Computerized Tongue Diagnosis Based on Bayesian Networks

Bo Pang, David Zhang*, *Senior Member, IEEE*, Naimin Li, and Kuanquan Wang, *Member, IEEE*

Abstract—Tongue diagnosis is an important diagnostic method in traditional Chinese medicine (TCM). However, due to its qualitative, subjective and experience-based nature, traditional tongue diagnosis has a very limited application in clinical medicine. Moreover, traditional tongue diagnosis is always concerned with the identification of syndromes rather than with the connection between tongue abnormal appearances and diseases. This is not well understood in Western medicine, thus greatly obstruct its wider use in the world. In this paper, we present a novel computerized tongue inspection method aiming to address these problems. First, two kinds of quantitative features, chromatic and textural measures, are extracted from tongue images by using popular digital image processing techniques. Then, Bayesian networks are employed to model the relationship between these quantitative features and diseases. The effectiveness of the method is tested on a group of 455 patients affected by 13 common diseases as well as other 70 healthy volunteers, and the diagnostic results predicted by the previously trained Bayesian network classifiers are reported.

Index Terms—Bayesian network, computerized tongue diagnosis, TCM modernization.

I. INTRODUCTION

TONGUE diagnosis [1], [2] is one of the most valuable and widely used diagnostic methods in traditional Chinese medicine (TCM). The beauty of tongue diagnosis lies in its simplicity and immediacy: whenever there is a complex disorder full of contradictions, examination of the tongue instantly clarifies the main pathological process. Therefore, it is of great value in both clinic applications and self-diagnosis. Moreover, tongue diagnosis is one of the few diagnostic techniques that accord with the most promising direction in the 21st century: no pain and no injury.

As tongue diagnosis has played such a prominent role in the diagnosis and the subsequent treatment of diseases, it has attracted an increasing amount of attention both in clinical medicine and in biomedicine. However, traditional tongue diagnosis has inevitable limitations that impede its medical applications. First, the clinical competence of tongue diagnosis is determined by the experience and knowledge of the practitioners. Second, tongue diagnosis is usually based on the detailed visual discrimination. Therefore, it depends on the subjective analysis of the examiners, so that the diagnostic results may be unreliable and inconsistent. Finally, traditional tongue diagnosis is intimately related to the identification of syndromes (also called patterns) [2], and it is not very well understood in Western medicine and modern biomedicine. Therefore, it is necessary to build an objective and quantitative diagnostic standard for tongue diagnosis, and explore the relations between features and diseases.

Recently, researchers have been developing various methods and systems [3]–[9] to circumvent these problems. Despite considerable progress in the standardization and quantification of tongue diagnosis, there are significant problems with the existing approaches. First, some methods are only concerned with the identification of syndromes that are expressed in sophisticated terms from TCM; consequently they will not be widely accepted. Second, many of the developed models are only dedicated to the recognition of pathological features defined in traditional tongue diagnosis, and the mapping from images of the tongue to diseases is not considered. This will undoubtedly limit the applications of these approaches in clinical medicine.

In this paper, we propose a computerized tongue inspection method based on quantitative features and Bayesian networks. Different from the existing approaches, our method is dedicated to the classification of 14 diagnostic categories (13 common diseases and healthy) instead of the identification of syndromes. Also, rather than trying to find numeric representations of those qualitative features that originate from traditional tongue diagnosis, we extract two ordinary kinds of quantitative features from tongue images, namely chromatic and textural features, using popular image processing techniques. One direct benefit is that the subjectivity of evaluation will be eliminated.

B. Pang is with the Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, P.R. China (e-mail: bpang@hit.edu.cn).

*D. Zhang is with the Department of Computing, the Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

N. Li is with the Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, P.R. China.

K. Wang is with the Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, P.R. China (e-mail: wangkq@hope.hit.edu.cn).

II. TONGUE DIAGNOSIS USING BAYESIAN NETWORKS

Uncertainty is an inherent issue in nearly all medical problems. The prevailing methods for managing various forms of uncertainty are formalized within a probabilistic framework. The corresponding Bayesian statistics provides a compelling theoretical foundation that coherent subjective beliefs of human experts should be expressible in a probabilistic framework. Bayesian network models provide a practical tool to create and maintain such probabilistic knowledge bases.

