperms, including 501 healthy sperms and 104 unhealthy sperms. Before feature extraction by PCA, the training set and the testing set should be selected in the database above. Even though the number of the training set may be different, the ratio of healthy type to unhealthy type should be 5:1, which is the ratio in the whole database.

As mentioned above, we cropped every sperm image to 60×60 image; the input of this step is a preprocessed 3600×1 vector. We used these vectors to estimate the covariance matrix. After estimation of the covariance matrix, significant eigenvectors of the covariance matrix are calculated. Number of eigenvector depend on our application and accuracy that we need, it is clear that if we compute large number of eigenvectors accuracy of the method improved but computational complexity increased in this step and next step. Figure 3 demonstrated operation done in this stage (covariance estimation and computation of significant eigenvectors).

We repeated our experiment with different number of the training set, such as 60, 90, 120, 150, 180, 210 and 240, which are selected with a fixed step in the sperm database, and compare the performance of the proposed sperm health diagnosis method.

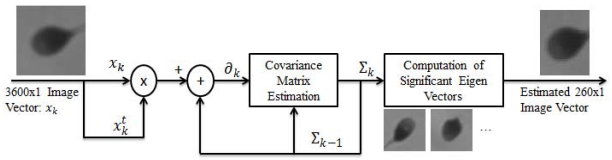

Fig. 3. Process of PCA: covariance estimation and computation of significant eigenvectors.

C. K-nearest Neighbor Classification

K-nearest neighbor (KNN) classification algorithm is a more mature approach in theory, but it is also one of the simplest machine learning algorithms, presented by the Cover and Hart[17] in 1968. This decision rule provides a simple nonparametric procedure for the assignment of a class label to the input pattern based on the class labels represented by the K-closest (say, for example, in the Euclidean sense) neighbors of the vectors.

KNN algorithm is simple and effective and it is a lazy-learning algorithm. It does not require the training set for training, the training time complexity is 0. Its computational complexity is proportional to the number of training samples, that is, if the total number of samples in the training set is n, then the KNN classification time complexity is O (n).

In the experiment, after extracting the PCA feature, the KNN is used to diagnose the status of the sperm, meanwhile, the BPNN[18, 19] method is also executed on the PCA features as the control groups.

Neural network[20] learning method provide a robust approach to approximating real-valued, discrete-valued, and vector-valued target functions. For certain type of problems, such as learning to interpret complex real-world sensor data, artificial neural networks are among the most effective learning methods currently known. For example, the backpropagation algorithm described has proven surprisingly successful in many practical problems such as learning to recognize handwritten characters and learning to recognize spoken words. The backpropagation algorithm learns the weights for a multilayer network, given a network with a fixed set of units and interconnections. The learning problem is to search a large hypothesis space defined by all possible weight values for all the units in the network. It employs gradient descent to attempt to minimize the squared error between the network output values and the target values for these outputs in order to find the best hypothesis in the hypothesis space mentioned above.

V. EXPRIMENTAL RESULTS DISCUSSION

In this study, the PCA and the KNN system is the main experimental method. As mentioned above, the training set should be selected by us, and the ratio of healthy type to unhealthy type should be 5:1, that is, if we selected 30 unhealthy sperms in the training set, the number of the healthy sperms is 180. In this group with 180 training samples, the accuracy rate is 87.53%. And the concrete classification is shown in Table I.

TABLE I. CLASSIFICATION RESULT (180 TRAINING; 425TESTING)

Category	Healthy	Unhealthy
Healthy	336	15
Unhealthy	38	36

In Table I, we can conclude that the diagnosis of health status is significantly better than that of unhealthy status of sperm. The accuracy of healthy diagnosis is 95.73%, while that of unhealthy diagnosis is 51.35%. So if we want to improve the result, the low accuracy of the latter should be solved.

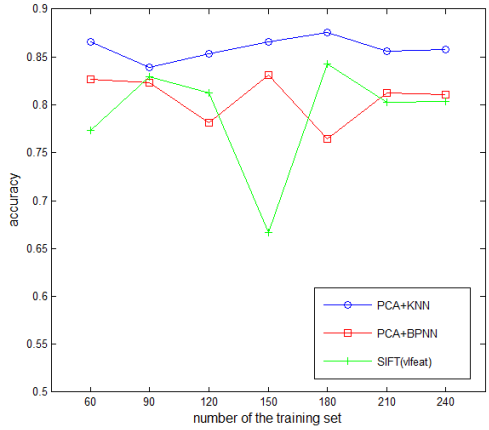

Fig. 4. Accuracy comparison among PCA+KNN, PCA+BPNN and SIFT (vifeat).

Next, we will compare the different feature extraction methods and classification methods, that is, PCA+KNN, PCA+BPNN, and SIFT (vlfeat[15]) as is shown in Figure 4. The obvious comparison can be concluded from Figure 4, that the classification accuracy in PCA+KNN experiment is severely higher than that in PCA+BPNN and SIFT (vlfeat) experiments regardless of the number of training set.

