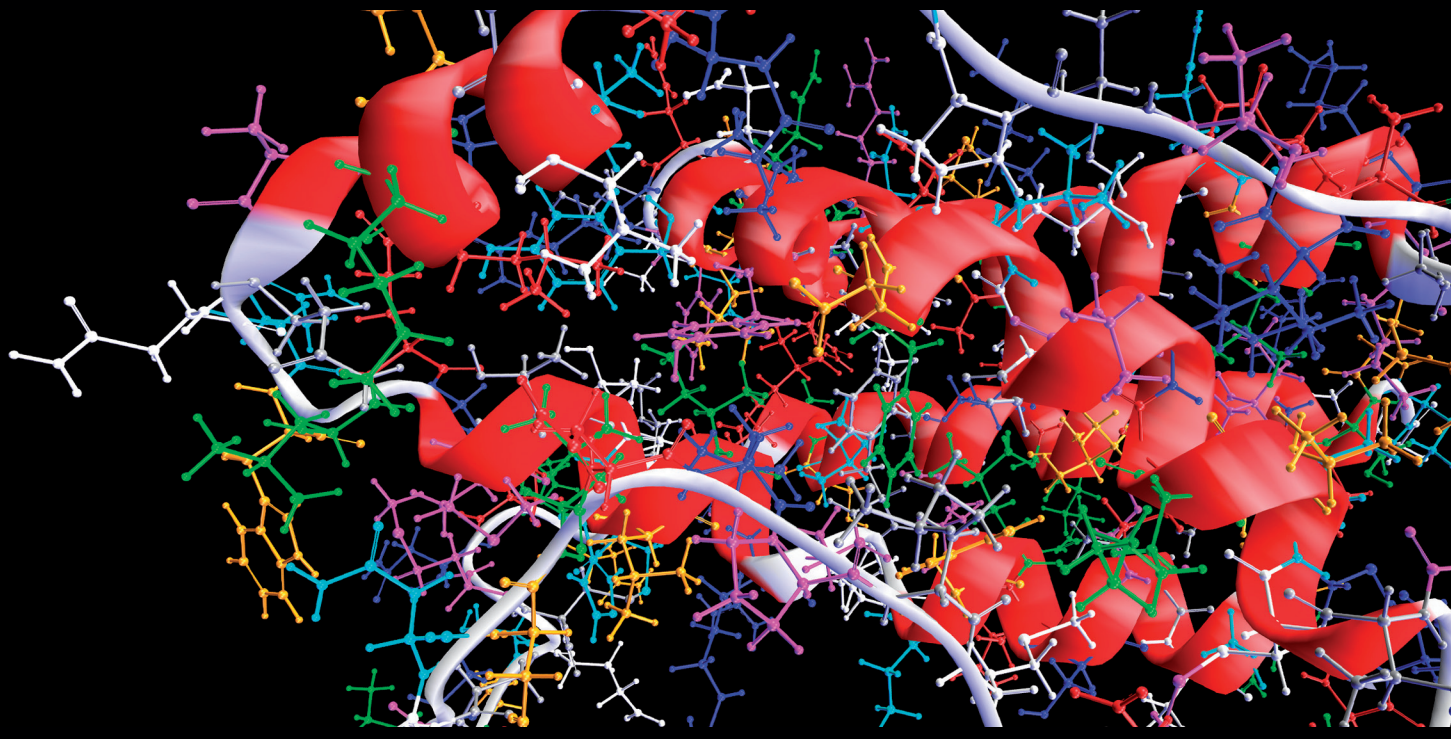
Computational and Mathematical Methods in Medicine

Data Preprocessing and Model Design for Medicine Problems

Guest Editors: Alberto Guillén, Amaury Lendasse, and Guilherme Barreto



Data Preprocessing and Model Design for Medicine Problems

Computational and Mathematical Methods in Medicine

Data Preprocessing and Model Design for Medicine Problems

Guest Editors: Alberto Guillén, Amaury Lendasse, and Guilherme Barreto

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 625623, 1 page
http://dx.doi.org/10.1155/2013/625623

Editorial

Data Preprocessing and Model Design for Medicine Problems

Alberto Guillén,[1] Amaury Lendasse,[2,3,4,5] and Guilherme Barreto[5]

[1] *Department of Computer Technology and Architecture, University of Granada, Granada, Spain*
[2] *Department of Information and Computer Science, Aalto University School of Science, Espoo, Finland*
[3] *IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain*
[4] *Computational Intelligence Group, Computer Science Faculty, University of the Basque Country, Paseo Manuel Lardizabal 1, Donostia/San Sebastián, Spain*
[5] *Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza, Brazil*

Correspondence should be addressed to Alberto Guillén; aguillen@ugr.es

Received 7 February 2013; Accepted 7 February 2013

Machine-learning disciplines including model design and data preprocessing are crucial in order to obtain a good performance in terms of accurate results and interpretability. However, they are not usually treated simultaneously and, when a model is evaluated, the origin and preprocessing of the data are ignored. Medicine and biomedical research provide a wide variety of problems where machine-learning can be very helpful in decision support, telemedicine, and the discovery of interactions.

These facts motivated the elaboration of this special issue; therefore, it is focused on methods and applications where machine learning could be applied holistically encompassing all stages to solve the problem. The papers included in the special issue go through the intersection between the medical field of application and theoretical models. For example, generalized estimating equations which are a common approach are compared against quadratic inference functions when applied to a lipid and glucose study. It is common in the field of medicine to be suspicious to predictions made by models, so it is interesting also to read another paper presenting the application of machine-learning techniques as a support decision tool that will not replace the expert judgment. This special issue not only considers the application of the models to improve the classification or prediction accuracy but also presents papers where the data are analysed properly. In medicine problems, it is quite common to have continuous and discrete variables in order to show how to deal with these situations; the paper entitled "*Let continuous outcome variables remain continuous*" shows how to apply a popular regression method without dichotomising the variables as this procedure could end up in the lost information.

We hope that the reading of this special issue will help medicine researchers to be aware of new methods and machine-learning techniques as well as to see how they could be applied. We also hope that the machine learning community can see here a wide variety of problems where the models and algorithms they create could be applied providing useful results.

Alberto Guillén
Amaury Lendasse
Guilherme Barreto

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 972037, 12 pages
doi:10.1155/2012/972037

Research Article

Extraction of Lesion-Partitioned Features and Retrieval of Contrast-Enhanced Liver Images

Mei Yu,[1,2] Qianjin Feng,[1] Wei Yang,[1] Yang Gao,[1] and Wufan Chen[1]

[1] *School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China*
[2] *Shandong Medical College, Linyi 276000, China*

Correspondence should be addressed to Qianjin Feng, qianjinfeng08@gmail.com and Wufan Chen, wufanchen@gmail.com

Received 22 March 2012; Revised 24 June 2012; Accepted 16 July 2012

Academic Editor: Guilherme de Alencar Barreto

The most critical step in grayscale medical image retrieval systems is feature extraction. Understanding the interrelatedness between the characteristics of lesion images and corresponding imaging features is crucial for image training, as well as for features extraction. A feature-extraction algorithm is developed based on different imaging properties of lesions and on the discrepancy in density between the lesions and their surrounding normal liver tissues in triple-phase contrast-enhanced computed tomographic (CT) scans. The algorithm includes mainly two processes: (1) distance transformation, which is used to divide the lesion into distinct regions and represents the spatial structure distribution and (2) representation using bag of visual words (BoW) based on regions. The evaluation of this system based on the proposed feature extraction algorithm shows excellent retrieval results for three types of liver lesions visible on triple-phase scans CT images. The results of the proposed feature extraction algorithm show that although single-phase scans achieve the average precision of 81.9%, 80.8%, and 70.2%, dual- and triple-phase scans achieve 86.3% and 88.0%.

1. Introduction

Computed tomographic (CT) is a primary imaging technique for the detection and characterization of focal liver lesions. Currently, CT is widely used for the diagnosis of liver tumors. A vast amount of information can be obtained from CT; however, even experienced radiologists or physicians have difficulty interpreting all the images in a certain cases within short duration. Moreover, the interpretation among radiologists shows substantial variation [1, 2], and its accuracy varies widely given the increasing number of images [3].

Studies on CT images retrieval have precedents [4]; however, existing medical image processing technologies are not sufficiently mature. Thus, diagnostic results are often less than ideal. Nationally, along with the developments in image processing and artificial intelligence, designing and developing systems for computer-aided diagnosis to characterize liver lesions have received considerable attention over the past years, because these systems can provide diagnostic assistance to clinicians for the improvement of diagnosis [5, 6]. Organically combining the key technologies of image processing and medical imaging has become a main research goal to provide scientific, convenient, and accurate medical means and to support diagnostic recommendations for radiologists. Such systems are implemented by image retrieval systems that enable radiologists to search for radiology patients in database and return the cases that are similar in terms of shared imaging features with their current cases. Currently, many image retrieval applications are used in the medical field. These applications are not only capable of retrieval of similar anatomical region [7–9], but also of similar lesions [10–13].

The most critical step in image retrieval systems is feature extraction, especially for grayscale medical images. Although low-level features, such as gray, texture, and shape [14, 15] are commonly used for visual perception of radiologic images, they cannot express the image or distinguish lesions adequately. Unfortunately, clinical diagnostic decisions are generally made based on medical imaging behavior of lesions. Therefore, the understanding of interrelatedness

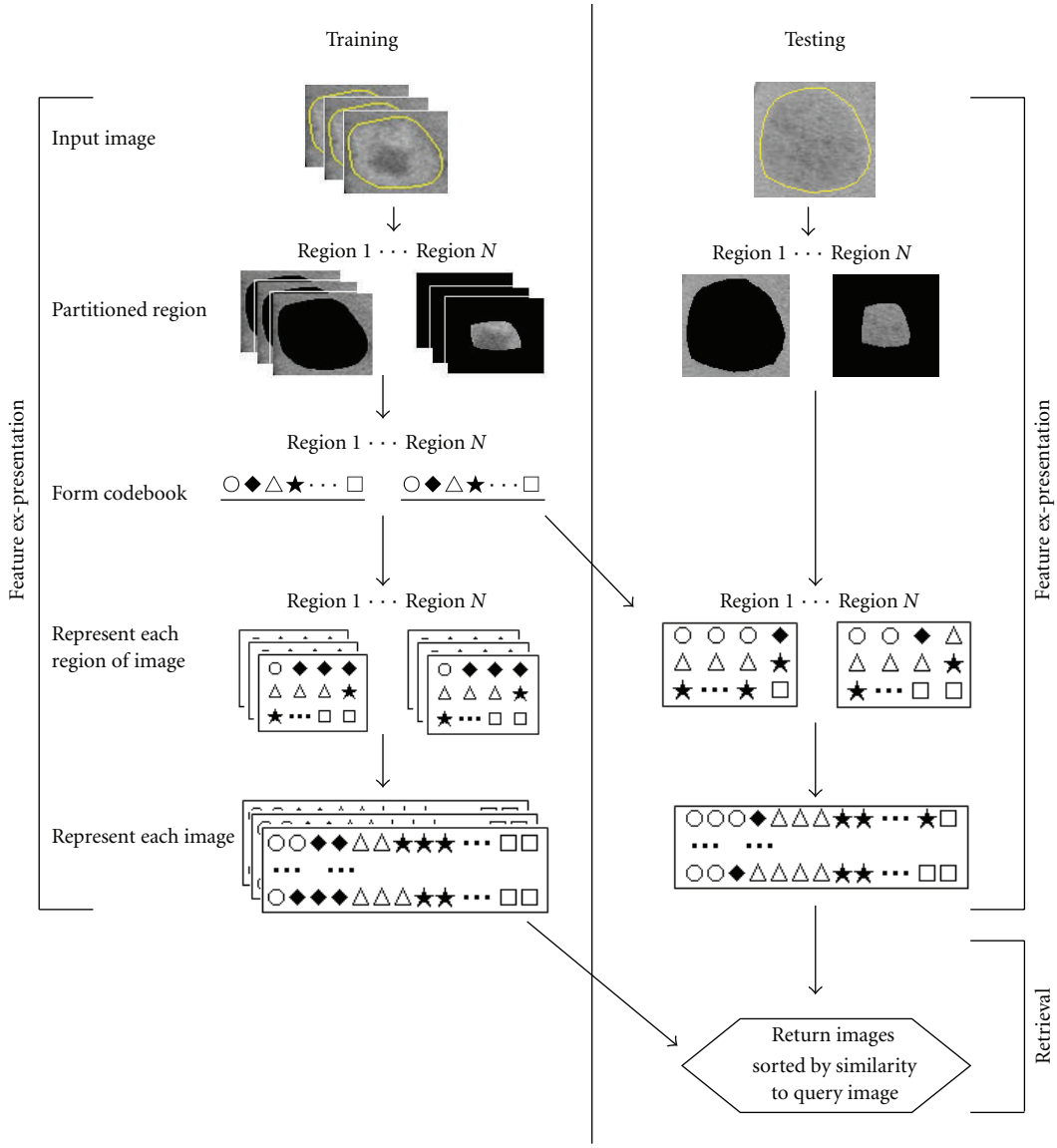


Figure 1: Flow chart of system based on our algorithm.

among the characteristics of lesion images and corresponding imaging features is critical for image training [16] as well as for features extraction. Several radiological studies have recently reported the relationship among the correlations [17].

The aims of the current study are three. (1) A feature extraction algorithm of hepatic lesions is provided considering the views of radiologists concerning diagnosis in triple-phase CT; (2) a content-based image retrieval (CBIR) system is developed. This system can facilitate the retrieval of radiology images that the lesions have with similar appearing to the query patient, and (3) a basis evaluation of this system is implemented. Hepatocellular carcinoma (HCC), hemangiomas, and cysts are the most common malignant and benign liver tumors. The proposed feature algorithm is derived from distinct imaging characteristics of lesion images and the surrounding liver parenchyma in triple-phase CT images, which comprise the diagnosis perspective of clinicians or radiologists for three types of tumor patients.

2. Methods

Figure 1 presents a summary of the current system based on the proposed feature extraction. The specific development and implementation are detailed below.

2.1. Liver Lesions. Triple-phase contrast-enhanced CT scans play an important role in the diagnosis of liver tumors, because triple-phase images fully display the characteristics of blood supply richly of HCC (Figure 2). In the arterial phase, most of lesions with rich blood supply appear hyperenhancement. Density is significantly higher than that of normal hepatic parenchyma, because hepatic parenchyma has not reached the enhanced peak. In the portal venous

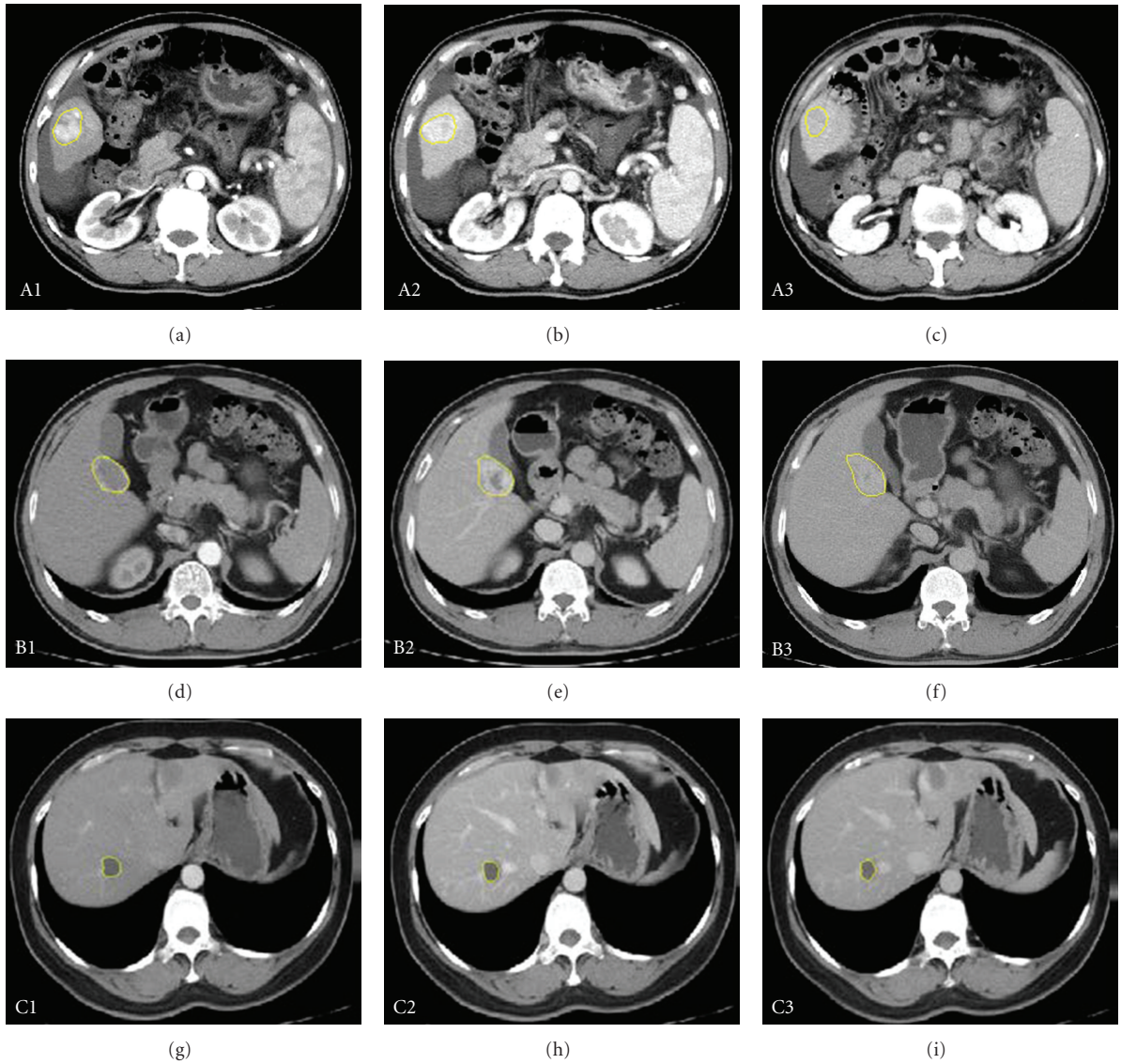


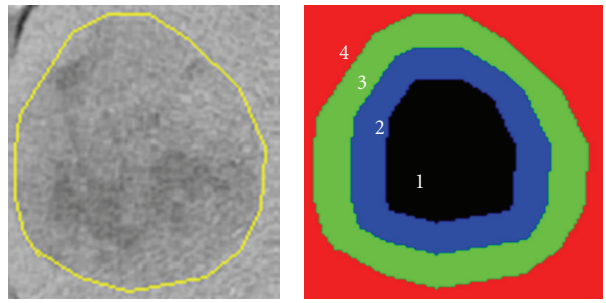
Figure 2: Triple-phase contrast-enhanced CT images. The row is liver cancers, hemangiomas, cysts. The vertical column is arterial phase, portal venous, and delayed-phase scans.

phase, the parenchyma reaches its peak, whereas the lesion almost joins the blood supply. The tumor is characterized by low-density nodules relative to parenchyma. "Fast in and fast out" is the most characteristic movement of HCC with rich blood supply.

The CT scan is the preferred imaging methods for hepatic hemangiomas. The enhanced characteristic of a hemangioma is as follows. The edge of the lesion in the arterial phase usually appears heavily enhanced, and the contrast agent gradually enters the lesion, traveling from the edge to the centre over time, which provides a reliable basis for diagnosing HCC and hemangioma. "Fast in and slow out" is the most characteristic movement for hemangiomas. Therefore, the density of the lesion is higher than that of parenchyma in the arterial phase and is lower than that of parenchyma in the portal venous lesion, which is also the typical behavior that distinguishes HCC and hemangiomas.

Liver cysts are commonly benign, and triple-phase enhanced CT scans of such cysts appear as single or multiple, round or oval, and with a smooth edge and uniform low density. The value of CT is close to water. Images of liver cysts are subjected to no further enhancement after contrast enhancement.

Two facts summarize the characteristics of triple-phase contrast-enhanced CT images. First, most lesions of HCC and hepatic hemangiomas have special characteristic changes, whereas no change occurs in that of cysts. Second, the surrounding liver parenchyma information of a lesion is important because of the discrepancy in density between the lesion and the adjacent normal parenchyma in triple-phase scans. Thus, according to the above analysis, a feature extraction algorithm of lesions is proposed considering the specific behavior of focal liver lesions and their surrounding liver parenchyma after enhancement.



(a) (b)

Figure 3: Partition of hepatic lesion. (a) is the lesion with external neighborhood and (b) is 4 regions divided.

2.2. Computer-Generated Features. Representation of Bag of Visual Words Combined with Distance Transformation. The proposed feature extraction algorithm aims to meet the requirements of radiologic diagnosis views, as derived formed from the analysis in Section 2.1. The algorithm includes mainly two processes: (1) distance transformation, which is used to divide the lesion into distinct regions and represents the spatial structure distribution and (2) the representation of bag of visual words (BoW) based on regions, which is the key step. Generally, the lesion is divided into three regions in the experiments, to best fit the imaging analysis of radiologists in triple phases for the diseases described above. The effect of number selection on the performance of CBIR is discussed later in this document. The algorithm is described below.

2.2.1. Partition of the Lesion through Distance Transformation. The concept of distance transformation has been widely used in image analysis, computer vision, and pattern recognition since its introduction by Rosenfeld and Pfaltz [18] in 1966.

Distance transformation is conducted against binary images to produce a grayscale image, such as the distance image. The gray values of every pixel point in the distance image are the distances between the pixel and its nearest background pixels. In two-dimensional space, a binary image contains only two kinds of pixels: target pixels and background pixels. The value of a target pixel is 1, and the value of a background pixel is 0. Currently, a variety of distance transformation algorithms is used, and these algorithms adopt mainly two types of distance: non-Euclidean distance and Euclidean distance. The former method commonly includes city-block, chessboard, and chamfer. City block distance transformation is used in this paper because the distance value after transformation is an integer, which is more convenient for the subsequent partition of lesions.

Then, the distance transformation image of the binary image is obtained. Set p, q as the quotient and the remainder, respectively, resulting from the division of the number of layers by three. Divide the lesion into three regions, and the number of layers in each region become p, q and $p + q$, respectively, from the boundary of the lesion to the center.

Apart from considering the lesion changes, the density discrepancy between the adjacent normal liver parenchyma and the lesion in triple phase scans is also a foundation for the diagnosis of radiologists. Therefore, the surrounding liver parenchyma of lesions is considered as the fourth region (Figure 3). Assuming the bounding box of the lesion region is $K' \times L'$ if each side of the box has an extension of two pixels, a bounding box that includes the lesion with size $(K' + 4) \times (L' + 4)$ is finally obtained. Thus, the new box not only contains the tumor, but also its surrounding normal liver parenchyma.

2.2.2. Regional BoW. Typically, BoW representation involves four major steps: (1) patches of interest image regions are detected; (2) patches are locally described using feature vectors (local descriptor); (3) features are quantized and labeled in terms of a predefined dictionary, that is, the construct process of codebook; (4) histograms are constructed by accumulating the labels of the feature vectors of each image in database.

In the following experiments, the BoW approach, used generally, follows the traditional visual codebook method [19–22]. The approach is accomplished by selecting patches from images, characterizing them with the vectors of the local visual descriptors, and labeling the vectors using a learned visual codebook. The occurrence of each label is quantified to build a global histogram that summarizes the image content. The histogram is then subjected to distance metric methods to estimate the disease category label of the images. The patches are extracted from each pixel point of the tumor images. The codeword vocabulary is typically obtained by clustering the descriptors of the training images. The intensity values are adopted to characterize the local visual descriptors, which implicitly reflect the category of the lesion in CT images, thereby providing more important information. Unsupervised K-means clustering is chosen as the base codebook learner in the current paper.

The selected size of square patches is seven. This selection considers the limitation of too many images in the database and the size restriction of lesions like cysts. Basically, an image is represented by a histogram of word frequencies that describes the probability density over the code words in the codebook. The background gray value of a radiologic image is 0. Thus, the patches that contained more background pixels are removed to save time and simply computational.

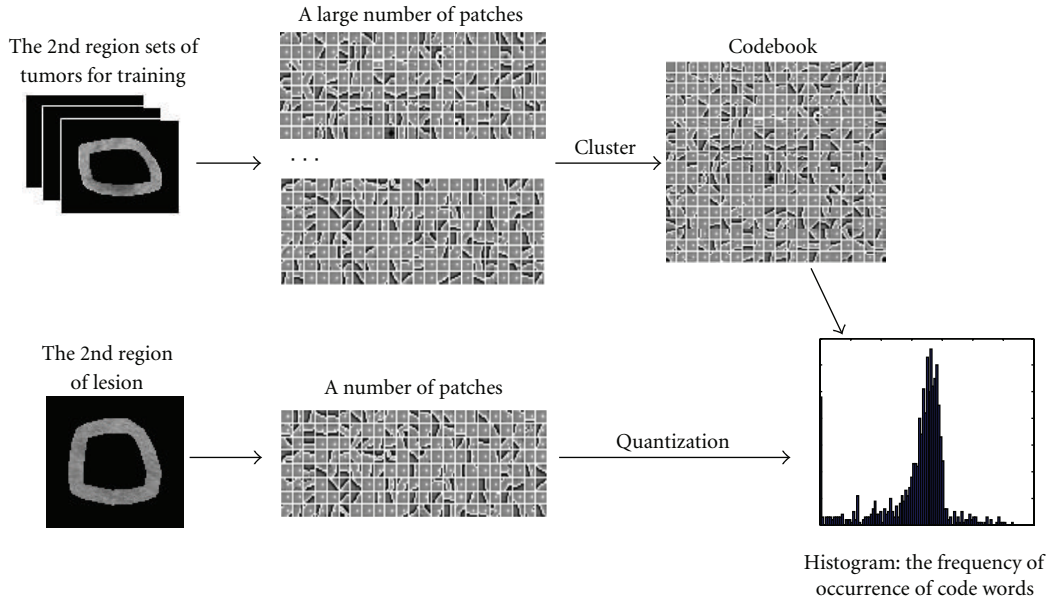


Figure 4: The BOW feature extraction based on regions of lesion.

The number of pixels with a value in patch not equal to 0 is set to be greater than 15 in this paper. For the vocabulary $V = \{v_1, v_2, \ldots, v_N\}$ with N code words, the traditional codebook model estimated the distribution of the code words in an image of r patches $\{p_1, p_2, \ldots, p_r\}$ by $x = [x_1, x_2, \ldots, x_N]^T$, where

$$x_i = \sum_{i=1}^{r} \begin{cases} 1 & \text{if } v_i = \arg\min_{u \in V} \operatorname{dist}(u, p_k), \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

denoting $\operatorname{dist}(\mu, p_k)$ to be the distance between code word μ in vocabulary and image patch.

The quantized vectors of each region of the lesion are obtained using the method described above. The final feature of the lesion is expressed as the arrangement. The process of BoW is shown in Figure 4. The features of each patient in different phases are calculated to determine the average of all corresponding images. For example, if an HCC patient has six images (one arterial phase image, three portal venous phase images, and two delayed phase images), the features used in retrieval of the portal venous phase, dual-phase (arterial phase plus portal venous phase), or triple-phase are the corresponding average of features of three images, four images, or all six images.

2.3. Common Low-Level Features. For each lesion, multiple features are computed within the lesion region of interest (ROI).

2.3.1. Intensity Features. The following five intensity features are calculated: mean, standard deviation, entropy, skewness, and kurtness of gray-level histogram [23].

2.3.2. Texture Features. Gray-level cooccurrence matrix (GLCM) [23–25] and Gabor [26, 27] describe the texture characteristics of each ROI. Sixteen GLCM features are calculated in the current experiments using contrast, homogeneity, energy, correlation for four angles (i.e., 0°, 45°, 90°, 135°), and distance of 1. Next, 48 Gabor features are computed from the mean and standard deviations of the energy in the frequency domain over four scales and six orientations. The mean and standard deviations of high-frequency coefficients of its three-level Daubechies4 wavelet decomposition are computed, resulting in 12 features.

2.3.3. Shape Features. The statistics of wavelet coefficients of the shape signature are used to characterize the shape of tumors. The one-dimensional shape signature $S(i)$, based on radial distance, is defined as follows:

$$S(i) = \sqrt{(x(i) - C_x)^2 + \left(y(i) - C_y\right)^2}, \quad (2)$$

where $x(i)$ and $y(i)$ are the coordinates of the ith point on the tumor boundary and c_x and c_y are the coordinates of the centroid of the tumor region. Twelve features are computed from the mean and variance of the absolute values of the wavelet coefficients in each subband by the five-level one-dimensional wavelet decomposition.

This computation yields a total of 93 features.

2.4. Similarity Distance Measure. When the feature vectors containing detailed imaging information of lesions are computed, the system calculates similar distance measures between them, that is, similarity between the corresponding images. The similarity of lesions is defined as the distance between the corresponding elements of the respective feature vectors that describe the lesions. Previous studies have shown that well-designed distance metrics can result in better retrieval or classification performance compared with Euclidean distance [28–31]. The goal of distance metric

learning is to determine a linear transformation matrix to project the features into a new feature space that can optimize a predefined objective function. The distance metric learning algorithms used in this paper are L1 distance, L2 distance, regularized linear discriminant analysis (RLDA) [32–34], and linear discriminant projections (LDP) [35–37].

The distance in the L1 norm is known as Manhattan distance. The L2 norm distance is called the familiar Euclidean distance. The L1 and L2 distance are described as follows:

$$\begin{aligned} \text{Distance-L1}\left[x^l, x^q\right] &= \sum_{i=1}^{N} \left| x_i^l - x_i^q \right|, \\ \text{Distance-L2}\left[x^l, x^q\right] &= \left(\sum_{i=1}^{N} \left(x_i^l - x_i^q \right)^2 \right)^{1/2}, \end{aligned} \tag{3}$$

where x^q, x^l represent the features of the query cases and the cases in the database and N is the dimension of features.

The RLDA was first presented in [32]. Ye and Wang then proposed an efficient algorithm to compute the solution for RLDA [33]. The performance of RLDA exceeds that of ordinary linear discriminant analysis (LDA) methods [34]. The numbers of samples is set to M. When M and the feature dimension N are large, applying RLDA is not feasible because of the memory limit. Considering N is large in the current paper, the dimension is first reduced to $M - 1$ using principal component analysis (PCA), after which RLDA is applied. Parameter α controls the smoothness of the estimator in RLDA. The value of α is set to 0.001.

The details of the LDP have been inferred from the previous papers. The LDP approach has three advantages. First, LDP can be adapted to any dataset and any descriptor, and may be directly applied to the descriptors. Second, LDP is not sensitive to noise and is thus suitable for work on hepatic CT images. Third, LDP can be trained much faster because the k-nearest of each sample point does not need to be determined [37]. Moreover, LDP has been proven to produce better results than some other approaches.

2.5. Lesion Database. All the imaging data of hepatic CT images for experiments were acquired from the General Hospital of Tianjin Medical University between February 2008 and October 2010. CT examinations were performed with a 64-detector helical scanner (LightSpeed VCT; GE Medical Systems, Waukesha, Wis). The following parameters were used: 120 kVp, 200–400 mAs, 2.5–5 mm section thickness, and a spatial resolution of 512 × 512 pixels. The imaging data included three diseases: HCC, hemangiomas, and cysts. In all, 1248 DICOM lesion images (498 HCC, 481 hemangiomas, and 269 cysts) were found in 187 patients (89 HCC, 54 hemangiomas, and 44 cysts) wherein each patient corresponded to 2–10 images. All images were classified into arterial phase, portal venous phase, and delayed phase, wherein the number is 388, 443, and 417, respectively. All the images in the database were manually delineated using semiautomatic segmentation to ensure the effectiveness of the CBIR system, and some inaccurate results were reevaluated by medical imaging experts blinded to the final diagnosis to obtain more precise lesion data.

In the current study, a patient is a query case. Each patient has more than one lesion. For example, he/she may have some cysts or have got hemangioma as well as cysts. However, only one typical lesion is selected from each image of each patient, and the lesions in each patient are the same. All the images of each patient are used for query. The features of the patient are the average value of features of the images in single-, dual-, and triple-phase scans.

2.6. Evaluation Measures. Precision and recall are common criteria used in evaluating the effectiveness of CBIR. Precision indicates the accuracy of retrieval, that is, how exclusively the relevant images are retrieved. Precision and recall ratio can be defined as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{Number of relevant images retrieved}}{\text{Total number of image retrieved}}, \\ \text{Recall} &= \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant image}}. \end{aligned} \tag{4}$$

The higher the value of these two indicators, the better the retrieval system. The two indicators are usually mutually contradictory. In theory, as precision increases, recall decreases, and vice versa. Therefore, generic retrieval systems that optimally balance these two indicators achieve better retrieval performance. Generally, the ultimate goal of the proposed CBIR system is to achieve retrieval results that better reflect the actual categories of the query case. The CBIR system retrieves similar cases and thereby calculates a decision value (i.e., similar distance) that describes the similarity to the query case. Therefore, precision is needed. The higher the precision, the more relevant cases are retrieved, which indicates that the CBIR system has important clinical applications. The evaluation measurement is mainly the average precision, which is defined as the average ratio of the number of relevant images returned over the total returned images. Therefore, in the current experiment the following measures are used to evaluate the CBIR system.

(i) Here, P(10), P(20), and P(n), the average precisions after the top 10, 20, and n patients are returned when lesion images are ranked according to similarity to a query lesion.

(ii) Mean average precision (MAP) is the mean of the average precisions when the number of images returned is varied from 1 to the total number of images.

(iii) Precision versus recall graph.

2.7. Training and Evaluation. All the experiments are based on 187 patients, using the K-fold cross-validation (K-CV) method. K-CV is used for the allocation of the samples. All the samples are evenly divided into K, where in K–1 samples are chosen to training, and the remainder performs the validation alternately. In this paper, the sampling plan for K-CV is as follows: 187 cases are evenly divided into K, and

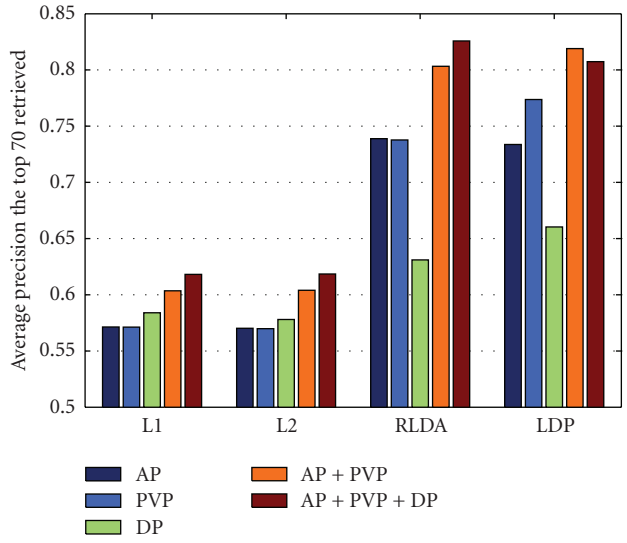


Figure 5: The average precision histogram of staging retrieval-based 3 regions using four distance metrics.

each is used for testing set, whereas the other $K - 1$ samples are used for training sets. Thus, an K-CV experiment needs to establish K-models, that is, perform the K tests. Generally in practice, the value of K needs to be sufficiently large to enable a sufficient number of training samples, which enables the distribution features of images in training sets to be sufficient for describing the distribution features of the entire image sets. Thus, the distribution features of images in the entire database are not significantly influenced when some images of a new patient are added. A K value equal to 10 is considered adequate; hence the value is set to 10 in the experiment. Each patient in each test is searched as a query case. Thus, the average precision of each test and the MAP of 10 tests are obtained.

3. Experiment and Results

The effectiveness of the proposed feature extraction method was verified by retrievals of single-, dual-, and triple-phase scans to maximize the average precision of a large hepatic CT image dataset. Four experiments were performed: (1) verification of the proposed algorithm given different conditions; (2) comparison between general low-level features and high-level BoW features, (3) exploration of the selection of region number, and (4) identification of the amount of clusters influence. PCA was used for dimension reduction because of the initial huge dimension first. Arterial phase, portal venous phase, and delayed phase are abbreviated as AP, PVP, and DP.

3.1. Retrieval Results of Regional BoW. The retrieval performance of the proposed feature extraction algorithm in single-, dual-, and triple-phase scans is shown in this experimental. The dual-phase not only refers to arterial plus portal venous phase, but also to portal venous plus delayed phase. Figure 5 provides the retrieval results based on three regions of lesions using the four distance metric methods mentioned above in terms of P(70). The figure shows that the results of single-phase scan are lower than the results of two dual-phase and triple-phase scans, and that the use of RLDA and LDP generate better results than the use of L1 and L2.

Table 1 shows the retrieval performance based on three regions, together with their surrounding liver parenchyma in triple phase scans in terms of MAP, P(10) and P(20). The estimated P(20) of single-phase scans using RLDA and LDP is lower than 85.8%, whereas the P(20) of dual-phase and triple-phase scans is higher than 91.2%, except for PVP + DP. Figure 5 and Table 1 indicate that the dual-phase and triple-phase scans are more precise than the single-phase scans, because regional BoW-based features greatly express the characteristics of the three tumors in the triple phases, that is, HCC and hemangiomas mostly exhibit characteristic changes, whereas no change occurs in cysts. Therefore, the proposed feature extraction method agrees with the diagnosis of the radiologist for the three lesions. In short, although single-phase scans may play an important role in diagnosis or detection, dual-phase and triple-phase scans also ensured more accurate diagnosis than single-phase scans. The founding demonstrates that arterial and portal venous phase scans play a major role in diagnosis, and explains why radiologists directly diagnosed some hepatic diseases only through dual-phase scans (i.e., artery plus portal venous phase).