A Bayesian network [or Bayesian belief network (BBN)] [10], [11] for a problem domain, which is just a set of variables

$\{x_1, \ldots, x_n\}$, is a causal probabilistic network that compactly represents a joint probability distribution (JPD) over those variables. The representation consists of a set of local conditional probability distributions, combined with a set of assertions of conditional independence (CI) that allow us to construct the global joint distribution from the local distributions. The decomposition is based on the chain rule of probability, which dictates that

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, x_1, \ldots, x_{i-1}). \qquad (1)$$

For each variable x_i, let $\Pi_i \subseteq \{x_1, \ldots, x_{i-1}\}$ (a parent set) be a set of variables that renders x_i and $\{x_1, \ldots, x_{i-1}\}$ conditionally independent. That is

$$p(x_i \,|\, x_1, \ldots, x_{i-1}) = p(x_i \,|\, \Pi_i). \qquad (2)$$

Given these sets, a Bayesian network can be described as a directed acyclic graph (DAG) such that each variable $x_1, \ldots, x_n$ corresponds to a node in that graph and the parents of the node corresponding to x_i are the nodes corresponding to the variables in Π_i. Associated with each node x_i are the conditional probability distributions $p(x_i \,|\, \Pi_i)$—one distribution for each instance of Π_i. Combining (1) and (2), we see that any Bayesian network for $\{x_1, \ldots, x_n\}$ uniquely determines a joint probability distribution for those variables. That is

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, \Pi_i). \qquad (3)$$

Bayesian networks have several advantages for data analysis [20]. First, since the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Second, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Third, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Fourth, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over-fitting of data. Finally, it is found [12] that diagnostic performance with Bayesian networks is often surprisingly insensitive to imprecision in the numerical probabilities. Knowledge that high levels of precision are not necessary should greatly improve acceptance of these techniques. Many researchers have criticized the use of Bayesian networks because of the need to provide many numbers to specify the conditional probability distributions. However, if rough approximations are adequate, then these criticisms may lose their sting. Thanks to these unique characteristics, BBNs have been widely used in many of machine learning applications, also in medical diagnosis [13], [14].

As for computerized tongue diagnosis, two points should be mentioned when using a BBN as a diagnostic model. First, although Bayesian networks provide a natural and efficient way of representing prior knowledge, we do not employ any such information when constructing our diagnostic model. Consequently, both the graphical structure and the conditional probability tables of the BBN must be estimated from patient case data using statistical algorithms. The reason is twofold. First, humans are often inconsistent in their assessment of probabilities, and demonstrate many forms of bias in their judgments [15]. Similarly, experts are not able to provide probability distributions for a large number of variables in a consistent fashion, although they are usually good at identifying important dependencies that exist across variables in the domain. Thus, it is argued that for the computerized tongue diagnosis application, which involves a large number of variables, obtaining probability estimates from an existing database is often more reliable than eliciting them from human experts. Second, expert knowledge in traditional tongue diagnosis is always concerned with the identification of syndromes instead of with the relationship between tongue appearances and diseases. Therefore, prior knowledge of the relationship between symptoms (tongue abnormal appearances) and diseases in terms of probability distributions is actually unavailable.

The second point concerning the use of a BBN as a diagnostic model is that, all of the nodes except for the root node (class node) in our model (a Bayesian network classifier) represent quantitative chromatic and textural features obtained by using image processing techniques, which are not directly related to qualitative pathological features employed in tradition tongue diagnosis. This is consistent with the original intention of our method: the quantification and objectivization of traditional tongue diagnosis.

The outline of a computerized tongue diagnosis system that uses a Bayesian network as the feature-matching model is illustrated in Fig. 1. Our method is dedicated to the feature extraction and matching processes.

III. QUANTITATIVE PATHOLOGICAL FEATURES EXTRACTION

As mentioned above, the main aim of our method is to diagnose diseases from a set of quantitative features that are extracted using image processing algorithms. However, traditional tongue diagnosis theories on the pathological features are all qualitative, thus subjective, using descriptions such as "reddish purple tongue," "white, thin, and slippery coating," and so on [2]. Therefore, how to develop appropriate objective features that are meaningful for diagnosis is an important issue. The most direct way is to find out a set of objective measurements, each of which corresponds to a specific qualitative feature in traditional tongue diagnosis, just as many existing methods [4], [6]–[8] for computerized tongue image analysis usually do. Although direct and simple, these methods suffer from difficulties concerning evaluation standards, since they are evaluated by physicians. This leads to a strange situation: methods are purposely devised to replace qualitative features so to avoid subjectivity, while they are evaluated subjectively.