Table II presents the classification results in the PCA+KNN, PCA+BPNN and SIFT (vlfeat) experiments. Though the result of the PCA+KNN is much better than the other two, its accuracy of all is not more than 90% and the largest accuracy is obtained when the number of training set is 180, including 30 unhealthy sperms and 150 healthy sperms. So the sperm diagnosis research still need to continue to study in order to seek better feature extraction method and classification method for getting better performance.

TABLE II. ACCURACY RESULTS IN THE THREE EXPERIMENTS: PCA+KNN, PCA+BPNN AND SIFT (VLFEAT)

Number\Accuracy	PCA+KNN	PCA+BPNN	SIFT(vlfeat)
60	86.61%	82.67%	77.29%
90	83.88%	82.33%	82.95%
120	85.36%	78.14%	81.28%
150	86.59%	83.08%	66.67%
180	87.53%	76.47%	84.27%
210	85.57%	81.27%	80.30%
240	85.75%	81.10%	80.33%

In this paper, two feature extraction methods and three classification methods, PCA+KNN, PCA+BPNN and SIFT (vlfeat), are researched on the problem, sperm health status diagnosis. Experiment based on PCA and KNN has better performance than the other two methods. However, all the accuracy results are less than 90%, further research still need.

ACKNOWLEDGMENT

I would particularly like to thank Kuo-Kun Tseng, who has suggested numerous improvements to both the content and presentation of this paper. In addition, I would like to thank many others for their suggestions and supporting sperm database. This research was supported by the institute of Innovative Information Industry Research Center.

REFERENCES

[1] J. MacLeod and R. Gold, "The male factor in fertility and infertility. IV. Sperm morphology in fertile and infertile marriage," Fertility and sterility, vol. 2, p. 394, 1951.

[2] V. Bartak, "Sperm count, morphology and motility after unilateral mumps orchitis," Journal of Reproduction and Fertility, vol. 32, pp. 491-494, 1973.

[3] V. Bartak, "Sperm quality in adult diabetic men," International journal of fertility, vol. 24, p. 226, 1979.

[4] R. Davis and C. Gravance, "Consistency of sperm morphology classification methods," Journal of andrology, vol. 15, pp. 83-91, 1994.

[5] M. Freund, "Standards for the rating of human sperm morphology. A cooperative study," International journal of fertility, vol. 11, p. 97, 1966.

[6] J. Moruzzi, A. Wyrobek, B. Mayall, and B. Gledhill, "Quantification and classification of human sperm morphology by computer-assisted image analysis," Fertility and sterility, vol. 50, pp. 142-152, 1988.

[7] R. Davis, D. Bain, R. Siemers, D. Thal, J. Andrew, and C. Gravance, "Accuracy and precision of the CellForm-Human automated sperm morphometry instrument," Fertility and sterility, vol. 58, pp. 763-769, 1992.

[8] J. Yániz, S. Vicente-Fiel, S. Capistrós, I. Palacín, and P. Santolaria, "Automatic evaluation of ram sperm morphometry," Theriogenology, vol. 77, pp. 1343-1350, 2012.

[9] L. Fan, Z. Yuan, X. Han, and W. Hua, "Overview of Content-based Image Feature Extraction Methods," 2013.

[10] A. H. Sahoolizadeh, B. Z. Heidari, and C. H. Dehghani, "A new face recognition method using PCA, LDA and neural network," International Journal of Computer Science and Engineering, vol. 2, pp. 218-223, 2008.

[11] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and intelligent laboratory systems, vol. 2, pp. 37-52, 1987.

[12] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," JOSA A, vol. 14, pp. 1724-1733, 1997.

[13] D. Zhang, X.-Y. Jing, and J. Yang, "Linear discriminant analysis," D. Zhang, X. Jing, Jing, & J. Yang (Eds.), Biometric Image Discrimination Technologies: Computational Intelligence and its Applications Series, pp. 41-64, 2006.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, pp. 91-110, 2004.

[15] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in Proceedings of the international conference on Multimedia, 2010, pp. 1469-1472.

[16] H. Wang, K. Yang, F. Gao, and J. Li, "Normalization Methods of SIFT Vector for Object Recognition," in Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2011 Tenth International Symposium on, 2011, pp. 175-178.

[17] T. Cover and P. Hart, "Nearest neighbor pattern classification," Information Theory, IEEE Transactions on, vol. 13, pp. 21-27, 1967.

[18] G. Bortolan, R. Degani, and J. Willems, "ECG classification with neural networks and cluster analysis," in Computers in Cardiology 1991, Proceedings., 1991, pp. 177-180.

[19] S. Sivathasan, F. Cecelja, and W. Balachandran, "ECG Diagnosis using neural network and fuzzy expert system," in Instrumentation and Measurement Technology Conference, 2000. IMTC 2000. Proceedings of the 17th IEEE, 2000, pp. 988-992.

[20] T. M. Mitchell, "Machine learning. 1997," Burr Ridge, IL: McGraw Hill, vol. 45, 1997.