Figure 6 shows the precision versus recall curves of the three regions, as well the region with their surrounding liver parenchyma in triple phases. The figure shows that the retrieval performance of the latter is better than performance of the former, regardless of single-, dual-, or triple-phases scans. Thus, the results that consider the surrounding liver of the lesion are more accurate because the discrepancy in density between the lesions and their adjacent liver parenchyma in the triple phase scans is also considered by radiologists as the basis for HCC, hemangiomas, and cysts diagnoses.

The validation of our proposed algorithm can be seen from different perspectives. Figures 7 and 8 separately compare the performance among three regions of lesion and the whole lesion, as well as the comparison between two cases with surrounding liver parenchyma. Figures 7 and 8 show that the results based on the regions always outperform the whole lesion with or without surrounding liver tissues in triple phases. This result can be attributed to the imaging characteristics of the enhanced images, quantitatively expressed by the proposed feature extraction algorithm.

3.2. Comparison of Common Low-Level Features and High-Level BoW Features. Common low-level features were compared with the proposed feature extraction algorithm in terms of precision versus recall curves. Figure 9 provides the results using shape alone (denoted as S), combination of shape and intensity (denoted as In+S), combination of shape, intensity, and texture (denoted as In+S+T), BoW alone, and combination of all features mentioned in Section 2.3 in PVP, dual-phase and triple-phase scans. As shown, the combination of intensity and shape outperforms shape

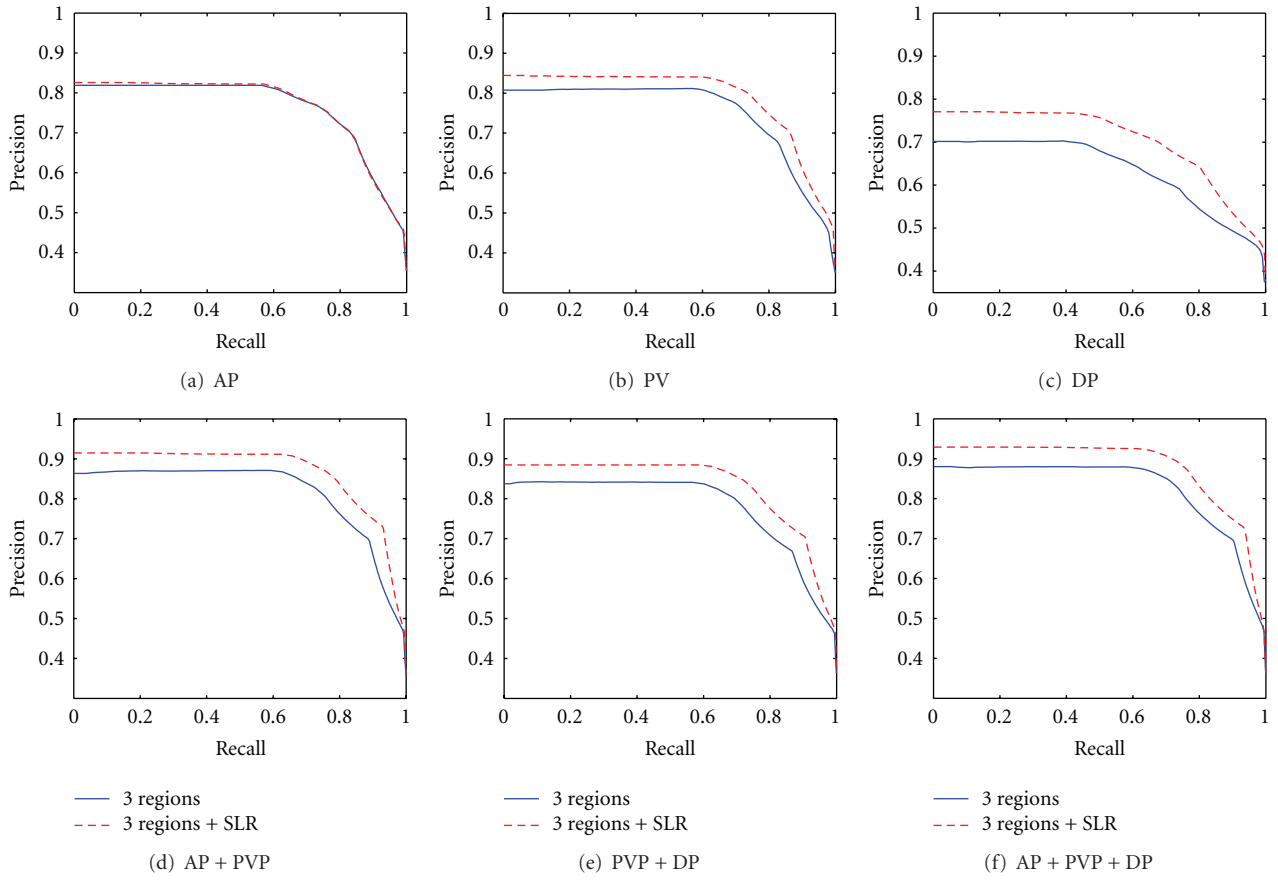


Figure 6: Precious versus recall curve in triple phases based on 3 Regions and 3 Regions with neighborhood, and SLR represents surrounding liver parenchyma.

Table 1: Retrieval result based on 3 regions with surrounding liver parenchyma.

Staging scan	MAP		P(10)		P(20)	
	RLDA	LDP	RLDA	LDP	RLDA	LDP
AP	0.6160	0.6135	0.8259	0.8192	0.8259	0.8190
PVP	0.6288	0.6333	0.8445	0.8575	0.8445	0.8575
DP	0.5972	0.5960	0.7707	0.7723	0.7707	0.7723
AP + PVP	0.6650	0.6651	0.9146	0.9144	0.9141	0.9144
PVP + DP	0.6522	0.6536	0.8845	0.8900	0.8845	0.8900
AP + PVP + DP	0.6755	0.6681	0.9292	0.9127	0.9292	0.9123

alone, whereas the combination of intensity, shape, and texture yields better results than the combination of intensity and shape. BoW alone outperforms the combination of common features, whereas the combination of all features mentioned is superior to BoW alone. Notably, the CBIR system based on our algorithm is better than the system based on other different feature extraction algorithms from Figure 9 due to the fact that our proposed approach can express the imaging characteristics of lesions in triple phases.

3.3. Discussion of the Number of Regions for Lesions. The number of regions that the lesion is divided into is set as parameter s. The effects of parameter s on our retrieval system are discussed in this section. Table 2 shows the average precision after retrieving the top 20 cases with the parameter s from 2–5. For convenience, only the dual-phase and triple-phase scans were used. When s is 2 and 5, the results of all multiple phases are below 90%; when s is 3 and 4, the results of some multiple phases are better than 90%; when s equal

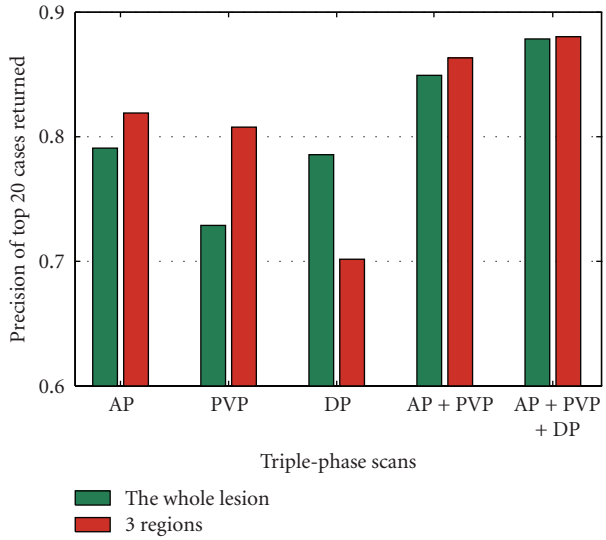


Figure 7: The retrieval performance comparison based on whole lesion and 3 regions.

to 3, the best retrieval performance is achieved of all the s values, as shown in Table 2.

3.4. Influence of the Amount of Clusters. Figure 10 shows the effects of the amount of clusters (i.e., codebook size) using distance metric L1, L2, RLDA, LDP. The plots show that performance increases with the number of codebook sizes; however, a large dictionary results in high computation cost. Thus, the codebook size is set to1024 in our paper.

4. Conclusion

We have developed a regional BoW feature extraction algorithm for lesion images. The proposed algorithm is mainly based on the imaging characteristics of lesion visible on contrast-enhanced triple-phase CT images. Our CBIR system, which incorporates the proposed feature extraction algorithm, can practically retrieve three types of liver lesions that appear similar. The accurate assessment of our approach shows reasonable retrieval results that are in accordance with the diagnoses of radiologists. Our system can aid in decision-making related to the diagnosis of hepatic tumors and support radiologists in multiphase contrast-enhanced CT images by showing them similar patients in lesions.

5. Discussion

The development of a feature extraction algorithm for lesion images is presented in this paper. Our experiments show that a CBIR system incorporated with this algorithm can yield excellent retrieval results. The development of this algorithm considers the imaging characteristics of three lesions and their surrounding normal liver parenchyma in contrast-enhanced triple-phase CT images. This algorithm combines feature vectors with the characteristic of the ROI, which is very essential in retrieval systems. Thus, our algorithm is

Table 2: Retrieval results with different number of regions.

	s	AP + PVP	PVP + DP	AP + PVP + DP
2	RLDA	0.8795	0.8670	0.8969
	LDP	0.8849	0.8724	0.8968
3	RLDA	0.9141	0.8845	0.9292
	LDP	0.9144	0.8900	0.9123
4	RLDA	0.9080	0.8836	0.8906
	LDP	0.9136	0.8895	0.9079
5	RLDA	0.8621	0.8443	0.8742
	LDP	0.8621	0.8382	0.8799

powerful and more advantageous compared with common low-level feature vectors. The system may serve as useful aided diagnosis system for inexperienced or experienced radiologists in searching databases of radiologic imaging and obtaining good retrieval results.

A number of studies have been conducted on various hepatic tumor imaging technologies [23, 24, 26]. Mougiakakou et al. [23] defined an aided diagnosis system for normal liver, hepatic cyst, hemangioma, and HCC in nonenhanced CT scans. Zhang and his colleagues [24] used an aided diagnosis system to segment and diagnose enhanced CT and MR images of HCC. More recently, a CBIR system that closely resembles the current study was presented [26]. In this system, metastases, hemangiomas, and cysts were all visible on portal venous phase images. However, the images used common low-level features such as intensity, texture, and shape, and did not consider the imaging characteristics of lesions in multiphase scans. In our opinion, multiphasic imaging is central to current clinical diagnosis.

In the present paper, the use of BoW was verified to be effective. Although existing image retrieval technologies achieved some good performances, they still have some limitations. Most of image retrieval technologies are based on the underlying characteristics of the images and used low-level features. Therefore, they are unable to resolve the semantic gap problem, which is the inconsistency between the low-level visual features and the high-level semantic features. BoW, as a high-level feature [38], has obtained great success in text retrieval problem, because of its speed and efficiency, and has been gaining recognition in its use in problems such as object recognition and image retrieval from large databases [22]. The BoW framework ignores the spatial configuration between visual words (i.e., the link between the characteristics and location) and can cause information loss. However, this framework can quickly and easily build a design model. In the proposed feature extraction algorithm, the spatial structure information of images is considered by dividing the lesion into three regions, which compensates for the lack of BoW. Thus, BoW successfully represents the regional features.

Semantic features are not considered in our study because of two factors. First, if the query patient remains undiagnosed, semantic features cannot be used because of lack of radiology reports. Second, radiologists may use different terminology to describe the same observation [39, 40].

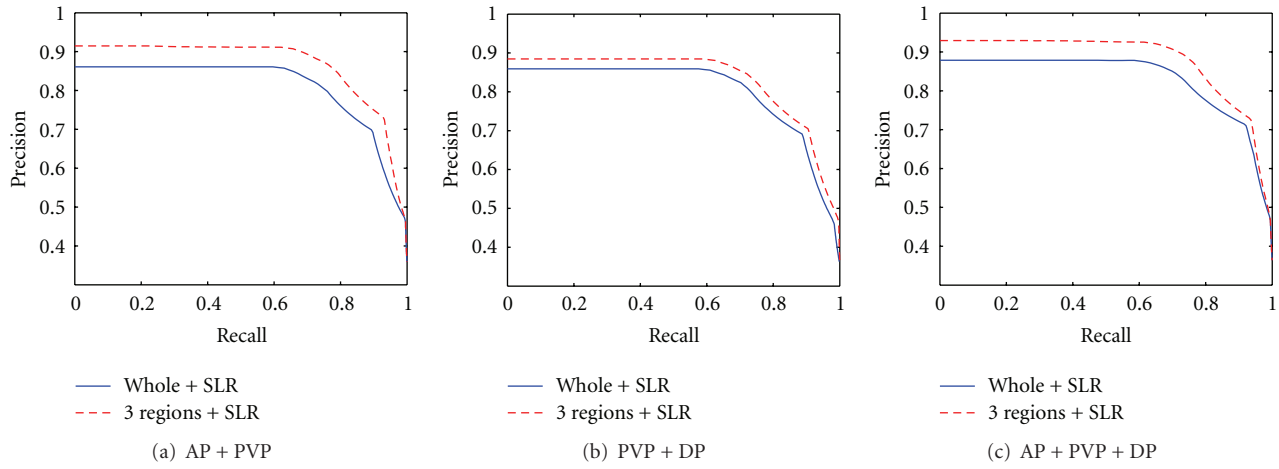


Figure 8: Precision-recall curve of dual-phase and three-phase scanning retrieval, and SLR represents surrounding liver parenchyma.

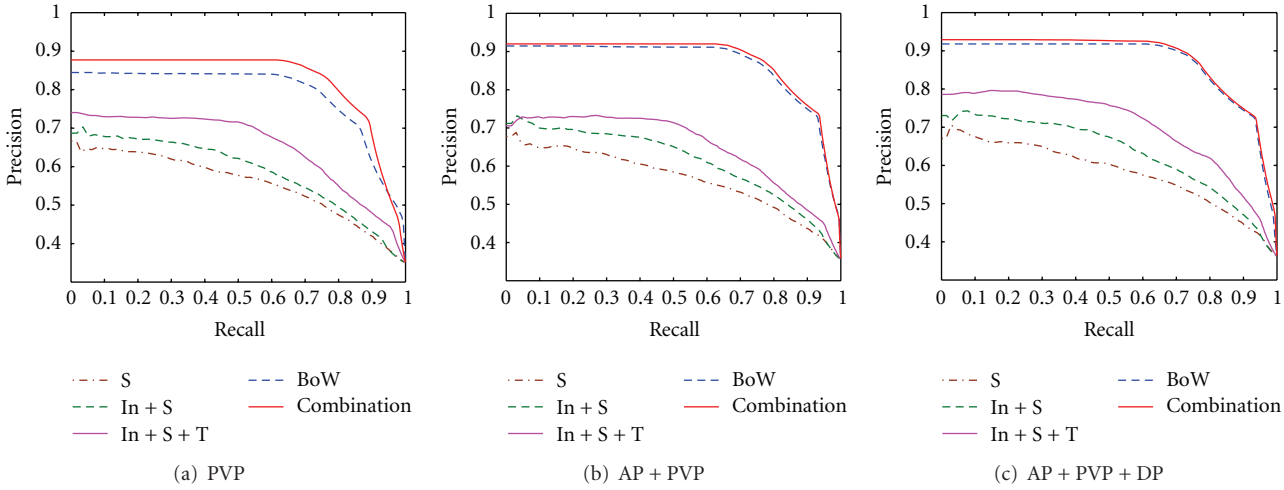


Figure 9: Precision-recall curve using different features in PVP, dual-phase and triple-phase scans.

Thus, some inevitable subject factors exist in the semantic annotation.

The number of regions that the lesions are divided into is set mainly due to the triple phase scans of the three tumors. The optimal number of regions is 3, as revealed in our previous some experiments, and we have verified that retrieval performance is the best of the number from 1 to 5. Dividing the lesion into three regions fits the best imaging behavior of lesion in triple phases described in Section 2.1. Thus, our selected number of regions is theoretically conducive to our proposed feature extraction algorithm.

Our retrieval system was implemented and evaluated according to the patients, namely, the images were grouped according to the patient and the patient is the primary unit of the query and retrieval. This scheme is very different from the traditional CBIR system, where single image or single slice is used as query and retrieval. This patient-based fashion is more helpful for the diagnosis aid, because the multiphase images from the retrieved patient obviously could supply more information than just one image for making decision for current query patient. The development of our feature extraction algorithm is focused on the views of imaging findings of the three lesions in triple-phase scans. Therefore, the retrieval process in our experiments follows single-, dual-, and triple-phase scans to verify the practicality of our retrieval system based on the proposed method.

Our study mainly has two main limitations. The first is number of lesions types used, which was limited to three. The proposed feature extraction algorithm was developed based on the imaging characteristics of three lesions (HCC, hemangiomas, and cysts) in contrast-enhanced triple-phase images. HCC is a common malignant tumor, whereas hemangiomas and cysts are the common benign cells. Studies on hepatic lesions and lesions in other body areas can be extended in future work to encourage continued development of relevant feature extraction methods. The second limitation

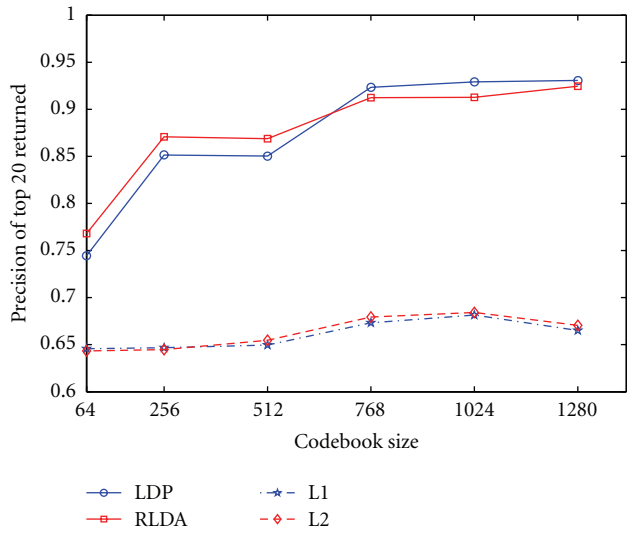


Figure 10: The average precision of top 20 cases returned in triple-phase scan with different dictionary sizes.

is the segmentation of lesions used in our system. Lesions should first be segmented from abdominal CT images because lesions contained important imaging information for image retrieval. Several segmentation algorithms [24] have been proposed to achieve automatic or semiautomatic segmentation in medical image analysis. However, because of the complexity of medical images and lesion infiltration, no standard method can generate satisfactory segmentation results for all hepatic-enhanced images. Consequently, manual segmentation is employed by imaging experts to obtain more accurate lesion images.

In conclusion, the CBIR system based on our proposed feature extraction algorithm has practical application in aided diagnosis, which can help radiologists to retrieve images that contain similar appearing lesions.

References

[1] S. G. Armato, M. F. McNitt-Gray, A. P. Reeves et al., "The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans," *Academic Radiology*, vol. 14, no. 11, pp. 1409–1421, 2007.

[2] W. E. Barlow, C. Chi, P. A. Carney et al., "Accuracy of screening mammography interpretation by characteristics of radiologists," *Journal of the National Cancer Institute*, vol. 96, no. 24, pp. 1840–1850, 2004.

[3] P. J. A. Robinson, "Radiology's Achilles' heel: error and variation in the interpretation of the Rontgen image," *British Journal of Radiology*, vol. 70, pp. 1085–1098, 1997.

[4] K. Yuan, Z. Tian, J. Zou, Y. Bai, and Q. You, "Brain CT image database building for computer-aided diagnosis using content-based image retrieval," *Information Processing and Management*, vol. 47, no. 2, pp. 176–185, 2011.

[5] Y. L. Huang, J. H. Chen, and W. C. Shen, "Computer-aided diagnosis of liver tumors in non-enhanced CT images," *Journal of Medical Physics*, vol. 9, pp. 141–150, 2004.

[6] E. L. Chen, P. C. Chung, C. L. Chen, H. M. Tsai, and C. I. Chang, "An automatic diagnostic system for CT liver image classification," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 6, pp. 783–794, 1998.

[7] H. Greenspan and A. T. Pinhas, "Medical image categorization and retrieval for PACS using the GMM-KL framework," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 2, pp. 190–202, 2007.

[8] D. K. Iakovidis, N. Pelekis, E. E. Kotsifakos, I. Kopanakis, H. Karanikas, and Y. Theodoridis, "A pattern similarity scheme for medical image retrieval," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 442–450, 2009.

[9] M. M. Rahman, B. C. Desai, and P. Bhattacharya, "Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion," *Computerized Medical Imaging and Graphics*, vol. 32, no. 2, pp. 95–108, 2008.

[10] Y. L. Huang, S. J. Kuo, C. S. Chang, Y. K. Liu, W. K. Moon, and D. R. Chen, "Image retrieval with principal component analysis for breast cancer diagnosis on various ultrasonic systems," *Ultrasound in Obstetrics and Gynecology*, vol. 26, no. 5, pp. 558–566, 2005.

[11] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 373–378, 2003.

[12] J. E. E. de Oliveira, A. M. C. Machado, G. C. Chavez, A. P. B. Lopes, T. M. Deserno, and A. D. A. Araújo, "MammoSys: a content-based image retrieval system using breast density patterns," *Computer Methods and Programs in Biomedicine*, vol. 99, no. 3, pp. 289–297, 2010.

[13] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: application to digital mammography," *IEEE Transactions on Medical Imaging*, vol. 23, no. 10, pp. 1233–1244, 2004.

[14] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, 2001.

[15] M. Bober, "MPEG-7 visual shape descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 716–719, 2001.

[16] F. Farzanegan, "Keep AFIP," *JACR Journal of the American College of Radiology*, vol. 3, no. 12, p. 961, 2006.

[17] L. Kreel, M. M. Arnold, and Y. F. Lo, "Radiological-pathological correlation of mass lesions in the liver," *Australasian Radiology*, vol. 35, no. 3, pp. 225–232, 1991.

[18] A. Rosenfeld and J. L. Pfaltz, "Sequential operations in digital picture processing," *Journal of ACM*, vol. 13, no. 4, pp. 471–494, 1996.

[19] F. F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 524–531, June 2005.

[20] H. L. Luo, H. Wei, and L. L. Lai, "Creating efficient visual codebook ensembles for object categorization," *IEEE Transactions on Systems, Man, and Cybernetics Part A*, vol. 41, no. 2, pp. 238–253, 2010.

[21] J. C. van Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Comparing compact codebooks for visual categorization," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 450–462, 2010.

[22] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1243–1256, 2008.

[23] S. G. Mougiakakou, I. K. Valavanis, A. Nikita, and K. S. Nikita, "Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers," *Artificial Intelligence in Medicine*, vol. 41, no. 1, pp. 25–37, 2007.

[24] X. Zhang, H. Fujita, T. Qin et al., "CAD on liver using CT and MRI," in *Proceedings of the 2nd International Conference on Medical Imaging and Informatics (MIMI '07)*, vol. 4987 of *Lecture Notes in Computer Science*, pp. 367–376, Beijing, China, 2008.

[25] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[26] S. A. Napel, C. F. Beaulieu, C. Rodriguez et al., "Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results," *Radiology*, vol. 256, no. 1, pp. 243–252, 2010.

[27] C. G. Zhao, H. Y. Cheng, Y. L. Huo, and T. G. Zhuang, "Liver CT-image retrieval based on Gabor texture," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '04)*, pp. 1491–1494, September 2004.

[28] O. Chapelle and M. Wu, "Gradient descent optimization of smoothed information retrieval metrics," *Information Retrieval*, vol. 13, no. 3, pp. 216–235, 2010.

[29] T. Qin, T. Y. Liu, and H. Li, "A general approximation framework for direct optimization of information retrieval measures," *Information Retrieval*, vol. 13, no. 4, pp. 375–397, 2010.

[30] M. Taylor, J. Guiver, S. Robertson, and T. Minka, "SoftRank: optimizing non-smooth rank metrics," in *Proceedings of the International Conference on Web Search and Data Mining (WSDM '08)*, pp. 77–85, February 2008.

[31] H. Chang and D. Y. Yeung, "Kernel-based distance metric learning for content-based image retrieval," *Image and Vision Computing*, vol. 25, no. 5, pp. 695–703, 2007.

[32] J. H. Friedman, "Regularized discriminant analysis," *Journal of Amercian Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[33] J. Ye and T. Wang, "Regularized discriminant analysis for high dimensional, low sample size data," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pp. 454–463, August 2006.

[34] D. Cai, X. He, and J. Han, "SRDA: an efficient algorithm for large scale discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 1, pp. 1–12, 2008.

[35] G. Hua, M. Brown, and S. Winder, "Discriminant embedding for local image descriptors," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[36] K. Mikolajczyk and J. Matas, "Improving descriptors for fast tree matching by optimal linear projection," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[37] H. Cai, K. Mikolajczyk, and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 338–352, 2011.

[38] C. W. Ngo, Y. G. Jiang, X. Y. Wei et al., "Experimenting VIREO-374: bag-of-visual-words and visual-based ontology for semantic video indexing and search," in *Proceedings of the TRECVID Workshop*, November 2007.

[39] J. L. Sobel, M. L. Pearson, K. Gross et al., "Information content and clarity of radiologists' reports for chest radiography," *Academic Radiology*, vol. 3, no. 9, pp. 709–717, 1996.

[40] M. J. Stoutjesdijk, J. J. Fütterer, C. Boetes, L. E. Van Die, G. Jager, and J. O. Barentsz, "Variability in the description of morphologic and contrast enhancement characteristics of breast lesions on magnetic resonance imaging," *Investigative Radiology*, vol. 40, no. 6, pp. 355–362, 2005.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 528781, 12 pages
doi:10.1155/2012/528781

Research Article

An Automated Optimal Engagement and Attention Detection System Using Electrocardiogram

Ashwin Belle, Rosalyn Hobson Hargraves, and Kayvan Najarian

Department of Computer Science, School of Engineering, Virginia Commonwealth University, 401 West Main Street, P.O. Box 843019, Richmond, VA 23284-3019, USA

Correspondence should be addressed to Ashwin Belle, bellea@vcu.edu

Received 1 May 2012; Accepted 18 June 2012

Academic Editor: Alberto Guillén

This research proposes to develop a monitoring system which uses Electrocardiograph (ECG) as a fundamental physiological signal, to analyze and predict the presence or lack of cognitive attention in individuals during a task execution. The primary focus of this study is to identify the correlation between fluctuating level of attention and its implications on the cardiac rhythm recorded in the ECG. Furthermore, Electroencephalograph (EEG) signals are also analyzed and classified for use as a benchmark for comparison with ECG analysis. Several advanced signal processing techniques have been implemented and investigated to derive multiple clandestine and informative features from both these physiological signals. Decomposition and feature extraction are done using Stockwell-transform for the ECG signal, while Discrete Wavelet Transform (DWT) is used for EEG. These features are then applied to various machine-learning algorithms to produce classification models that are capable of differentiating between the cases of a person being attentive and a person not being attentive. The presented results show that detection and classification of cognitive attention using ECG are fairly comparable to EEG.

1. Introduction

In today's high-paced, hi-tech, and high-stress environment, a common sufferer is our cognitive processing and capacity. Cognitive psychology primarily deals with people's ability to acquire, process, and retain information which is a fundamental necessity for task execution [1]. Quality of task performance largely depends on the individual's capacity to inculcate and sustain high levels of engagement and attention during cognitive activities. However, considering the perils of modern lifestyles such as extended work hours, long to-do lists, and neglected personal health coupled with repetitious nature of daily activities and professions, sleep deprivation and fluctuating attention levels as well are becoming a commonplace issue that needs to be tackled. Momentary or prolonged lapse of attention for certain critical professions such as doctors, pilots, defense personnel, and road transportation drivers can be catastrophic and sometimes deadly.

Studying alertness and drowsiness is not a new domain in scientific research. Numerous research areas are actively studying the concepts of attention, alertness, distraction, and drowsiness. Many of these researches focuses on nonsensory mechanisms to identify and quantify levels of attention in individuals [2–5] such as user's daily routine, schedules, activities, with self-reports from users describing patterns in activities and attention levels and so forth. More recently researchers have begun using biosignals to understand the complex implication of cognitive processing on physiological parameters. Electroencephalogram (EEG) is a popular example of a physiological signal that researchers use extensively in understanding cognitive functioning [6–8]. The use of EEG for detecting and identifying attention/focus in individuals is an established concept. Several concepts have been developed for improving concentration and other cognitive functions of both attention-related disorder and head trauma patients [9–11]. However, there are some fundamental issues regarding the procedure of collecting EEG. It requires the individual to wear a head gear which can be disruptive and troublesome for long-duration usage. The EEG electrode sensors also need to be moistened with electrode gel which can be uncomfortable for the user at the contact points on the scalp.

Also, the EEG collection device is usually not designed to be portable; they tend to be slightly large fixed devices which make the collection of EEG confined to a set of environmental contingencies. Furthermore, the EEG signal itself is highly sensitive to noise. Movement of the muscles around the scalp, movement of the subject, talking, blinking, and so forth can induce various unwanted artifacts into the signal thereby disrupting the quality of neuroelectric information contained within the signal.

For this reason, this research is attempting to use Electrocardiogram (ECG) for detecting cognitive attention in individuals. The ECG is a fundamental physiological signal which can be collected easily with a tiny wearable and portable monitor. Since the collection device is portable and has a small footprint on the body, it allows the capture of ECG signals from individuals in various situations in a noninvasive manner. The portability of such a data collection unit allows a more realistic study of human cognitive activities during task execution under various circumstances. The research presented in this paper is attempting to establish a correlation between cognitive attention and its implications on ECG. By being able to identify a pattern and correlation between the two it becomes possible to predict well in advance, an individual's potential loss of attention and ingression of sleepiness during a task execution. This also provides the ability for preemptive feedback to the user upon identifying diminishing attention levels and thereby improving the individuals' overall performance.

The rest of this paper is organized as follows: Section 2 describes the experimental setup, followed by a description of methods in Section 3. Section 4 describes the results and conclusion of this research.

2. Experimental Setup

An essential aspect of this research has been the collection of the data itself. Extensive search revealed that there was no dataset available, freely or otherwise, which catered to the exact needs to this particular study. Since the study is about utilizing ECG collected via a portable armband to detect the presence or lack of attention/focus in an individual, the dataset had to be collected specifically based on the requirements of this research.

In the designed experiment, volunteer subjects were individually asked to watch a series of preselected video clips during which two physiological signals, that is, the ECG and EEG, were acquired. Based on their content, the chosen video clips fell in either of two categories that is either "interesting" or "noninteresting," requiring high and low levels of viewer engagement, respectively. The average length of each selected video clips was about 4-minute long. For each category the respective video clips were put together to form a video montage of about 20-minute viewing duration. The first category of the video montage named "interesting" included engaging scenes from documentaries, popular movie scenes, high-speed car chases, and so forth. which were intended to keep the viewers attentive and engaged with its content. The second video montage named "noninteresting" contained videos which were repetitive and monotonous in nature such as a clock ticking and still images shown for extended periods of time. These were intended to induce boredom in subjects and thereby reduce their attentiveness. Viewing the two categories of video montages one after the other required contrasting levels of engagement and focus from the participant, thereby ensuring (as far as possible) that the subjects were interested and paid attention to the interesting video set and the subjects were subsequently bored and lost focused attention during the noninteresting videos.

Figure 1: Two leads ECG collection from Armband.

During the experiment the ECG signal was collected using the SenseWear-Pro armband developed by Bodymedia Inc. This armband is capable of collecting ECG data at 128 Hz [12].

As shown in Figure 1, two leads from the armband are attached to the subject using ECG adhesive electrodes patches. One lead of the leads is placed on the side of the arm and the other lead is fastened on the bridge between the neck and shoulder.

The EEG signal was collected from the subjects using MP150: EEG-100C a product by Biopac Inc. With this system an EEG cap is provided that fits snug on the head of the subject and it collects the EEG signal at a sampling rate of 1000 Hz. Signals were collected from the forehead or the frontal cortex (fp1 and fp2) with a ground reference from the ear lobe. The frontal cortex is primarily responsible for attention and higher-order functions including working memory, language, planning, judgment, and decision-making [13]. The entire setup is completely noninvasive and only utilizes surface contact sensors. The data collection has been conducted with required IRB approval.

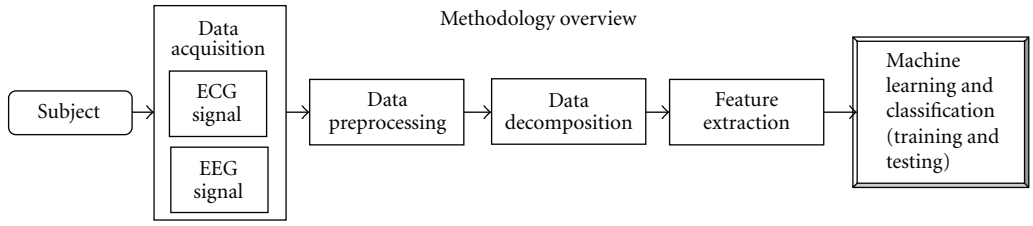


Figure 2: Methodology overview.

3. Methods

The schematic diagram in Figure 2 illustrates the overall method of this study. As shown the two physiological signals ECG and EEG are acquired from the subject during the experiment.

The acquired raw signals are first preprocessed to remove unwanted artifacts presented within the signals. Next the preprocessed signals are decomposed using various decomposition and analysis methods. In the next step valuable and informative features are extracted from the decomposed components of the signal. These extracted features are finally fed to the machine-learning step where classification models are developed to classify the feature instances to either of two cases "attention" or "nonattention."

3.1. Data Preprocessing. The acquired raw ECG signal contains some inherent unwanted artifacts that need to be dealt with before any analysis can be performed on it. The cause of these artifacts, which is usually frequency noise or baseline trend, could be due to a number of reasons such as subjects' movement causing motion artifacts, breathing patter artifact, loose skin contact of the electrodes, and electric interference (usually found around 55 Hz). Therefore a preprocessing step has been designed to ensure that the signal is as clean and artifact free before analysis.

3.1.1. ECG Preprocessing. The preprocessing steps for the ECG signal are shown in Figure 3. Since each signal has to be filtered differently based on the type of inherent noise, the raw ECG signal is first filtered using "SGolay" filtering method. The "SGolay" filter was developed by Savitzky-Golay. This filter is a digital polynomial filter based on least square smoothing mechanism. The SGolay filters are typically used to smooth out a noisy signal with a large frequency span. They perform better than standard averaging FIR filters, since these filters tend to retain a significant portion of the signals high-frequency content while removing only the noise [14].

Next, the filtered ECG data is sent through a baseline drift removal step. Typically baseline drift is observed in ECG recordings due to respiration, muscle contraction, and electrode impedance changes due to subject's movement [15]. To remove the baseline drift first the regression line that best fits the samples within a window of size equal to the sampling rate is determined.

Given n points of the ECG signal (x_1, y_1),(x_2, y_2),... , (x_n, y_n), the best fit line associated with these points can be computed as follows:

$$m = \frac{n\left(\sum_1^n xy\right) - \left(\sum_1^n x\right)\left(\sum_1^n y\right)}{n\left(\sum_1^n x^2\right) - \left(\sum_1^n x\right)^2},$$
$$b = \frac{\sum_1^n y - m\left(\sum_1^n x\right)}{n}, \quad (1)$$
$$y = mx + b,$$

where y is a point on the line, m is the slope of the line, and b is the intercept. The computed best fit line for each window is then subtracted from the original signal window to obtain a baseline drift-free signal.

After the raw ECG signal has been filtered of noise and baseline drift, the signal is then split into two portions based on the acquisition and experiment framework. The two portions of signals, namely, "interesting" and "noninteresting" are extracted from the original signal using timestamps that are recorded and indexed during signal acquisition. Splitting and analyzing the two sections of data separately facilitate supervised learning mechanism during the training phase in the machine learning step.

3.1.2. EEG Preprocessing. The EEG signal is comprised of a complex and nonlinear combination of several distinct waveforms which are also called band components. Each of the band components is categorized by the frequency range that they exist in. The state of consciousness of the individuals may make one frequency range more pronounced than others [16]. As shown in Figure 4, the different band components are extracted from the raw EEG signal using Butterworth bandpass filters. Five primary bands of the EEG signal are extracted, namely, Delta (0.2–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–55 Hz).

3.2. ECG Decomposition: Using Stockwell Transform. The *S*-transform was proposed by Stockwell and his coworkers in 1996. The distinction of *S*-transform is that it produces decomposition of frequency-dependant resolution in the time-frequency domain while entirely retaining the local phase information. In other words, the *S*-transform not only estimates the local power spectrum, but also the local phase spectrum, which is highly desirable in studying complex physiological signals such as the ECG.