Actually, many descriptive features in traditional tongue diagnosis indicate some implicit relations to color and texture related features (for example "reddish purple," "white," "thin," and "slippery"). In order to remain consistent with the original intention of our method, we employ several general chromatic
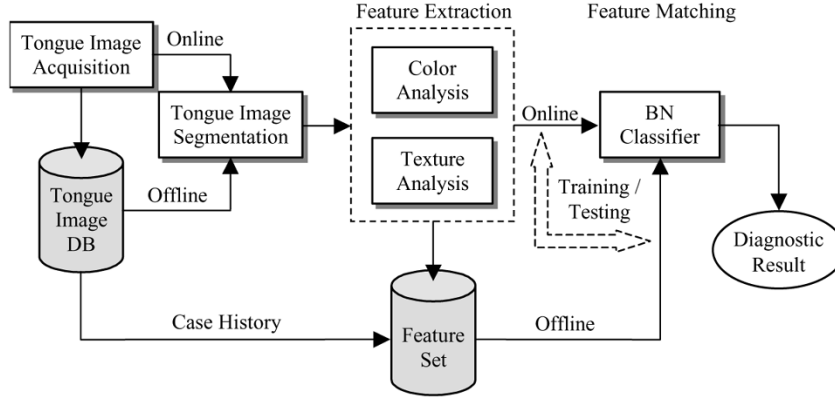
Fig. 1. The outline of the computerized tongue diagnosis system.

and textural measurements [18], [19] and take no considerations of whether these measurements correspond to specific qualitative features explicitly. Note that some of these features may be neither visible nor understandable by tongue diagnosis practitioners. The correlation and effectiveness of these features in diagnosis are identified in a statistical manner during the training of Bayesian network classifiers.

A. Quantitative Color Features

A color is always to be given in relation to a specific color space, and the extraction of color features can be performed in different color spaces [18]. A color space is a method by which we can specify, create and visualize color. A color is usually specified using three coordinates, or parameters. These parameters describe the position of the color within the color space being used. Familiar color spaces frequently used in image processing include RGB, HSV, CIEYxy, CIELUV, and CIELAB.

RGB (Red Green Blue) is an additive color system based on tri-chromatic theory, which often found in systems that use a CRT to display images. RGB is easy to implement but nonlinear with visual perception. RGB is frequently used in most computer applications since no transform is required to display information on the screen. Consequently, RGB is commonly the basic color space for most applications.

The CIE (the International Commission on Illumination) system is based on the description of color as a luminance component Y, and two additional components X and Z. The magnitudes of the XYZ components are proportional to physical energy, i.e., any color is represented by a positive set of values. The CIEXYZ color space is usually used as a reference color space and is as such an intermediate device-independent color space. Practically, it is often convenient to discuss "pure" color in the absence of brightness. The CIE defines a normalization process to compute two chromaticity coordinates: $x = X/(X + Y + Z)$ $y = Y/(X + Y + Z)$. Thus, a color can be specified by its chromaticity and luminance, in the form of a xyY (CIEYxy) triple.

The CIEXYZ and RGB systems are far from exhibiting perceptual uniformity. So the CIE standardized two systems based on CIEXYZ, CIELUV, and CIELAB, whose main goal was to provide a perceptually equal space. This means that the Euclidian distance between two colors in the CIELUV/CIELAB color space is strongly correlated with the human visual perception. CIELUV and CIELAB are device independent but suffer from being quite unintuitive despite the L parameter having a good correlation with perceived lightness.

Different from other color spaces, the HSV color space is an intuitive system in which a specific color is described by its hue, saturation and brightness values. It is a linear transformation from the RGB space. However, HSV has discontinuities in the value of hue around red, which make this system noise-sensitive. As a result, we use the other four color spaces (RGB, CIEYxy, CIELUV, and CIELAB) for the extraction of quantitative color features.

The color-related measurements used in our method are the mean and standard deviation of the colors of pixels within the whole region of the tongue, in all the four color spaces. Since both of the L channels in CIELUV and CIELAB indicate the sensation of the lightness in human vision system, we use it only once. Thus, there are a total of 22 different measures as follows. $\mathrm{CR}_i (i = 1, 2, \ldots, 11)$: Means of each color plane in the four color spaces, and $\mathrm{CR}_j (j = 12, 13, \ldots, 22)$: Standard deviations of each color plane in the four color spaces.

B. Quantitative Texture Features

Among all statistical methods, the most popular one, which is based on the estimation of the second-order statistics of the spatial arrangement of the gray level values, are the gray level co-occurrence matrices. A co-occurrence matrix [22] is a square matrix whose elements correspond to the relative frequency of occurrence of pairs of gray level values of pixels separated by a certain distance in a given direction. Formally, the elements of a $G \times G$ gray level co-occurrence matrix $P_{\mathbf{d}}$ for a displacement vector $\mathbf{d} = (dx, dy)$ is defined as

$$P_{\mathbf{d}}(g_1, g_2) = \left| \left\{ \begin{array}{l} (a, b) \in N \times N \colon I(a, b) = g_1, \\ I(a + dx, b + dy) = g_2, \\ (a + dx, b + dy) \in N \times N \end{array} \right\} \right| \quad (4)$$

where $I(\cdot, \cdot)$ denotes an image of size $N \times N$ with G gray values, g_1 and g_2 are two gray level values, and $|\cdot|$ is the cardinality of a set.

