Future Generation Computer Systems 115 (2021) 59–75

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

An automatic multi-view disease detection system via Collective Deep Region-based Feature Representation☆

Jianhang Zhou, Qi Zhang, Bob Zhang*

PAMI Research Group, Department of Computer and Information Science, University of Macau, Macau

ARTICLE INFO

Article history:
Received 30 January 2020
Received in revised form 1 August 2020
Accepted 26 August 2020
Available online 3 September 2020

Keywords:
Disease detection system
Multi-view learning
Feature representation
Medical biometrics
Image segmentation

ABSTRACT

With today's growing requirements in disease diagnosis, we are constantly looking for better solutions. To meet the current demands, a disease detection system being highly effective as well as efficient is required. Existing and popular medical biometrics methods mainly focus on the local features extracted from raw medical image data, rather than study them globally. Meanwhile, prior knowledge is predefined in these methods so that procedures are inconsistent and require more manual operations. To address these, we present an automatic multi-view disease detection system, which contains a series of automatic procedures. The system first takes a tuple of images containing the face, tongue, and sublingual vein as the multi-view input, before directly outputting the predicted class label. To perform multi-view disease diagnosis, we propose a collective deep region-based feature representation. In summary, there are three real innovations in this paper: (1) Automated end-to-end medical biometrics system, (2) Deep region-based feature representation, (3) Multi-view multi-disease medical biometrics diagnosis. Extensive experiments were conducted on four diseases and one healthy control group using binary classification, showing both the effectiveness and efficiency of the proposed system. The average accuracy achieved was 95.8%, 96.49%, 96%, and 96.8% for breast tumor, heart disease, fatty liver, and lung tumor versus healthy control group taking 0.0031s, 0.003s, 0.0046s, and 0.0033s to process each sample respectively.

1. Introduction

Recently, methods in medical biometrics have been developing at an accelerated rate and have already achieved promising performances [1–8]. Some diseases for example, such as diabetes mellitus, significant achievements have been made by incorporating pattern recognition and image processing techniques [9] along with Traditional Chinese Medicine (TCM) theory to form a detection method known as medical biometrics [10–12]. Different from modern disease diagnostic methods like the FPG test (in the case of diabetes mellitus), medical biometrics performs non-invasive detection, which does not require bodily fluid extraction or injections. Therefore, medical biometrics methods can reduce the pain of patients as much as possible. According to TCM, changes in a person's condition caused by diseases will be reflected on the organs such as the face, tongue, or sublingual vein. Based on this idea, the applications of various artificial intelligence methods ranging from feature engineering to the classification of medical image profiles is a feasible way to perform disease detection.

In reality, medical biometrics (in terms of analyzing the human surface characteristics such as the face, tongue, and sublingual vein) borrows from the idea of TCM diagnosis, making judgments jointly in terms of the characteristics from different surfaces of the human body, which can be seen as a multi-view disease diagnosis problem between a particular disease and healthy individuals. To computerize these features, there are different devices designed specifically for capturing the images of these specific organs [2,13]. However, previous methods either focus on diabetes mellitus only [4,5] or used a single view [1,14,15], indicating the lack of breadth. Furthermore, these methods run in a manual or semi-automatic way.

As mentioned above, since multiple views of the human body are jointly considered when performing disease diagnosis in medical biometrics, the multi-view learning strategy is involved in our proposed system. We used the late fusion strategy to fuse the information of three views (face, tongue, and sublingual vein) in order to disambiguate the result made by one single view. We designed a coding procedure integrating the highest probability

☆ This document is the result of a research project funded by the University of Macau (MYRG2018-00053-FST).

* Corresponding author.
E-mail addresses: yc07424@um.edu.mo (J. Zhou), yc07485@um.edu.mo (Q. Zhang), bobzhang@um.edu.mo (B. Zhang).
URLs:
http://sites.google.com/view/zhou-j-h-s-personal-page/%E9%A6%96%E9%A1%B5 (J. Zhou), http://www.cis.umac.mo/bobzhang (B. Zhang).

https://doi.org/10.1016/j.future.2020.08.038
0167-739X/

class label from each view in a coding vector for further decision making. Since the coding vector integrates information from the three views, we termed it as the collective representation.