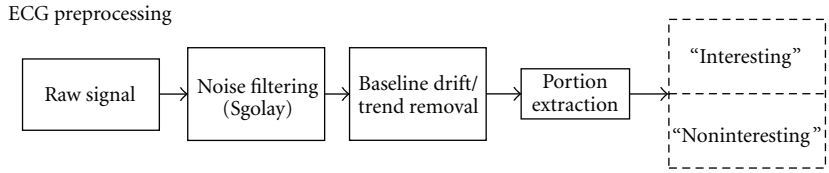


Figure 3: ECG preprocessing.

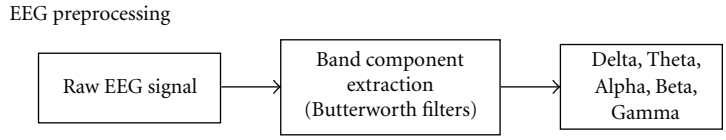


Figure 4: EEG preprocessing steps.

When it comes to analyzing dynamic spectrum or local spectral nature of nonstationary observations such as the ECG some of the popular methods include Short-Time Fourier Transform (STFT) [17], Gabor transform [18], complex demodulation [19] which produces a series of band pass filtered voices and is also related to the filter bank theory of wavelets and so forth. Some methods represent the transformation in a combination of time and frequency domain such as the Cohen class [20] of generalized time-frequency distributions (GTFD), Cone-Kernel distribution [21], Choi-Williams distribution [22] as well as the smoothed pseudo Wigner distribution (PWD) [23]. One of the more popular methods for decomposition and analysis in time-frequency domain is Wavelet Transform. Discrete Wavelet Transform or DWT performs decomposition of a signal that provides excellent time resolution while maintaining key spectral information or frequency resolution [24, 25].

Although *S*-transform is similar to wavelet transform in having progressive resolution, unlike wavelet transform, the *S*-transform retains absolutely referenced phase information. Absolutely referenced phase implies that the phase information calculated by the *S*-transform is referenced to time $t = 0$, which is also true for the phase given by the Fourier transform. The only difference being the *S*-transform provides the absolute referenced phase information for each sample of the time-frequency space.

3.2.1. Mathematical Formulation of S-Transform. There are two varieties of *S*-transform, continuous and discrete. The continuous *S*-transform [26] is essentially an extension of the STFT. It can also be seen as a phase-corrected format of the Continuous Wavelet Transform (CWT).

The STFT of a signal *h(t)* is defined as

$$\mathrm{STFT}(\tau, f) = \int_{-\infty}^{\infty} h(t)\, g(\tau - t) e^{-j2\pi f t} dt, \tag{2}$$

where

(i) τ is the time of spectral localization,

(ii) f is the Fourier frequency,

(iii) $g(t)$ denotes a window function.

The *S*-transform can be derived from the above STFT equation simply by substituting the window function $g(t)$ the Gaussian function:

$$g(t) = \frac{|f|}{\sqrt{2\pi}} e^{-(t^2 f^2)/2}. \tag{3}$$

Therefore the *S*-transform be mathematically defined as follows:

$$S(\tau, f) = \int_{-\infty}^{\infty} h(t)\, \frac{|f|}{\sqrt{2\pi}} e^{-((\tau - t)^2 f^2)/2} e^{-j2\pi f t} dt. \tag{4}$$

Since *S*-transform essentially functions with the Gaussian window during decomposition, it can be deduced that with a wider window in the time domain the transformation can provide better resolution for lower frequency, and with a narrow Gaussian window the resolution for higher frequency is better accentuated.

For application of *S*-transform in this study, the continuous *S*-transform does not prove to be a practical choice. Simply because the acquisitions of the ECG signal itself were performed with discrete sampling and also a continuous decomposition of this signal for all frequencies can be extremely time consuming, thereby not computationally pragmatic. Hence a Discrete version of the *S*-transform has been adopted for the decomposition of the ECG signal.

The discrete *S*-transform can be presented as follows.

Let $h[kT]$ be the discrete time series signal to be investigated, where $k = 0, 1, \ldots, N - 1$, and T is the time sampling interval. The discrete format of the Fourier transform can be shown as follows:

$$H\left[\frac{n}{NT}\right] = \frac{1}{N} \sum_{k=0}^{N-1} h[kT] e^{2j\pi nk/N}. \tag{5}$$

Using the continuous *S*-transform equation and the above equation, the time series, $h[kT]$'s *S*-transform can be represented as follows: (making $f \to n/NT$ and $\tau \to jT$)

$$S\left[jT, \frac{n}{NT}\right] = \sum_{m=0}^{N-1} H\left[\frac{m+n}{NT}\right] e^{2\pi^2 m^2/n^2} e^{2j\pi mj/N}, \quad n \neq 0, \tag{6}$$

where j, m, and $n = 0, 1, \ldots, N - 1$.

3.2.2. Application of S-Transform. Figure 5 shows the different steps involved in the decomposition of the ECG signal using *S*-transform. First, the preprocessed ECG signal is sent through a windowing mechanism. In this mechanism, the preprocessed ECG signal is partitioned into tiny windows. These windows are nonoverlapping and contain ECG data of 10 sec interval (128 Hz $*$ 10 sec = 1280 data-points/window).

After the windowing step, each of the 10 seconds windows is decomposed using *S*-transform. The output of the *S*-transform is a complex 2-dimensional matrix with rows representing the frequencies and the columns represent the time values. The *S*-transform algorithm applied in this study is tuned to produce a stepwise frequency range with step size being 1 Hz and the time interval between samples in the result is 1 step unit.

An example output of a 5-second window of an ECG data after *S*-transform is given in Figure 6 .

Figure 6 shows the exact point-to-point representation of the original (Figure 6(b)) signal in the *S*-transforms time-frequency domain. The *S*-transform output matrix has been shown in a contour map display (Figure 6(a)).

3.2.3. Feature Extraction. The output of each window is a frequency-time represented matrix. Each instance of the matrix is frequency point and a time point (by the row and column position, resp.). So the entire output matrix can be presented as follows: ST(x, y), where x is the frequency (row) location and y is the time (column) location.

The extraction of features from the derived output matrix of ST is performed in two steps. In the first step the output matrix is reduced from two dimensions to a single dimension. This is done by computing certain statistical measures along the frequency dimension x, while retaining the discreteness in the time dimension y as is. The computed statistical measures along frequencies (f) are as follows:

(i) mean of frequencies (f),

(ii) sum of frequencies (f),

(iii) product of frequencies (f),

(iv) standard Deviation of frequencies (f),

(v) range (f).

At the end of the first step we get an array of features from the frequency domain as follows:

$$\text{Freq}_{\text{fets}} = [\text{mean}(f), \text{sum}(f), \text{product}(f), \text{std}(f), \text{range}(f)]. \quad (7)$$

The next step is to compute statistical features along the time domain.

(i) Mean:

$$\text{mean}(\text{ST}) = \text{mean}(f_i), \quad \text{where } f_i \in \text{Freq}_{\text{fets}}. \quad (8)$$

(ii) Sum:

$$\text{sum}(\text{ST}) = \text{sum}(f_i), \quad \text{where } f_i \in \text{Freq}_{\text{fets}}. \quad (9)$$

(iii) Mean of autocovariance:

$$\text{mean}(\text{autocovariance}(\text{ST})) = \text{mean}(\text{autocovariance}(f_i)), \quad (10)$$

where $f_i \in \text{Freq}_{\text{fets}}$.

(iv) Sum of cross-correlation:

$$\text{sum}(\text{autocorrilation}(\text{ST})) = \text{sum}(\text{autocorrelation}(f_i)), \quad (11)$$

where $f_i \in \text{Freq}_{\text{fets}}$.

(v) Log_2 of Variance:

$$\text{Log}_2(\text{variance}(\text{ST})) = \text{Log}_2(\text{variance}(f_i)), \quad (12)$$

where $f_i \in \text{Freq}_{\text{fets}}$.

Two additional features are calculated from the initially obtained ST matrix.

(i) Mean of max frequencies:

$$\text{mean}(\max(\text{ST})) = \text{mean}\left(\max\left(\text{ST}_{1,y}, \text{ST}_{2,y}, \ldots, \text{ST}_{x,y}\right)\right). \quad (13)$$

(ii) Mean absolute deviation of frequencies:

$$\text{mean}(\text{abs}(\text{ST})) = \text{mean}(\text{abs}(\text{ST} - \text{mean}(\text{ST}))). \quad (14)$$

After the feature extraction has been performed, the total feature set for the *S*-Transform step will contain (5 (features in step 1) $*$ 5 (features in step 2)) + 2 (additional noniterative features) = 27 (features columns per window).

3.3. EEG Decomposition and Analysis: Using Wavelet Transform. The EEG signal exhibits complex behavior and nonlinear dynamics. In the past wide range of work has been done in understanding the complexities associated with the brain through multiple windows of mathematics, physics, engineering and chemistry, physiology, and so forth [27, 28]. The intention of acquiring and analyzing EEG in this research is to develop a benchmark of sorts for attention recognition. The key point of this study is to see if the ECG signal that can be collected from a portable armband can be comparably efficient in recognizing an individual's attention and focus.

The small yet complex varying frequency structure found in scalp-recorded EEG waveforms contains detailed neuroelectric information about the millisecond time frame of underlying processing systems, and many studies indicate that waveform structure at distinct scales holds significant basic and clinical information [29, 30]. Small-scale neural rhythms, in particular event-related oscillation EROs, have been regarded as fundamental to perception and cognition [29]. Wavelet analysis provides a powerful method of isolating such rhythms for study. There are several applications of wavelet transform on EEG analysis. It has been used in removal of noise from raw EEG waveforms since wavelet

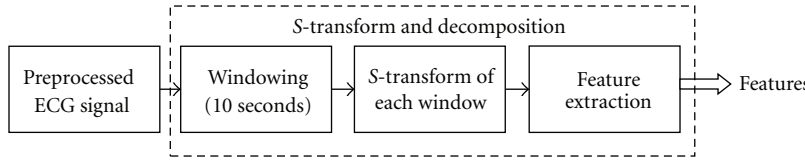


Figure 5: S-Transform application on ECG signal.

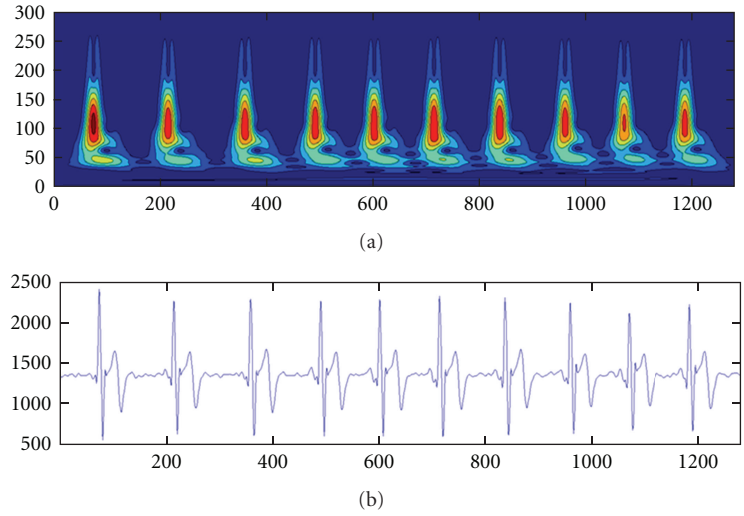


Figure 6: (a) shows the contour-based visualization of frequency spectrum along time, based on the S-transform of the signal window. (b) shows the original signal window.

coefficients facilitate the precise noise filtering mechanism by zeroing out or attenuating any coefficients associated primarily with noise before reconstructing the signal with wavelet synthesis [31–33]. Wavelet analysis of EEG has also been extensively used for signal processing applications in intelligent detection systems for use in clinical settings [34, 35]. Wavelet transform has also been used for compression EEG signals. Wavelet compression techniques have been shown to improve neuroelectric data compression ratios with little loss of signal information [36, 37]. It can also be seen for component and event detection as well as spike and transient detection within the EEG waveforms. Wavelet analysis has proven quite effective in many research studies [33–38].

3.3.1. Mathematical Formulation of Wavelet Transform. Wavelet transforms essentially exist in two distinct types: the Continuous Wavelet Transform (CWT) and the Discrete Wavelet Transform (DWT). In this study for the analysis of the EEG signal the DWT method has been employed. The advantages of using DWT is that it allows the analysis of signals by applying only discrete values of shift and scaling to form the discrete wavelets. Also, if the original signal is sampled with a suitable set of scaling and shifting values, the entire continuous signal can be reconstructed from the DWT (using Inverse-DWT). A natural way of setting up the parameters a (scaling) and b (shifting) is to use a logarithmic discretization of the "a" scale and link this, respectively, to the step size taken between "b" locations or shifts. To link "b" to "a" discrete steps are taken to each location "b," which are proportional to the "a" scale. This kind of mother wavelet can be shown in the following form.

Discrete mother wavelet representation:

$$\Psi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}}\left(\frac{t - nb_0a_0^m}{a_0^m}\right), \tag{15}$$

where

(i) integer's m and n control the wavelet shifting and scaling, respectively,

(ii) a_0 is a specified fixed dilation step parameter set at a value greater than 1,

(iii) b_0 is the location parameter which must be greater than zero.

Analysis equation (DWT):

$$W_{mn} = \int_{-\infty}^{+\infty} x(t)\Psi_{mn}^*(t)dt. \tag{16}$$

Synthesis equation (inverse DWT):

$$x(t) = c\sum_m\sum_n W_{mn}\Psi_{mn}(t), \tag{17}$$

where c is a constant associated with the mother wavelet.

3.3.2. Application of DWT on EEG. In this study, Discrete Wavelet Transform or DWT is applied to the EEG band components which are extracted in the preprocessing step.

As shown in Figure 7, each of the extracted band components is sent through the "windowing" step. In this step the interesting and boring portions of the band components based on the timestamps of the original EEG are extracted and sent through a windowing mechanism. In this mechanism, each band component signal is partitioned into tiny windows. The windows are 10-second long and are nonoverlapping. The EEG signal is acquired at a sampling rate of 1000 Hz, so each window will have 1000 Hz $*$ 10 sec = 10000 data points.

Each window is then decomposed using DWT. Performance of the Wavelet transform depends on the mother wavelet chosen for decomposition of the signal. A common heuristic is to choose one similar to the shape of the signal of interest. So for the set of band components that is extracted from the original EEG signal different mother wavelets that suit different bands are applied during decomposition.

As shown in Figure 8, the analysis of the Gamma wave component, the mother wavelet chosen is the "bior3.9" from the bi-orthogonal family of wavelets. Delta, Theta, and Alpha wave components are decomposed using "db4" as their mother wavelet from the Daubechies family of wavelets. Finally Beta waves are decomposed using "coif3" as the mother wavelet from the Coiflets wavelet family. These wavelets were chosen not only based on the shape and complexity but also because they seemed to be commonly used for such application in related research.

The decomposition process in wavelet transform can be performed iteratively into several levels. The number of levels chosen for decomposition is application specific and also depends on the complexity of the signal. For window of the EEG signal band components, 5 levels of decomposition seemed to provide all the required useful information; further decomposition did not yield a better result. The detailed coefficients of all the stages from 1 through 5 and the approximation coefficient of level 5 are retained for feature extraction step.

3.3.3. Feature Extraction Step. The features computed from these coefficients are as follows. (Here, $(x_1, x_2, \ldots, x_n)$ represents the values of each coefficient from the 10 sec window.)

(i) Standard deviation:

$$\text{std} = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \tag{18}$$

(ii) Entropy: entropy is a statistical measure of randomness. It is very useful in evaluating the information present within a signal:

$$\text{entropy} = -\text{sum}(p * \log 2(p)), \tag{19}$$

where p is the histogram of the signal.

(iii) Log of variance: let the probability mass function of each element be as follows $x_1 \mapsto p_1, \ldots, x_n \mapsto p_n$, then

$$\text{Variance} = \sum_{i=1}^{n} p_i * (x_i - \mu)^2, \tag{20}$$

where μ is the expected value, that is,

$$\mu = \sum_{i=1}^{n} p_i * x_i. \tag{21}$$

Therefore, Log of variance $= \log_2(\text{Variance}(x))$.

(iv) Mean of frequencies (discrete Fourier domain):

$$\text{dft}(x_k) = \sum_{k=1}^{N-1} X(j) e^{j(2\pi/N)kn}, \tag{22}$$

where a net of N time samples, $\text{dft}(x_k)$, represents the magnitude of sine and cosine components in the samples given by $e^{j(2\pi/N)kn}$:

$$\text{mean of fourier domain} = \text{mean}(\text{dft}(x)). \tag{23}$$

(v) Variance of probability distribution:

$$\text{Probability Distribution Function} = P[a \le x \le b] = \int_a^b f(x)dx \tag{24}$$

Variance of distribution = variance(P).

(vi) Sum of autocorrelation:

$$\text{Autocorrelation function} = R(s,t) = \frac{E[(x_t - \mu) * (x_{t+r} - \mu)]}{\sigma_t \sigma_s}, \tag{25}$$

where s and t are different times in the time series, μ is the mean of X, σ is the standard deviation of X, and "E" is the expected value operator:

$$\text{Sum of AutoCorrelation} = \text{sum}(R(s,t)). \tag{26}$$

(vii) Mean of autocovariance:

$$C(s,t) = E[(x_t - \mu_t) * (x_s - \mu_s)], \tag{27}$$

where s and t are different times in the time series, μ is the mean of X, and "E" is the expected value operator:

$$\text{mean of autocorrelation} = \text{mean}(C(s,t)). \tag{28}$$

After the feature extraction has been performed, the total feature set for the wavelet transform step will

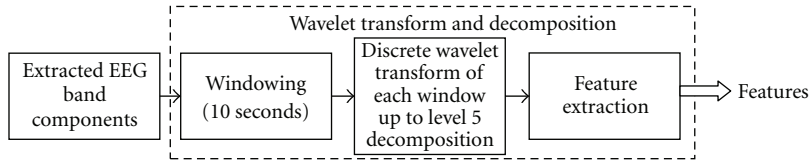


Figure 7: EEG decomposition and analysis steps using wavelet transform.

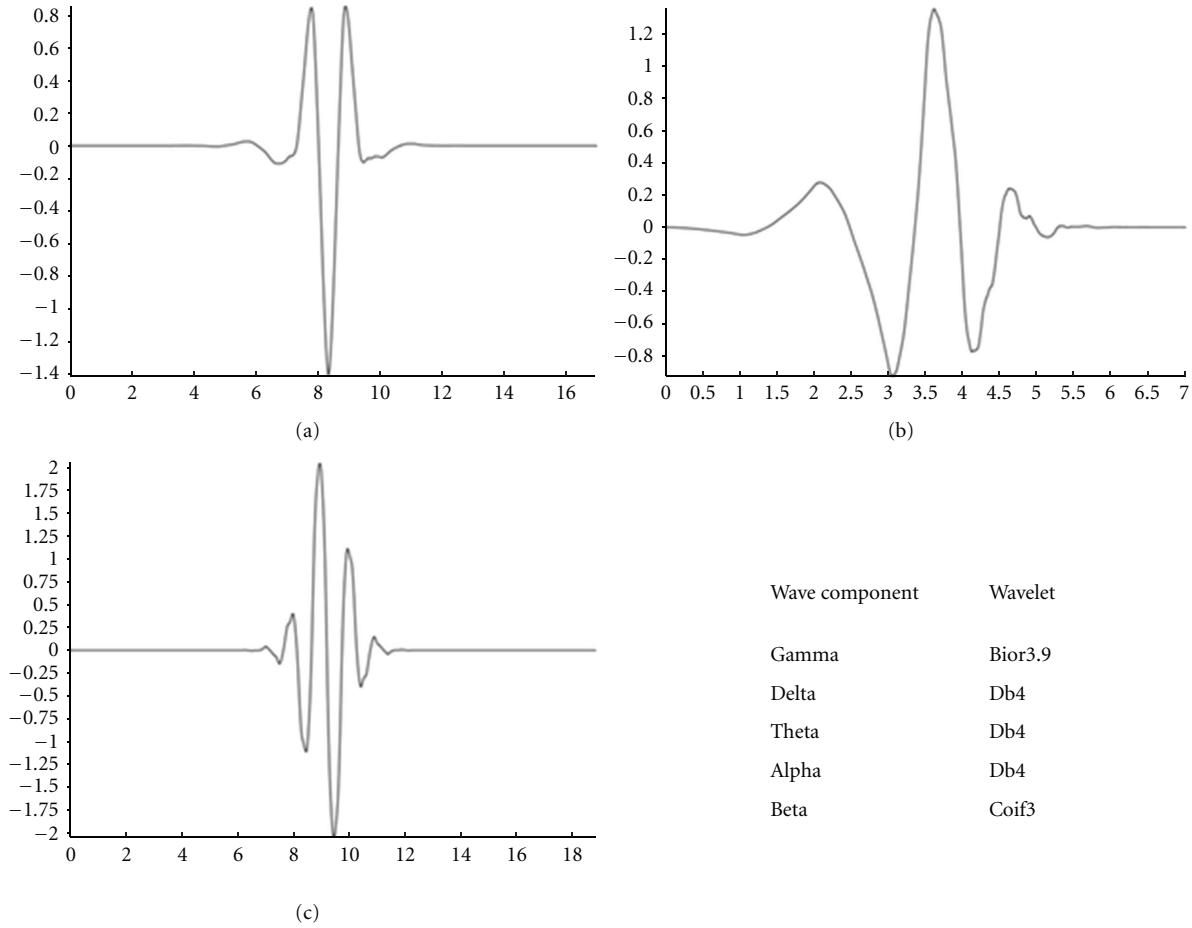


Figure 8: (a) "COIF3" wavelet, (b) "DB4" wavelet, and (c) "BOIR3.9" wavelet.

contain; 6 coefficients (5 detailed + 1 approximation) ∗ 7 (features per coefficient) = 42 (features columns per band component). In total there are 5 extracted band components, so, 42 (features per band component) ∗ 5 (different band components) = 210 (total features from EEG). These computed features are then sent to the machine learning stage for classification, training, and testing.

3.4. Machine Learning and Classification Model. In this application the result after signal processing on various acquired psychological signals is a large set of features. Since the data was collected in a systematic and controlled environment, the features extracted from respective portions of the signals can be classified under the two presumed categories: "attention" and "nonattention." Hence supervised learning method is used for this study to developed classification heuristics.

Three different machine learning algorithms have been implemented and tested for this experiment. These are as follows.

3.4.1. Classification via Regression. There are different models for predicting continuous variables or categorical variables from a set of continuous predictors and/or categorical factor effects such as General Linear Models (GLMs) and General Regression Models (GRMs). Regression-type problems are those where attempt is made to predict the values of a continuous variable from one or more continuous and/or categorical predictor variables [28, 38, 39]. This is a nonparametric approach meaning that no distribution assumptions are made about the data whereas in GLM it

is either known or assumed that the data follows a specific linear model such as binomial or Poisson. In regression-based classifiers, splits for the decision trees are made based on the variables that best differentiate between the categories of the target classification label variables. Here the decision splits are composed based on regression trees. In regression trees each node is split into two child nodes. As the regression tree grows certain stopping rules are applied to stop the tree growth.

In more general terms, the purpose of the analyses via tree-building algorithms is to determine a set of if-then logical (split) conditions that permit accurate prediction or classification of cases. Tree classification techniques, when applied correctly, produce accurate predictions or predicted classifications based on few logical if-then conditions. Their advantage of regression tree-based classifier over many of the alternative techniques is that they produce simplicity in the output classifier results. This simplicity not only is useful for purposes of rapid classification of new observations but can also often yield a much simpler "model" for explaining why observations are classified or predicted in a particular manner. The process of computing classification and regression trees can be characterized as involving four basic steps: specifying the criteria for predictive accuracy, selecting splits, determining when to stop splitting, and selecting the "right-sized" tree.

3.4.2. C4.5 Classification Method. C4.5 is also a decision-tree-based classification algorithm, developed by Quinlan [39, 40]. It has been developed based on the fundamentals of the ID3 machine-learning algorithm [41]. The C4.5 computes the input data to form a decision tree based on a divide-and-conquer strategy. In C4.5 each node in the tree is associated with a set of cases. Every case is assigned weights to deal with unknown attribute values. At first the entire training set is started off as a root where the weights assigned to all cases are 1.0. From here the tree computes the information gain presented by each attribute of the training set. For discrete attributes the information gain is relative to the splitting of case at every node with distinct values. The attribute with the highest information gain is selected as a test node. After this the divide-and-conquer approach consists of recursively splitting the attributes at each node to form children node based on the information gain of the attribute at each node. C4.5 has been used for several applications in healthcare informatics [42, 43].

3.4.3. Random Forest. Breiman developed random forest classification method which is basically an ensemble classifier that consists of multiple decision trees [44]. It is a very accurate classifier which displays great success with multiple datasets. It is especially useful with data mining extremely large datasets and databases. Unlike the other two mentioned tree-based classifiers random forest uses multiple trees or a forest to develop decisions and classifications. Although in this study it is being used to develop models based on supervised data, random forest can be used for unsupervised data learning as well [45, 46]. Random forest is also popular for applications in biosignal and biomedicine [46].

Table 1: *S*-transform feature classification results of ECG.

S-transform feature classification result ECG	Accuracy (average)	Specificity (average)	Sensitivity (average)
C4.5	74.22%	67.31%	81.13%
Classification via regression	71.63%	63.11%	80.15%
Random forest	**76.96**%	66.73%	87.20%

All of the above-mentioned machine-learning methods are known to have comparable performance to methods such as neural networks in physiological and medical applications [47]. Moreover, methodologies such as neural networks, when analyzed using statistical learning theory, are shown to be susceptible to the issue of overfitting [48–50], hence further encouraging the use of the methods described above, in particular when the number of data or subjects used for training and testing is limited.

In the machine learning step, the three mentioned classifiers are independently implemented on the extracted features of ECG and EEG and the results of each of these classifiers are compared. This is based on a setup developed earlier during initial stages of this experiment. For this experiment ECG signal from 21 subjects and EEG signal from 12 subjects have been collected.

4. Results and Conclusion

The classification model for each of the classifiers is developed using "by-subject" or "leave one subject out" based training and test sets. In this type of training and testing, out of the given number of subject say x, $x - 1$ subjects are subjects used for training and developing the classification model, while the xth subject's data is used for testing the developed model. This procedure is repeated in a round robin fashion until each of the subject's data in the total collected data has been tested with a classification model developed exclusively for it. In this section for each type of classification method used, the average accuracies and other statistics have been presented over all the subjects.

4.1. Classification Results of ECG Using S-Transform. The results obtained from the analysis and classification of the computed features from Stockwell transform (ST) from the ECG signal are presented.

Table 1 presents the overall average accuracies, specificities, and sensitivities of the three classification algorithms for ECG testing and training models across all subjects.

It can be seen that overall accuracy of random-forest-based classification model was more successful than both C4.5 and classification via regression models with a classification accuracy of nearly 77%.

4.2. Classification Results of EEG Using Discrete Wavelet Transform. The features computed from the analysis of the EEG signal using discrete wavelet transform is used

Table 2: DWT features classification results of EEG.

DWT feature classification result EEG	Accuracy (average)	Specificity (average)	Sensitivity (average)
C4.5	80.93%	81.11%	80.96%
Classification via regression	82.5%	76.74%	88.26%
Random forest	**85.70**%	79.74%	91.66%

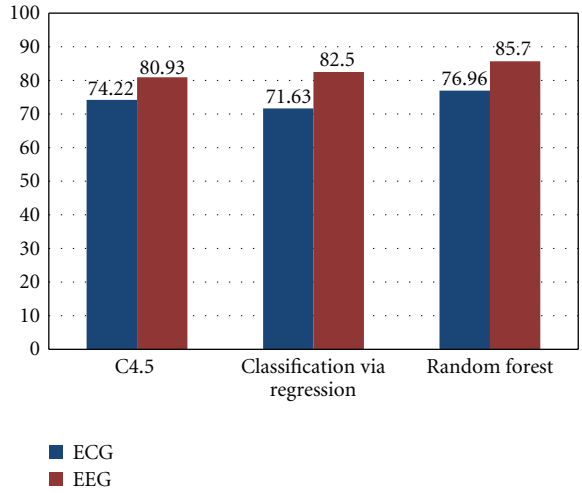


Figure 9: ECG versus EEG classification comparison.

to develop different classification models based on the three described classification methods. The results of these classification are presented in Table 2 .

Table 2 presents the overall average accuracies, specificities, and sensitivities of the three classification algorithms for EEG testing and training models across all subjects. It can be seen that overall accuracy of random-forest-based classification model was more successful than both C4.5 and classification via regression models with a classification accuracy of nearly 86% for the EEG feature set.

4.3. ECG versus EEG Classification Comparison. The results from the ECG feature classification of all three classifier are compared against the classification results of the EEG.

From Figure 9 it can be seen that although EEG inherently has more information to indicate the presence of attention or the lack of it, ECG signal analysis and classification are not very far behind. Random Forest seems to work best for both modalities given an average accuracy of 77% for ECG and 86% for EEG.

5. Conclusion

The analysis of the EEG signals is primarily to set a benchmark against which the analysis of the physiological features from the armband can be compared. This system as it has been proposed primarily focuses on the electrocardiogram (ECG) signal and various methods of decomposition are performed on it. The following are the conclusive statements that can be deduced from the systems performance so far.

(i) It can be seen that to a reasonable level of accuracy the system is able to identify cognitive attention in comparison with that detected by the EEG collected in the same experiment. The focus of this proposal was entirely on ECG alone, and with just this signal it was demonstrated that its classification accuracy was comparable to that of EEG.

(ii) Amongst the various machine learning methods investigated, "classification via regression" seems to perform the best on the combined feature set. However, it was also demonstrated that "random-forest-" based classification works on the subset of features for each different decomposition and analysis method.

(iii) This study also establishes that ECG alone can be used in analyzing cognitive attention and that the fluctuation of attention does have a translated impact on the Cardiac rhythm of an individual.

Here are some of the future work planned to improve the system's classification and prediction performance.

(i) A larger data set is needed to further validate this experiment. A larger data set is expected to provide a more robust classifier model.

(ii) More novel features are going to be developed and tried for the feature extraction step after decomposition. Having a more diverse base of features usually provides insight into some connate characteristics of the signal which might not be openly evident.

(iii) Feature pruning and other classification methods need to be tried for increasing the accuracy.

Acknowledgments

The authors would like to acknowledge BodyMedia Advanced Development (BodyMedia) for providing the armbands for this research. This study was designed and conducted in collaboration with Dr. Paul Gerber, Professor of Dyslexia Studies, Special Education and Disability Policy, VCU. The authors wish to also acknowledge the subjects who volunteered for this study.

References

[1] J. Locke, *An Essay Concerning Human Understanding*, TE Zell, 1847.

[2] E. Horvitz and J. Apacible, "Learning and reasoning about interruption," in *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI '03)*, pp. 20–27, November 2003.

[3] D. S. McCrickard and C. M. Chewar, "Attuning notification design to user goals and attention costs," *Communications of the ACM*, vol. 46, no. 3, pp. 67–72, 2003.

[4] C. Roda, A. Angehrn, T. Nabeth, and L. Razmerita, "Using conversational agents to support the adoption of knowledge sharing practices," *Interacting with Computers*, vol. 15, no. 1, pp. 57–89, 2003.

[5] B. P. Bailey, P. D. Adamczyk, T. Y. Chang, and N. A. Chilson, "A framework for specifying and monitoring user tasks,"

Computers in Human Behavior, vol. 22, no. 4, pp. 709–732, 2006.

[6] S. K. L. Lal and A. Craig, "A critical review of the psychophysiology of driver fatigue," *Biological Psychology*, vol. 55, no. 3, pp. 173–194, 2001.

[7] F. Mamashli, M. Ahmadlu, M. R. H. Golpayegani, and S. Gharibzadeh, "Detection of attention using chaotic global features," *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 22, no. 2, article E20, 2010.

[8] S. K. L. Lal and A. Craig, "Driver fatigue: electroencephalography and psychological assessment," *Psychophysiology*, vol. 39, no. 3, pp. 313–321, 2002.

[9] A. R. Clarke, R. J. Barry, R. McCarthy, and M. Selikowitz, "EEG analysis in attention-deficit/hyperactivity disorder: a comparative study of two subtypes," *Psychiatry Research*, vol. 81, no. 1, pp. 19–29, 1998.

[10] J. F. Lubar, "Discourse on the development of EEG diagnostics and biofeedback for attention-deficit/hyperactivity disorders," *Biofeedback and Self-Regulation*, vol. 16, no. 3, pp. 201–225, 1991.

[11] T. P. Tinius and K. A. Tinius, "Changes after EEG biofeedback and cognitive retraining in adults with mild traumatic brain injury and attention deficit hyperactivity disorder," *Journal of Neurotherapy*, vol. 4, pp. 27–44, 2000.

[12] A. Al-Ahmad, M. Homer, and P. Wang, *Accuracy and Utility of Multi-Sensor Armband ECG Signal Compared to Holder Monitoring*, Arrhythmia Technologies Retreat, Chicago, Ill, USA, 2004.

[13] J. M. Stern and J. Engel, *Atlas of EEG Patterns*, Lippincott Williams & Wilkins, 2004.

[14] S. J. Orfanidis, *Introduction to Signal Processing*, Prentice-Hall, 1995.

[15] G. M. Friesen, T. C. Jannett, M. Afify Jadallah, S. L. Yates, S. R. Qu int, and H. Troy Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 1, pp. 85–98, 1990.

[16] J. B. Ochoa, *Eeg Signal Classification for Brain Computer Interface Applications*, Ecole Polytechnique Federale De Lausanne, 2002.

[17] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 55–59, 1980.

[18] D. Gabor, "Theory of communication. Part 1: the analysis of information," *Electrical Engineers-Part III*, vol. 93, pp. 429–441, 1946.

[19] P. Bloomfield, *Fourier Analysis of Time Series: An Introduction*, Wiley-Interscience, 2004.

[20] L. Cohen, "Time-frequency distributions—a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, 1989.

[21] Y. Zhao, L. E. Atlas, and R. J. Marks, "Use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1084–1091, 1990.

[22] H. I. Choi and W. J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 6, pp. 862–871, 1989.

[23] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Magazine*, vol. 9, no. 2, pp. 21–67, 1992.

[24] A. Belle, R. Hobson, and K. Najarian, "A physiological signal processing system for optimal engagement and attention detection," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '11)*, pp. 555–561, 2011.

[25] A. Belle, S. Y. Ji, S. Ansari, R. Hakimzadeh, K. Ward, and K. Najarian, "Frustration detection with electrocardiograph signal using wavelet transform," in *Proceedings of the 1st International Conference on Biosciences (BioSciencesWorld '10)*, pp. 91–94, March 2010.

[26] R. G. Stockwell, L. Mansinha, and R. P. Lowe, "Localization of the complex spectrum: the S transform," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 998–1001, 1996.

[27] A. Babloyantz, J. M. Salazar, and C. Nicolis, "Evidence of chaotic dynamics of brain activity during the sleep cycle," *Physics Letters A*, vol. 111, no. 3, pp. 152–156, 1985.

[28] K. Natarajan, U. R. Acharya, F. Alias, T. Tiboleng, and S. K. Puthusserypady, "Nonlinear analysis of EEG signals at different mental states," *BioMedical Engineering Online*, vol. 3, article 7, 2004.

[29] E. Başar, C. Başar-Eroglu, S. Karakaş, and M. Schürmann, "Brain oscillations in perception and memory," *International Journal of Psychophysiology*, vol. 35, no. 2-3, pp. 95–124, 2000.

[30] V. J. Samar, A. Bopardikar, R. Rao, and K. Swartz, "Wavelet analysis of neuroelectric waveforms: a conceptual tutorial," *Brain and Language*, vol. 66, no. 1, pp. 7–60, 1999.

[31] E. A. Bartnik, K. J. Blinowska, and P. J. Durka, "Single evoked potential reconstruction by means of wavelet transform," *Biological Cybernetics*, vol. 67, no. 2, pp. 175–181, 1992.

[32] O. Bertrand, J. Bohorquez, and J. Pernier, "Time-frequency digital filtering based on an invertible wavelet transform: an application to evoked potentials," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 1, pp. 77–88, 1994.

[33] L. J. Trejo and M. J. Shensa, "Feature extraction of event-related potentials using wavelets: an application to human performance monitoring," *Brain and Language*, vol. 66, no. 1, pp. 89–107, 1999.

[34] T. Kalayci, O. Ozdamar, and N. Erdol, "Use of wavelet transform as a preprocessor for the neural network detection of EEG spikes," in *Proceedings of the IEEE Creative Technology Transfer-A Global Affair (Southeastcon '94)*, pp. 1–3, April 1994.

[35] D. M. Tucker, "Spatial sampling of head electrical fields: the geodesic sensor net," *Electroencephalography and Clinical Neurophysiology*, vol. 87, no. 3, pp. 154–163, 1993.

[36] S. J. Schiff, J. G. Milton, J. Heller, and S. L. Weinstein, "Wavelet transforms and surrogate data for electroencephalographic spike and seizure localization," *Optical Engineering*, vol. 33, no. 7, pp. 2162–2169, 1994.

[37] J. Raz, L. Dickerson, and B. Turetsky, "A wavelet packet model of evoked potentials," *Brain and Language*, vol. 66, no. 1, pp. 61–88, 1999.