In this paper, two measures of textural features, which are derived from the co-occurrence matrix, are used to extract different textural features from tongue images. These two descriptors are

the second-order moment and the contrast measures of the matrix, which are shown as follows:

$$W_M = \sum_{g_1} \sum_{g_2} p^2(g_1, g_2)$$

$$W_C = \sum_{g_1} \sum_{g_2} |g_1 - g_2| p(g_1, g_2) \qquad (5)$$

where $p(g_1, g_2)$ is a normalized co-occurrence matrix. That is $p(g_1, g_2) = P_\mathbf{d}(g_1, g_2)/S$ where S is the total number of pixel pairs (g_1, g_2) across all g_1 and g_2, i.e., $S = \sum_{g_1=0}^{G-1} \sum_{g_2=0}^{G-1} P_\mathbf{d}(g_1, g_2)$. W_M measures the smoothness or homogeneity of an image, which will reach its minimum value when all of the $p(g_1, g_2)$ have the same value. W_C is the first-order moment of the differences in the values of the gray level between the entries in a co-occurrence matrix. Both of the textural descriptors are calculated quantitatively. Notice that they have little correlation with the sensation of the human vision system [19]. For all of the textural measures in the following experiments, we take 64 gray levels (i.e., $G = 64$) and $\mathbf{d} = (6, 6)$.

It is believed [2] that different parts of the tongue correspond to different internal organs. The tip of the tongue, for example, reveals heart and lung conditions, and the middle tongue conditions of the spleen and stomach. It does not matter whether this theory is true; the important fact is that there are usually abnormal changes in texture in different parts of the tongue when various diseases are present. Therefore, to represent these possible pathological changes, we calculate the above textural measures for each partition of a tongue. For convenience, we denote each partition of a tongue using a digit: 1—Tip of the tongue; 2—Left edge of the tongue; 3–Center of the tongue; 4–Right edge of the tongue; and 5–Root of the tongue. Thus, we obtain a set of textural measurements for each tongue, which contains a total of 10 texture measures as follows:

$$\begin{cases} \mathrm{TR}_i = W_{M,i} \\ \mathrm{TR}_{i+5} = W_{C,i} \end{cases} \qquad (i = 1, 2, \ldots, 5) \qquad (6)$$

where $W_{M,i}$ and $W_{C,i}$ denote the measurements of W_M and W_C for partition i, respectively.

IV. EXPERIMENTAL RESULTS

We use the Bayesian Network PowerPredictor, developed by Cheng [16], [17], to train and test the tongue diagnosis models. The PowerPredictor takes a database as input and constructs the Bayesian network classifier, both structure and parameters, as output. The construction process is based on dependence analysis using information theory. The dependency relationships among nodes are measured by using some kind of CI test. The learning algorithm uses a three-phase construction mechanism, at the same time a wrapper algorithm is applied to fight the overfitting problem. Moreover, a natural method for feature subset selection is introduced in the learning process, which can often produce a much smaller Bayesian network classifier without compromising the classification accuracy.

A total of 525 subjects, including 455 patients and 70 healthy volunteers, are involved in the following experiments. There are totally 13 common internal diseases included (see Table I). The

TABLE I
LIST OF THE 13 COMMON DISEASES AND HEALTHY SUBJECTS

Disease ID	Disease	Number of subjects
D00	Healthy	70
D01	Intestinal infarction	11
D02	Cholecystitis	21
D03	Appendicitis	43
D04	Pancreatitis	41
D05	Nephritis	17
D06	Diabetes mellitus	49
D07	Hypertension	65
D08	Heart failure	17
D09	Pulmonary heart disease	21
D10	Coronary heart disease	71
D11	Hepatocirrhosis	25
D12	Cerebral infarction	30
D13	Upper respiratory infection	44

Fig. 2. Four tongue image samples of patients suffering intestinal infarction (upper left), cholecystitis (upper right), appendicitis (lower left), and pancreatitis (lower right).

patients are all in-patients mainly from five different departments at the Harbin 211 Hospital, and the healthy volunteers are chosen from the campus students of Harbin Institute of Technology. We take a total of 525 digital tongue images, exactly one for each subject, as the experimental samples. Four typical tongue image samples are shown in Fig. 2.

A. Several Issues

In many projects concerning tongue diagnosis in TCM, the straightforward way to label the samples is to ask TCM doc-

tors to judge. However, the judgment of tongue diagnosis doctors is always related to syndromes or qualitative features, rather than medical diseases. Therefore, these approaches do not solve the problem of the individual empiricism inherent in the tongue diagnosis, and the diagnostic results given in this way are too sophisticated to understand. In this research, we use the diagnostic results obtained by using the clinical differential diagnosis methodology as the labels of the tongue images. Since all the subjects in the experiment are in-patients, the diagnoses are highly reliable. During the testing process, the diagnostic results obtained by querying the Bayesian networks are compared with the corresponding labels of the tongue images. This forms an objective evaluation basis for our method.