Given the advantages of multi-view disease detection and deep learning, in this paper we propose a multi-view disease detection system which takes a tuple of images containing the face, tongue, and sublingual vein as input and directly outputs the predicted label (healthy or a specific disease). To realize the automatic detection, we build a structure of View Classification–Deep Region Segmentation–Feature Representation–Coding-Decision Making in this system. To extract effective feature representation with more complete information, we propose a Deep region-based feature representation. Furthermore, to perform the multi-view disease diagnosis, we propose a **C**ollective **D**eep **R**egion-based **F**eature **R**epresentation method, termed as CDRFR. The pipeline of our proposed system is shown in Fig. 1. In our proposed system, an image instance (multi-view images containing views of the face, tongue, and the sublingual vein) is first input. Since the system receives one image at a time, we place a View Classification network in order to ensure the system can automatically extract the features from an arbitrary image without any prior knowledge of the view information. Besides this, View Classification avoids any possible human errors caused by an operator manually classifying each image. The View Classification network will perform classification precisely on each image of a given instance and later transmit it to a corresponding Deep Region Segmentation network for further processing. As previous state-of-the-art works mentioned above [1,3,4] used local information from each view by cropping pre-defined blocks. Here, we make a distinction by extracting a 'region' which contains more information than 'blocks' for each view. The segmentation networks (FaceNet, TongueNet, and SublingualNet) are in charge of extracting the 'deep region' from the corresponding view image. The deep learning architecture can learn the most informative regions for disease diagnosis. After all the region extraction works are done, we perform Feature Representation by considering every single view to output the effective representation. First, we extract the color, texture, and geometry features based on the deep region using conventional feature extraction methods. Then, the extracted features will be concatenated together to form a single feature vector. Afterwards, in the collective representation step, the coding procedure will produce a coding vector whose entries are the highest probability class label from the three views based on the feature vectors of all three views (face, tongue, and sublingual vein). Next, to fuse the information from the three views, a coding vector (which only contains 0 or 1) whose entries are from the highest probability class label of the three views will be generated. After that, a classification algorithm is designed in the Decision Making unit and will be introduced in the subsequent Section 3. Finally, the system will output the Prediction Result of the given instance. Overall, there are three real innovations in this paper:

(1) **Automated end-to-end medical biometrics system.** This work proposes an automatic end-to-end system to perform disease detection in an automatic way all through the necessary procedures (including segmentation of ROIs, feature extraction, and classification) based on the medical biometrics.

(2) **Deep region-based feature representation.** In this work, we proposed a deep region-based feature representation for disease detection. We produced the feature representation based on the 'region' of each view extracted by the deep segmentation networks.

(3) **Multi-view multi-disease medical biometrics diagnosis.** In this work, we focused on multi-view multi-disease diagnosis based on the medical biometrics rather than a single view and/or a single disease detection.

Our proposed system achieved high efficiency and state-of-the-art disease detection results. The remaining parts of the paper are organized as follows. In Section 2, we describe the medical images and the dataset used. In Section 3, we present the automatic multi-view disease detection system as well as the proposed CDRFR method. In Section 4, extensive experiments are performed, which includes testing on View Classification, Deep Region Segmentation, and diseases diagnosis, to evaluate the effectiveness and efficiency of the proposed system and proposed method. In Section 5, we reach some conclusions.

2. Background

2.1. Related work

Many works have been done on computer-aided disease detection through the analysis of human surface characteristics (such as the face, tongue, and sub-lingual vein) recently [4,7,13,16]. Li et al. [4] used two views to detect fatty liver disease, showing the significance and the benefits of considering two views simultaneously. Zhang et al. [14] proposed a deep learning architecture to detect multiple diseases, where the accuracy they achieved is not promising. On the other hand, Shu et al. [7,15,17] acquired promising detecting results in diabetes mellitus by extracting features from the facial blocks which is a local region on the human face. Wang et al. [18] and Zhang et al. [2] extract features (e.g., color, texture, and geometry) from tongue images to perform disease diagnosis, achieving progressive outcomes.

Deep learning methods are popular in image processing areas and have shown effective performances. After the first convolutional neural network architecture was proposed by [19], many effective architectures have been designed year after year, such as VGG16, VGG19 [20], ResNet [21], DenseNet [22] and so on. Deep learning has strong transportability due to transfer learning. After learning plenty of images, the network is able to deal with the unseen image as it keeps common features in memory. With transfer learning [23], it is possible to classify large-scale datasets by training them with a small subset of the dataset. In addition to image classification, deep learning also acquires a leading performance in image segmentation. For example, FCN [24], U-net [25] are proposed to perform precise and efficient image segmentation compared to the traditional pixel prediction methods.

A multi-view learning strategy is employed to fuse information from different sources. There are three ways to fuse information from multiple views: early fusion [26], late fusion [27], and combined-style fusion [28]. In early fusion, features from multiple views are concatenated together before recognition, while in the late fusion, results from separate classifiers trained on different view are fused. The combined-style fusion strategy takes both strategies above, by enforcing the separately trained classifiers to learn from each other. For example, MV-LSSVM [29] jointly trained multiple LSSVMs from each view with a coupling parameter to share the error variables of the multiple views. The deep multi-view feature representation [30,31] is a popular method family, which contains early and late fusion strategies. DCCAE [30] is an example of a late fusion strategy trained using multiple autoencoders first, before performing optimization on the separately learned features from each view. The method proposed by [31] shows an early fusion strategy by extracting the shared representation of the two views before classification. Information fusion from multiple sources is helpful to learn better representations [30].