[38] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten, "Using model trees for classification," *Machine Learning*, vol. 32, no. 1, pp. 63–76, 1998.

[39] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, Morgan kaufmann, 1993.

[40] J. R. Quinlan, "Bagging, boosting, and C4.5," in *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI '96)*, pp. 725–730, August 1996.

[41] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[42] C. A. Frantzidis, C. Bratsas, M. A. Klados et al., "On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach

for healthcare applications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 309–318, 2010.

[43] C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders," *Biomedical Signal Processing and Control*, vol. 6, no. 3, pp. 261–268, 2011.

[44] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, pp. 18–22, 2002.

[46] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, pp. 65–75, 2012.

[47] S. Y. Ji, R. Smith, T. Huynh, and K. Najarian, "A comparative analysis of multi-level computer-assisted decision making systems for traumatic injuries," *BMC Medical Informatics and Decision Making*, vol. 9, no. 1, article 2, 2009.

[48] K. Najarian, "Learning-based complexity evaluation of radial basis function networks," *Neural Processing Letters*, vol. 16, no. 2, pp. 137–150, 2002.

[49] K. Najarian, G. A. Dumont, M. S. Davies, and N. Heckman, "PAC learning in non−linear FIR models," *International Journal of Adaptive Control and Signal Processing*, vol. 15, pp. 37–52, 2001.

[50] K. Najarian, "A fixed-distribution PAC learning theory for neural FIR models," *Journal of Intelligent Information Systems*, vol. 25, no. 3, pp. 275–291, 2005.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 750151, 6 pages
doi:10.1155/2012/750151

Research Article

Machine Learning Approach to Extract Diagnostic and Prognostic Thresholds: Application in Prognosis of Cardiovascular Mortality

Luis J. Mena,[1] Eber E. Orozco,[1] Vanessa G. Felix,[1] Rodolfo Ostos,[1] Jesus Melgarejo,[2] and Gladys E. Maestre[2,3]

[1] *Department of Computer Engineering, Polytechnic University of Sinaloa, 82199 Mazatlan, SIN, Mexico*
[2] *Institute for Biological Research and Cardiovascular Institute, Faculty of Medicine, University of Zulia, Maracaibo 4002, Venezuela*
[3] *Departments of Psychiatry and Neurology, and the Gertrude H. Sergievsky Center, Columbia University, New York, NY 10032, USA*

Correspondence should be addressed to Luis J. Mena, lmena@upsin.edu.mx

Received 30 March 2012; Revised 25 June 2012; Accepted 3 July 2012

Academic Editor: Guilherme de Alencar Barreto

Machine learning has become a powerful tool for analysing medical domains, assessing the importance of clinical parameters, and extracting medical knowledge for outcomes research. In this paper, we present a machine learning method for extracting diagnostic and prognostic thresholds, based on a symbolic classification algorithm called REMED. We evaluated the performance of our method by determining new prognostic thresholds for well-known and potential cardiovascular risk factors that are used to support medical decisions in the prognosis of fatal cardiovascular diseases. Our approach predicted 36% of cardiovascular deaths with 80% specificity and 75% general accuracy. The new method provides an innovative approach that might be useful to support decisions about medical diagnoses and prognoses.

1. Introduction

Machine learning (ML) disciplines provide computational methods and learning mechanisms that can help generate new knowledge from large databases. Applications of ML are useful for constructing approaches to solving problems of classification, prediction, recognition patterns, and knowledge extraction, where the data take the form of a set of examples, and the output takes the form of prediction of new examples [1, 2]. In this sense, ML can provide techniques and tools that help solve diagnostic and prognostic problems in medical domains, where the input is a dataset with characteristics of the subjects, and the output is a diagnosis or prognosis of a specific disease [3]. Although diagnosis and prognosis are relatively straightforward ML problems, clinical decision-making using ML applications is not yet widely used by the medical community [4], because such a complex task requires not only accuracy, but also the confidence of physician specialists about the functional use of ML approaches in the medical field.

To successfully implement an ML application in problems related to clinical decisions, it is necessary to consider some specific requirements [4, 5]. For example, the prediction of disease progression is generally associated with the evolution of certain risk factors; in the case of some chronic diseases (e.g., cancer, cardiovascular diseases, and diabetes), the risk factors include nonchangeable characteristics, such as age or gender. The use of such nonchangeable qualities to predict the onset of a disease might not be as useful for avoiding evolution of the disease, because currently there is no medical treatment for modifying these biological characteristics. Thus, ML applications usually focus on changeable qualities, which make the prognostic task more difficult and complex.

Another important aspect to consider is the need to obtain interpretable approximations, in order to provide medical staff with useful information about the given problem. This is typically achieved using symbolic learning methods (e.g., decision trees and rules systems), which allow decisions to be explained in an easily comprehensible manner. However, the use of a symbolic learning algorithm

to obtain a more comprehensible model frequently sacrifices accuracy in the prediction.

Another problem that often hinders high overall performance in the analysis of medical datasets is that generally these exhibit an unbalanced class distribution [6], which include a majority or negative class of healthy people (normal data) and a minority or positive class of sick people (the important class) with higher cost of erroneous classification. The latter usually has a higher rate of misclassification, because the performance of standard ML algorithms tends to be overwhelmed by the majority class, ignoring the minority class examples and obtaining results with acceptable accuracy and specificity (healthy subjects diagnosed correctly), but low sensitivity (sick subjects diagnosed correctly).

In addition to developing ML approaches that result in good overall performance and provide medical staff with interpretable prognostic information, providing the ability to support decisions and to reduce the number of medical tests for a reliable prognosis are also desirable. A measure of reliability of the diagnosis or prognosis is also important, because this would give medical staff sufficient confidence to put the new approach into practice. On the other hand, it is also desirable to have an approach that can provide reliable predictions based on a small amount of information about the patient, because collection of that information is often expensive, possibly subject to privacy issues, time consuming, and possibly harmful to the patient [4].

The present study focused on the implementation of a ML method to support medical decisions in the prognosis of fatal cardiovascular diseases, which are ranked among the top ten in the global disease burden [7]. The goal was to solve previously identified problems, through interdisciplinary work that included the collection and preprocessing of data from an ambulatory blood pressure (ABP) monitoring study [8], the implementation of a current ML algorithm with specific application to medical diagnosis and prognosis [9], and the identification of new prognostic thresholds for risk factors of cardiovascular mortality.

2. Methods

2.1. Ambulatory Blood Pressure Monitoring. Currently available ABP monitors are fully automatic and portable devices (Figure 1) that can record BP for 24 hours or longer, while patients go about their normal daily activities [10]. This BP measurement technique provides a better estimate of risk in an individual patient than the traditional method, because it removes variability among individual observers, avoids the "white coat" effect (the transient but variable elevation of BP in a medical environment) [11] and the "masked hypertension" (normotensive by clinic measurement and hypertensive by ambulatory measurement) [12] and includes the inherent variability of BP [13]. Detailed descriptions of the ABP measurement methods are provided in previous reports of the Maracaibo Aging Study (MAS) [8, 14, 15].

2.2. Subjects. The MAS is an ongoing population-based, longitudinal study that includes 2500 subjects older than 55 years, residing in the Santa Lucia County, Maracaibo, Venezuela. All participants underwent extensive clinical and laboratory examinations and randomly selected individuals also underwent ABP monitoring. Informed consent was obtained from the subjects who agreed to participate, and from a close family member when doubts existed about the competence of the subject. The ethical review board of the Institute of Cardiovascular Diseases of the University of Zulia approved the protocol.

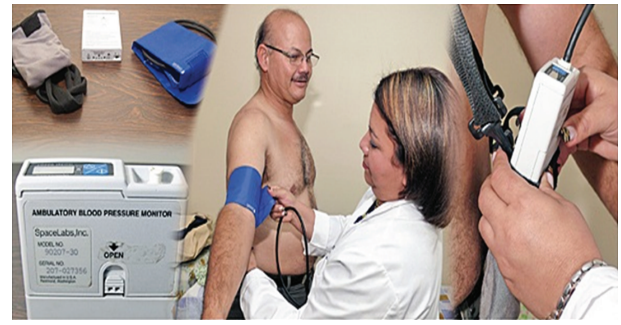


Figure 1: Ambulatory blood pressure monitoring procedure.

2.3. Cardiovascular Risk Factors. The leading global risk factor for mortality is high BP, which is responsible for 13% of deaths globally. Eight changeable risk factors (alcohol use, tobacco use, high BP, high body mass index, high cholesterol, high blood glucose, low fruit and vegetable intake, and physical inactivity) account for 61% of cardiovascular deaths. Combined, these same risk factors account for over three quarters of ischaemic heart disease, the leading cause of death worldwide [16].

However, investigators continue to look for new and emerging risk factors for cardiovascular disease. Recent ABP monitoring studies using a novel variability index [14] reported significant relationships between high BP variability (BPV) and cardiovascular outcomes [17–19]. BPV is a multifaceted phenomenon, influenced by the interaction between external emotional stimuli, such as stress and anxiety, and internal cardiovascular mechanisms that can vary from heartbeat to heartbeat. However, the complexity of BPV makes analysis difficult, and its independent contribution as a predictor of cardiovascular outcomes is not yet clear [20]. The present study aimed to identify new prognostic thresholds of risk factors for cardiovascular mortality, including high BP (the most significant cardiovascular predictor) and abnormal BPV (a potential independent predictor).

To estimate 24-hour BP level, we computed the weighed mean of valid BP readings (WBP) using the time interval between successive valid measurements as weighting factors [18]. In the case of BPV over 24 hours, we calculated the Average Real Variability (ARV) index [14] using (1):

$$\mathrm{ARV} = \frac{1}{\sum w_k} \sum_{k=1}^{n-1} w_k \times |\mathrm{BP}_k - \mathrm{BP}_{k-1}|, \tag{1}$$

where n is the number of valid BP readings, k ranges from 1 to $n-1$, and w_k is the time interval between BP_k and BP_{k-1}.

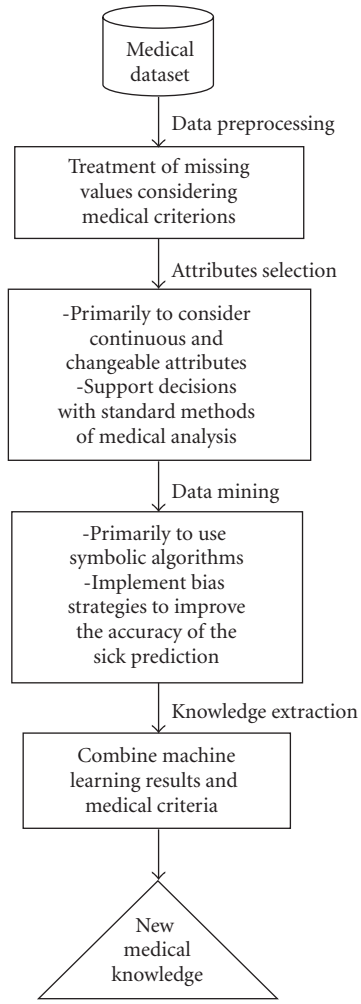


Figure 2: Machine learning method proposed.

2.4. Machine Learning Approach. We implemented an interdisciplinary ML method that encompassed all stages of knowledge extraction from databases (data preprocessing, attribute selection, data mining, and knowledge extraction), to examine the application of ML to support clinical decisions (Figure 2).

To improve the accuracy of predictions for affected subjects (positive class), we used the Rule Extraction for MEdical Diagnosis (REMED) algorithm [9], a symbolic one-class classification approach that implements internal bias strategies during the learning process [21]. REMED employs three main procedures in the knowledge extraction process: (1) selection of attributes, (2) selection of initial partitions, and (3) construction of classification rules.

First, REMED attempts to select the best combination of relevant attributes, using a simple logistic regression model. This is a standard method of analysis in medical research that uses the odds ratio metric [22] to determine if there is a significant association ($P < 0.01$) between a considered attribute and the positive class. REMED then begins to build initial partitions (exclusionary and exhaustive conditions) to maximize sensitivity and maintain acceptable accuracy without significantly decreasing specificity. Finally, REMED uses the respective partitions for each selected attribute to construct a system of rules that includes m conditions (one for each selected attribute) in the following way:

If Condition 1 <relation> p_1

and Condition 2 <relation> p_2

and Condition j <relation> p_j **and** $\cdots\cdots\cdots$

and Condition m <relation> p_m

then class = 1

Else class = 0,

where <relation> is either $\geq$ or $\leq$ depending on whether j is positively or negatively associated with the positive class through p_j (partition for attribute j).

To avoid overfitting during the training and testing phase, REMED implements the k-fold cross validation technique, which is based on randomly shuffling sample vectors among training and testing spaces [23]. REMED also maintains the approximate imbalance of the original dataset through the k iterations.

2.5. Data Preprocessing and Attributes Selection. Based on current medical guidelines [24], we only included participants that had ABP recordings of good technical quality. Therefore, subjects with <40 BP readings during the 24-hour ABP period were excluded. Systolic BP readings values >260 mmHg or <70 mmHg, and diastolic BP readings >150 mmHg or <40 mmHg were considered outliers or erroneous values and discarded. The treatment of missing values was addressed with predictive techniques, specifically multiple linear regression analysis considering medical criterions.

Only continuous and changeable attributes were considered in the knowledge extraction process. Continuous attributes have a higher degree of uncertainty than discrete attributes, because discrete attributes are usually binary in the clinical environment (e.g., smoker versus nonsmoker), and their associations with specific diseases are almost always well known. We also excluded age, which is a nonchangeable attribute. The attributes considered in the initial ML analysis were body mass index (BMI), serum cholesterol level, 24-hour heart rate, and systolic and diastolic 24-hour WBP and ARV.

3. Results

3.1. Dataset. The minable dataset was composed of 551 observations with 7 attributes, with only 43 missing values (1.1%) in the serum cholesterol attribute. The missing data were estimated from the regression slope on sex and age, according to the criteria of physician specialists. The sample included 374 women (67.8%) and 170 patients (30.9%) undergoing treatment with antihypertensive drugs (Table 1). The average number of BP readings was 65.1 (5th to 95th percentile = 51.5−77.5), indicating good quality ABP recordings. Mean age was 67.1 ± 8 years. At enrolment, 61 participants (11.1%) had a history of cardiovascular disease;

Table 1: Baseline characteristics.

	Frequency in percent or median
Demographic variables	
Men, % (n)	32.1 (177)
Age, years	67.1 ± 8
Race, % (n)	
Mixed	73.1 (404)
Caucasian	22.2 (122)
African-Venezuelan	4 (22)
Natives	0.5 (3)
Use of antihypertensive drugs, % (n)	30.9 (170)
Use of anti-diabetic drugs, % (n)	11.1 (61)
History of cardiovascular disease, % (n)	11.5 (63)
Diagnosis of diabetes mellitus, % (n)	18.1 (100)
Lifestyle, physical and lipid factors	
Smoking current status, % (n)	15.6 (86)
Drinking current status, % (n)	31.6 (174)
Body max index, kg/m^2	27.1 ± 5.6
Total serum cholesterol, mmol/L	5.5 ± 1.3
24-hour ambulatory measurements	
Systolic blood pressure, mm Hg	133.8 ± 16.6
Diastolic blood pressure, mm Hg	76.1 ± 10
Heart rate, bpm	73.7 ± 9.8

100 (18.1%) had a history of diabetes mellitus, of whom 59 (59%) were undergoing diabetes treatment; 86 (15.6%) were current smokers; 174 (31.6%) reported intake of alcohol. The average total cholesterol level was 5.5 ± 1.3 mmol L^{-1}, and BMI averaged 27.1 ± 5.6 kg m^{-2}. Mean 24-hour systolic WBP was 133.8 ± 16.6 mmHg, and diastolic WBP was 76.1 ± 10 mmHg. Average heart rate was 73.7 ± 9.8 bpm.

The median follow-up period was 7.1 ± 3.7 years (5th to 95th percentile = 1.7−12.3 years). Only the participants that died from cardiovascular diseases ($n = 61$) were classified as positive examples. Cardiovascular mortality included 10 strokes and 51 cardiac deaths for a high event rate of 15.5 per 1000 person-years. The imbalance ratio between the positive (affected) and negative (unaffected) class was approximately of 1 : 9.

3.2. Machine Learning Process

3.2.1. Selection of Attributes. Using the simple logistic regression model, REMED found only two attributes significantly associated with the positive class: systolic WBP ($P = 0.008$) and ARV ($P = 0.0001$). However, other well-known cardiovascular risk factors, such as serum cholesterol level, BMI, and diastolic WBP [16, 25], were considered in further analyses.

3.2.2. Rule System. To provide medical staff with more information and comprehensible models, we used REMED to build several simple rule systems, which included individual and combined predictions of the more significant attributes (systolic WBP and ARV), as well as the combined predictions with the additional risk factors.

Table 2: Confusion matrix of REMED predictions.

		Predictive class	
		Positive	Negative
Actual class	Positive	22	39
	Negative	98	392

Table 3: Performance of classifiers.

Classifiers	Sensitivity	Specificity	Accuracy
If systolic ARV ≥ 9.6 **then** 1 **Else** 0	55.7%	60.4%	59.9%
If systolic WBP ≥ 134.6 **then** 1 **Else** 0	52.5%	58.8%	58.08%
If systolic ARV ≥ 9.6 **and** systolic WBP ≥ 137 **then** 1 **Else** 0	36.1%	80.0%	75.1%
If systolic ARV ≥ 9.6 **and** systolic WBP ≥ 138.6 **and** cholesterol ≥ 5.5 **then** 1 **Else** 0	8.2%	93.3%	83.8%
If systolic ARV ≥ 10.4 **and** systolic WBP ≥ 139.8 **and** BMI ≥ 27.3 **then** 1 **Else** 0	9.8%	93.3%	84.0%
If systolic ARV ≥ 9.6 **and** systolic WBP ≥ 137 **and** diastolic WBP ≥ 78.4 **then** 1 **Else** 0	22.9%	87.5%	80.4%
Naïve Bayes	11.48%	95.92%	86.57%

3.3. Performance. The confusion matrix from the predictions of the system rule, combining only high systolic ARV and WBP and using 10-fold cross-validation, indicated that REMED performed at 0.36 sensitivity, correctly diagnosing more than 35% of the cardiovascular deaths (Table 2). REMED focuses on improving sensitivity over specificity, because in the case of medical diagnosis/prognosis, the cost of misclassification of false negatives (FN, i.e., sick subjects diagnosed incorrectly) is higher than that of false positives (FP, healthy subjects diagnosed incorrectly), because more specific medical tests could discover the FP error, but an FN could cause a life-threatening condition and possibly lead to death [26]. Additionally, to compare the performance of our approach in terms of reliable prediction, we selected from the WEKA framework [2] the ML approach that better performed with our dataset: the Naïve Bayes classifier, which is one of the most effective and efficient classification algorithms and has been successfully applied to many medical problems [27, 28]. The performance of all classifiers is showed in Table 3.

4. Discussion

Use of the REMED algorithm selecting only the more significant attributes provided some of the desired features for solving medical diagnosis/prognosis problems: (1) good overall performance for imbalanced datasets, with 36.1% of sensitivity, 80% specificity, and 75.1% general accuracy; (2) comprehensible prognostic information, based on a rule system with a high degree of abstraction (only one rule to predict positive class examples, independent of the number of instances and initial attributes); (3) the ability to provide the medical staff with sufficient confidence to use the rule system in practice, because it was based on attributes with high confidence levels (>99%), estimated with a standard method of medical analysis; (4) the ability to reduce the number of medical tests necessary to obtain a reliable diagnosis/prognosis, because a simple logistic regression model was used to select attributes strongly associated with the specific disease.

The ML approach generated a new prognostic threshold for cardiovascular mortality: systolic WBP $\geq$ 137 mmHg, which is lower than the currently proposed by hypertension guidelines ($\geq$140 mmHg) and in agreement with recent ABP studies [29, 30], but with the advantage that our analysis was fully automated and had a smaller sample. Moreover, our ML approach generated a new prognostic threshold for abnormal systolic ARV ($\geq$9.6 mmHg). Together, these new thresholds could provide improved predictions of cardiovascular mortality.

Both systolic WBP and ARV were independent predictors of cardiovascular mortality, performed >50% of sensitivity, but sacrificed significantly in specificity and general accuracy ($\leq$60%). The addition of other well-known cardiovascular risk factors decreased considerably the accuracy in the prediction of affected subjects (<23%). Therefore, the use of logistic regression for the selection of significant attributes (>99%) could be an effective strategy in this stage of ML analysis in medical datasets.

Undoubtedly, one of the most important goals of the application of ML in the medical field is to generate new knowledge, providing the medical community with tools to develop novel points of view about any given problem. In our case, for example, although previous medical studies determined possible ranges of a low and high BPV measured whit ARV through statistical methods (median and quartiles analysis) [17, 18], our work is pioneer proposing a prognostic threshold for abnormal systolic ARV ($\geq$9.6 mmHg). This threshold has a good performance as an independent a composed predictor of fatal cardiovascular events. The use of this threshold should facilitate new fields of investigation regarding BPV and its prognostic relevance.

We do not claim that our ML analysis using REMED is the ultimate solution for medical diagnosis/prognosis problems from unbalanced datasets, because it is necessary to implement modifications that improve REMED's predictive capacity in terms of sensitivity ($\geq$50%) without significantly deteriorating its specificity. However, we obtained better results than the Naïve Bayes classifier (11.48%), which is considered as a benchmark algorithm that in any medical domain has to be tried before any other advanced method [27]. Therefore, we believe that our approach could improve performance in these medical tasks, and increase the confidence of the medical community in the use of ML approaches to support clinical decisions.

Acknowledgments

The authors are grateful to the referees for their detailed review on the paper and thoughtful comments. This paper was supported by the Secretaria de Educación Pública, México DF, México (PROMEP/103-5/11/4145). The Maracaibo Aging Study was funded by the Venezuelan Grant FONACIT G-97000726, FundaConCiencia, and by Award no. R01AG036469 from the National Institute on Aging.

References

[1] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, Cambridge, Mass, USA, 2nd edition, 2010.

[2] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Burlington, Mass, USA, 3rd edition, 2011.

[3] S. Karpagavalli, K. S. Jamuna, and M. S. Vijaya, "Machine learning approach for preoperative anaesthetic risk prediction," *International Journal of Recent Trends in Engineering*, vol. 1, no. 2, pp. 19–22, 2009.

[4] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[5] Z. Bosnić and I. Kononenko, "Estimation of individual prediction reliability using the local sensitivity analysis," *Applied Intelligence*, vol. 29, no. 3, pp. 187–203, 2008.

[6] N. V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.

[7] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[8] G. E. Maestre, G. Pino-Ramírez, A. E. Molero et al., "The maracaibo aging study: population and methodological issues," *Neuroepidemiology*, vol. 21, no. 4, pp. 194–201, 2002.

[9] L. Mena and J. A. Gonzalez, "Symbolic one-class learning from imbalanced datasets: application in medical diagnosis," *International Journal on Artificial Intelligence Tools*, vol. 18, no. 2, pp. 273–309, 2009.

[10] T. G. Pickering, D. Shimbo, and D. Haas, "Ambulatory blood-pressure monitoring," *New England Journal of Medicine*, vol. 354, no. 22, pp. 2316–2374, 2006.

[11] T. G. Pickering, G. D. James, C. Boddie, G. A. Harshfield, S. Blank, and J. H. Laragh, "How common is white coat hypertension?" *Journal of the American Medical Association*, vol. 259, no. 2, pp. 225–228, 1988.

[12] A. Frattola, G. Parati, C. Cuspidi, F. Albini, and G. Mancia, "Prognostic value of 24-hour blood pressure variability," *Journal of Hypertension*, vol. 11, no. 10, pp. 1133–1137, 1993.

[13] T. G. Pickering, K. Davidson, W. Gerin, and J. E. Schwartz, "Masked hypertension," *Hypertension*, vol. 40, no. 6, pp. 795–796, 2002.

[14] L. Mena, S. Pintos, N. V. Queipo, J. A. Aizpúrua, G. Maestre, and T. Sulbarán, "A reliable index for the prognostic

significance of blood pressure variability," *Journal of Hypertension*, vol. 23, no. 3, pp. 505–511, 2005.

[15] L. Mena, J. D. Melgarejo, C. Chavez et al., "Relevance of blood pressure variability among the elderly: findings from the Maracaibo aging study," *Journal of Hypertension*, vol. 29, Article ID e312, 2011.

[16] World Health organization, *Global Health Risks-Mortality and Burden of Disease Attributable to Selected Major Risk*, World Health Organization, Geneva, Switzerland, 2009.

[17] S. D. Pierdomenico, M. Di Nicola, A. L. Esposito et al., "Prognostic value of different indices of blood pressure variability in hypertensive patients," *American Journal of Hypertension*, vol. 22, no. 8, pp. 842–847, 2009.

[18] T. W. Hansen, L. Thijs, Y. Li et al., "Prognostic value of reading-to-reading blood pressure variability over 24 hours in 8938 subjects from 11 populations," *Hypertension*, vol. 55, no. 4, pp. 1049–1057, 2010.

[19] P. Veerabhadrappa, K. M. Diaz, D. L. Feairheller et al., "Enhanced blood pressure variability in a high cardiovascular risk group of African Americans: FIT4Life Study," *Journal of the American Society of Hypertension*, vol. 4, no. 4, pp. 187–195, 2010.

[20] T. W. Hansen, Y. Li, and J. A. Staessen, "Blood pressure variability remains an elusive predictor of cardiovascular outcome," *American Journal of Hypertension*, vol. 22, no. 1, pp. 3–4, 2009.

[21] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: a review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[22] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explorations*, vol. 6, pp. 80–89, 2004.

[23] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Conference on Artificial Intelligence*, pp. 1137–1143, 1995.

[24] G. Mancia, G. De Backer, A. Dominiczak et al., "2007 guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC)," *Journal of Hypertension*, vol. 25, no. 6, pp. 1105–1187, 2007.

[25] G. M. Weiss, "Mining with rarity a unifying frame-work," *ACM SIGKDD Explorations*, no. 1, pp. 7–19, 2004.

[26] Y. Cui, R. S. Blumenthal, J. A. Flaws et al., "Non-high-density lipoprotein cholesterol level as a predictor of cardiovascular disease mortality," *Archives of Internal Medicine*, vol. 161, no. 11, pp. 1413–1419, 2001.

[27] K. M. Al-Aidaroos, A. A. Bakar, and Z. Othman, "Medical data classification with Naive Bayes approach," *Information Technology Journal*, vol. 11, no. 9, pp. 1166–1174, 2012.

[28] M. Kukar and C. Grošelj, "Reliable diagnostics for coronary artery disease," in *Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems*, pp. 7–12, June 2002.

[29] M. Kikuya, T. W. Hansen, L. Thijs et al., "Diagnostic thresholds for ambulatory blood pressure monitoring based on 10-year cardiovascular risk," *Circulation*, vol. 115, no. 16, pp. 2145–2152, 2007.

[30] T. W. Hansen, M. Kikuya, L. Thijs et al., "Diagnostic thresholds for ambulatory blood pressure moving lower: a review based on a meta-analysis-clinical implications," *Journal of Clinical Hypertension*, vol. 10, no. 5, pp. 377–381, 2008.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 943431, 11 pages
doi:10.1155/2012/943431

Research Article

Investigating Properties of the Cardiovascular System Using Innovative Analysis Algorithms Based on Ensemble Empirical Mode Decomposition

Jia-Rong Yeh,[1] Tzu-Yu Lin,[2,3] Yun Chen,[4,5] Wei-Zen Sun,[6] Maysam F. Abbod,[7] and Jiann-Shing Shieh[2]

[1] *Research Center for Adaptive Data Analysis & Center for Dynamical Biomarkers and Translational Medicine, National Central University, Jhongli 3200, Taiwan*
[2] *Department of Mechanical Engineering, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li, Taoyuan 320, Taiwan*
[3] *Department of Anesthesiology, Far Eastern Memorial Hospital, Taipei 220, Taiwan*
[4] *Department of Surgery, Far Eastern Memorial Hospital, Taipei 22060, Taiwan*
[5] *Department of Chemical Engineering & Materials Science, Yuan Ze University, Taoyuan 320, Taiwan*
[6] *Department of Anesthesiology, College of Medicine, National Taiwan University, Taipei 100, Taiwan*
[7] *School of Engineering and Design, Brunel University, London UB83PH, UK*

Correspondence should be addressed to Jiann-Shing Shieh, jsshieh@saturn.yzu.edu.tw

Received 3 March 2012; Revised 31 May 2012; Accepted 15 June 2012

Academic Editor: Amaury Lendasse

Cardiovascular system is known to be nonlinear and nonstationary. Traditional linear assessments algorithms of arterial stiffness and systemic resistance of cardiac system accompany the problem of nonstationary or inconvenience in practical applications. In this pilot study, two new assessment methods were developed: the first is ensemble empirical mode decomposition based reflection index (EEMD-RI) while the second is based on the phase shift between ECG and BP on cardiac oscillation. Both methods utilise the EEMD algorithm which is suitable for nonlinear and nonstationary systems. These methods were used to investigate the properties of arterial stiffness and systemic resistance for a pig's cardiovascular system via ECG and blood pressure (BP). This experiment simulated a sequence of continuous changes of blood pressure arising from steady condition to high blood pressure by clamping the artery and an inverse by relaxing the artery. As a hypothesis, the arterial stiffness and systemic resistance should vary with the blood pressure due to clamping and relaxing the artery. The results show statistically significant correlations between BP, EEMD-based RI, and the phase shift between ECG and BP on cardiac oscillation. The two assessments results demonstrate the merits of the EEMD for signal analysis.

1. Introduction

Arterial stiffness is a powerful physiological marker of cardiovascular morbidity and mortality. However, the cardiovascular system is a complicated system which has effects of multiple underlying mechanisms. Correlations among systolic arterial pressure (SAP), arterial stiffness, and systemic resistance are significant topics for cardiovascular system. Moreover, since a cardiovascular system is nonlinear and nonstationary, the characteristics of the system should be assessed by suitable algorithms based on innovative signal processing techniques for such a nonlinear system. Therefore, two methods were developed to assess the arterial stiffness and systemic resistance of a cardiovascular system based on ensemble empirical mode decomposition (EEMD) technique. EEMD is an innovative signal processing algorithm developed to decompose intrinsic mode functions from a nonlinear and nonstationary time series [1].

In this study, for the purpose of obtaining a sequence of changes in the blood pressure, such as increasing then steady high blood pressure for SAP, arterial stiffness, and systemic resistance in a cardiovascular system, an experimental surgical operation has been conducted on a healthy young pig. In such an experiment, the clamping of intestine artery

stimulated an acute rising of SAP and the relaxing of arterial clamping reversed the reaction to arterial clamping. Changes in SAP stimulated corresponding changes on arterial stiffness and systemic resistance of the cardiovascular system [2, 3]. This procedure has provided the material for the investigation so that a better understanding of the connections between SAP, arterial stiffness, and systemic resistance of the cardiovascular system can be realized.

Previous studies have shown that augmentation index (AIx) and reflection index (RI) provide as good indicators for aortic stiffness [4–6], which can be calculated as the ratios between the amplitudes of forward wave, reflected wave and systolic peak. AxI is determined by both the magnitude and timing of the reflected wave [6]. Furthermore, a more accurate measurement can be obtained after separating the BP signal into its forward and reflected components, which requires an extra measurement of aortic flow. Previously, Westerhof et al. presented a new method to quantify the magnitude of reflection independent of the time of the reflected wave. In his method, a triangular shape of the flow wave was assumed to determine the timing features of arterial pressure. Hence, the reflection index (RI) derived by Westerhof's method can be calculated via BP only [6].

On the other hand, pulse wave velocity (PWV) is another popular method for the quantification of aortic stiffness [7]. The most widely used method for determining PWV is to measure the time delay between characteristic points on two pressure waveforms that are a known distance apart. Recently, an innovative analysis algorithm of multimodal pressure flow (MMPF) was proposed to trace the interaction between BP and blood flow using the phase shift of spontaneous oscillations [8–10]. In this study, it is assumed that the ECG can present the activating potential of heart beating and it is measured as the driving signal for the cardiovascular system [11]. In addition, BP performs as the output signal of the cardiac cycle, which reflects complicated responses of the overall cardiovascular system. Thus, a new application of multimodal analysis was proposed to investigate the interactive phase shift between ECG and BP during a cardiac cycle. The assumption made in this study is that the phase shift between intrinsic components of cardiac oscillations extracted from recordings of ECG and BP reflects the systemic resistance of a cardiovascular system. Therefore, signal processing techniques for decomposing the intrinsic components from ECG and BP signals are critical for these new applications.

Methodologically, there are many different signal processing methods that perform high-efficiency signal decomposition, such as independent component analysis (ICA) [12] and wavelet decomposition [13]. ICA contributes to the applications of blind signal separations based on statistical characteristics of the signals, which reflect linear combinations of different signal sources. Wavelet decomposition offers simultaneous interpretation of the signal in both time and frequency that allows local, transient, intermittent components to be calculated. However, such traditional signal processing method is based on linear assumption. The components derived by wavelet decomposition are often obscured due to the inherent averaging. In 1998, Huang et al. proposed the innovative algorithm of EMD signal decomposition, in which the components are decomposed adaptively to the nature of signals but not the base of transformation [14]. Theoretically, each intrinsic mode function (IMF) decomposed by EMD reflects the response actuated by the corresponding activity of a particular underlying physiological mechanism. In practices, the unpredictable intermittent turbulences damage the consistencies of IMFs. This phenomenon is noted as mode mixing. Recently, an ensemble empirical mode decomposition (EEMD) has been introduced which is considered as an enhanced algorithm of EMD, which solves the problem of mode mixing in the original EMD [1]. In this pilot study, it is assumed that the reflected waves of BP can be derived as a particular intrinsic component (i.e., IMF) by EEMD. Hence, a new EEMD-based calculation of RI can be achieved. Moreover, EEMD also works to decompose the cardiac oscillations from ECG and BP in the new application of multimodal analysis. Phase shift between the cardiac oscillations of ECG and BP is considered to be a phase delay between the driving signal (i.e., ECG) and the output signal (i.e., BP) of the cardiovascular system. It is considered as a new assessment of systemic impedance of the cardiovascular system which is the second EEMD-based assessment presented in this study.

Finally, Pearson's correlation coefficient was applied to check the correlations between SAP, EEMD-based RI, and the phase shift (between ECG and BP on cardiac oscillation). According to the results of the correlation analysis, EEMD-based RI acts as an indicator of arterial stiffness, showing significant positive correlation with SAP and significant negative correlation with the phase shift between ECG and BP on cardiac oscillation. The phase shift between ECG and BP on cardiac oscillation also acts as another indicator for systemic resistance of a cardiovascular system, which has a negative correlation with SAP. These two indicators show two different profiles of the cardiovascular system and have significant negative correlations with each other. Moreover, correlations between SAP (a direct measurement of BP), RI (a secondary parameters depends on the waveform of BP), and phase shift between ECG and BP (a phase delay between two different signals) show different profiles of the cardiovascular system and significant connections among them.

2. Material

In this investigation, the study material (i.e., ECG and BP recordings) was recorded during an animal experiment, which was approved by the Animal Research Ethics Review Committee of the Far Eastern Memorial Hospital in Taiwan. In this experiment, a male Lanyu-50 pig with body weight of around 10–15 kg was the subject. After intramuscular injection of Zoletil (Zoletil 50 Vet; Virbac S.A., Carros, France) 3–5 mg/kg, an intravenous line was established in the vein behind the ear. An oximeter was applied on the tail. Other monitored biosignals included body temperature and ECG. Body temperature was maintained by a heating blanket and warm air. Additional Zoletil was prepared to achieve immobility before intubation. After intubation and confirming the

position of the endotracheal tube (size 5.0–5.5 mm internal diameter), 4 mg pancuronium was injected intravenously. Subsequently, 5 mg/kg Zoletil and 4 mg pancuronium were given hourly. The pig was anesthetized following the same procedures above, with additional central venous catheter (20G-22G-22G, BD) at the right internal jugular vein and an arterial catheter (20G) at the left femoral artery under cut-down procedure. Lactate Ringer's solution, Hespander, and whole blood (donated from other pigs) were administered to maintain adequate volume status (central venous pressure >5 mmHg) and hemoglobin level (>8 g/dL). Norepinephrine or epinephrine (bolus or continuous infusion) can be administered as required to maintain systolic blood pressure >100 mmHg, especially after graft reperfusion. At the end of the surgery, if the hemodynamic profile was stable, weaning from ventilator support can be attempted.

To generate the ECG and BP recordings during clamping-relaxing-clamping-relaxing of the intestinal artery, the pig's intestinal artery was blocked by clamping briefly (e.g., one minute) and then relaxing the clamping to produce successive time series recording under different situations and transition state between them. This designed process was run twice consecutively to derive four-minute recordings of ECG and BP. The raw data of ECG and BP were measured by IntelliVue MP60 (Philips), an multichannel physiological monitoring system usually equipped in surgical operation rooms and intensive care units. The data was measured and stored at sampling rate of 1000 Hz and length of 240,000 sample points. No preprocessing algorithms had been applied to the raw data recorded by the MP 60 before further analysis.