Another issue concerns the relative small sample size. The weak point of this study, and very probably of most attempts to provide medical statistics is the difficulty of gathering both a sufficient number of cases and reliable data for each case. To circumvent this problem, we utilized a stratified k-fold cross-validation (CV) technique [21] (for all of the experiments, k equals to 10) in all of the following experiments to estimate the accuracy of classifiers. K-fold CV technique partitions a pool of labeled data, S, into k approximately equally sized subsets. Each subset is used as a test set for a classifier trained on the remaining $k-1$ subsets. The empirical accuracy is given by the average of the accuracies of these k classifiers. When employing a stratified partitioning in which the subsets contain approximately the same proportion of classes as S, we get a stratified k-fold CV, which can reduce the estimate's variance.

Next, we use discrete Bayesian networks in all of the following experiments. The discretization method [17] of fields (attributes) is "Equal Width," and the number of intervals is set to 5 for all of the fields. Thus, the network parameters—local conditional probability distributions—are actually local conditional probabilities. At the same time, in order to demonstrate the superiority of Bayesian network classifiers in tongue diagnosis, we implement a nearest-neighbor classifier (NNC) as comparison. Similarly, a stratified k-fold CV ($k = 10$) is used for the evaluation of its accuracy. Thus, each sample in a test subset is assigned to the same label as its nearest labeled sample belonging to the rest $(k-1)$ subsets. This process is repeated k times on each subset to estimate the overall accuracy.

Finally, in the following experiments, we use a misclassification cost table assuming that all types of misclassification are equally important. This may be not correct in practice, but it does not invalidate the demonstration of the effectiveness of our method in diagnosing diseases.

B. Bayesian Network Classifier Based on Textural Features

In the first experiment, we trained a Bayesian network classifier based on textural features, which is called a "texture BNC" (T-BNC). The graphical structure of the learned T-BNC is illustrated in Fig. 3. Note that, this BNC structure corresponds to the highest scoring out of the tenfold CV iterations. A subset of 5 textural features out of the original feature set (containing a total of 10 textural features) is selected by an underlying feature selection function integrated in the training algorithm. Two of the five surviving features (namely, TR_2 and TR_5) are the
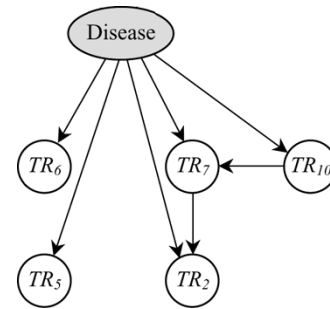

Fig. 3.   Structure of the T-BNC.

measurements of W_M for the left side and root of the tongue; the other three are the measurements of W_C for the tip, left side and root of the tongue, respectively. Obviously, textural measurements related to the tip, sides (although only the measurements for the left side of the tongue are selected, it is found that both sides have the similar values), and root of the tongue are most discriminating for the classification.

The diagnostic results given by the T-BNC are shown in the first column of Table II. The average true positive rate (TPR) is about 26.1%, which demonstrates that the textural features utilized in this study are not very discriminating in diagnosing these diseases. Nevertheless, for the identification of appendicitis (D03), pancreatitis (D04), and coronary heart disease (D10), these textural features are more meaningful. Due to the low sensitivity values, we do not list the positive predictive values (PPV) for the T-BNC.

C. Bayesian Network Classifier Based on Chromatic Features

This section evaluates the suitability of chromatic measures for the classification. The graphical structure of the trained model, called a "color BNC" (C-BNC), is shown in Fig. 4. Again, the learned structure also corresponds to the highest scoring in the CV iterations. A subset of 12 chromatic features is selected from the original feature set containing 22 features. Among these surviving measurements, there are 6 that are directly connected to the class node (Disease), which are called "contribution" nodes thereafter. Note that, each of the four color spaces used in our experiment includes at lease one "contribution" node, and the mean and standard deviation measurements have similar significance for the classification.

The diagnostic results of the color BNC are listed in the second column of Table II in terms of sensitivity and the first column of Table III in terms of PPVs. Obviously, the diagnostic classification capability of the color BNC is significantly better than that of texture BNC: the average TPR of the C-BNC reaches 62.3%. It should be noticed that the C-BNC makes a very accurate and relative reliable diagnosis of pancreatitis (D04) with 90.2% TPRs and 68.5% PPVs. The reason for this is that it is found that patients with pancreatitis usually have a distinctively bluish tongue (see Fig. 2).