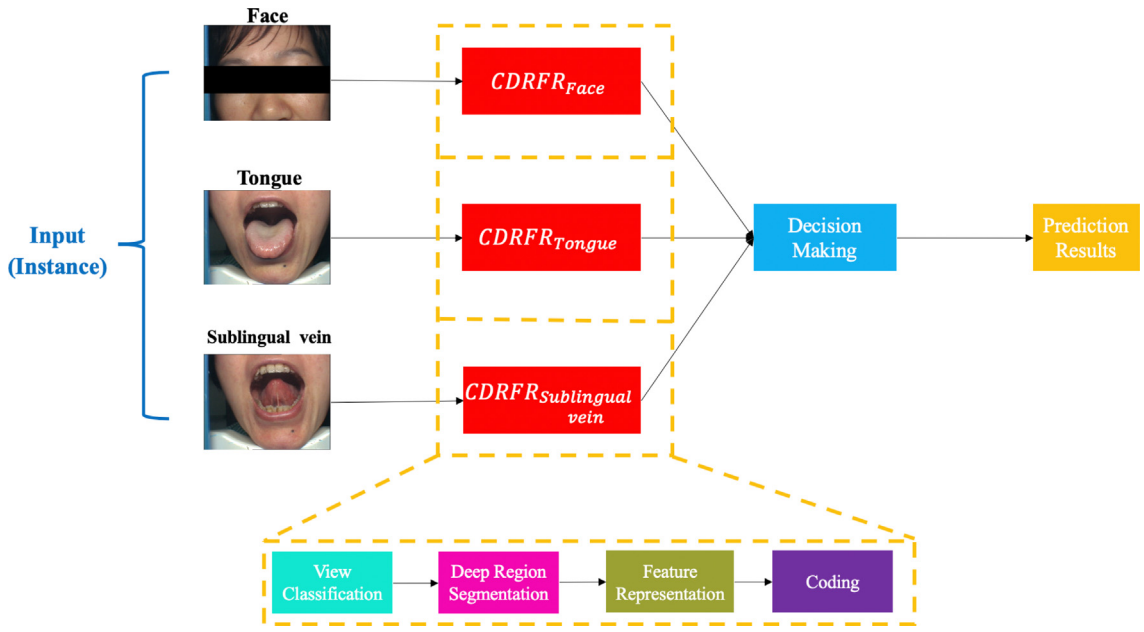


Fig. 1. The pipeline of the proposed multi-view disease detection system. The input is an image tuple consisting of the face, tongue, and sublingual vein coming from the same patient. Then, the image of each view is converted to a novel representation via our proposed CDRFR method. Inside the CDRFR, there are four sub-procedures (View classification, Deep region segmentation, Feature representation, and Coding). Afterwards, a decision making procedure is performed based on the combination of the three representations. Finally, the system outputs the Prediction results.

2.2. Face, tongue, and sublingual vein

n this paper, we chose the face, tongue, and sublingual vein as the three views to perform disease detection. For the view of the face, we borrowed and were inspired from Traditional Chinese Medicine (TCM) [32,33], where different regions on the human face can reveal the health status of the internal organs. For example, in [2], facial blocks were extracted to accurately detect diabetes mellitus, which proves the effectiveness of analyzing one's face. Besides this, there are lots of microvascular vessels on the human face, reflecting changes in the human body. For the view of the tongue (which is similar to the face), it is an important surface that is widely used in TCM [34] and Western Medicine [35] to index one's health status. For the view of the sublingual vein, changes to the body condition can be indicated in the appearance of the vessels on the sublingual vein [34,36]. In addition to the reasons above, the fact that all of face, tongue, and sublingual vein reflect body conditions through appearance, which means they can be analyzed through digital image. The device shown in Fig. 2 can simultaneously take images of each of these three views, improving the efficiency of diagnosis.

2.3. Medical images and dataset

In our medical image dataset, the samples of patients we collected and labeled were according to their diseases. This label was given by medical professionals using commonly adopted diagnostic techniques (Western medicine). For each patient, there are three views of the images captured, which are the face, tongue, and sublingual vein. The overview of the image capture procedure and device are shown in Fig. 2.