3. Methods

3.1. Empirical Mode Decomposition (EMD). Empirical mode decomposition (EMD) performs an adaptive method to remove oscillation successively though repeatedly subtraction of the envelope means [14]. To a signal $x(t)$, the EMD algorithm consists of the following steps.

(1) Connect the sequential local maxima (respective minima) to derive the upper (respective lower) envelop using cubic spline.

(2) Derive the mean of envelope, $m(t)$, by averaging the upper and lower envelopes.

(3) Extract the temporary local oscillation $h(t) = x(t) - m(t)$.

(4) Repeat the steps of 1–3 (i.e., the sifting process) on the temporary local oscillation $h(t)$ until $m(t)$ is close to zero. Then, $h(t)$ is an IMF noted as $c(t)$.

(5) Compute the residue $r(t) = x(t) - c(t)$.

(6) Repeat the steps from (1) to (5) using $r(t)$ for $x(t)$ to generate the next IMF and residue.

Therefore, the original signal $x(t)$ can be reconstructed using the following formula:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t), \tag{1}$$

where $c_i(t)$ is the ith IMF (i.e., local oscillation) and $r_n(t)$ is the nth residue (i.e., local trend).

As the algorithm uses all the local extremes to construct the envelopes, the mode mixing would be inevitable when the signal contains intermittent processes. As discussed by Wu and Huang [1], the intermittence would cause the resulting true physical processes to be obscured by the fragmentation of a given signal.

3.2. Ensemble Empirical Mode Decomposition (EEMD). EMD is an iterative signal processing algorithm which decomposes the IMFs from the signal by the iterative sifting processes [14]. The essential algorithm of EMD is associated with a major difficulty of mode mixing. Figure 1 shows first 8 IMFs decomposed from a pig's BP recording by the original technique of EMD. Significant phenomenon of mode mixing can be observed in IMF 4–6, which perform inconsistencies in mode functions. Recently, Wu and Huang proposed EEMD as a noise-assisted data analysis method to overcome mode mixing problem [1]. In EEMD, white noise is added into the original signal to generate the mixtures for decompositions by EMD. Ensemble IMFs can be derived by averaging the IMFs decomposed from the mixtures. Since the intermittent fluctuations, which cause mode mixing problem, are coupled with the added white noise to be filtered, the problem of mode mixing has been effectively solved in EEMD. Figure 2 shows first 8 IMFs decomposed from the same recording by the noise-assisted technique of EEMD. The problem of mode mixing was solved and IMFs present consistencies in mode functions.

3.3. Monte Carlo Verification and Noise Removal. Monte Carlo simulation is a computational algorithm that relies on repeated random sampling to compute their results. In the confidential test of EMD, the repeated numerical simulations to characterize the properties of random noises applied to EMD can be based on the application of Monte Carlo simulation. Then, the confidential zone of IMFs decomposed from random noises can be defined by Monte Carlo simulations. An IMF with properties out of the confidential zone can be verified as a dominant component of the signal. This approach for verifying the dominant components of the signals is noted as Monte Carlo verification [2, 15]. Monte Carlo verification works to verify the IMFs contributed by noise or the dominant signal. The high-frequency noise of real-world signals can be reconstructed via the noisy components verified by the Monte Carlo verification, and the main waveform of signals can be reconstructed by the rest of intrinsic components and residual.

In the Monte Carlo verification, two parameters of energy density and averaged period for each IMF should be calculated using the following equations [16]:

$$E_n = \frac{1}{N}\sum_{j=1}^{N}[C_n(j)]^2,$$
$$\overline{T}_n = \int S_{\ln T,n}\, d\ln T \left(\int S_{\ln T,n}\frac{d\ln T}{T}\right)^{-1}, \tag{2}$$

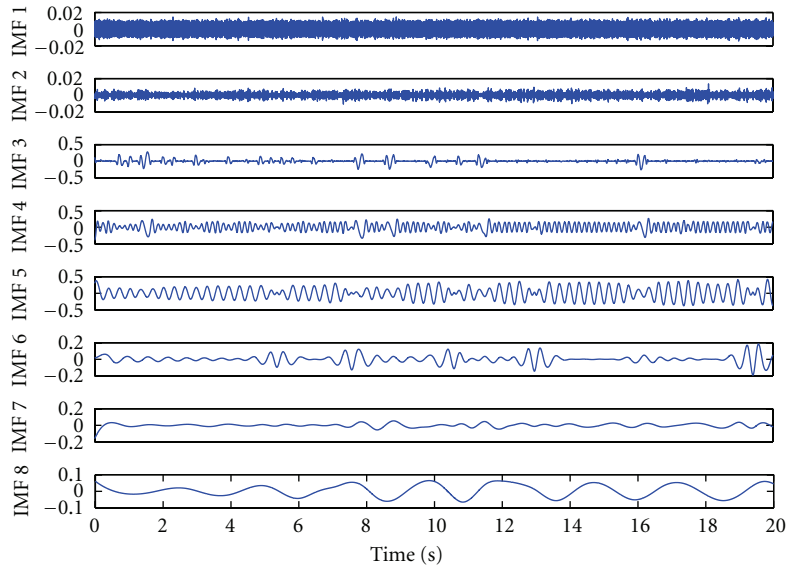


Figure 1: First 8 IMFs derived from a 12-second recording of a pig's BP by the original technique of EMD. Significant mode shifting can be observed in IMFs 4-5, which reflect inconsistencies in mode functions.

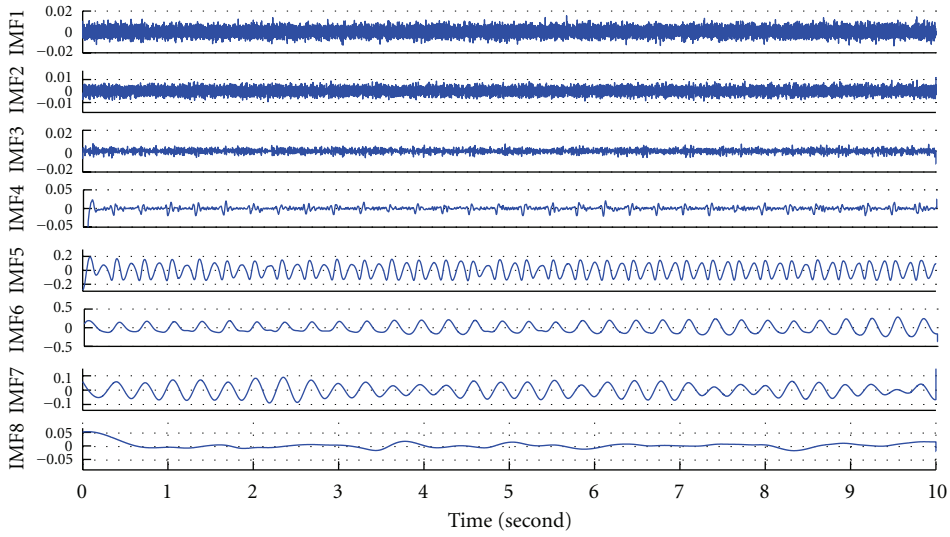


Figure 2: First 8 IMFs derived from a 12-second recording of a pig's BP by the noise-assisted technique of EEMD.

where $C_n(j)$ is the jth sample of the nth IMF, E_n is the energy density of the nth IMF, $S_{\ln T,n}$ is the Fourier spectrum of the nth IMF as a function of $\ln T$, T is the period, and $\overline{T}_n$ is the averaged period of the nth IMF.

On the logarithmic energy density/averaged period plot as shown in Figure 3, the first 3 IMFs can be fitted by a straight line with negative slope. According to the characteristics of white noise and fractal Gaussian noise derived by EMD [16–18], logarithmic energy density/averaged period plot for IMFs decomposed from a Gaussian noise is similar to a straight line with negative slope. Thus, the high-frequency noisy components are considered as the first n IMFs, which have a distribution of logarithmic energy densities and averaged periods similar to a straight line with negative slope value in the Monte Carlo verification. In Figure 3, the first 3 IMFs are verified as the noisy components of blood pressure signals. Moreover, IMF 8 has an averaged frequency of 0.46 Hz, which is induced by the activity of an unidentified physiological mechanism with lower frequency band than that of the basic cardiac cycle. Hence, the main waveform of blood pressure signal can be constructed via IMFs 4–7. In Figure 4, the reconstructed pulses of BP have main waveforms similar to the original pulses but excluding high-frequency noise and baseline shifting.

3.4. The EEMD-Based Calculation for RI. Augmentation index (AIx) is an assessment of wave reflection and an indicator of aortic stiffness [4, 5]. Unfortunately, the inflection points on systolic peaks are not distinguishable, and

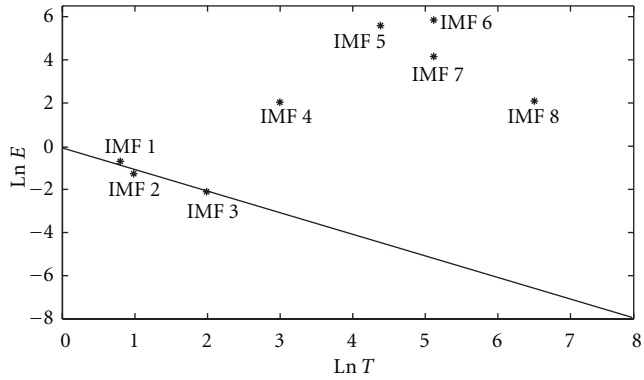


Figure 3: Logarithmic energy density-averaged period plot for the first 8 IMFs decomposed from pig's blood pressure signal.

so AIx cannot be obtained easily in this study. Recently, Westerhof demonstrated a new quantification method for wave reflection in the human aorta [6]. They assumed a triangular wave to simulate the extra measurement of aortic blood flow, with duration equal to ejection time, and to get approximations of the inflection points of BP using the time point of 30% of ejection time. In this study, the calculation of RI using the assumption of 30% ejection time is noted as the referred calculation of RI. Magnitudes of forward wave (P_f) and reflected backward wave (P_b) are separated using the magnitude of BP at the inflection point and the secondary rising magnitude of BP. Then, the reflection index (RI) is defined as

$$\mathrm{RI} = \frac{P_b}{P_b + P_f}. \tag{3}$$

In the EEMD-based calculation of RI, IMFs 1–3 decomposed from BP had been verified as high-frequency noisy components using the Monte Carlo verification [15]. Thus, complete pulses of the pig's BP can be reconstructed via IMFs 4–7. IMF 4 contributes a high-frequency part of BP with small amplitude. Sometimes, an intrinsic component of the original signal coupling with different added white noises can be decomposed into two different IMFs in EEMD. Then, two IMFs present a very high value of Pearson's correlation coefficient and can be merged together as single IMF. In this investigation, Pearson's correlation coefficient between IMFs 6 and 7 is 0.825 and the averaged frequencies are similar. Therefore, these 2 IMFs can be combined as an intrinsic component. Moreover, IMF 5 presents double in the number of peaks compared to the number of heart beats, as the same number of fluctuating cycles of the combination of IMFs 6 and 7. Half of the peaks of IMF 5 accompany the systolic peak of BP, and the other half accompany the dicrotic peaks of BP. Theoretically, the decomposition of EEMD is adaptive to the waveform of the signal; the separation between IMF 5 and its corresponding residue is sensitive to the discontinuous point on the systolic peak of BP as the inflection point. Therefore, the reconstructed wave via IMFs 4, 6, and 7 presents the basic fluctuation pattern of BP. And IMF 5 contributes the appended part of BP as the combination of reflection wave and dicrotic wave. In this investigation, the reconstructed wave via IMFs 4, 6, and 7 is considered as the forward wave as shown in Figure 5(a). It is also assumed that IMF 5 contributes the reflected wave and the dicrotic wave as two riding waves on the forward wave of BP. Figure 5(b) illustrates the forward wave only, and Figure 5(c) illustrates IMF 5, which contains the reflected wave and dicrotic wave. The forward wave follows the same rhythm as the heartbeat and presents the main cardiac oscillation of BP. IMF 5 contains the reflected wave and the dicrotic wave and shows an averaged frequency of oscillation twice that of the cardiac oscillation. Thus, the magnitude of the reflected wave (P_b) was defined as the amplitude of the reflected wave in IMF 5. In addition, the magnitude of forward wave (P_f) was measured using the amplitude of the reconstructed forward wave.

3.5. Phase Shift between ECG and BP on Cardiac Oscillation. Cerebral autoregulation controls dilatation and contributes to the constriction of the arterioles to maintain blood flow in response to changes of systemic blood pressure [19]. Therefore, a multimodal analysis algorithm was used to assess autoregulation mechanism by quantifying nonlinear phase interactions between spontaneous oscillation in blood pressure and flow velocity [8, 9]. Multimodal analysis acts to trace the phase delay between the spontaneous oscillations extracted from two different physiological signals (i.e., blood pressure and blood flow in the pioneering application).

In this investigation, ECG and BP are treated as the driving and output signals of the cardiovascular system. As a system defined in the field of digital signal processing, system impedance causes the decay ratio and phase delay between the output and the input. Phase shift between ECG and BP reflects the phase delay between the input and output of a human cardiovascular system. Peaks of IMF 6 decomposed from ECG present the R points of ECG signal, and peaks of IMF 6 decomposed from BP present the peaks of systolic wave. Therefore, phase shift between ECG and BP also presents a ratio between the pulse transit time (i.e., transit time between R peaks of ECG and peaks of systolic blood pressure) and heartbeat interval. It is assumed that the interactive phase shift (phase delay) between ECG and BP on the cardiac oscillation reflects the phase delay caused by the systemic impedance of the cardiovascular system. To determine the intrinsic components (i.e., IMFs) which reflect the cardiac oscillations of BP and ECG, the pig's ECG and BP recordings are decomposed into the first 9 IMFs. Table 1 shows the averaged frequencies of IMFs 5–9 for ECG and BP. Average frequency of IMF contributes as a clue to find the corresponding physiological mechanism for each component. In contrast to the human heartbeat rhythm, a young pig's heartbeat is much quicker than that of a human. Average frequency of a pig's heartbeat is around 3 Hz. Therefore, the cardiac oscillations were identified as the 6th IMFs for both ECG and BP. Furthermore, Hilbert transform was used to derive the time-amplitude-phase distribution from the cardiac oscillations [8–10]. Figure 6 illustrates the evaluated phase shift between ECG and BP

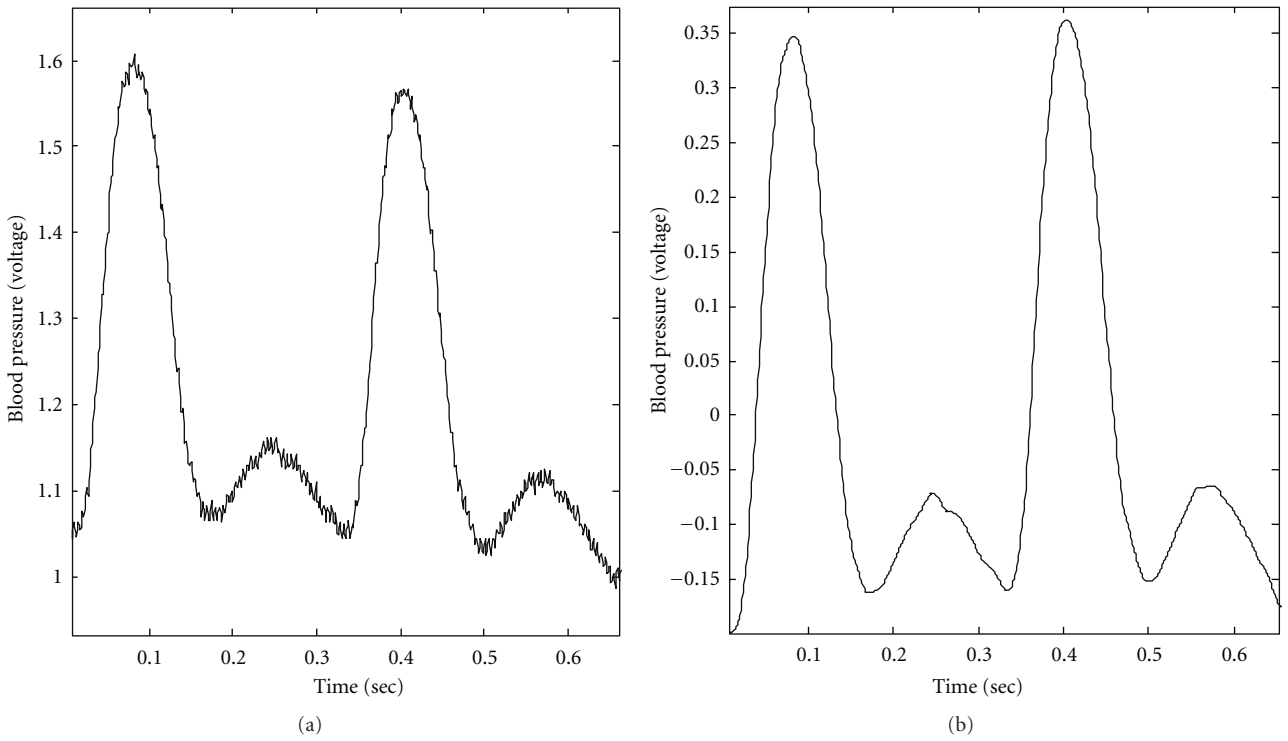


Figure 4: The original pulses and the reconstructed pulses of BP. (a) The original pulses of a pig's BP. (b) The reconstructed pulses of a pig's BP, which are reconstructed via the IMFs 4–7.

Table 1: Averaged frequencies of IMFs 5–9 decomposed from a pig's ECG and blood pressure by EEMD.

	ECG	Blood pressure
IMF 5	7.19 Hz	6.20 Hz
IMF 6	3.10 Hz	3.09 Hz
IMF 7	2.16 Hz	2.98 Hz
IMF 8	1.12 Hz	0.46 Hz
IMF 9	0.47 Hz	0.24 Hz

on cardiac oscillation. The cardiac oscillation of ECG was defined as the IMF with rhythm similar to heart beating, as IMF 6 derived from ECG. And the cardiac oscillation of BP was defined as the IMF with rhythm similar to the occurrence rhythm of systolic peak, as IMF 6 derived from BP. Then, the accumulative time-phase distributions can be via the time-phase distributions shown in Figure 6. Therefore, the phase shift is defined as the difference between the accumulative phases for every time point.

3.6. Pearson's Correlation Coefficient. Pearson's product-moment correlation coefficient is a measurement to identify the linear relationship between two variables [20]. In Pearson's correlation coefficient, the value of 1 indicates a perfect linear relationship between two variables and a negative correlation is indicated by the value of −1.

The traditional interpretation of a correlation coefficient uses five "rules of thumb" to interpret the correlation between two variables as follows [21]:

$0.20 > |r| > 0$ as *negligible correlation,*

$0.40 > |r| > 0.20$ as *low correlation,*

$0.60 > |r| > 0.40$ as *moderate correlation,*

$0.80 > |r| > 0.60$ as *significant correlation,*

$1.00 > |r| > 0.80$ as *high correlation.*

A positive value of correlation coefficient represents a positive correlation between two variables and a negative one presents a negative correlation. In this study, the value of correlation coefficient is interpreted using such interpretation rules.

4. Results

The analysis results of EEMD-based RI and progression of SAP as well as the magnitude of the forward wave of BP during the simulated surgical operation are shown in Figure 7. According to the results, it is shown that SAP rises and then remains steady on a high level during the period of artery clamping then falling during the period of arterial relaxing as shown in Figure 7(a). Moreover, it is also shown that there are cyclic changes in SAP and in the magnitude of forward wave. To verify the underlying physiological mechanism causing the cyclic changes, the number of cycles were counted and found that the average

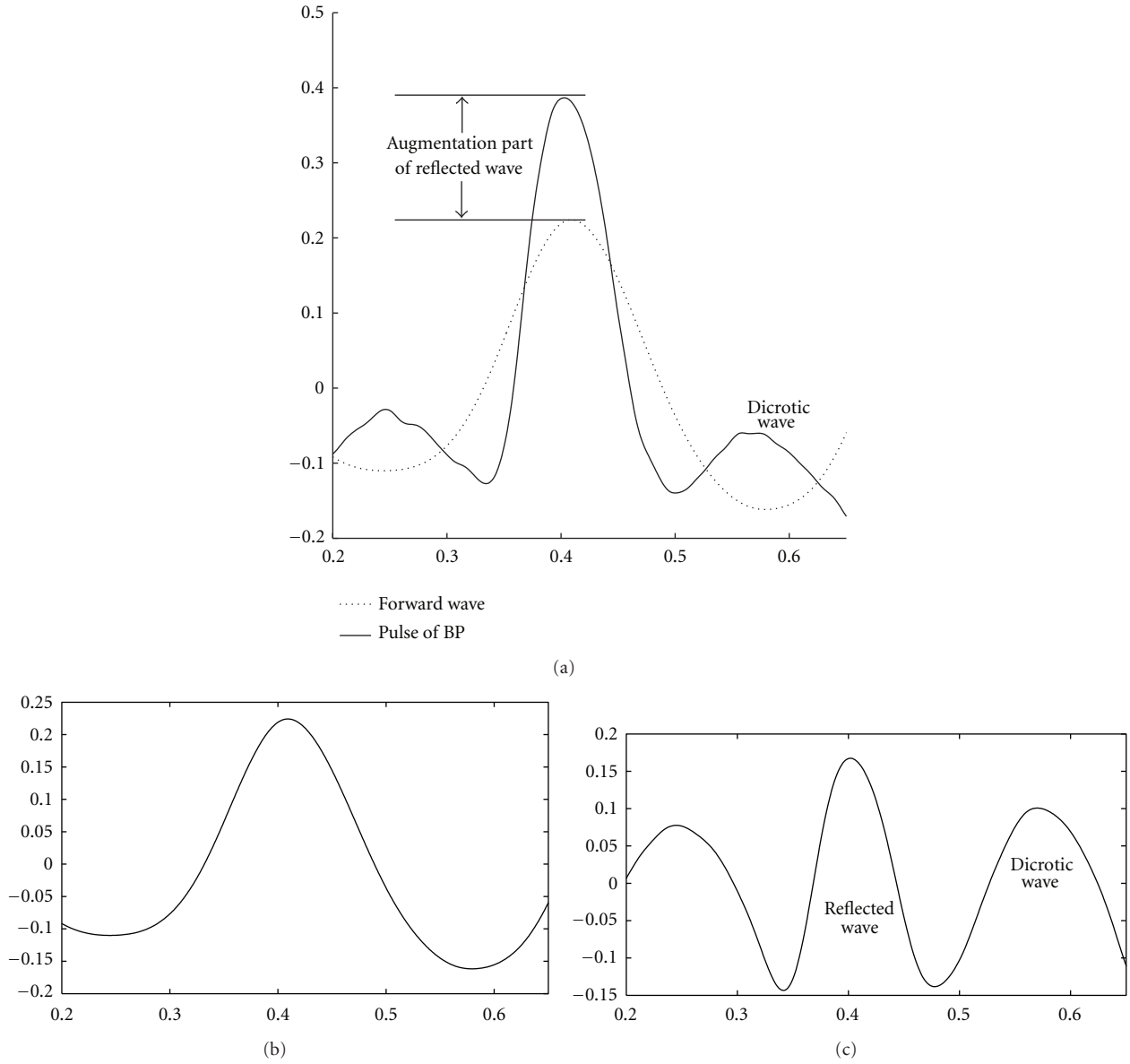


Figure 5: Illustration of a reconstructed pulse of a pig's BP. A whole pulse is assumed to be the ensemble of the forward wave and two riding waves (i.e., reflected wave and dicrotic wave). (a) Solid line shows the reconstructed forward wave of a pig's BP and the dash line shows complete waveform. (b) Assumed forward wave, which was reconstructed via IMFs 4, 6, and 7; (c) IMF 5 contains the reflected wave and dicrotic wave.

period of the cyclic change of SAP is 2.92 seconds (with average frequency of 0.34 Hz), which performs a rhythm similar to the respiration rate according to our observation. Moreover, the cyclic changes in SAP and in the magnitude of the forward wave also affect the values of RI, which also contains cyclic changes in values. To eliminate the effect caused by the interaction between respiration and the heartbeat, EEMD-based RI was filtered using a moving average filter (9 samples have been used for the moving average filter, since the average number of heartbeats during a cyclic change of SAP is around 9). Figure 7(b) shows the original and filtered EEMD-based RI. Furthermore, the same calculations of RI were repeated using the referred algorithm proposed by Westerhof et al., and compared to the EEMD-based results. In Figure 8(a) the two different RI are presented by time-sequence plots. Furthermore, the distribution of the two different RIs is shown in Figure 8(b). A positive correlation has been observed between the two RIs ($r = 0.759$).

In addition, multimodal analysis was conducted to investigate the systemic resistance in the cardiovascular system using the phase shift between ECG and BP on the cardiac oscillation. Due to the sensitivity of the Hilbert spectrum, the phase shift between two cardiac oscillations is not constant. Therefore, phase shift was also filtered by a moving average filter (the number of points used for moving average filter is 100, which is the equivalent cut-off frequency of 10 Hz for the sampling rate of 1000 Hz). The phase shift between

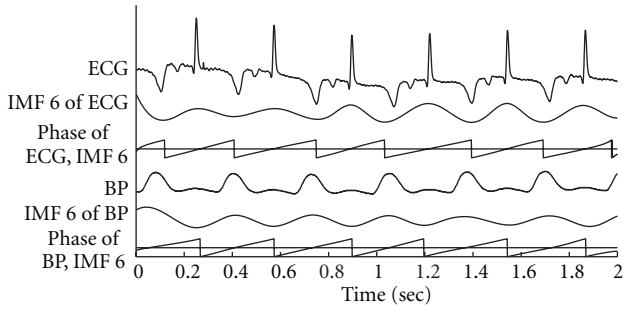


Figure 6: Illustration of phase shift between cardiac oscillations extracted from ECG and blood pressure. Plots from top to bottom are ECG signal, IMF 6 derived from ECG, time-phase distribution of ECG cardiac oscillation, BP, IMF 6 derived from BP, time-phase distribution of BP cardiac oscillation. Phase shift can be observed as the difference between the accumulative time-phase distributions of cardiac oscillations for ECG and BP.

ECG and BP on cardiac oscillation is shown in Figure 9(a). For the purpose of comparison with the analysis results of phase shift between ECG and BP, pulse transit time (PTT) between ECG and BP was analyzed. Figure 9(b) shows the analysis result using PTT. Phase shift is found to be more sensitive to the changes of manual control conditions (i.e., actions of clamping and relaxing). Analysis results of PTT are shown in Figure 9(b). PTT reflects the time delay between the R peak of ECG and the systolic peak of BP. The phase shift presented by the phase delay is different from the time delay presented by PTT. According to the plots shown in Figure 9, the phase shift is found to be more sensitive to the manual control actions than that presented by PTT.

For further comparisons among the original physiological signals (i.e., SAP) and the physiological indexes (i.e., phase shift between ECG and BP on cardiac oscillation and two different RIs derived by the EEMD-based and the referred algorithms), correlation coefficients were used to evaluate the correlations between the two different physiological signal/index. Table 2 shows the values of correlation coefficients for correlations of one-to-one comparisons. According to the results shown in Table 2, the two different assessments of RI have a positive correlation since they similarly perform as indicators for arterial stiffness. RI has a positive correlation with SAP, and phase shift between ECG and BP has a negative correlation with SAP. Furthermore, Figure 10 shows interesting correlations among SAP, RI, and phase shift between ECG and BP on cardiac oscillation.

5. Discussions and Conclusions

In previous studies, there were many different physiological parameters (such as pulse transit time, augmentation index, and reflection index) developed to investigate humans' cardiovascular systems using traditional algorithms based on linear assumption. However, since the human cardiovascular system is nonlinear and nonstationary, 4 necessary conditions (i.e., complete, orthogonal, local, and adaptive) should be considered in system analysis. Recently, EEMD proposed as an innovative analysis algorithm, which had been developed to satisfy the 4 conditions, is considered as a better solution to develop new assessments for cardiovascular system. Therefore, this approach has been considered to develop EEMD-based algorithms for cardiovascular system evaluation. This study did not provide satisfiable number of cases to prove any clinical findings. However, the EEMD-based analysis algorithm is computing extensive and time consuming. Hundred times of EMD are required in an EEMD decomposition to diminish the residue of added white noises. Therefore, EEMD-based analysis algorithms are hard to implement in an embedded system and applied to online monitoring system.

In the practical applications of EMD and EEMD, which algorithm fits the requirements of decomposition to nonlinear and nonstationary signals is still a critical issue. IMFs decomposed by the original EMD can conserve the characteristics of nonlinearity well in mode functions. However, mode mixing is a weakness of EMD in applications for extracting any mode functions with particular physical or physiological meanings. In contrast EEMD works to solve the problem of mode mixing. But characteristics of nonlinearity for mode functions can be destroyed in the ensemble form of IMFs. In this study, extracting intrinsic components with consistent characteristics in modulation is more important than conserving the characteristics of nonlinearity in the mode functions. Therefore, EEMD was applied in this investigation. What kind of characteristics should be conserved in the IMFs determines the use of EMD or EEMD.

In this study, an animal experiment was conducted for simulating changes in the cardiovascular system using a designed process to generate study material. In this one-animal experiment, relationships among different parameters are considered purely and directly. Influences caused by individual can be ignored in this investigation. Moreover, SAP is considered as a directly physiological measurement, EEMD-based RI is a secondarily derived parameter from BP, and phase shift between ECG and BP is a correlated phase delay between two physiological measurements. Therefore, connections among SAP, EEMD-based RI, and phase shift between ECG and BP are considered to reflect interactions of different physiological mechanisms in the human cardiovascular system.

According to the results, EEMD-based RI and phase shift between ECG and BP are significantly correlated with SAP. Furthermore, the correlation between these two parameters is also significant. It contributes an evidence for interactions among SAP, arterial stiffness, and systemic resistance of cardiovascular system. Moreover, this pilot study aims to present the functions of these two presented analysis techniques basedon EEMD but not the physiological findings. Hence, in order to make a contribution for understandings of underlying mechanisms of humans' cardiovascular systems, further study should be conducted with a sufficient number of animal experiments in the future works. Furthermore, mutual information analysis provides a powerful tool to verify the dependence between the two variables [22]. For the purpose of detailing the connections and dependencies

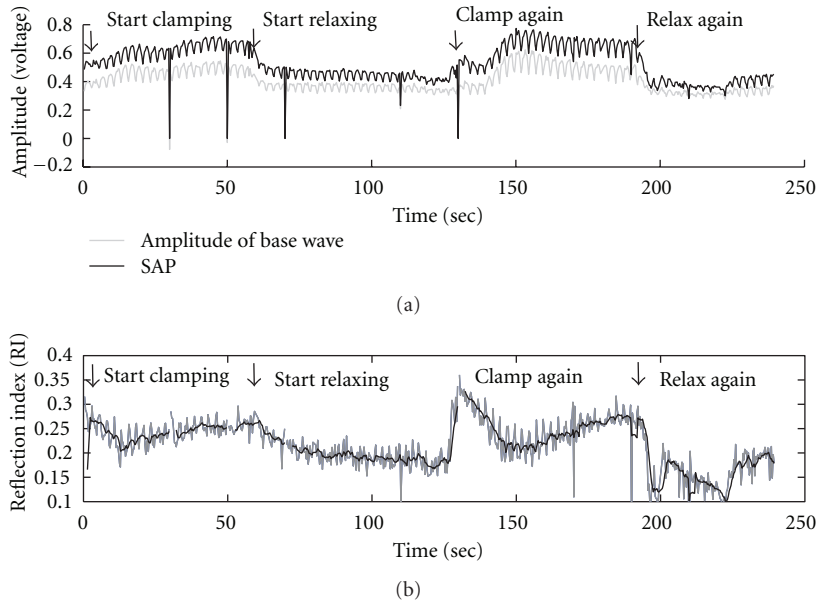


Figure 7: The analysis results of EEMD-based RI. (a) SAP and the magnitude of forward wave. (b) The original and filtered EEMD-based RI.

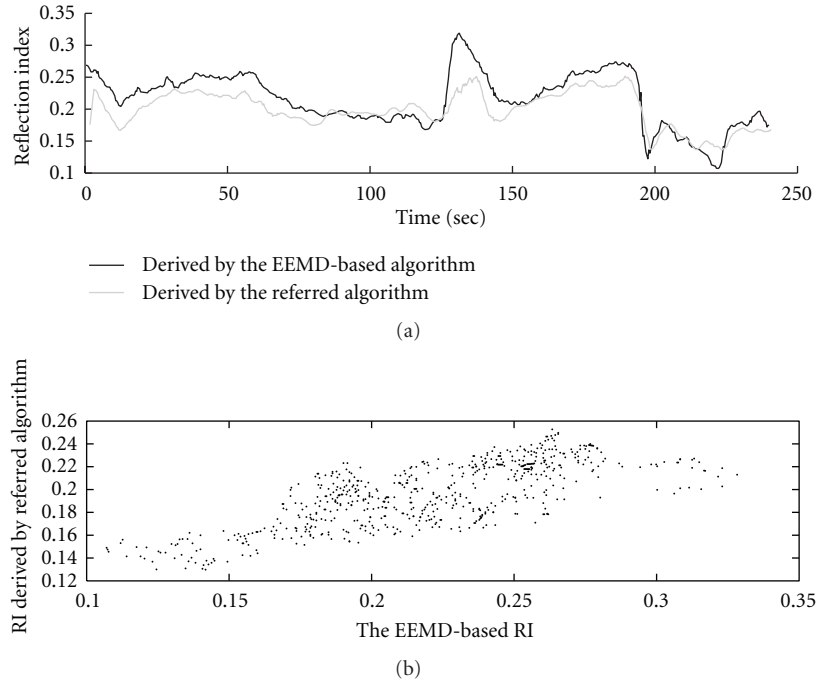


Figure 8: Comparisons between the analysis results of RI using different algorithms. (a) The time-sequence plot of RI during the simulated surgical operation. (b) The distribution of the referred RI against the EEMD-based RI.

among those parameters, the mutual information criteria should be considered and applied in future work.

Moreover, both the referred and the EEMD-based algorithms of RI evaluation present an interesting phenomenon during the period of artery clamping as shown in Figure 8. The value of RI eruptively increases at the instant of artery clamping and falls down at the first 20 seconds during artery clamping. Then, the RI arises again and becomes steady. During the periods of artery relaxing, the changes of RI values present much smoother patterns than those during the clamping periods.

In addition, IMFs decomposed by EEMD are ensembles of many EMD decompositions to mixtures of the signal and different added white noises. A complicated signal, which contains many intrinsic mode functions, coupled with different added white noise to generate different combinations

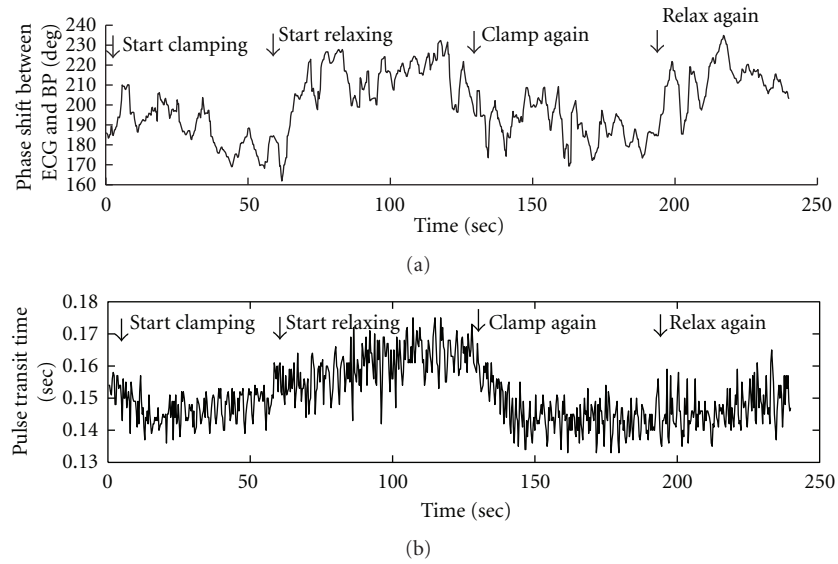


Figure 9: Phase shift between ECG and BP on cardiac oscillation.

Table 2: Correlations among the SAP, RI, and phase shift. According to the interpretation rules used in this study, $0.6 > r > 0.4$ represents a moderate correlation and $0.8 > r > 0.6$ represents a significant correlation.

Physiological signal/index		Correlation coefficient	Correlation
EEMD-based RI	Referred RI	0.759	Positive and significant
EEMD-based RI	Phase shift	−0.707	Negative and significant
Referred RI	Phase shift	−0.543	Negative and moderate
SAP	EEMD-based RI	0.708	Positive and significant
SAP	Referred RI	0.731	Positive and significant
SAP	Phase shift	−0.693	Negative and significant

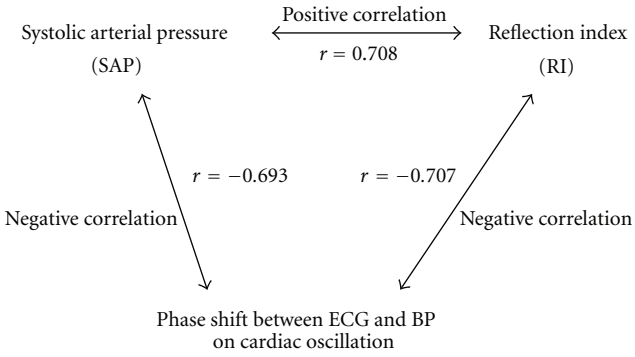


Figure 10: Illustration of the correlations among SAP, RI, and phase shift between ECG and BP on cardiac oscillation.

of IMFs in EEMD. Therefore, an intrinsic component of the original signal may appear with different orders in different EMD decompositions because of coupling with different added noises. Two IMFs sharing the same frequency can be resulted when an intrinsic component is decomposed into two IMFs evenly in EEMD. In Figure 2, IMFs 6 and 7 sharing the same frequency are a good example for this phenomenon. This is not a difficult problem to deal with. An orthogonal test to two successive IMFs is helpful to verify this phenomenon. The two IMFs sharing the same frequency can be merged together as a single IMF.