D. Bayesian Network Classifier Based on Combined Features

Finally, we use both chromatic and textural features to construct a joint BNC (J-BNC) for the classification of these diseases. The graphical structure of the trained J-BNC is illustrated

TABLE II
DIAGNOSTIC RESULTS (IN PERCENTAGE) OF VARIOUS BAYESIAN NETWORK
CLASSIFIERS IN TERMS OF SENSITIVITY (TRUE POSITIVE RATE)

Disease ID	Texture BNC	Color BNC	Joint BNC
D00	20.0	50.0	77.1
D01	9.1	45.5	63.6
D02	4.8	42.9	61.9
D03	53.5	86.0	93.0
D04	70.7	90.2	100
D05	5.9	17.6	23.5
D06	4.1	53.1	65.3
D07	3.1	61.5	75.4
D08	5.9	35.3	35.3
D09	4.8	47.6	71.4
D10	64.8	90.1	93.0
D11	12	48.0	64.0
D12	13.3	60.0	80.0
D13	20.5	56.8	70.5
Average	26.1	62.3	75.8

TABLE III
POSITIVE PREDICTIVE VALUES (PPV) OF THE C-BNC AND THE
J-BNC (IN PERCENTAGE)

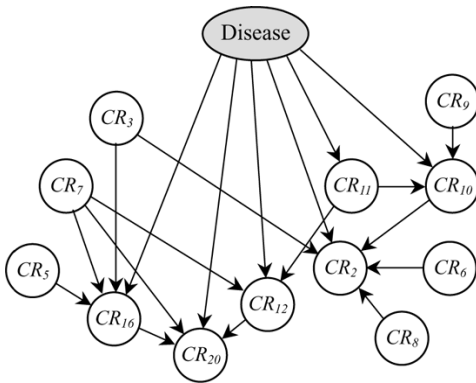| Disease ID	Color BNC	Joint BNC
D00	79.5	85.7
D01	100	87.5
D02	100	92.9
D03	60.7	72.7
D04	68.5	75.9
D05	100	80
D06	81.3	84.2
D07	66.7	80.3
D08	100	100
D09	83.3	88.2
D10	37.9	54.5
D11	100	88.9
D12	66.7	75.0
D13	80.6	93.9

Fig. 4.   Structure of the C-BNC.

Fig. 5.   Structure of the J-BNC.

in Fig. 5. As well, this structure corresponds to the highest classification scoring in the CV iterations. Out of the 32 features originally taken into consideration, 14 are finally selected and among which 9 (5 chromatic features and 4 textural features) are the "contribution" nodes, which are both discriminative and independent. Note that, the four surviving textural features are the measurements of W_M for the tip (TR_1) and right side (TR_4) of the tongue, and the measurements of W_C for the tip (TR_6) and root (TR_{10}) of the tongue. Similar to the T-BNC, textural measurements related to the tip, sides, and root of the tongue are most relevant to the diagnostic classification for the J-BNC. It is interesting that this apparent relationship between diseases and parts of the tongue that is learned purely from statistics has a high degree of accordance with the beliefs of traditional tongue diagnosis.

Tables II and III show the diagnostic results given by the J-BNC in terms of TPR and PPV. The estimate prediction accuracy is 75.8%, which outperforms both the T-BNC and the C-BNC. Note that, as for the diagnosis of D00 (healthy), D04
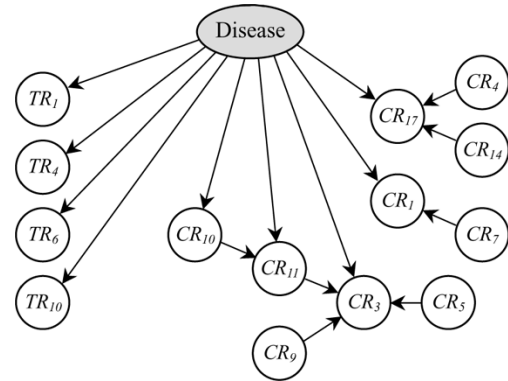
(pancreatitis), D07 (hypertension), and D12 (cerebral infarction), the TPRs and PPVs of the J-BNC are all higher than 75%. For the convenience of illustration, the confusion matrix of the J-BNC is given in Table IV.

For comparison, we employ a NNC on the combined feature set and utilize the stratified k-fold $(k = 10)$ CV technology to evaluate the classification accuracy. The diagnostic results are given in Table V in terms of sensitivity. It can be seen that Bayesian network classifiers are much more superior to the NNC with respect to the classification accuracy of tongue images.

V. CONCLUSION

In this paper, we propose a computerized tongue diagnosis method aimed at eliminating the subjective and qualitative characteristics of traditional tongue diagnosis and establishing the relationship between tongue appearance and diseases. Bayesian network classifiers based on quantitative features, namely chromatic and textural measurements, are employed as the decision models for diagnosis. Experiments are carried out on a total of

TABLE IV
CONFUSION MATRIX OF THE J-BNC

							Predicted								
		D00	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13
Actual	D00	54			2	3			2			9			
	D01	1	7			1		2							
	D02	1		13		3						4			
	D03	1			40							1			1
	D04					41									
	D05	3			1		4		1			7		1	
	D06	1			3	1		32	3			5		4	
	D07				3				49		1	10	2		
	D08	1		1	2			1	1	6		5			
	D09				1				1		15	3			1
	D10	1			1			1	2			66			
	D11				1	1						5	16	2	
	D12					1			1	1	1	2		24	
	D13		1		1	3	1	1	1			4		1	31

TABLE V
COMPARISON OF J-BNC AND NNC ON COMBINED FEATURES IN
TERMS OF SENSITIVITY

Disease ID	Joint BNC	NNC
D00	77.1	28.6
D01	63.6	54.5
D02	61.9	33.3
D03	93.0	44.2
D04	100	56.1
D05	23.5	35.3
D06	65.3	18.4
D07	75.4	18.5
D08	35.3	52.9
D09	71.4	52.4
D10	93.0	46.5
D11	64.0	40.0
D12	80.0	20.0
D13	70.5	43.2
Average	75.8	36.2

that are very popular in TCM. Instead, it establishes a mapping from quantitative features to diseases. Consequently, the method is actually independent of TCM, except the fact that it is motivated by the art of traditional tongue diagnosis. Second, the underlying validity of our method is based on diagnostic results using Western medicine techniques. The measurements of the chromatic and textural properties of a tongue, which are extracted via image processing procedures, are connected with the corresponding diagnostic results obtained by using Western medicine, instead of the judgment of a TCM doctor. This forms an objective basis of the method and such an approach could expedite its use in clinical applications.

ACKNOWLEDGMENT

455 in-patients affected by 13 common internal diseases and 70 healthy volunteers. The estimate prediction accuracy of the joint BNC is up to 75.8%. In particular, the diagnosis of four groups: healthy, pancreatitis, hypertension, and cerebral infarction have both TPRs and PPVs higher than 75%. The experimental results reasonably demonstrate the effectiveness of the method described in this paper, thus establishing the potential usefulness of computerized tongue diagnosis in clinical medicine.

Compared with the existing approaches, the method presented in this paper has several outstanding characteristics. First, it is not concerned with the identification of syndromes

REFERENCES

[1] B. Kirschbaum, *Atlas of Chinese Tongue Diagnosis*. Seattle, WA: Eastland, 2000.
[2] G. Maciocia, *Tongue Diagnosis in Chinese Medicine*. Seattle, WA: Eastland, 1995.
[3] C. Yang, "A novel imaging system for tongue inspection," in *Proc. IEEE-IMTC*, 2002, pp. 159–163.
[4] C. H. Li and P. C. Yuen, "Tongue image matching using color content," *Pattern. Recogn.*, vol. 35, no. 2, pp. 407–419, Feb. 2002.
[5] K. Wang, D. Zhang, N. Li, and B. Pang, "Tongue diagnosis based on biometric pattern recognition technology," in *Pattern Recognition from Classical to Modern Approaches*, 1st ed, S. K. Pal and A. Pal, Eds, Singapore: World Scientific, 2001, pp. 575–598.
[6] C. C. Chiu, "A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue," *Comput. M. PR.*, vol. 61, no. 2, pp. 77–89, Feb. 2000.
[7] P. C. Yuen, Z. Y. Kuang, W. Wu, and Y. T. Wu, "Tongue texture analysis using opponent color features for tongue diagnosis in traditional Chinese medicine," in *Proc. TAMV*, 1999, pp. 21–27.
[8] M. Takeichi and T. Sato, "Computerized color analysis of 'Xue Yu' (blood stasis) in the sublingual vein using a new technology," *Amer. J. Chinese Med.*, vol. 25, pp. 213–219, 1997.

[9] C. C. Chiu, H. S. Lin, and S. L. Lin, "A structural texture recognition approach for medical diagnosis through tongue," *Biomed. Eng. Appl. Basis Commun.*, vol. 7, no. 2, pp. 143–148, 1995.

[10] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artif. Intell.*, vol. 29, pp. 241–288, 1986.

[11] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learn.*, vol. 29, pp. 131–163, 1997.

[12] M. Pradhan, M. Henrion, G. Provan, B. D. Favero, and K. Huang, "The sensitivity of belief networks to imprecise probabilities: An experimental investigation," *Artif. Intell.*, vol. 85, pp. 363–397, 1996.

[13] A. Tsymbal and S. Puuronen, "Ensemble feature selection with the simple Bayesian classification in medical diagnostics," in *Proc. IEEE-CBMS*, 2002, pp. 225–230.