Finally, the referred algorithm of RI analysis is based on the triangular method to separate reflective and forward waves of BP. This method was derived and validated in central aorta but not femoral aorta. In this investigation, EEMD was considered to perform an adaptive algorithm in intrinsic component separation. Reflective and forward waves of BP are considered as two intrinsic components of BP with slight phase delay and difference in waveforms. EEMD works to separate these two components adaptively to the waveform of BP. The referred algorithm is considered to be as a criterion of inflection point determination without validation for BP signals derived from femoral aorta. The analysis results by the referred algorithm were used to be compared with the analysis results by EEMD-based method. In practical applications, the referred algorithm based on triangular method in femoral BP analysis should be validated.

Acknowledgments

The authors wish to thank the National Science Council (NSC) of Taiwan (Grant number NSC 99-2221-E-155-046-MY3) for supporting this research. This research was also supported by the Centre for Dynamical Biomarkers

and Translational Medicine, National Central University, Taiwan which is sponsored by National Science Council (Grant number: NSC 100–2911-I-008-001). Moreover, it was supported by the Chung-Shan Institute of Science & Technology in Taiwan (Grant numbers: CSIST-095-V101 and CSIST-095-V102).

References

[1] Z. Wu and N. E. Huang, "Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, pp. 1–41, 2009.

[2] J. R. Yeh, T. Y. Lin, J. S. Shieh et al., "Investigating complex patterns of blocked intestinal artery blood pressure signals by empirical mode decomposition and linguistic analysis," *Journal of Physics: Conference Series*, vol. 96, no. 1, Article ID 012153, 2008.

[3] K. S. Heffernan, S. R. Collier, E. E. Kelly, S. Y. Jae, and B. Fernhall, "Arterial stiffness and baroreflex sensitivity following bouts of aerobic and resistance exercise," *International Journal of Sports Medicine*, vol. 28, no. 3, pp. 197–203, 2007.

[4] G. F. Mitchell, Y. Lacourcière, J. M. O. Arnold, M. E. Dunlap, P. R. Conlin, and J. L. Izzo, "Changes in aortic stiffness and augmentation index after acute converting enzyme or vasopeptidase inhibition," *Hypertension*, vol. 46, no. 5, pp. 1111–1117, 2005.

[5] M. Vyas, J. L. Izzo, Y. Lacourcière et al., "Augmentation Index and Central Aortic Stiffness in Middle-Aged to Elderly Individuals," *American Journal of Hypertension*, vol. 20, no. 6, pp. 642–647, 2007.

[6] B. E. Westerhof, I. Guelen, N. Westerhof, J. M. Karemaker, and A. Avolio, "Quantification of wave reflection in the human aorta from pressure alone: a proof of principle," *Hypertension*, vol. 48, no. 4, pp. 595–601, 2006.

[7] B. M. Pannier, A. P. Avolio, A. Hoeks, G. Mancia, and K. Takazawa, "Methods and devices for measuring arterial compliance in humans," *American Journal of Hypertension*, vol. 15, no. 8, pp. 743–753, 2002.

[8] V. Novak, A. C. C. Yang, L. Lepicovsky, A. L. Goldberger, L. A. Lipsitz, and C. K. Peng, "Multimodal pressure-flow method to assess dynamics of cerebral autoregulation in stroke and hypertension," *BioMedical Engineering Online*, vol. 3, article no. 39, 2004.

[9] K. Hu, C. K. Peng, M. Czosnyka, P. Zhao, and V. Novak, "Nonlinear assessment of cerebral autoregulation from spontaneous blood pressure and cerebral blood flow fluctuations," *Cardiovascular Engineering*, vol. 8, no. 1, pp. 60–71, 2008.

[10] K. Hu, C. K. Peng, N. E. Huang et al., "Altered phase interactions between spontaneous blood pressure and flow fluctuations in type 2 diabetes mellitus: nonlinear assessment of cerebral autoregulation," *Physica A*, vol. 387, no. 10, pp. 2279–2292, 2008.

[11] P. W. Macfarlane and T. D. W. Lawrie, Eds., *Comprehensive Electrocardiology:Theory and Practice in Health and Disease (Vols. 1, 2, 3)*, Pergamon Press, New York, NY, USA, 1989.

[12] P. Comon, "Independent component analysis, A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[13] P. S. Addison, "Wavelet transforms and the ECG: a review," *Physiological Measurement*, vol. 26, no. 5, pp. R155–R199, 2005.

[14] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A*, vol. 454, no. 1971, pp. 903–995, 1998.

[15] K. T. Coughlin and K. K. Tung, "11-Year solar cycle in the stratosphere extracted by the empirical mode decomposition method," *Advances in Space Research*, vol. 34, no. 2, pp. 323–329, 2004.

[16] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proceedings of the Royal Society A*, vol. 460, no. 2046, pp. 1597–1611, 2004.

[17] P. Flandrin, P. Concalves, and G. Rilling, "EMD Equivalent Filter Banks, from Interpretation to Applications," in *Hilbert-Huang Transform: Introduction and Applications*, N. E. Huang and S. S. P. Shen, Eds., pp. 57–74, World Scientific, Singapore, 2003.

[18] P. Flandrin, G. Rilling, and P. Gonçalvés, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[19] L. A. Lipsitz, S. Mukai, J. Hamner, M. Gagnon, and V. Babikian, "Dynamic regulation of middle cerebral artery blood flow velocity in aging and hypertension," *Stroke*, vol. 31, no. 8, pp. 1897–1903, 2000.

[20] S. A. Glantz, *Primer of Biostatistics*, The McGraw-Hill, Singapore, 6th edition, 2005.

[21] A. Franzblau, *A Primer of Statistics For Non-Statisticians*, Harcourt, Brace & World, 1958.

[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 898430, 12 pages
doi:10.1155/2012/898430

Research Article

Hemorrhage Detection and Segmentation in Traumatic Pelvic Injuries

Pavani Davuluri,[1] Jie Wu,[2] Yang Tang,[3] Charles H. Cockrell,[3] Kevin R. Ward,[4,5] Kayvan Najarian,[2,5] and Rosalyn H. Hargraves[1]

[1] *Department of Electrical and Computer Engineering, Virginia Commonwealth University, Richmond, VA 23284, USA*
[2] *Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA*
[3] *Department of Radiology, Virginia Commonwealth University, Richmond, VA 23298, USA*
[4] *Department of Emergency Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA*
[5] *Virginia Commonwealth University Reanimation and Engineering Science Center (VCURES), Richmond, VA 23298, USA*

Correspondence should be addressed to Pavani Davuluri, davulurip@vcu.edu

Received 30 April 2012; Accepted 14 June 2012

Academic Editor: Guilherme de Alencar Barreto

Automated hemorrhage detection and segmentation in traumatic pelvic injuries is vital for fast and accurate treatment decision making. Hemorrhage is the main cause of deaths in patients within first 24 hours after the injury. It is very time consuming for physicians to analyze all Computed Tomography (CT) images manually. As time is crucial in emergence medicine, analyzing medical images manually delays the decision-making process. Automated hemorrhage detection and segmentation can significantly help physicians to analyze these images and make fast and accurate decisions. Hemorrhage segmentation is a crucial step in the accurate diagnosis and treatment decision-making process. This paper presents a novel rule-based hemorrhage segmentation technique that utilizes pelvic anatomical information to segment hemorrhage accurately. An evaluation measure is used to quantify the accuracy of hemorrhage segmentation. The results show that the proposed method is able to segment hemorrhage very well, and the results are promising.

1. Introduction

Hemorrhage is the leading cause of death in patients with traumatic pelvic fractures. These fractures are most often associated with motor vehicle accidents, falling from heights, and with crush injuries. The mortality rate for pelvic fractures range from 5% to 15%, and the mortality rate for pelvic fracture patients with hemorrhagic shock ranges from 36% to 54% [1, 2]. The majority of deaths caused due to hemorrhage occur within the first 24 hours after the injury [1, 3]. Hence, it is very important to quickly and accurately identify the source of bleeding and control the hemorrhage in a very short period.

The bleeding sites in the pelvic region originate from the fractured bone, venuous plexus, major pelvic veins, and/or damaged arteries [4, 5]. In recent years, contrast-enhanced computed tomography (CT) has been widely used by the radiologists for the examination of hemorrhage and characterization of fractures in traumatic pelvic injuries [2–4, 6]. However, depending on the CT slice thickness, it is rather time consuming for the radiologists to examine all the images, and it is often difficult to identify bleeding sites in the first review of these images. As time is a crucial factor in emergency medicine, there is a need for automated detection of hemorrhage. Identification of the bleeding site alone is not sufficient to assess the bleeding severity. Therefore, it is valuable to segment the detected hemorrhage to see if angiography is needed or not.

Detection and segmentation of hemorrhage in the pelvic region is very challenging due to the injury severity, variation in bleeding contrast from patient to patient, variation in size and shape of the bone, and the presence of several arteries in the region that may be injured. Due to the location of bones and arteries in various locations within the

image, the entire image must be searched for hemorrhage. In addition, hemorrhage cannot be characterized by a single gray level. The gray levels of hemorrhage depend on the phase of CT scan. In the arterial phase (phase in which the pelvic region is scanned soon after the injection of contrast enhancer), the arteries in pelvic region are highlighted and if any hemorrhage is present, it is also differentiable from the soft tissues due to the contrast enhancer. But in the veinal phase (phase in which the pelvic region is scanned with some delay after the injection of contrast enhancer), the hemorrhage is not much differentiable from the soft tissues as the soft tissues start absorbing the enhancer. In general, the hemorrhage gray levels vary from patient to patient in a way that if a patient is bleeding heavily then the hemorrhage is highlighted more than in the patient where the bleeding is slow. Identification of hemorrhage boundary is not easy as the variation in gray level between the hemorrhage and the soft tissues does not vary much. Also, the hemorrhage gray level is not constant throughout the region. The gray level of hemorrhage is much higher around the center of the hemorrhage and fades out around the edges. Another important challenge is, the hemorrhage can occur due to the fractured bones. Hence, it is important to segment the hemorrhage region accurately when near bone. To overcome these challenges, anatomical information must be incorporated in the segmentation process.

Very few researchers have developed techniques for hemorrhage segmentation in the pelvic region [7]. Previous studies utilized a threshold-based method to segment hemorrhage. Furthermore, the method is only able to segment hemorrhage located in one particular region in the image. Even though there are very few studies on hemorrhage segmentation in pelvic region, there are several studies on medical image segmentation for various applications such as vascular segmentation, bone segmentation, hemorrhage segmentation, and so forth [8, 9]. Some of the existing methods are threshold based methods, region growing methods, clustering, markov random field (MRF) models, artificial neural networks, deformable models, atlas-based methods, level set methods, and so forth.

Threshold-based methods are one of the simplest methods that are used for segmentation. In this method, the pixels in the image are classified into groups based on a threshold value. Though this method is simple, it is sensitive to noise and intensity inhomogeneities, as it does not account for spatial characteristics of an image [10, 11]. Region-growing techniques are used to segment regions based on some similarity criteria. In this technique, a single seed is selected initially, and all the pixels around it are selected based on some predefined criteria. The limitation of this method is that it is susceptible to noise and partial volume effects [12, 13]. Clustering techniques like fuzzy c-means algorithms, K-means clustering, Kernel based methods, and so forth are unsupervised techniques developed for segmentation [14]. Though these techniques are computationally fast, they are either sensitive to noise or intensity inhomogeneities as they do not consider spatial context or depends on initialization.

Some researchers have used artificial neural networks for the segmentation [15, 16]. Artificial neural networks are parallel networks of processing elements that simulate biological learning. These networks have high-parallel ability and high interaction among the processing units enabling it to model any kind of process. However, these networks need to be trained beforehand, and the amount of time taken for training may be very long, and the results of these networks are influenced by initialization.

Deformable model techniques are other techniques that are used for segmentation [17, 18]. These techniques use closed parametric curves or surfaces that deform under the influence of internal and external forces. These techniques incorporate a smoothness constraint that provides robustness to noise and spurious edges. However, the disadvantages include poor convergence to concave boundaries and sensitivity to initialization. Level-set methods are other techniques that are based on a moving contour as the zero-level set of a time-evolving scalar function over a regular grid [19, 20]. The curve is deformed according to a given set of partial differential equations. Atlas-based methods are based on a standard template or atlas [21, 22]. The atlas is created based on the information of the anatomy that requires segmentation. The created atlas is then used as a reference for segmenting new images. The atlas-based methods are useful only for the segmentation of structures that do not exhibit great variation and are not extremely detailed.

Along with these segmentation techniques, there are other techniques such as watershed techniques that use concepts from edge detection and mathematical morphology to partition image into homogeneous regions [23]. These techniques suffer from over segmentation. However, recent studies have developed improved methods to overcome some of the drawbacks to segmentation [24, 25].

Some of these above mentioned techniques use a specific criterion to segment regions which are not usually adaptable to images with poor quality. However, incorporation of anatomical information makes the approach more adaptable to each and every image as the gray levels vary from image to image within the same patient. This paper presents a novel heuristic approach to segment hemorrhage which utilizes artery and bone information to initially detect the hemorrhage and then segments hemorrhage in multistages through hemorrhage matching, rule optimization, and region growing.

The rest of the paper is organized as follows. Section 2 describes the methodology used for the study. The results section gives the results obtained using the described methods along with the data used for the study. This section also discusses the obtained results. Finally, the conclusion summarizes the work done and presents the future work for the study.

2. Methods

Automated detection of the presence and extent of hemorrhage is extremely important for assessing injury severity and for fast accurate decision making and treatment planning. Hence, it is very crucial to utilize the artery and bone information in order to detect and segment the hemorrhage.

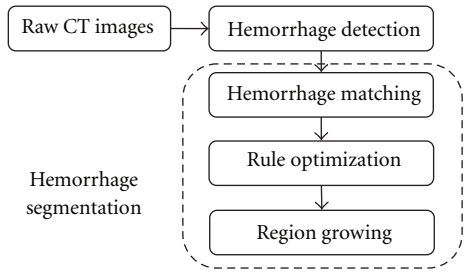


Figure 1: Schematic diagram of hemorrhage detection and segmentation.

Figure 1 provides the schematic diagram of hemorrhage detection and segmentation.

The proposed hemorrhage segmentation technique involves locating the hemorrhage, hemorrhage matching, support vector machine (SVM) based rule optimization for determining hemorrhage regions under different cases, and finally region growing to determine the hemorrhage pixels missed even after the optimization. Each step in the process is explained in detail in the following subsections.

2.1. Hemorrhage Detection. Hemorrhage detection is vital in pelvic trauma to assess the injury severity and is the preparation step for hemorrhage segmentation. Our previous work focused on the hemorrhage detection from pelvic CT images [26, 27]. This work is a continuation of our previous work on hemorrhage detection. Figure 2 shows the schematic setup for hemorrhage detection. A brief description of our previous work is provided below.

2.1.1. Preprocessing. The first step in the hemorrhage detection is to remove any artifacts such as tables, hands, cables, and so forth from the pelvic CT images and extract the pelvic region. This is achieved using morphologic operations and blob analysis [26]. The next stage of hemorrhage detection is to segment bone.

2.1.2. Bone Segmentation and Masking. Once the pelvic region is extracted, the pelvic bones are segmented. Figure 3 below shows the setup for bone segmentation. This involves bone mask formation, edge detection, shape matching and object recognition, edge merging, bone segmentation, and masking. The bone mask is formed by setting a threshold in order to separate bone regions from nonbone regions. However, nonbone regions with gray levels greater than the threshold may also be determined as bone regions at this stage. These false bone regions are later eliminated in the shape matching and object recognition phase. Canny edge detection technique is used to determine the edges of the obtained mask. This technique is used because of its ability to detect true strong and weak edges. Once the bone edges are determined, seed growing technique is used to select pixels closer to the true edge of the bone region. This gives the initial segmented bone image. Later, shape matching is used to determine the best templates that match these segmented regions in each image. These templates are obtained from Visible Human Project dataset manually and offline. A total of 73 templates are used for the study. The best template detection helps determine the position of arteries in the pelvic region, explained later. This process eliminates the nonbone objects from the image by determining the shape matching cost [28–32]. Hence, initial bone regions are segmented.

After segmenting the bone regions, the edges of the bones are determined using canny edge detection technique. In some cases, the edges of the bones may not be fully connected. In order to ensure better masking of the bone, the edges of the bone in the current slice are merged with the bone in the previous and the next slice. Since the study is not about fracture detection, bone merging will have minimal effect on the hemorrhage detection. The next step is final bone segmentation. This is done in a way similar to that of the initial bone segmentation using seed growing technique. The final segmented bone is masked by setting its gray level values to zero.

2.1.3. Artery Detection and Masking. The major arteries in the pelvic region are aorta and its branches (common iliac arteries). Since arteries and bleeding are of similar gray levels, the detection of arteries will help estimate the bleeding gray levels. Hence, the next step is to detect arteries in the pelvic region. The aorta, common iliac arteries, and the external iliac arteries are determined using template matching and from segmented bone location [26, 29–31, 33]. The internal iliac arteries are determined from the position of the external iliac arteries. These detected arteries are then masked to avoid any false hemorrhage detection.

2.1.4. Hemorrhage Detection. After masking the major arteries, the image is searched for unwanted objects other than hemorrhage. The unwanted objects are residual bone pixels or any pixels that are left even after masking the bone and arteries other than the hemorrhage pixels. They are removed by using morphologic operations. After the filtration of unwanted objects, the region in the image that falls within the gray-level range of arteries is considered as hemorrhage and its center coordinates are identified as the centroid of the hemorrhage region [26, 30].

The hemorrhage detected may not be the complete region of hemorrhage especially during the veinal phase. If some of the hemorrhage pixels gray levels are similar to that of soft tissues, especially during the veinal phase, then those pixels would have been eliminated during the filtration of unwanted objects. In addition, the gray levels of hemorrhage that lie within artery gray levels and higher are considered as hemorrhage. However to identify the hemorrhage severity, the entire hemorrhage region must be known.

2.2. Hemorrhage Segmentation. Another important challenge is the identification of bleeding next to the bone, as the hemorrhage can occur due to the fractured bones. Hence, it is important to segment the hemorrhage region accurately when present next to the bone. The proposed segmentation process consists of hemorrhage matching, rule optimization,

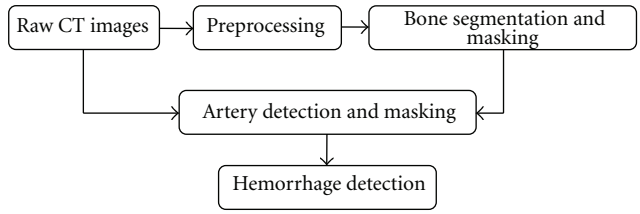


Figure 2: Schematic setup for hemorrhage detection.

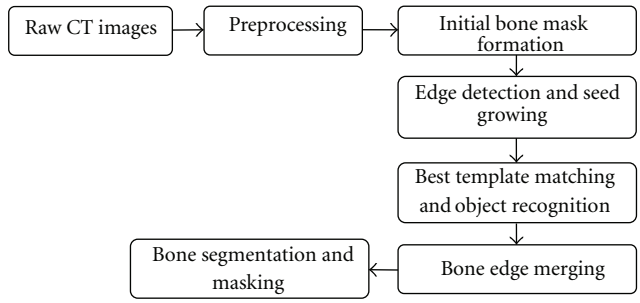


Figure 3: Bone segmentation setup.

and region growing, which are described in detail in the following sub sections.

2.2.1. Hemorrhage Matching by Mutual Information Maximization. The first step of hemorrhage segmentation is hemorrhage matching. The hemorrhage region detected using the previously mentioned method does not contain all the hemorrhage pixels especially the boundaries of the hemorrhage. Hemorrhage matching helps identify the threshold, that is, the optimum minimum gray level G_{opt} for segmenting the hemorrhage region. This is accomplished using the mutual information maximization (MIM). First, a window of size $q \times q$ in the preprocessed CT image is selected as a region of interest (ROI) S around the centroid of the detected hemorrhage. The range $[G_{\text{min}}, G_{\text{max}}]$ of the hemorrhage gray levels are then determined from the detected hemorrhage. Then a gray level G_{mi}, where $G_{\text{min}} \leq G_{\text{mi}} \leq G_{\text{max}}$ is chosen as the minimum gray level and all the pixels in ROI S that lie within $[G_{\text{mi}}, G_{\text{max}}]$ are chosen as hemorrhage pixels. Morphologic operations are performed to eliminate any nonhemorrhage regions in each of these determined hemorrhage images. This obtained hemorrhage image is individually compared to the initial detected hemorrhage image using mutual information (MI) technique in order to find the amount of information each image contains about the detected hemorrhage [34]. This MI is calculated between the previously detected hemorrhage image and the hemorrhage images obtained for different gray level ranges. The cut-off gray level that contains the maximum information about the detected hemorrhage is considered as the optimum minimum gray level G_{opt} at this stage. The mutual information in this process is determined in the following manner. Let C_d be the detected hemorrhage image from the previous section, and let $\{B_1, \ldots, B_i, \ldots, B_m\}$, where $i = 1, 2, \ldots, m$ be the hemorrhage regions obtained with the initial cutoff that ranges within $[G_{\text{min}}, G_{\text{max}}]$. The mutual information between images C_d and B_i is determined using

$$\text{MI}\ (C_d, B_i) = H\ (C_d) + H\ (B_i) - H\ (C_d, B_i), \tag{1}$$

where $H(C_d)$, and $H(B_i)$, are the entropies of images C_d and B_i, and $H(C_d, B_i)$ is their joint entropy, and are computed as follows:

$$\begin{aligned} H\ (C_d) &= -\sum_c P_{C_d}(c) \log P_{C_d}(c), \\ H\ (B_i) &= -\sum_b P_{B_i}(b) \log P_{B_i}(b), \\ H\ (C_d, B_i) &= -\sum_{c,b} P_{C_d,B_i}(c,b) \log P_{C_d,B_i}(c,b), \end{aligned} \tag{2}$$

where, $P_{C_d}(c)$, $P_{B_i}(b)$ denote individual probability distributions. $P_{C_d,B_i}(c,b)$ denotes the joint probability distribution of the images.

The cut-off gray level for which the mutual information between C_d and B_i is maximum, is the optimum gray level G_{opt} and the image is the optimum image at this stage of segmentation. This process is called mutual information maximization. The pixels within the image that lie within $[G_{\text{opt}}, G_{\text{max}}]$ are considered as hemorrhage pixels, and G_{opt} is considered as the minimum hemorrhage gray level from now on. However, G_{opt} may not be the actual minimum gray level of hemorrhage as these cut-off gray levels are from the detected hemorrhage and may not include all the hemorrhage pixels such as boundary pixels which might have gray levels less than G_{opt}. From now on, the hemorrhage region is denoted by R. These undetermined hemorrhage pixels are segmented using the method explained in the following subsection.

2.2.2. Support Vector Machine-Based Rule Optimization for Hemorrhage Segmentation. The utilization of pixel gray levels alone is not enough to determine whether a pixel is hemorrhage or not. Hence, there is a need for incorporation of pixel information such as location, gradient, and so forth around the detected hemorrhage region to properly classify hemorrhage pixels from the nonhemorrhage pixels. This incorporation must be adaptable depending on whether the hemorrhage pixel is in the neighborhood of all hemorrhage pixels or soft tissue pixels. This study incorporates pixel gray levels, distance of the pixel from the hemorrhage foci (the pixel with maximum gray level), the gray level variation within the selected window, and the magnitude of the gradient of each pixel within the selected window in order to achieve better segmentation.

(1) *Rule Generation.* Let B_{opt} be the hemorrhage region image obtained using MIM technique. Let T_{opt} be the boundary of the hemorrhage region in image B_{opt} and $p(x_i, y_j)$ be the hemorrhage pixel of T_{opt}. A window W of size $m \times m$ ($m < q$) is selected around pixel $p(x_i, y_j)$. There are three cases that need to be considered for an optimum segmentation: (1) the selected window W contains all hemorrhage pixels with gray levels within $[G_{\text{opt}}, G_{\text{max}}]$, (2) the majority of the pixels in W being hemorrhage pixels and with gray levels $\geq G_{\text{opt}}$, and (3) the majority of the pixels in W (being hemorrhage or soft tissue pixels) with gray levels $< G_{\text{opt}}$. Therefore, heuristic rules need to be generated for each case in order to optimally segment hemorrhage from nonhemorrhage pixels. The rule for each case is given as follows.

Case 1. W containing all hemorrhage pixels with gray levels within $[G_{\text{opt}}, G_{\text{max}}]$.

If the window contains all pixels with gray levels within $[G_{\text{opt}}, G_{\text{max}}]$, then all these pixels are hemorrhage pixels and can be added to the hemorrhage region R. So the rule in this case is that the pixel must satisfy the below condition in order to be added to region R.

$$R = \left\{\text{pixel} : p_{(x_r, y_s)} \mid G_{\text{opt}} \leq p(x_r, y_s) \leq G_{\text{max}}\right\} \tag{3}$$

Case 2. W containing a majority of hemorrhage pixels, that is, more pixels with gray levels $\geq G_{\text{opt}}$.

If the window contains a majority of (i.e., >50%) hemorrhage pixels with gray levels $\geq G_{\text{opt}}$, then the probability of the rest of the pixels within the neighborhood being hemorrhage is high. As a result, the neighborhood will be dominant with hemorrhage pixels. As the neighborhood is dominant with hemorrhage pixels, pixel gray level and the distance of the pixel from the foci are incorporated into the rule in this case. These parameters are only considered because the variation in magnitude of the gradient and the variation between the pixel gray levels will not add any advantage in differentiating hemorrhage pixels from soft tissue pixels. Each of the parameters used will have a certain weightage which needs to be incorporated for determining hemorrhage pixels. Therefore, the rule is if the pixel satisfies the condition given in (4), then it is considered as hemorrhage pixel and is added to region R.

$$\begin{aligned} R = \Big\{&\text{pixel} : p_{(x_r, y_s)} \mid w_1 \times p(x_r, y_s) \\ &+ w_2 \times D(x_r, y_s) + b > 0\Big\}, \end{aligned} \tag{4}$$

where $D(x_r, y_s)$ is the distance between the pixel in the window and the foci (x_f, y_g), and is given by

$$D(x_r, y_s) = \sqrt{\left(x_f - x_r\right)^2 + \left(y_g - y_s\right)^2} \tag{5}$$

and w_1 and w_2 are the weights and b is the bias.

In order to achieve proper segmentation, these weights need to be optimized. An SVM-based dual Lagrangian technique is used to determine the optimized weights and bias. This optimization technique is explained in the later subsections.

Case 3. W containing a majority of pixels (soft tissue or hemorrhage) with gray levels $< G_{\text{opt}}$.

If the window contains more (i.e., >50%) pixels (soft tissue or hemorrhage) with gray levels $< G_{\text{opt}}$, then the probability of the rest of the pixels within the neighborhood being hemorrhage is lower. Hence, it is required for the algorithm to be more restrictive in this case when compared to the other two cases. Hence, inclusion of magnitude of gradient and the gray level variation within the window along with the pixel gray level and its distance from the foci will help avoid oversegmentation which is crucial. Therefore, the rule associated with this case is

$$\begin{aligned} R = \Big\{&\text{pixel} : p_{(x_r, y_s)} \mid w_3 \times p(x_r, y_s) + w_4 \times D(x_r, y_s) \\ &+ w_5 \times V(x_r, y_s) + w_6 \\ &\times \left| \nabla f_{(x_r, y_s)} \right| + b_1 > 0\Big\}, \end{aligned} \tag{6}$$

where,

$$V(x_r, y_s) = p(x_b, y_b) - p(x_r, y_s), \tag{7}$$

where $p(x_b, y_b)$ is the gray level of the center coordinate of window W, $V(x_r, y_s)$ is the difference in gray level of the center coordinate and the gray level of the pixel in the window. The magnitude of the gradient of each pixel is given in

$$\left| \nabla f_{(x_r, y_s)} \right| = \sqrt{\left(\frac{\partial f}{\partial x_r}\right)^2 + \left(\frac{\partial f}{\partial y_s}\right)^2}. \tag{8}$$

If a pixel in the selected window satisfies the above mentioned condition, then it is considered as hemorrhage and is added to the existing hemorrhage region R.

The weightage of the parameters given in (6) must be determined for each image as these can vary among different images. The weights w_3 through w_6 and the bias b_1 are later optimized using SVM-based dual Lagrangian optimization technique.

(2) *SVM Based Rule Optimization.* The weights used in the previously mentioned rules must be optimized to ensure proper segmentation. These weights must be optimized for each image as these can vary from image to image within the same patient. An SVM-based Lagrangian function in the dual space is used to optimize the weights and the bias. The optimization is solved by the saddle point of Lagrange function in the dual space. For optimization, the data for soft tissue pixels is selected outside the boundary T_{opt}, and the data for hemorrhage pixels is selected from the pixels within the boundary. The selection of these pixels outside the boundary and within the boundary will facilitate the process of identifying the gray level of the boundary pixels. A tenfold cross-validation is used for training and testing the data in order to determine the optimum weights and bias for each of the parameters used in the study. The size of the data set for training and testing depends on the size of the boundary of the hemorrhage in each image. The weights and the bias are optimized separately for each case. For solving with the Lagrangian in dual space, Karush-Kuhn Tucker conditions for the optimum of a constraint function are considered in the study [35].

With those conditions, the dual Lagrangian is given as follows:

$$L_d(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j x_i x_j, \tag{9}$$

where, α_i are the Lagrange multipliers, and x and y are the inputs and the labels and n is the dimensionality of the input.

The inputs in this study are pixel gray level, distance of pixel from the foci, magnitude of the gradient, and the gray level variation. If it is Case 2, there are only 2 input variables. The labels are the classes. In this study, there are two classes: hemorrhage and nonhemorrhage class.

This standard quadratic optimization problem is expressed in matrix notation and formulated as follows:

$$\begin{aligned} &\text{Maximize} \quad L_d(\alpha) = -0.5\alpha^T H \alpha - 1^T \alpha, \\ &\text{subject to} \quad y^T \alpha = 0, \quad 0 \le \alpha \le C, \end{aligned} \tag{10}$$

where H is the Hessian matrix ($H_{ij} = y_i y_j x_i x_j$), C is the penalty parameter, and 1 is a unit vector $1 = [1 \;\; 1 \;\; \cdots \;\; 1]^T$. C is chosen as the upper bound of α because with C the influence of training data points that remain on the "wrong" side of a separating nonlinear hypersurface is limited. Also, the width of the soft margin is controlled by a corresponding C. Large C leads to small number of misclassifications, smaller margin and vice versa. In our study, C is considered to be greater than zero and less than infinity for feasibility. The penalty parameter is optimized using 10-fold cross-validation technique. Solution α_0 from the above equation determines the parameters of the optimal hyperplane w_0 and b_0 as given in

$$\begin{aligned} w_0 &= \sum_{i=1}^{N_{\text{sv}}} \alpha_{0i} y_i x_i, \\ b_0 &= \frac{1}{N_{\text{fsv}}} \left(\sum_{s=1}^{N_{\text{fsv}}} \left(\frac{1}{y_s} - x_s^T w_0 \right) \right), \end{aligned} \tag{11}$$

where w_0 and b_0 are the optimized weights and bias, N_{sv} denotes the number of support vectors, and N_{fsv} denotes the number of free support vectors.

In (11), the support vectors are only used because the Lagrange multipliers are zero for nonsupport vectors. Finally, with the optimal weights and bias, the decision hyperplane $d(x)$ is determined using

$$d(x) = \sum_{i=1}^{n} w_{0i} x + b_0, \tag{12}$$

where x is the test data.

The output of the test data is determined by using an indicator function given in

$$i_F = \text{sign}(d(x)). \tag{13}$$

The number of wrongly classified pixels are determined by comparing the test output with the desired output. The obtained optimized weights and bias are used to determine if a pixel is a hemorrhage pixel or not. The optimized weights are used in the rules, and the pixels in each window W are considered as hemorrhage if they satisfy the optimized rules. However, there is a slight chance of missing the hemorrhage pixels which are outside the boundary and are not located in the selected window. Hence, it is required to include these pixels in the hemorrhage region. Region growing process is used to grow the region around the already determined hemorrhage region R to determine any hemorrhage pixels that are missed during the optimization process. This is described in the following subsection.

2.2.3. Region Growing. The region growing process is the final phase of hemorrhage segmentation. This process is used to determine any missed hemorrhage pixels that are located outside the boundary of R. Figure 4 shows the region growing process used in this study. The region growing process consists of several steps. First, the boundary of the segmented hemorrhage R from the previous phase is used to select a window of size $m \times m$ around each boundary pixel. If the percent of total number of pixels within that window that satisfy the conditions described earlier are $> \eta$, the pixel factor, then the threshold t_1 for the window is determined using

$$t_1 = \text{me}_1 + \text{std}_1, \tag{14}$$

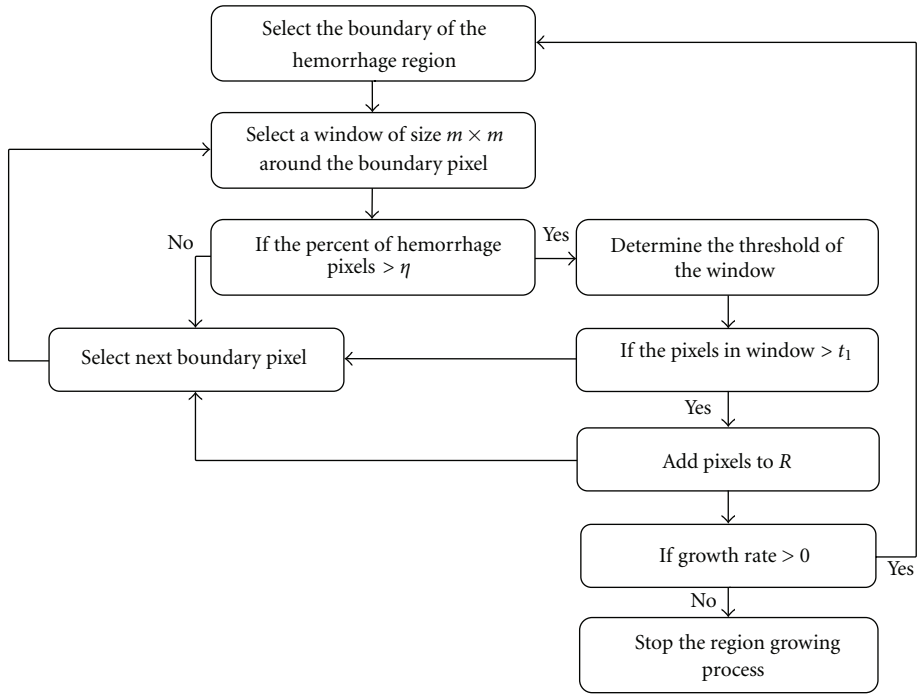


Figure 4: Region growing process.

where me_1 and std_1 are the mean and standard deviation of the gray levels of all the nonbackground pixels in the window and are given by

$$\text{me}_1 = \frac{\sum_{x=1}^{m}\sum_{y=1}^{m} f(x,y)}{m \times m - \text{Card}(S)},$$
$$\text{std}_1 = \sqrt{\frac{\sum_{x=1}^{m}\sum_{y=1}^{m} (f(x,y) - m_1)^2}{m \times m - \text{Card}(S)}}, \quad (15)$$

and $S = \{(x,y) \mid f(x,y) = 0\}$ is the set of pixels located in the background having zero gray level. $\text{Card}(S)$ denotes the cardinality of set S.

If any of the pixels that lie outside the boundary and within the window satisfy t_1, then they are considered as hemorrhage pixels and are added to the existing hemorrhage region R. This entire region growing process is repeated for all the boundary pixels. This complete process constitutes one epoch. If the growth rate of hemorrhage region is >0 in the current epoch, then the entire process is repeated starting from selecting the boundary of the hemorrhage region, else the region growing process is stopped. The growth rate in each epoch is calculated using

$$\text{Growth rate} = \frac{E_c - E_p}{E_c} \times 100, \quad (16)$$

where E_c is the total area of the hemorrhage by the end of current epoch, and E_p is the total area of the hemorrhage by the end of previous epoch. The total region-grown by the end of the region growing process is considered as the final segmented hemorrhage.