[14] O. Ogunyemi, J. R. Clarke, and B. Webber, "Using Bayesian networks for diagnostic reasoning in penetrating injury assessement," in *Proc. IEEE-CBMS*, 2000, pp. 115–120.

[15] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, pp. 1124–1131, 1974.

[16] J. Cheng, R. Greiner, J. Kelly, D. A. Bell, and W. Liu, "Learning Bayesian networks from data: An information theory based approach," *Artif. Intell.*, vol. 137, pp. 43–90, 2002.

[17] J. Cheng. (2000) PowerPredictor System software. [Online]. Available: http://www.cs.ualberta.ca/~jcheng/bnpp.htm

[18] I. Pita, *Digital Image Processing Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 1993, pp. 23–40.

[19] T. R. Reed and J. M. H. DuBuff, "A review of recent texture segmentation and feature extraction techniques," in *Proc. CVGIP: Image Understanding*, vol. 57, May 1993, pp. 359–372.

[20] D. Heckerman, "Bayesian networks for data mining," *Data M. K. D.*, vol. 1, pp. 79–119, 1997.

[21] M. Mullin and R. Sukthankar, "Complete cross-validation for nearest neighbor classifiers," in *Proc. ICML*, 2000, pp. 639–646.

[22] R. M. Haralick, K. Shanmugan, and I. Dinstein, "Textural features for image classification," in *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, 1973, pp. 610–621.

David Zhang (M'89-SM'95) graduated in computer science from Peking University in 1974 and received his M.Sc. and Ph.D. degrees in computer science and engineering from the Harbin Institute of Technology (HIT), Harbin, P.R. China, in 1983 and 1985, respectively. He received the second Ph.D. degree in electrical and computer engineering at the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Postdoctoral Fellow at Tsinghua University, Beijing, China, and became an Associate Professor at Academia Sinica, Beijing, China. Currently, he is a Professor with the Hong Kong Polytechnic University, Hong Kong. He is Founder and Director of Biometrics Research Centers supported by the Government of the Hong Kong SAR (UGC/CRC) and the National Nature Scientific Foundation (NSFC) of China, respectively. He is also Founder and Editor-in-Chief of the *International Journal of Image and Graphics* (IJIG), Book Editor, The Kluwer International Series on Biometrics, and an Associate Editor of several international journals, such as IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS AND IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, and *Pattern Recognition*. His research interests include automated biometrics-based authentication, pattern recognition, biometric technology and systems. As a principal investigator, he has finished many biometrics projects since 1980. So far, he has published over 200 papers and ten books.

Naimin Li graduated from the Department of Medical Treatment in the Traditional Chinese Medicine (TCM) training class, Shenyang, China, in 1962.

He is a Fellow Professor at the Harbin Institute of Technology, Harbin, China. He began the study and exploration of the application of tongue diagnosis in modern disease diagnosis and treatment since 1965, and put it forward to the fields of internal medicine, pediatrics, epidemiology and gynecology. During 1970 and 1972, he was appointed as a Leader of Experts Group of TCM by the Chinese government in the Middle-Europe and Albania. In 1989, the first tongue image laboratory in China was set up under his guidance. So far, he has authored and co-authored over 250 papers and 14 books around his research areas.
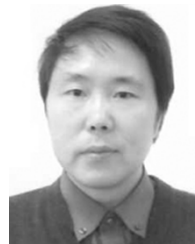
Kuanquan Wang (M'01) received the B.E. and M.E. degrees in computer science from Harbin Institute of Technology (HIT), Harbin, China, and the Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 1985, 1988, and 2001, respectively.

From 1988 to 1998, he worked in the Department of Computer Science, Southwest Normal University, Chongqing city, China, as a Tutor, Lecturer, and Associate Professor, respectively. Since 1998, he has been working in the Biocomputing Research Centre of Computer Science and Engineering Department, Harbin Institute of Technology, China. Meanwhile, from 2000 to 2001 has been a Visiting Scholar at the Hong Kong Polytechnic University supported by Hong Kong Croucher Funding and from 2003 to 2004 he was a Research Fellow in the same university. Currently, he is a Professor and a supervisor of Ph.D. degree candidates in the Department of Computer Science and Engineering, and an Associate Director of the Biocomputing Research Centre at HIT. To date, he has published over 70 papers. His research interests include biometrics, image processing, and pattern recognition.

Prof. Wang is an editorial board member of the *International Journal of Image and Graphics*. He is also a reviewer for EEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS and *Pattern Recognition*.

Bo Pang received the B.Eng. and M.Eng. degrees in computer science and engineering from Harbin Institute of Technology, Harbin, P.R. China, in 1997 and 1999, respectively. He is currently working toward the Ph.D. degree in computing and biomedical engineering at the Biocomputing Research Center of Harbin Institute of Technology.

His research interests include image processing, pattern recognition, data mining, and biomedical engineering of Chinese medicine.