2.3. Evaluation Measure for Segmentation. Once the hemorrhage is segmented, a suitable measure is required to quantify the accuracy of segmentation. This study utilizes a measure called missegmented area. The missegmented area measure represents the uncommon area of segmented region (i.e., the pixels of segmented region that are not a true hemorrhage) compared to the gold standard area of segmented hemorrhage. If A_1 and A_2 are the areas of actual and the segmented region, the missegmented area of the two regions is defined as

$$\frac{\text{Cardinality } \{K\}}{\text{Cardinality } \{A_1\}} \times 100, \quad (17)$$

where

$$K = \{\text{pixels} : p \mid p \in A_1 \cup A_2, p \notin A_1 \cap A_2\}. \quad (18)$$

Based on this measure, the segmented hemorrhage will be classified into three categories: good, acceptable, and unacceptable through consultation with a trauma physician and a radiologist, who identified actual hemorrhage contour as the ground truth.

The segmented regions with missegmented area <10% will be classified as good, and regions with missegmented area between 10% and 20% will be considered as acceptable, and finally any region with missegmented area greater than 20% will be considered as unacceptable. These ranges for good, acceptable, and unacceptable are used in the study based on the discussion with expert radiologists who utilize these ranges to determine if a region is properly segmented or not and how severe the bleeding is. The numerical values of K itself are not considered in this study as the radiologists

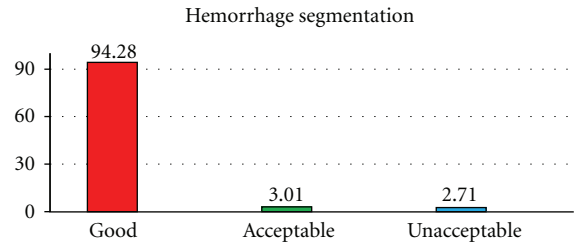


Figure 5: Proposed method performance for hemorrhage segmentation.

are not concerned about the numerical values because these values do not provide any additional information to radiologists about the injury severity.

3. Results and Discussion

3.1. Dataset. The dataset for the study is obtained from Carolinas Health System and Virginia Commonwealth University Medical Center. The data is collected from twelve pelvic trauma patients with each scan consisting of 30 to 70 images with a total of 515 images. These twelve patients exhibit very mild to severe hemorrhage and these patients are selected at random. From the discussion with expert radiologists, it has been found that these number of images selected are sufficient to validate the performance of the proposed method. A statistical t-test is conducted in addition to see if the total number of images used in the study is statistically significant or not. A P value < 0.05 is considered as statistically significant, and a greater value is considered statistically not significant. These images chosen are axial CT images with 5 mm slice thickness.

3.2. Results and Discussion. The proposed method is tested on twelve pelvic trauma patients who exhibit mild to severe bleeding. The total number of images used for the study from these twelve patients is 515 images. The dimensions of each image are 512×512 pixels. A P value of 0.0029 is obtained using the t-test showing that the selected number of images is statistically significant to test the proposed method. The CT scan include both images taken during arterial phase and the veinal phase. The hemorrhage is more distinguishable in the arterial phase than in the veinal phase.

The ROI size $q \times q$ in the hemorrhage matching section is chosen as 100. This value is chosen because a smaller window size may not contain the entire hemorrhage region and if a larger size is chosen, then the nonhemorrhage tissues might be present along with the hemorrhage region making hemorrhage segmentation much complicated. During the rule optimization, the values chosen initially for the penalty parameter C are 0.1, 0.01, and 0.001. The optimal C value obtained is different for each image in the patient. It is dependent on the accuracy of classification. The penalty parameter for which the accuracy is maximum is chosen as the optimal penalty parameter. For the region growing process, the window size m is chosen as 3. The pixel factor η is chosen as 50. This value is selected because, in order for the algorithm to be restrictive in region growing, it is required to consider a window that is dominated by the hemorrhage pixels. If the value is chosen lower than this, the probability of oversegmentation might increase, and if the value is chosen higher than this value, then the algorithm becomes too restrictive and might leave hemorrhage pixels out affecting the segmentation.

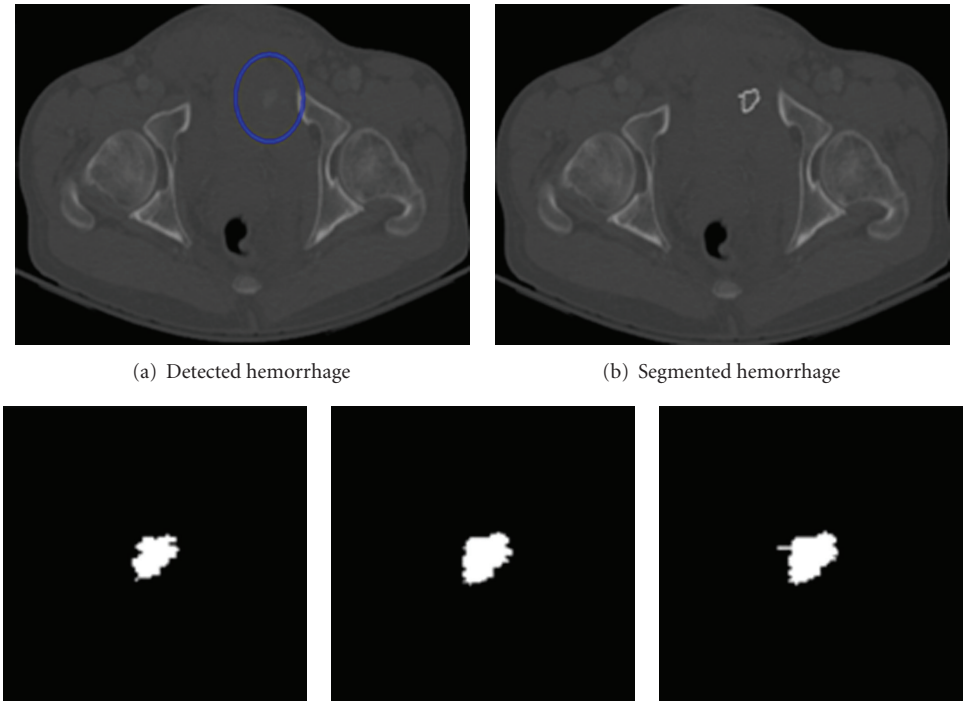
Figure 5 shows the hemorrhage segmentation results. The proposed method is able to segment the hemorrhage very well for 94.28% of the cases used in the study. These cases are considered as good as the missegemented area is <10%. The overall average missegmented area is 5.3% For 3.01% of the cases, the segmented hemorrhage is acceptable. The average missegmented area in these acceptable cases is 14.47%. For the remaining 2.71% of the cases, the segmented hemorrhage is unacceptable, and the average missegmented area is 26.52%.

Figures 6 and 7 show the results of segmented hemorrhage. These are some of the cases where hemorrhage is very well segmented. The results show that the proposed method has segmented hemorrhage very well. Figures 6(c) through 6(e) gives the segmentation results at various stages of segmentation, that is, segmentation results after hemorrhage matching using MIM, rule optimization, and region growing. The percentile of hemorrhage area grown from the results of MIM technique to optimization technique is 24.6% and the percentile of hemorrhage area grown from the optimization to region growing is 3.53%. These results show that the rule optimization helps in determining hemorrhage accurately, and the region growing helps determine the missing hemorrhage pixels.

In the case of patient in Figure 7, the percentile of hemorrhage area grown from the results of MIM technique to optimization technique is 22.3% and the hemorrhage area is not grown during the region growing process as all the hemorrhage pixels are identified in the earlier stage itself.

Figure 8 shows the segmentation results of hemorrhage located next to the bone. This segmentation is considered as acceptable. As hemorrhage is located next to the bone, the gray levels of the faded bone edges might be similar to hemorrhage gray levels. The use of distance information and gray level variation information helped in differentiating the hemorrhage from the bone regions for majority of the pixels. However for few pixels, the proposed method is unable to differentiate between the hemorrhage and bone pixels. Figures 8(c) through 8(e) shows the performance of proposed method at various stages. In these figures, the percentile of hemorrhage area grown from the results of MIM technique to optimization technique is 25.42%. And the percentile of hemorrhage area grown from the optimization to region growing is 0.56%. The hemorrhage area grown through region growing is much less in this case. It can be observed from this that the rule optimization has segmented most of the hemorrhage pixels.

The results are validated on the basis of assessment and evaluation made by the radiologists on the CT images. The proposed method is able to segment hemorrhage very well for majority of the cases. The segmentation is unacceptable in few cases which may be due to the bridging



(a) Detected hemorrhage

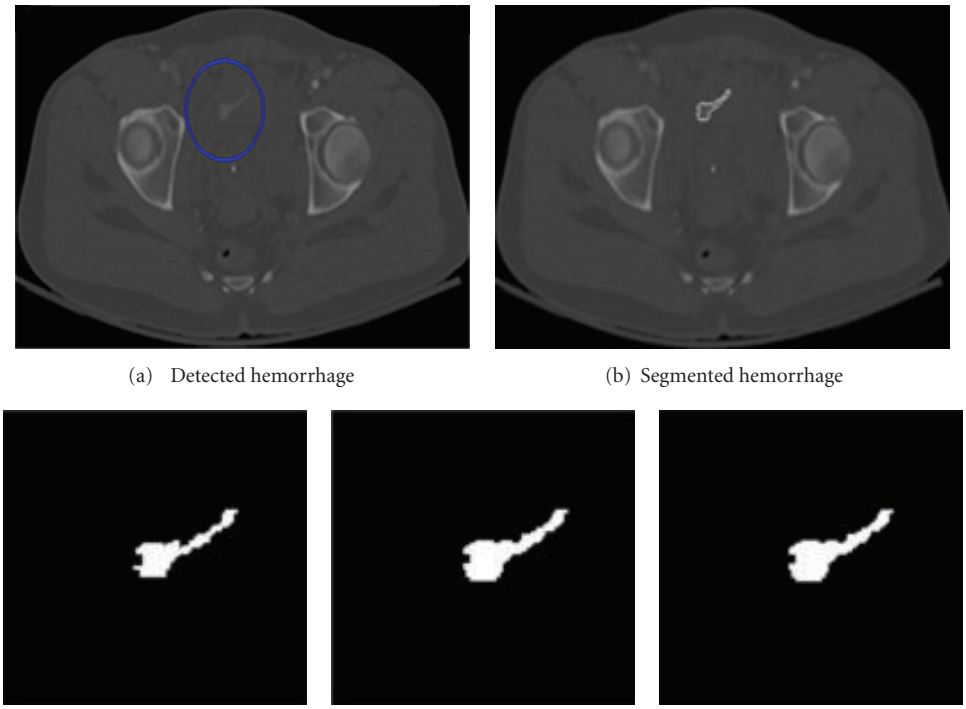
(b) Segmented hemorrhage

(c) Segmented image after hemorrhage matching

(d) Segmented image after rule optimization

(e) Segmented image after Region growing

Figure 6: Sample hemorrhage segmentation results.



(a) Detected hemorrhage

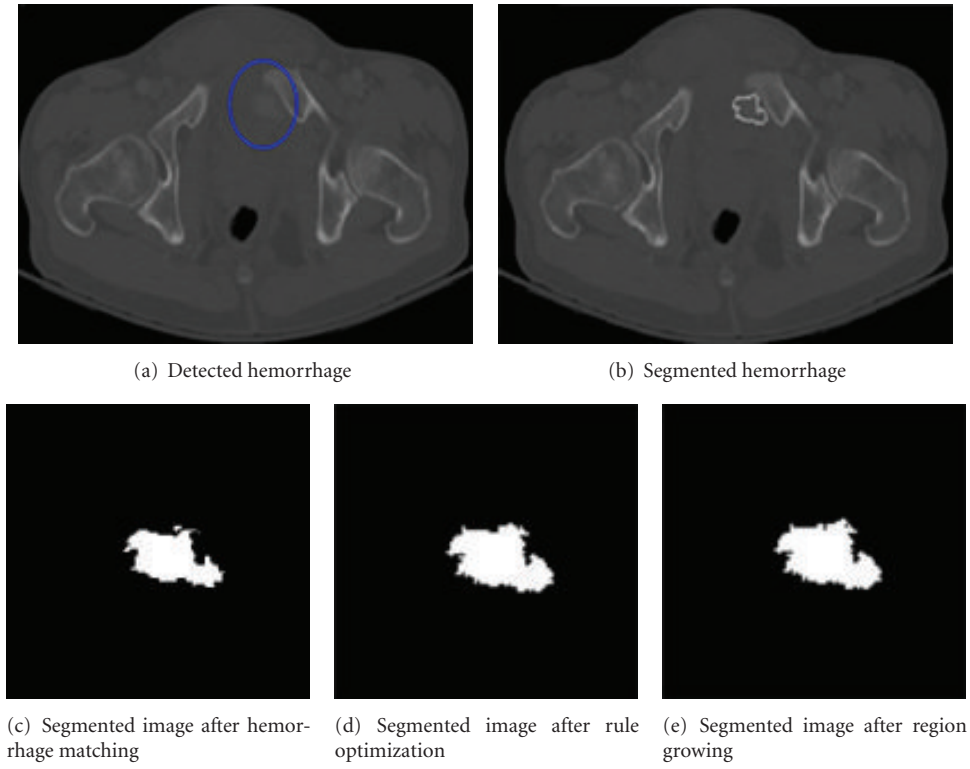
(b) Segmented hemorrhage

(c) Segmented image after hemorrhage matching

(d) Segmented image after rule optimization

(e) Segmented image after region growing

Figure 7: Sample hemorrhage segmentation results.



(a) Detected hemorrhage

(b) Segmented hemorrhage

(c) Segmented image after hemorrhage matching

(d) Segmented image after rule optimization

(e) Segmented image after region growing

Figure 8: Sample segmentation results for hemorrhage located next to bone.

of hemorrhage pixels through soft tissue pixels. Hence, these few pixels are left out during the segmentation. Increasing the size of selected window might help segment these pixels. However, the tradeoff is, it might lead to oversegmentation. Incorporating pixel information into the rule optimization helps to differentiate the hemorrhage from soft tissue and bone region. The optimization technique is able to segment hemorrhage edges very well. The region growing process is able to determine the missed hemorrhage pixels. In addition, the proposed method is able to segment hemorrhage edges that may not be measurable through visual inspection. The overall processing time of hemorrhage detection and segmentation for each slice in a scan is a few seconds when run on a Intel(R)Core(TM)i7-2600 CPU@3.40 GHz machine. This is much faster than the manual hemorrhage detection that takes more than a minute for each slice. The entire process is fully automated. Automated detection with relatively high speed helps physicians make fast and accurate diagnostic decisions and treatment planning which is very crucial for traumatic pelvic injuries.

4. Conclusions and Future Work

This paper presents a fully automated hemorrhage segmentation technique that consists of hemorrhage matching, rule optimization, and region growing. These techniques incorporate the pixel gray level information, magnitude of the gradient, distance measure, and the gray level variation for segmentation. The results show that the proposed method is capable of segmenting hemorrhage well. Automated hemorrhage segmentation, once verified with more data, will be an important component of computer-assisted decision making system. Future work will focus on the quantitative measurement of hemorrhage such as determining hemorrhage volume, identifying the location of hemorrhage with respect to the bone, and so forth on the basis of larger data set.

Conflict of Interests

The authors report no actual or potential conflict of interest in relation to this paper.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant no. IIS0758410. The authors would like to thank Carolinas Health System and Virginia Commonwealth University Medical Center for providing data for the study.

References

[1] K. Eckroth-Bernard and J. W. Davis, "Management of pelvic fractures," *Current Opinion in Critical Care*, vol. 16, no. 6, pp. 582–586, 2010.

[2] J. A. Requarth and P. R. Miller, "Aberrant obturator artery is a common arterial variant that may be a source of unidentified

hemorrhage in pelvic fracture patients," *Journal of Trauma*, vol. 70, no. 2, pp. 366–372, 2011.

[3] J. Uyeda, S. W. Anderson, J. Kertesz, and J. A. Soto, "Pelvic CT angiography: application to blunt trauma using 64MDCT," *Emergency Radiology*, vol. 17, no. 2, pp. 131–137, 2010.

[4] W. Yoon, J. K. Kim, Y. Y. Jeong, J. J. Seo, J. G. Park, and H. K. Kang, "Pelvic arterial hemorrhage in patients with pelvic fractures: detection with contrast-enhanced CT," *Radiographics*, vol. 24, no. 6, pp. 1591–1605, 2004.

[5] H. C. Jeske, R. Larndorfer, D. Krappinger et al., "Management of hemorrhage in severe pelvic injuries," *Journal of Trauma*, vol. 68, no. 2, pp. 415–420, 2010.

[6] A. Furlan, S. Fakhran, and M. P. Federle, "Spontaneous abdominal hemorrhage: causes, CT findings, and clinical implications," *American Journal of Roentgenology*, vol. 193, no. 4, pp. 1077–1087, 2009.

[7] S. Vasilache, *Image segmentation and analysis for automated classification of traumatic pelvic injuries*, Ph.D. thesis, 2010.

[8] N. Pérez, J. Valdés, M. Guevara, and A. Silva, *Advances in Computational Vision and Medical Image Processing*, Springer, Amsterdam, The Netherlands, 2009.

[9] Z. Ma, J. M. R. S. Tavares, R. N. Jorge, and T. Mascarenhas, "A review of algorithms for medical image segmentation and their applications to the female pelvic cavity," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 13, no. 2, pp. 235–246, 2010.

[10] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[11] T. Heimann, B. Van Ginneken, M. A. Styner et al., "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, Article ID 4781564, pp. 1251–1265, 2009.

[12] R. B. Dubey, M. Hanmandlu, S. K. Gupta, and S. K. Gupta, "Region growing for MRI brain tumor volume analysis," *Indian Journal of Science and Technology*, vol. 2, no. 9, pp. 26–31, 2009.

[13] Z. Peter, V. Bousson, C. Bergot, and F. Peyrin, "A constrained region growing approach based on watershed for the segmentation of low contrast structures in bone micro-CT images," *Pattern Recognition*, vol. 41, no. 7, pp. 2358–2368, 2008.

[14] N. A. M. Isa, S. A. Salamah, and U. K. Ngah, "Adaptive fuzzy moving K-means clustering algorithm for image segmentation," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 4, pp. 2145–2153, 2009.

[15] Z. Dokur, "A unified framework for image compression and segmentation by using an incremental neural network," *Expert Systems with Applications*, vol. 34, no. 1, pp. 611–619, 2008.

[16] G. Ertaş, H. Ö. Gülçür, O. Osman, O. N. Uçan, M. Tunaci, and M. Dursun, "Breast MR segmentation and lesion detection with cellular neural networks and 3D template matching," *Computers in Biology and Medicine*, vol. 38, no. 1, pp. 116–126, 2008.

[17] L. He, Z. Peng, B. Everding et al., "A comparative study of deformable contour methods on medical image segmentation," *Image and Vision Computing*, vol. 26, no. 2, pp. 141–163, 2008.

[18] J. V. Stough, R. E. Broadhurst, S. M. Pizer, and E. L. Chaney, "Clustering on local appearance for deformable model segmentation," in *Proceedings of the 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '07)*, pp. 960–963, April 2007.

[19] G. Chung and L. A. Vese, "Image segmentation using a multilayer level-set approach," *Computing and Visualization in Science*, vol. 12, no. 6, pp. 267–285, 2009.

[20] C. Li, R. Huang, Z. Ding, J. C. Gatenby, D. N. Metaxas, and J. C. Gore, "A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2007–2016, 2011.

[21] Z. Li, E. A. Hoffman, and J. M. Reinhardt, "Atlas-driven lung lobe segmentation in volumetric X-ray CT images," *IEEE Transactions on Medical Imaging*, vol. 25, no. 1, pp. 1–16, 2006.

[22] I. Išgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. Van Ginneken, "Multi-atlas-based segmentation with local decision fusion-application to cardiac and aortic segmentation in CT scans," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1000–1010, 2009.

[23] R. Shojaii, J. Alirezaie, and P. Babyn, "Automatic lung segmentation in CT images using watershed transform," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '05)*, pp. 1270–1273, September 2005.

[24] M. Frucci and G. Sanniti di Baja, "Oversegmentation reduction in watershed-based grey-level image segmentation," *International Journal of Signal and Imaging Systems Engineering*, vol. 1, no. 1, pp. 4–10, 2008.

[25] X. Xie, C. Ma, X. Yu, and R. Du, "Liver image segmentation using improved watershed method," *Applied Mechanics and Materials*, vol. 58–60, pp. 1311–1316, 2011.

[26] P. Davuluri, J. Wu, K. R. Ward, C. H. Cockrell, K. Najarian, and R. S. Hobson, "An automated method for hemorrhage detection in traumatic pelvic injuries," in *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '11)*, pp. 5108–5111, Boston, Mass, USA, 2011.

[27] S. Vasilache, K. Ward, C. Cockrell, J. Ha, and K. Najarian, "Unified wavelet and gaussian filtering for segmentation of CT images; Application in segmentation of bone in pelvic CT images," *BMC Medical Informatics and Decision Making*, vol. 9, supplement 1, article S8, 2009.

[28] S. Vasilache, W. Chen, K. Ward, and K. Najarian, "Hierarchical object recognition in pelvic CT images," in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '09)*, pp. 3533–3536, September 2009.

[29] J. Wu, P. Davuluri, A. Belle et al., "Fracture detection and quantitative measure of displacement in pelvic CT images," in *Proceedings of the IEEE International Conference onBioinformatics and Biomedicine Workshops (BIBMW '11)*, pp. 600–606, Atlanta, Ga, USA, 2011.

[30] P. Davuluri, J. Wu, A. Belle et al., "A hybrid approach for hemorrhage segmentation in pelvic CT scans," in *Proceedings of the IEEE International Conference onBioinformatics and Biomedicine Workshops (BIBMW '11)*, pp. 548–554, Atlanta, Ga, USA, 2011.

[31] W. Chen, C. Cockrell, K. R. Ward, and K. Najarian, "Intracranial pressure level prediction in traumatic brain injury by extracting features from multiple sources and using machine learning methods," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '10)*, pp. 510–515, Hong Kong, Hong Kong, December 2010.

[32] J. Wu, P. Davuluri, K. R. Ward, C. Cockrell, R. Hobson, and K. Najarian, "Fracture detection in traumatic pelvic CT images," *International Journal of Biomedical Imaging*, vol. 2012, Article ID 327198, 10 pages, 2012.

[33] J. Wu, P. Davuluri, K. Ward, C. Cockrell, R. Hobson, and K. Najarian, "A new hierarchical method for multi-level segmentation of bone in pelvic CT scans," in *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '11)*, pp. 3399–3402, Boston, Mass, USA, 2011.

[34] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[35] T. Huang, V. Kecman, and I. Kopriva, *Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-Supervised, and unSupervised Learning (Studies in Computational Intelligence)*, Springer, New York, NY, USA, 2006.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 639124, 12 pages
doi:10.1155/2012/639124

Research Article

Let Continuous Outcome Variables Remain Continuous

Enayatollah Bakhshi,[1] Brian McArdle,[2] Kazem Mohammad,[3] Behjat Seifi,[4] and Akbar Biglarian[1]

[1] *Department of Statistics and Computer, University of Social Welfare and Rehabilitation Sciences, Tehran 1985713834, Iran*
[2] *Department of Statistics, The University of Auckland, Private Bag 92010, Auckland, New Zealand*
[3] *Department of Biostatistics, School of Public Health and Institute of Public Health Research, Tehran University of Medical Sciences, Tehran, Iran*
[4] *Department of Physiology, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran*

Correspondence should be addressed to Enayatollah Bakhshi, bakhshi@razi.tums.ac.ir

Received 8 November 2011; Revised 21 February 2012; Accepted 29 February 2012

Academic Editor: Alberto Guillén

The complementary log-log is an alternative to logistic model. In many areas of research, the outcome data are continuous. We aim to provide a procedure that allows the researcher to estimate the coefficients of the complementary log-log model without dichotomizing and without loss of information. We show that the sample size required for a specific power of the proposed approach is substantially smaller than the dichotomizing method. We find that estimators derived from proposed method are consistently more efficient than dichotomizing method. To illustrate the use of proposed method, we employ the data arising from the NHSI.

1. Introduction

Recently, logistic regression has become a popular tool in biomedical studies. The parameter in logistic regression has the interpretation of log odds ratio, which is easy for people such as physicians to understand. Probit and complementary log-log are alternatives to logistic model. For a covariate X and a binary response variable Y, let $\pi(X) = P(Y = 1 \mid X = x)$. A related model to the complementary log-log link is the log-log link. For it, $\pi(x)$ approaches 0 sharply but approaches 1 slowly. When the complementary log-log model holds for the probability of a success, the log-log model holds for the probability of a failure [1].

These models use a categorical (dichotomous or polytomous) outcome variable. In many areas of research, the outcome data are continuous. Many researchers have no hesitation in dichotomizing a continuous variable, but this practice does not make use of within-category information. Several investigators have noted the disadvantages of dichotomizing both independent and outcome variables [2–10]. Ragland [11] showed that the magnitude of odds ratio and statistical power depend on the cutpoint used to dichotomize the response variable. From a clinical point of view, binary outcomes may be preferred for some reasons such as (1) setting diagnostic criteria for disease, (2) offering a simpler interpretation of common effect measures from statistical models such as odds ratios and relative risks. However, all advantages come at the lost information. From a statistical point of view, this loss of information means more samples which are required to attain prespecified powers.

Moser and Coombs [12] provided a closed-form relationship that allows a direct comparison between the logistic and linear regression coefficients. They also provided a procedure that allows the researcher to analyze the original continuous outcome without dichotomizing. To date, a method that applies the complementary log-log model without dichotomizing and without loss of information has not been available.

We aim to (a) provide a method that allows the researcher to estimate the coefficients of the complementary log-log model without dichotomizing and without loss of information, (b) show that the coefficient of the complementary log-log model can be interpreted in terms of the regression coefficients, (c) demonstrate that the coefficient estimates from

this method have smaller variances and shorter confidence intervals than the dichotomizing method.

2. Methods

2.1. Model. Let $y_1, y_2, \ldots, y_n$ be n independent observations on y, and let $x_1, x_2, \ldots, x_{p-1}$ be $p-1$ predictor variables thought to be related to the response variable y. The multiple linear regression model for the ith observation can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + E_i \quad i = 1, 2, \ldots, n, \tag{1}$$

or

$$y_i = x_i \beta + E_i \quad i = 1, 2, \ldots, n, \tag{2}$$

where

$$x_i = \left(1, x_{i1}, x_{i2}, \ldots, x_{ip-1}\right). \tag{3}$$

To complete the model, we make the following assumptions:

(1) $E(E_i) = 0$ for $i = 1, 2, \ldots, n$,

(2) $\operatorname{var}(E_i) = \sigma^2$ for $i = 1, 2, \ldots, n$,

(3) the independent E_i follows an extreme value distribution for $i = 1, 2, \ldots, n$.

Writing the model for each of the n observations, in matrix form, we have

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 x_{11} & x_{12} & \ldots & x_{1p-1} \\ 1 x_{21} & x_{22} & \ldots & x_{2p-1} \\ . & & & \\ . & & & \\ 1 x_{n1} & x_{n2} & \ldots & x_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ . \\ . \\ E_n \end{bmatrix}, \tag{4}$$

or

$$y = X\beta + E. \tag{5}$$

The preceding three assumptions on E_i and y_i can be expressed in terms of this model:

(1) $E(E) = 0$,

(2) $\operatorname{cov}(E) = \sigma^2 I$,

(3) the E_i is extreme value $(0, \sigma^2)$ for $i = 1, 2, \ldots, n$.

2.2. (Largest) Extreme Value Distribution. The PDF and CDF of the extreme value distribution are given by

$$\begin{aligned} f(y \mid x\beta, \sigma) &= \frac{\pi}{\sigma\sqrt{6}} \\ &\quad \times \exp\left(-\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}} \right. \\ &\quad \left. - \exp\left(\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\ &\quad - \infty \langle x \langle \infty, \ \sigma \rangle 0, \\ P(y \le c) &= \exp\left(-\exp\left(-\frac{c - x\beta + k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\ &\quad - \infty \langle x \langle \infty, \sigma \rangle, \ k \approx 0.45. \end{aligned} \tag{6}$$

It is easy to check that

$$\begin{aligned} \omega_j = \frac{\ln \pi_1}{\ln \pi_2} &= \frac{\ln(p(y \le c \mid x))}{\ln\left(p\left(y \le c \mid x_{(-1,j)}\right)\right)} \\ &= \frac{-\exp(-((c - x'\beta + k\sigma)/\sigma) \times \pi/\sqrt{6})}{-\exp\left(-\left(\left(c - x'_{(-1,j)}\beta + k\sigma\right)/\sigma\right) \times \pi/\sqrt{6}\right)} \\ &= \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right) \Longrightarrow \pi_1 = \pi_2^{\exp((\pi/\sqrt{6}) \cdot (\beta_j/\sigma))}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} x &= \left(1, x_1, \ldots, x_j, \ldots, x_{p-1}\right), \\ x_{(-1,j)} &= \left(1, x_1, \ldots, x_j - 1, \ldots, x_{p-1}\right), \\ \beta &= \left(\beta_0, \ \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}\right)'. \end{aligned} \tag{8}$$

To return to a random sample of observations $(y_1, y_2, \ldots, y_n)$, we conclude that the PDF and CDF of each independent y_i are given by (6), and the corresponding equality (7) is given by

$$\frac{\ln \hat{\pi}_1}{\ln \hat{\pi}_2} = \exp\left(\frac{\pi}{\hat{\sigma}\sqrt{6}}\hat{\beta}_j\right), \tag{9}$$

where the estimate $\hat{\beta}_j$ is the $(j+1)$th element of vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_{p-1})'$. It is readily shown that the results also hold true for the smallest extreme value distribution (Appendix A).

2.3. The Proposed Confidence Intervals. Let

$$\begin{aligned}\hat{\beta} &= \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_{p-1}\right)' \\ &= (X'X)^{-1}X'Y \quad j = 0, \ldots, p-1, \\ \hat{\sigma}^2 &= \frac{Y'\left(I_n - X(X'X)^{-1}X'\right)Y}{(n-p)}.\end{aligned} \tag{10}$$

According to the preceding three assumptions on E_i and y_i, we obtain

$$\begin{aligned}E\left(\hat{\beta}\right) &= E\left[(X'X)^{-1}X'Y\right] \\ &= (X'X)^{-1}X'EY = (X'X)^{-1}X'X\beta = \beta, \\ E(\hat{\sigma}^2) &= \frac{1}{n-p}E\left(Y'\left(I_n - X(X'X)^{-1}X'\right)Y\right) \\ &= \frac{1}{n-p}\left\{\operatorname{tr}\left[\left(I_n - X(X'X)^{-1}X'\right)\sigma^2 I\right]\right. \\ &\qquad \left. + E(Y')\left[I_n - X(X'X)^{-1}X'\right]E(Y)\right\} \\ &= \frac{1}{n-p}\left\{\sigma^2 \operatorname{tr}\left[I_n - X(X'X)^{-1}X'\right]\right. \\ &\qquad \left. + \beta'X'\left[I_n - X(X'X)^{-1}X'\right]X\beta\right\} \\ &= \frac{1}{n-p}\left\{\sigma^2\left[n - \operatorname{tr}\left(X(X'X)^{-1}X'\right)\right]\right. \\ &\qquad \left. + \beta'X'X\beta - \beta'X'X(X'X)^{-1}X'X\beta\right\} \\ &= \frac{1}{n-p}\left\{\sigma^2\left[n - \operatorname{tr}\left(X(X'X)^{-1}X'\right)\right]\right. \\ &\qquad \left. + \beta'X'X\beta - \beta'X'X\beta\right\} \\ &= \frac{1}{n-p}\sigma^2[n - \operatorname{tr}(I_P)] = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2.\end{aligned} \tag{11}$$

Therefore, $\hat{\beta}$ and $\hat{\sigma}^2$ are unbiased estimators of β and σ^2.

We have assumed that E_i is distributed as an extreme value, and we use the approximation of the extreme value distribution of the errors E_i by the normal distribution. For normally distributed observations, $\hat{\beta}_j/(\hat{\sigma}\sqrt{\delta_j})$ follows a noncentral t distribution with $n-p$ degree of freedom and noncentrality parameter $-\infty < \beta_j/(\sigma\sqrt{\delta_j}) < \infty$,

$$\begin{aligned}1 - \alpha = P\left\{t_{1-(\alpha/2)}\left[n-p, \frac{\beta_j}{\left(\sigma\sqrt{\delta_j}\right)}\right]\right. \\ \left. < \frac{\hat{\beta}_j}{\left(\hat{\sigma}\sqrt{\delta_j}\right)} < t_{\alpha/2}\left[n-p, \frac{\beta_j}{\left(\sigma\sqrt{\delta_j}\right)}\right]\right\},\end{aligned} \tag{12}$$

where $t_{\alpha/2}[r, s]$ represents the $100(1-(\alpha/2))$ percentile point of a noncentral t distribution with r degrees of freedom and noncentrality parameter $-\infty < s < \infty$, and δ_j is the $(j+1)$st diagonal element of $(X'X)^{-1}$. We use the approximation of the percentiles of the noncentral t distribution by the standard normal percentiles [13], then

$$\begin{aligned}1 - \alpha = P\left\{\frac{\beta_j/\left(\sigma\sqrt{\delta_j}\right) - z_{\alpha/2}\left[1 + \left(\beta_j^2/\left(\sigma^2\delta_j\right) - z_{\alpha/2}^2\right)/2(n-p)\right]^{1/2}}{1 - \left(z_{\alpha/2}^2/2(n-p)\right)}\right. < \\ \left. \frac{\hat{\beta}_j}{\left(\hat{\sigma}\sqrt{\delta_j}\right)} < \frac{\beta_j/\left(\sigma\sqrt{\delta_j}\right) + z_{\alpha/2}\left[1 + \left(\left(\beta_j^2/\left(\sigma^2\delta_j\right) - z_{\alpha/2}^2\right)/2(n-p)\right)\right]^{1/2}}{1 - \left(z_{\alpha/2}^2/2(n-p)\right)}\right\},\end{aligned} \tag{13}$$

$$\left(\frac{\beta_j}{\sigma}\right)^U = \left\{\frac{\hat{\beta}_j}{\hat{\sigma}}\left[1 - \frac{z_{\alpha/2}^2}{2(n-p)}\right] + z_{\alpha/2}\left[\delta_j\left(1 + \left(\frac{\left(\hat{\beta}_j^2/\hat{\sigma}^2\delta_j\right) - z_{\alpha/2}^2}{2(n-p)}\right)\right)\right]^{1/2}\right\},$$

$$\left(\frac{\beta_j}{\sigma}\right)^L = \left\{\frac{\hat{\beta}_j}{\hat{\sigma}}\left[1 - \frac{z_{\alpha/2}^2}{2(n-p)}\right] - z_{\alpha/2}\left[\delta_j\left(1 + \left(\frac{\left(\hat{\beta}_j^2/\hat{\sigma}^2\delta_j\right) - z_{\alpha/2}^2}{2(n-p)}\right)\right)\right]^{1/2}\right\},$$

Thus, we obtain an approximate $100(1-\alpha)$ percent confidence interval for ω_j

$$\left\{\exp\left[\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^L\right], \exp\left[\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^U\right]\right\}. \tag{14}$$

3. Comparison of the Two Methods

Let Y_i be a continuous outcome variable. For fixed value of C, we define Y_i^* such that

$$Y_i^* = \begin{cases}1 & \text{if } Y_i \geq C, \\ 0 & \text{if } Y_i < C.\end{cases} \tag{15}$$

Suppose that $Y_1^*, \ldots, Y_n^*$ form a random sample of observations, and we fit a complementary log-log model

$$\begin{aligned} \pi_{i1} &= P(Y_i^* = 1 \mid x_i) = \exp(-\exp(x_i\theta)), \\ \pi_{i2} &= P(Y_i^* = 1 \mid x_{(-1,i)}) = \exp(-\exp(x_{(-1,i)}\theta)), \end{aligned} \tag{16}$$

where $x_i = (1, x_{i1}, \ldots, x_{i,p-1})'$ is the $P \times 1$ vector of covariates for the ith observation, and $\theta = (\theta_0, \ldots, \theta_{p-1})'$ is the $P \times 1$ vector of unknown parameters. The dichotomized ω_j^* parameter corresponding to the effect θ_j is

$$\begin{aligned} \omega_j^* &= \frac{\ln(\pi_1)}{\ln(\pi_2)} \\ &= \frac{\ln(P(Y^* = 1 \mid x))}{\ln\left(P\left(Y^* = 1 \mid x_{(-1,j)}\right)\right)} \\ &= \frac{(\exp(x\theta))}{\left(\exp\left(x_{(-1,j)}\theta\right)\right)} \\ &= \exp\left(\theta_j\right) \quad j = 0, \ldots, p-1. \end{aligned} \tag{17}$$

In general, maximum likelihood estimation (MLE) can be used to estimate the parameter $\theta = (\theta_0, \ldots, \theta_{p-1})$. Let $\hat{\theta} = (\hat{\theta}_0, \ldots, \hat{\theta}_{p-1})'$ be the $P \times 1$ ML estimate of θ, and let $\mathrm{COV}(\hat{\theta})$ be the $P \times P$ covariance matrix of $\hat{\theta}$. Using $\mathrm{COV}(\hat{\theta})$ from (23), one can construct confidence intervals. This matrix has as its diagonal the estimated variances of each of the ML estimates. The $(j+1)$th diagonal element is given by $\sigma^2_{\hat{\theta}_j}$. Therefore,

$$\hat{\omega}_j^* = \exp\left(\hat{\theta}_j\right), \tag{18}$$

and for large samples, $(\hat{\theta}_j^L, \hat{\theta}_j^U) = (\hat{\theta}_j - z_{\alpha/2}\hat{\sigma}_{\hat{\theta}_j}, \hat{\theta}_j + z_{\alpha/2}\hat{\sigma}_{\hat{\theta}_j})$ is a $100(1-\alpha)$ percent confidence interval for the true θ_j. Then $(\exp(\hat{\theta}_j^L), \exp(\hat{\theta}_j^U))$ is a $100(1-\alpha)$ percent confidence interval for the true ω_j^*.

We now compare the ω_j from (7) with the ω_j^* from (17)

$$\begin{aligned} \omega_j &= \frac{\ln(\pi_1)}{\ln(\pi_2)} \\ \omega_j^* &= \frac{\ln(\pi_1)}{\ln(\pi_2)} \end{aligned} \Longrightarrow \omega_j^* = \omega_j$$

$$\Longrightarrow \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right) = \exp\left(\theta_j\right)$$

$$\Longrightarrow \frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma} = \theta_j \quad \forall \beta_j, \theta_j, \sigma. \tag{19}$$

This show that the coefficient of the complementary log-log model, θ_j, can be interpreted in terms of the regression coefficients, β_j. Note that β are related to the responses through the general linear regression model

$$y_i = x_i\beta + E_i \quad i = 1, \ldots, n, \tag{20}$$

where the independent E_i are distributed as an extreme value with mean 0 and variance $\sigma^2 > 0$.

4. Covariance Matrix of Model Parameter Estimators

4.1. Derivation of $\mathrm{var}(\omega_j^*)$ for Large n. The information matrix of generalized linear models has the form $\int = X'WX$ [1], where W is the diagonal matrix with diagonal elements $w_i = (\partial\mu_i/\partial\eta_i)^2/(\mathrm{var}(y_i))$, y is response variable with independent observations $(y_1, \ldots y_n)$, and x_{ij} denote the value of predictor j,

$$\mu_i = E(y_i), \qquad \eta_i = g(\mu_i) = \sum_j \theta_j x_{ij}, \quad j = 0, 1, \ldots, p-1. \tag{21}$$

The covariance matrix of $\hat{\theta}$ is estimated by $(X'\widehat{W}X)^{-1}$.

Maximum likelihood estimation for the complementary log-log model is a special case of the generalized linear models. Let

$$\begin{aligned} \mu_i = \pi_i &= \exp\left(-\exp\left(\sum_j \theta_j x_{ij}\right)\right) \\ &\Longrightarrow \pi_i = \exp(-\exp(\eta_i)), \\ \frac{\partial\mu_i}{\partial\eta_i} &= (-\exp(\eta_i))' \exp(-\exp(\eta_i)) = \pi_i \ln\pi_i, \\ w_i &= \frac{(\pi_I \ln\pi_i)^2}{\pi_i(1-\pi_i)} = \frac{\pi_i(\ln\pi_i)^2}{1-\pi_i}, \end{aligned} \tag{22}$$

then

$$X'WX = \begin{bmatrix} \sum_{i=1}^n \frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \sum_{i=1}^n x_{i1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \cdots & \sum_{i=1}^n x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} \\ \sum_{i=1}^n x_{i1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \sum_{i=1}^n x_{i1}^2\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \cdots & \sum_{i=1}^n x_1 x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} \\ \vdots & & \vdots & \\ \sum_{i=1}^n x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \sum_{i=1}^n x_{1i}x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \cdots & \sum_{i=1}^n x_{i,p-1}^2\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} \end{bmatrix}. \tag{23}$$

It is readily shown that the results hold true for the largest extreme value distribution (Appendix A).

In large samples, $\operatorname{var}(\hat{\theta}_j)$ approaches $\sigma^2_{\theta_j}|_{\theta=\hat{\theta}}$ [14] which equals the $(j+1)$th diagonal element of $(X'WX)^{-1}$.

By applying the delta method, let $f(\hat{\theta}_j) = \exp(\hat{\theta}_j)$, then

$$\begin{aligned}\operatorname{var}\left(\hat{\omega}_j^*\right) \longrightarrow \operatorname{var}\left(\exp\left(\hat{\theta}_j\right)\right) &= \operatorname{var}\left(f\left(\hat{\theta}_j\right)\right)\\ &= \left(\frac{\partial f\left(\hat{\theta}_j\right)}{\partial \hat{\theta}_j}\Big|_{\hat{\theta}_j=\theta_j}\right)^2 \left(\operatorname{var}\left(\hat{\theta}_j\right)\right)\\ &= \left(\exp\left(\theta_j\right)\right)^2 \times \sigma^2_{\hat{\theta}_j}.\end{aligned} \tag{24}$$

4.2. Derivation of $var(\hat{\omega}_j)$ for Large n. In large samples, from (10) $\hat{\sigma}^2 \to \sigma^2$ [15]. Therefore,

$$\operatorname{var}\left(\hat{\omega}_j\right) = \operatorname{var}\left(\exp\left(\frac{\pi\hat{\beta}_j}{\hat{\sigma}\sqrt{6}}\right)\right) \longrightarrow \operatorname{var}\left(\exp\left(\frac{\pi\hat{\beta}_j}{\sigma\sqrt{6}}\right)\right). \tag{25}$$

In addition, $\operatorname{var}(\hat{\beta}_j) = \sigma^2\delta_j$.

By applying the delta method, let $g(\hat{\beta}_j) = \exp(\pi\hat{\beta}_j/(\sigma\sqrt{6}))$, then

$$\begin{aligned}\operatorname{var}\left(\hat{\omega}_j\right) &\longrightarrow \operatorname{var}\left(\exp\left(\frac{\pi\hat{\beta}_j}{\sigma\sqrt{6}}\right)\right)\\ &= \operatorname{var}\left(g\left(\hat{\beta}_j\right)\right)\\ &= \left(\frac{\partial g\left(\hat{\beta}_j\right)}{\partial \hat{\beta}_j}\Big|_{\hat{\beta}_j=\beta_j}\right)^2 \times \operatorname{var}\left(\hat{\beta}_j\right)\\ &= \left(\frac{\pi}{\sigma\sqrt{6}}\exp\left(\frac{\pi\beta_j}{\sigma\sqrt{6}}\right)\right)^2 \sigma^2\delta_j\\ &= \frac{\pi^2}{\sqrt{6}}\delta_j\left(\exp\frac{\pi\beta_j}{\sigma\sqrt{6}}\right)^2.\end{aligned} \tag{26}$$

5. Sample Sizes Saving

5.1. The Power for the Dichotomized Method. In large samples, $\hat{\sigma}_{\hat{\theta}_j}$ converges to $\sigma_{\hat{\theta}_j}$ almost surely [14]. Therefore, for a given value of $\omega_j = \exp\theta_j$ (i.e., $\ln\omega_j = \theta_j$), the power is given by

$$\begin{aligned}p\left(\omega_j\right) &= p\left\{\text{rejection of } \omega_j = 1 \mid \omega_j \neq 1\right\}\\ &= p\left\{\exp\left(\theta_j^L\right) > 1 \mid \theta_j\right\} + p\left\{\exp\left(\theta_j^U\right) < 1 \mid \theta_j\right\}\\ &= p\left\{\hat{\theta}_j > z_{\alpha/2}\sigma_{\hat{\theta}_j} \mid \theta_j\right\} + p\left\{\hat{\theta}_j < -z_{\alpha/2}\sigma_{\hat{\theta}_j} \mid \theta_j\right\}\\ &= p\left\{Z > \frac{z_{\alpha/2}\sigma_{\hat{\theta}_j} - \ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\}\\ &\quad + p\left\{Z < \frac{-z_{\alpha/2}\sigma_{\hat{\theta}_j} - \ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\}\\ &= p\left\{Z > z_{\alpha/2} - \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\} + p\left\{Z < -z_{\alpha/2} - \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\}\\ &= P\{Z > z_1^\bullet\} + P\{Z < -z_2^\bullet\},\end{aligned} \tag{27}$$

where

$$\left\{\begin{aligned}z_1^\bullet &= z_{\alpha/2} - \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\\ z_2^\bullet &= z_{\alpha/2} + \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\end{aligned}\right\}. \tag{28}$$

5.2. The Power for the Proposed Method. In large samples, $\hat{\sigma}$ converges to σ almost surely [15]. Therefore, for a given value of $\omega_j = \exp(\pi\beta_j/\sigma\sqrt{6})$ (i.e., $\beta_j = \sigma(\ln\omega_j\sqrt{6}/\pi)$), the power is given by

$$\begin{aligned}p\left(\omega_j\right) &= p\left\{\exp\left(\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^L\right) > 1 \mid \omega_j\right\}\\ &\quad + p\left\{\exp\left(\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^U\right) < 1 \mid \omega_j\right\}\\ &= P\left\{\beta_J^L > z_{\alpha/2}\sigma\sqrt{\delta_j} \mid \beta_j = \frac{\sigma\ln\omega_j\sqrt{6}}{\pi}\right\}\\ &\quad + P\left\{\beta_J^U < -z_{\alpha/2}\sigma\sqrt{\delta_j} \mid \beta_j = \frac{\sigma\ln\omega_j\sqrt{6}}{\pi}\right\}\\ &= p\left\{Z > \frac{z_{\alpha/2}\sigma\sqrt{\delta_j} - \left(\sigma\ln\omega_j\sqrt{6}/\pi\right)}{\sigma\sqrt{\delta_j}}\right\}\\ &\quad + p\left\{Z < \frac{-z_{\alpha/2}\sigma\sqrt{\delta_j} - \left(\sigma\ln\omega_j\sqrt{6}/\pi\right)}{\sigma\sqrt{\delta_j}}\right\}\end{aligned}$$

$$= p\left\{Z > z_{\alpha/2} - \frac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}}\right\} + p\left\{Z < -z_{\alpha/2} - \frac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}}\right\} = p\{Z > z_1\} + p\{Z < -z_2\}, \tag{29}$$

where

$$\left\{\begin{array}{l} z_1 = z_{\alpha/2} - \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \\ z_2 = z_{\alpha/2} + \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \end{array}\right\}. \tag{30}$$

Our proposed method, since it is based on continuous data rather than dichotomized, is likely to be more powerful. We show that the proposed method can produce substantial sample size saving for a given power. Let

(i) the number of parameters $p = 2$ (i.e., $\theta = (\theta_0, \theta_1)$),

(ii) $x_i = (1, x_{i1})', x_{i1} \in \{-a+(2an/(g-1)) \mid n = 0, \ldots, g-1\}$, that is, x_{i1} follows a discrete uniform distribution with range $(-a,\ a)$. For simplicity, $a = 2$.

(iii) Total samples are n and n^* for the proposed and dichotomized methods, respectively. These samples included k and k^* set of these g uniformly distributed points for the proposed and dichotomized methods, respectively. That is, $n = gk$ and $n^* = gk^*$, then

$$\delta_j = \left[k\sum_{i=1}^{g}(x_{1i} - \bar{x}_{1.})^2\right]^{-1}, \quad j = 1, \tag{31}$$

and from (23),

$$\sigma^2_{\hat{\theta}_j} = \frac{\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)}{(k^*)\left\{\sum_{i=1}^{g} x_{1i}^2\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right) - \left[\sum_{i=1}^{g} x_{1i}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\right]^2\right\}}. \tag{32}$$

We consider the same power for two methods:

$$\begin{array}{l} z_1 = z_1^* \\ z_2 = z_2^* \end{array} \Longrightarrow \left\{\begin{array}{l} z_{\alpha/2} - \dfrac{\ln \omega_j}{\sigma_{\hat{\theta}_j}} = z_{\alpha/2} - \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \\ z_{\alpha/2} + \dfrac{\ln \omega_j}{\sigma_{\hat{\theta}_j}} = z_{\alpha/2} + \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \end{array}\right. \Longrightarrow \frac{\pi}{\sqrt{6}}\sqrt{\delta_j} = \sigma_{\hat{\theta}_j}, \quad j = 1 \Longrightarrow \frac{\pi}{\sqrt{6}}\sqrt{\left[k\sum_{i=1}^{g}(x_{1i} - \bar{x}_{1.})^2\right]^{-1}} \tag{33}$$

$$= \sqrt{\frac{\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)}{(k^*)\left\{\sum_{i=1}^{g} x_{1i}^2\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right) - \left[\sum_{i=1}^{g} x_{1i}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\right]^2\right\}}}$$

relative sample size

$$\frac{n^*}{n} = \frac{k^*}{k} = \frac{6\sigma^2_{\hat{\theta}_j}}{\pi^2\delta_j} = \frac{\sum_{i=1}^{g}(x_{1i} - \bar{x}_{1.})^2 \times \sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)}{(\pi^2/6)\left\{\sum_{i=1}^{g} x_{1i}^2\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right) - \left[\sum_{i=1}^{g} x_{1i}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\right]^2\right\}}. \tag{34}$$

Table 1: Relative sample sizes required to attain any power for the dichotomizing method versus the proposed method.

$\omega^* = \exp(\theta)$	Average proportion of successes ($\overline{\pi}$)				
	0.1	0.2	0.3	0.4	0.5
0.25	23.7166	9.5092	7.4954	7.1996	6.8575
0.50	10.6719	5.4176	3.4215	2.5209	2.1784
0.75	7.7088	3.8713	2.5171	1.9380	1.5841

That is, (34) is independent of σ^2 and applies for any power, and any test size α.

Table 1 presents relative sample sizes n^*/n for a given fixed parameter ω_j^* and an average proportion of success $\overline{\pi}$. We consider the situations in which $\overline{\pi} = \sum_{i=1}^{g}(\pi_i/g) = 0.1, 0.2, 0.3, 0.4, 0.5$, $g = 9$, $\omega_j^* = 0.25, 0.50, 0.75$.

For given fixed ω_j^* and $\overline{\pi}$, the relative sample sizes in Table 1 can be computed by the following step:

(i) compute the value θ_j via the equation $\theta_j = \ln(\omega_j^*)$,

(ii) calculate the cut-off point C iteratively such that $\overline{\pi}$ attained the specified value for the values x_{i1}, using the value of θ_j in (i).

As can be seen from Table 1, all values are greater than 1. The values of n^*/n increase as the ω_j^* moves farther away from 1. Values of Table 1 immediately highlight the improvement accomplished by the proposed method.

6. Relative Efficiency of $\widehat{\omega}_j$ with $\widehat{\omega}_j^*$

Here, we examine the relative efficiency of the estimate $\widehat{\omega}_j$ to the estimate $\widehat{\omega}_j^*$.

Using (24) and (26), the relative efficiency is given by

$$\begin{aligned}\text{r.e.}\left(\widehat{\omega}_j, \widehat{\omega}_j^*\right) &= \frac{\text{var}\left(\widehat{\omega}_j^*\right)}{\text{var}\left(\widehat{\omega}_j\right)} \\ &= \frac{6\left(\exp\left(\theta_j\right)\right)^2 \times \sigma_{\widehat{\theta}_j}^2}{\pi^2\delta_j\left(\exp\left(\lambda\beta_j/\sigma\right)\right)^2} = \frac{6\sigma_{\widehat{\theta}_j}^2}{\pi^2\delta_j}.\end{aligned} \tag{35}$$

Note that the relative efficiency is independent of n and σ^2 and converges to a constant. Comparing (34) and (35), the relative efficiency equals the relative sample sizes. Therefore, as in Table 1, the proposed method is a consistent improvement over the dichotomizing method with respect to relative efficiencies.

It should be noted that these results hold true under the following assumptions:

(1) the responses y_i and β are related through the equation $y_i = x_i\beta + E_i$ where the independent E_i are distributed as an extreme value with mean 0 and variance $\sigma^2 > 0$,

(2) the independent variables x_i follow a discrete uniform distribution.

7. Odds Ratio

For values of π larger than 0.90, $-\ln(\pi)$ and $\pi/(1-\pi)$ are very close. Hence, for large values of π,

$$\frac{\ln(\pi_1)}{\ln(\pi_2)} \cong \frac{\pi_1/1-\pi_1}{\pi_2/1-\pi_2} = \text{OR}. \tag{36}$$

And from (7), odds ratio is given by

$$\text{OR} = \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right). \tag{37}$$

The parameters estimated from the linear regression can be interpreted as an odds ratio.

8. Simulation Study

It should be noted that, as in Table 1, the proposed method is a consistent improvement over the dichotomizing method with respect to relative efficiencies. These results hold true under the assumption that predictor variable has a discrete uniform distribution and that the random variables E_i follow an extreme value distribution. To demonstrate the robustness of this conclusion to changes in the distributions of predictor variables, simulations were run under different distributional conditions. The data were sampled 10000 times for three sample sizes $\{n = 250, 500, 1000\}$, three average proportions of successes $\{\overline{\pi} = 0.10, 0.50, 0.95\}$, and seven $\omega_j\{\omega_j = 0.75, 0.90, 1.1, 1.2, 1.3, 1.4, 1.5\}$. The simulated data are generated using the following algorithm

(1) Generate y_i, where $y_i = \beta_0 + \beta_1 x_i + E_i$, $\beta_1 = \sqrt{6}\ln\omega_j/\pi$ through (7) to produce the correct ω_j, and for simplicity $\beta_0 = 0$, $\sigma^2 = 1$.

(2) For fixed $\overline{\pi}$, generate cutoff point C using (15).

We simulated the data for two scenarios based on the distribution of the explanatory variable. In the first scenario, the independent variable follows a continuous uniform distribution and range $(-2, 2)$, and in the second, the independent variable follows a truncated normal distribution with mean 0 and range $(-2, 2)$. The relative mean square errors, relative interval lengths, absolute biases, and the probability of coverage were calculated.

Results of the simulations addressing the validity of the proposed method are displayed in Tables 2 and 3.

The simulations show that the relative mean square errors are all greater than 1, increasing with the average proportion of successes and when the ω_j moves farther away

Table 2: Simulated relative mean square errors, relative intervals lengths, coverage probabilities, and absolute biases for the proposed and dichotomizing methods (using a continuous uniform distribution for the explanatory variable and an extreme value distribution for the errors).

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.15[a]	1.07	1.09	1.14	1.24	1.47	1.71
		1.10[b]	1.03	1.03	1.07	1.14	1.23	1.35
	0.10	0.943[c]	0.948	0.949	0.949	0.945	0.938	0.933
		0.948[d]	0.947	0.949	0.947	0.951	0.947	0.953
		0.05[e]	0.04	0.12	0.14	0.10	0.15	0.11
		0.07[f]	0.01	0.17	0.13	0.24	0.34	0.58
		1.23	1.26	1.27	1.28	1.27	1.24	1.26
		2.16	1.13	1.23	1.14	1.15	1.17	1.19
1000	0.50	0.940	0.951	0.951	0.945	0.942	0.937	0.934
		0.951	0.949	0.951	0.950	0.948	0.947	0.948
		0.04	0.01	0.08	0.10	0.05	0.09	0.04
		0.05	0.04	0.15	0.12	0.09	0.12	0.13
		12.75	12.44	13.22	12.68	13.14	12.91	12.79
		3.67	3.57	3.58	3.63	3.69	3.76	3.84
	0.95	0.943	0.951	0.952	0.944	0.944	0.938	0.929
		0.952	0.954	0.952	0.952	0.951	0.951	0.951
		0.04	0.07	0.11	0.10	0.10	0.17	0.10
		0.75	0.68	0.86	1.01	1.21	1.45	1.24
		1.30	1.08	1.07	1.17	1.24	1.54	1.95
		1.16	1.03	1.04	1.08	1.15	1.25	1.39
	0.10	0.942	0.950	0.951	0.95	0.944	0.941	0.936
		0.951	0.950	0.949	0.951	0.954	0.954	0.953
		0.12	0.07	0.24	0.25	0.21	0.18	0.29
		0.23	0.08	0.33	0.39	0.41	0.73	1.21
		1.35	1.10	1.27	1.26	1.26	1.25	1.26
		1.26	1.03	1.13	1.14	1.16	1.17	1.20
500	0.50	0.940	0.949	0.947	0.948	0.943	0.940	0.933
		0.952	0.951	0.949	0.949	0.954	0.950	0.951
		0.23	0.34	0.27	0.23	0.26	0.25	0.38
		0.48	0.11	0.17	0.18	0.31	0.26	0.42
		13.04	13.17	13.8	13.90	14.45	14.48	14.47
		3.72	3.65	3.68	3.73	3.82	3.91	3.99
	0.95	0.942	0.947	0.951	0.949	0.947	0.938	0.935
		0.953	0.952	0.954	0.955	0.955	0.953	0.954
		0.05	0.11	0.08	0.08	0.24	0.32	0.27
		0.94	1.38	1.78	1.92	2.52	3.00	2.90
		13.41	14.46	1.12	1.28	1.52	1.96	2.33
		3.78	3.73	1.04	1.09	1.18	1.30	1.45
	0.10	0.942	0.949	0.949	0.945	0.942	0.942	0.933
		0.957	0.954	0.948	0.949	0.952	0.957	0.953
		0.02	0.20	0.38	0.33	0.42	0.41	0.66
		2.11	2.74	0.42	0.84	1.18	1.78	2.24

Table 2: Continued.

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.27	1.25	1.32	1.28	1.30	1.30	1.29
		1.16	1.13	1.13	1.14	1.16	1.18	1.20
250	0.50	0.941	0.948	0.952	0.947	0.945	0.943	0.933
		0.951	0.951	0.951	0.950	0.951	0.951	0.951
		0.12	0.13	0.35	0.44	0.41	0.53	0.55
		0.11	0.22	0.39	0.47	0.51	0.74	0.59
		12.98	14.6	15.64	15.46	17.05	16.89	18.33
		3.75	3.72	3.82	3.88	4.01	4.12	4.29
	0.95	0.945	0.955	0.946	0.948	0.940	0.937	0.932
		0.959	0.955	0.955	0.959	0.958	0.957	0.952
		0.02	0.16	0.39	0.22	0.46	0.47	0.51
		1.22	2.75	3.97	3.98	4.99	5.19	6.19

a: Relative mean square errors, b: Relative intervals lengths, c: Coverage probability (proposed), d: Coverage probability (dichotomized), e: % bias (proposed), f: % bias (dichotomized).

from 1. The results in Tables 1 and 2 demonstrate that the proposed method provides confidence intervals which successfully maintain their nominal 95 percent coverage. For the proposed method in first scenario, 51 out of 63 coverage probabilities fell within (0.94, 0.96), and all 63 coverage probabilities are greater than 0.93 and, in the second scenario, almost all coverage probabilities fell within (0.94, 0.96). The absolute biases for proposed method are never greater than a few percent. The proposed method is less biased than the dichotomizing method in 6 of 63 simulations in both two scenarios.

9. An Example

To illustrate the application of the proposed method presented in the previous section, we utilize the data arising from the National Health Survey in Iran. The other analyses using this data appear in many places [16].

In this study, 14176 women aged 20–69 years were investigated. BMI (body mass index), our dependent variable, was calculated as weight in kilograms divided by height in meters squared (kg/m^2). Independent variables included place of residence, age, smoking, economic index, marital status, and education level. The independent variables considered were both categorical and continuous. At first, BMI was treated as a continuous variable, and $\widehat{\omega}_j$ and 95 percent confidence intervals were calculated using the proposed linear regression method. Then subjects were classified into obese (BMI $\geq$ 30 kg/m^2) and nonobese (BMI <30 kg/m^2). A complementary log-log model was used for the binary analysis, with obese or nonobese used as the outcome measure. The $\widehat{\omega}_j^*$ and 95 percent confidence intervals were calculated using the dichotomized method. Table 4 presents the coefficient estimates, estimated confidence intervals, and relative confidence interval lengths. The proposed and dichotomizing methods produced different confidence intervals, although the $\widehat{\omega}_j$ and $\widehat{\omega}_j^*$ were similar only varying slightly. The $\widehat{\omega}_j$ estimate from the proposed method had smaller variances and shorter confidence intervals than the dichotomizing method. All relative confidence interval lengths were greater than 2.58.

10. Discussion

When assuming the errors E_i are distributed as an extreme value distribution, as noted before, the method has several advantages. First, the method allows the researcher to apply the complementary log-log model without dichotomizing and without loss of information. Second, the $\widehat{\omega}_j^*$ from the dichotomizing method is dependent on the chosen cutoff point C and will vary with c. However, the proposed $\widehat{\omega}_j$ is independent of the c since $\widehat{\omega}_j$ is a function of the continuous Y_i and not a function of the dichotomized Y_i^* defined through C. Third, we show that the coefficient of the complementary log-log model, θ_j, can be interpreted in terms of the regression coefficients, β_j. Fourth, when the independent variables x_i follow a discrete uniform distribution, the proposed method is a consistent improvement over the dichotomizing method with respect to relative efficiencies. The proposed method can provide sample size saving, smaller variances, and shorter confidence intervals than the dichotomized method. Fifth, when π is large, the parameters estimated from the linear regression can be interpreted as odds ratios.

Our results were consistent with the findings by Moser and Coombs [12] and Bakhshi et al. [16] showing the greater efficiency of parameter estimates from the regression method that avoids dichotomizing in comparison with a more traditional dichotomizing method using the logistic regression.

Our main recommendation is to let continuous response remain continuous. Do not throw away information by transforming the data to binary. This means that if the objective is to estimate and/or test coefficients when responses

Table 3: Simulated relative mean square errors, relative intervals lengths, coverage probabilities, and absolute biases for the proposed and dichotomizing methods (using a truncated normal distribution for the explanatory variable and an extreme value distribution for the errors).

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.17[a]	1.02	1.08	1.13	1.19	1.28	1.36
		1.11[b]	1.03	1.03	1.06	1.10	1.25	1.22
	0.10	0.942[c]	0.948	0.948	0.952	0.944	0.942	0.940
		0.951[d]	0.951	0.950	0.952	0.949	0.951	0.951
		0.08[e]	0.06	0.03	0.14	0.13	0.14	0.16
		0.10[f]	0.11	0.15	0.23	0.30	0.39	0.39
		1.26	1.24	1.26	1.28	1.28	1.25	1.28
		1.24	1.13	1.13	1.14	1.14	1.15	1.17
1000	0.50	0.944	0.948	0.952	0.947	0.947	0.944	0.941
		0.948	0.951	0.949	0.949	0.947	0.950	0.949
		0.02	0.09	0.08	0.07	0.18	0.16	0.13
		0.03	0.06	0.12	0.16	0.20	0.16	0.14
		12.33	13.12	13.03	12.71	12.86	12.55	12.88
		3.62	3.59	3.61	3.62	3.64	3.68	3.71
	0.95	0.944	0.951	0.948	0.948	0.945	0.945	0.946
		0.952	0.948	0.95	0.949	0.949	0.951	0.952
		0.10	0.04	0.11	0.04	0.16	0.16	0.20
		1.26	1.05	1.56	1.36	1.43	1.80	1.94
		1.18	1.09	1.06	1.75	1.23	1.32	1.58
		1.11	1.03	1.03	1.06	1.11	1.16	1.23
	0.10	0.945	0.95	0.951	0.951	0.949	0.943	0.944
		0.953	0.953	0.953	0.950	0.949	0.951	0.950
		0.04	0.13	0.31	0.18	0.33	0.36	0.37
		0.21	0.08	0.37	0.50	0.62	0.69	0.96
		1.25	1.27	1.27	1.29	1.27	1.29	1.25
		1.14	1.13	1.13	1.14	1.15	1.16	1.17
500	0.50	0.944	0.948	0.949	0.947	0.948	0.944	0.935
		0.951	0.951	0.951	0.948	0.951	0.948	0.949
		0.13	0.22	0.35	0.37	0.35	0.30	0.44
		0.16	0.19	0.39	0.48	0.44	0.41	0.54
		13.11	14.02	14.02	13.5	13.54	13.80	14.32
		3.73	3.71	3.73	3.75	3.77	3.81	3.86
	0.95	0.944	0.95	0.951	0.950	0.947	0.944	0.944
		0.954	0.95	0.951	0.953	0.948	0.956	0.953
		0.15	0.10	0.24	0.38	0.32	0.33	0.43
		2.50	2.70	2.92	3.10	2.92	3.36	3.89
		1.28	1.11	1.12	1.19	1.33	1.54	1.76
		1.11	1.03	1.04	1.08	1.13	1.19	1.28
	0.10	0.947	0.951	0.950	0.947	0.950	0.950	0.942
		0.951	0.950	0.950	0.952	0.954	0.952	0.951
		0.40	0.34	0.37	0.64	0.69	0.58	0.81
		0.26	0.06	0.69	1.08	1.30	1.55	2.22

Table 3: Continued.

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.32	1.30	1.27	1.33	1.31	1.33	1.31
		1.15	1.13	1.13	1.14	1.18	1.17	1.18
250	0.50	0.951	0.95	0.953	0.951	0.940	0.945	0.940
		0.949	0.951	0.952	0.948	0.948	0.950	0.948
		0.22	0.43	0.57	0.69	0.66	0.58	0.66
		0.38	0.53	0.64	0.89	0.91	0.82	0.91
		14.09	14.51	16.27	15.91	15.89	15.73	15.60
		3.86	3.87	3.93	3.92	3.98	4.04	4.11
	0.95	0.943	0.95	0.951	0.951	0.947	0.944	0.937
		0.953	0.95	0.953	0.956	0.953	0.956	0.952
		0.30	0.37	0.57	0.68	0.42	0.62	0.75
		4.98	5.52	6.547	5.91	6.17	6.88	7.72

a: Relative mean square errors, b: Relative intervals lengths, c: Coverage probability (proposed), d: Coverage probability (dichotomized), e: % bias (proposed), f: % bias (dichotomized).

Table 4: Adjusted $\widehat{\omega}_j^*$, $\widehat{\omega}_j$ for obesity and confidence intervals using two methods for the National Health Survey.

Covariates	$\widehat{\omega}_j(\widehat{\omega}_j^*)$	95% CI[a] (proposed)	95% CI (dichotomized)	Relative[b] length of CI
Place of residence	1.65 (1.97)[c]	1.58–1.74	1.79–2.18	2.43
Age	1.021 (1.019)	1.018–1.022	1.015–1.022	1.75
Years of education	0.99 (0.98)	0.985–0.997	0.971–0.994	1.92
Smoking	0.76 (0.68)	0.66–0.90	0.51–0.92	1.71
Marital status	1.16 (1.42)	1.10–1.22	1.27–1.58	2.58
Lower-middle economy index	1.24 (1.32)	1.14–1.32	1.18–1.48	1.67
Upper-middle economy index	1.21 (1.26)	1.14–1.29	1.12–1.42	2.0
High economy index	1.20 (1.21)	1.11–1.30	1.08–1.36	1.47

[a]Confidence interval, [b]dichotomized/proposed, [c]proposed (dichotomized).

are continuous, please resist dichotomizing your response variable.

Appendix

A. Largest Extreme Value Distribution

(a) *The PDF and CDF are Given by*

$$
\begin{aligned}
f(y \mid x\beta, \sigma) &= \frac{\pi}{\sigma\sqrt{6}} \\
&\quad \times \exp\left(-\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right. \\
&\quad \left. - \exp\left(\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\
&\quad - \infty \langle x \langle \infty, \sigma \rangle 0, \\
P(y \le c) &= 1 - \exp\left(-\exp\left(-\frac{c - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\
&\quad - \infty \langle x \langle \infty, \sigma \rangle 0,
\end{aligned}
\tag{A.1}
$$

where Y is a continuous outcome variable, $x = (1, x_1, \ldots, x_{p-1})$ is the $p \times 1$ vector of known independent variables, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is the $p \times 1$ vector of unknown parameters, and $k \approx 0.45$.

It is easy to check that

$$
\begin{aligned}
\omega_j &= \frac{\ln(1 - \pi_1)}{\ln(1 - \pi_2)} = \frac{\ln(1 - p(y \le c \mid x))}{\ln\left(1 - p\left(y \le c \mid x_{(-1,j)}\right)\right)} \\
&= \frac{-\exp(-((c - x'\beta - k\sigma)/\sigma) \times (\pi/\sqrt{6}))}{-\exp\left(-\left(\left(c - x'_{(-1,j)}\beta - k\sigma\right)/\sigma\right) \times (\pi/\sqrt{6})\right)} \\
&= \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right) \Longrightarrow 1 - \pi_1 \\
&= (1 - \pi_2)^{\exp((\pi/\sqrt{6}) \cdot (\beta_j/\sigma))},
\end{aligned}
\tag{A.2}
$$

where

$$
\begin{aligned}
x &= \left(1, x_1, \ldots, x_j, \ldots, x_{p-1}\right), \\
x_{(-1,j)} &= \left(1, x_1, \ldots, x_j - 1, \ldots, x_{p-1}\right), \\
\beta &= \left(\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}\right)'.
\end{aligned}
\tag{A.3}
$$

(b) *Suppose that E_i is distributed as a largest extreme value with mean 0 and variance $\sigma^2 > 0$.* We conclude that the PDF and CDF of each independent Y_i are given by (A.1), and the corresponding equality (A.2) is given by

$$\widehat{\omega}_j = \frac{\ln(1-\widehat{\pi}_1)}{\ln(1-\widehat{\pi}_2)} = \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\widehat{\beta}_j}{\widehat{\sigma}}\right). \tag{A.4}$$

(c) *Similar to largest extreme value distribution*

$$\begin{aligned} \mu_i = \pi_i &= 1 - \exp\left(-\exp\left(\sum_j \theta_j x_{ij}\right)\right) \\ \Longrightarrow \pi_i &= 1 - \exp(-\exp(\eta_i)), \end{aligned} \tag{A.5}$$

then

$$\begin{aligned} \frac{\partial \mu_i}{\partial \eta_i} &= -(-\exp(\eta_i))' \exp(-\exp(\eta_i)) \\ &= -(1-\pi_i)\ln(1-\pi_i) \\ w_i &= \frac{((1-\pi_i)\ln(1-\pi_i))^2}{\pi_i(1-\pi_i)} \\ &= \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i}, \end{aligned}$$

$$X'WX = \begin{bmatrix} \sum_{i=1}^{n} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \sum_{i=1}^{n} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \cdots & \sum_{i=1}^{n} x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} \\ \sum_{i=1}^{n} x_{i1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \sum_{i=1}^{n} x_{i1}^2 \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \cdots & \sum_{i=1}^{n} x_1 x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} \\ \vdots & & & \vdots \\ \sum_{i=1}^{n} x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \sum_{i=1}^{n} x_i x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \cdots & \sum_{i=1}^{n} x_{i,p-1}^2 \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} \end{bmatrix}. \tag{A.6}$$

Conflict of Interests

The authors have declared no conflict of interests.

References

[1] A. Agresti, *Categorical Data Analysis*, Wiley, New York, NY, USA, 2nd edition, 2002.

[2] L. P. Zhao and L. N. Kolonel, "Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies," *American Journal of Epidemiology*, vol. 136, no. 4, pp. 464–474, 1992.

[3] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, "On the practice of dichotomization of quantitative variables," *Psychological Methods*, vol. 7, no. 1, pp. 19–40, 2002.

[4] J. Cohen, "The cost of dichotomization," *Applied Psychological Measurement*, vol. 7, no. 3, pp. 249–253, 1983.

[5] S. Greenland, "Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis.," *Epidemiology*, vol. 6, no. 4, pp. 450–454, 1995.

[6] P. C. Austin and L. J. Brunner, "Inflation of the type I error rate when a continuous confounding variable is categorized in logistics regression analyses," *Statistics in Medicine*, vol. 23, no. 7, pp. 1159–1178, 2004.

[7] A. Vargha, T. Rudas, H. D. Delaney, and S. E. Maxwell, "Dichotomization, partial correlation, and conditional independence," *Journal of Educational and Behavioral Statistics*, vol. 21, no. 3, pp. 264–282, 1996.

[8] S. E. Maxwell and H. D. Delaney, "Bivariate median splits and spurious statistical significance," *Psychological Bulletin*, vol. 113, no. 1, pp. 181–190, 1993.

[9] D. L. Streiner, "Breaking up is hard to do: the heartbreak of dichotomizing continuous data," *Canadian Journal of Psychiatry*, vol. 47, no. 3, pp. 262–266, 2002.

[10] H. Chen, P. Cohen, and S. Chen, "Biased odds ratios from dichotomization of age," *Statistics in Medicine*, vol. 26, no. 18, pp. 3487–3497, 2007.

[11] D. R. Ragland, "Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint," *Epidemiology*, vol. 3, no. 5, pp. 434–440, 1992.

[12] B. K. Moser and L. P. Coombs, "Odds ratios for a continuous outcome variable without dichotomizing," *Statistics in Medicine*, vol. 23, no. 12, pp. 1843–1860, 2004.

[13] N. L. Johnson, H. Welch, and C. Z. Wei, "Application of the non-central t distribution," *Biometrika*, vol. 31, no. 3-4, pp. 362–389, 1940.

[14] R. J. Serfling, *Approximation Theory of Mathematical Statistics*, Wiley, New York, NY, USA, 1980.

[15] T. L. Lai, H. Robbins, and C. Z. Wei, "Strong consistency of least squares estimates in multiple regression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 75, no. 7, pp. 3034–3036, 1978.

[16] E. Bakhshi, M. R. Eshraghian, K. Mohammad, and B. Seifi, "A comparison of two methods for estimating odds ratios: results from the National Health Survey," *BMC Medical Research Methodology*, vol. 8, article 78, 2008.