In Fig. 2, a well-designed image capture device [2] (center) is used to capture images of the face, tongue, and sublingual vein from a patient (right). More information about this device can be found in [2]. The output of the device is a group of images showing the patient's face, tongue, and sublingual vein (shown in the right hand side of Fig. 2). Here, we consider three images from a patient's three views (one image for each view) as a complete instance. In this paper, instances from the same disease will be classified to the same class. Finally, different instances are categorized into different groups according to the diseases. All captured images need to go through a color correction operation in order to remove the influences of illumination and disturbance from the external environment. To accomplish this, two groups of RGB values in the target RGB space and source RGB space from the images are extracted. Then, transformation parameters for casting values from the source RGB space to the target RGB space are estimated, in terms of a regression algorithm [37]. At last, all images are calibrated according to the transformation parameters. After the color correction the images are stored in a database. The images in our dataset was collected from the from Prince of Wales Hospital, Hong Kong and the Guangdong Provincial Hospital of Traditional Chinese Medicine, China, respectively. Please note the data collection procedure followed the Declaration of Helsinki and that no person was diagnosed as having more than one illness.

As mentioned above, the face, tongue, and sublingual vein are the three views captured from a single patient. In this dataset, all images are in the RGB color space with a resolution of 576×768 and stored in the Bitmap (BMP) format. A description of the format of each image in the dataset is shown in Table 1. Fig. 3 shows some sample images of the face, tongue, and sublingual vein in the dataset, respectively. The face image in Fig. 3(a) shows that there are external artifacts around the face region such as hairs and the edge of the device. From Fig. 3(c), the sublingual vein in light blue is located in the center of the image, which is surrounded by Plica Fimbriata (three yellow dash lines). Compared with the Plica Fimbriata and teeth, the sublingual vein is vaguer, which is difficult to recognize sometimes. The tongue image in Fig. 3(b) shows a complete tongue region in the center [38]. Although the tongue region is a relatively complete organ compared to the face and sublingual vein, it is still hard to segment from its surroundings perfectly.

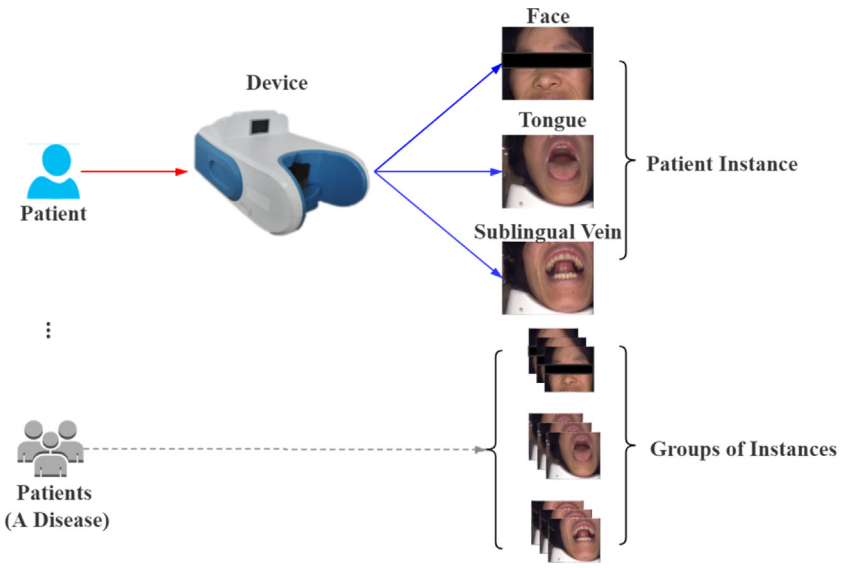


Fig. 2. An overview of image capture and the device. The device can only be occupied by one patient (left) at a time. The device (center) will capture three views (face, tongue, and sublingual vein) from each patient once. A tuple of images from each view (right) forms an instance.

Fig. 3. Samples of the face, tongue, and sublingual vein. (a) Face, (b) Tongue, and (c) Sublingual Vein. The yellow dashed lines in (c) is the Plica Fimbriata.

Table 1
The description of the image format in the dataset.

Resolution	Format	Bit depth	Color model
576 × 768	Bitmap	24	RGB

3. Methodology

In this section, we present our automatic multi-view disease detection system and our proposed **C**ollective **D**eep **R**egion-based **F**eature **R**epresentation (CDRFR) method. The system takes a multi-view instance (defined in Section 2.1) as input, then outputs the predicted label (specific disease or healthy) at the end. To extract features effectively as well as completely, CDRFR is proposed and utilized within the system. In the following part, we first introduce the overall architecture of the disease detection system. Then, we explain our proposed CDRFR in detail. Lastly, a classification algorithm used for disease detection method is presented.

The proposed CDRFR is composed of two general procedures: (1) Deep Region-based Feature Representation, (2) Coding. We used the Deep Region-based Feature Representation to extract effective representations through a sequence of deep neural networks with different functions and a series of feature extraction methods. The Coding procedure generates discriminative codes for each view of an individual respectively.

3.1. The overall architecture

We first provide an overall architecture to show our proposed system from a general perspective. As shown in Fig. 4, the whole method takes a multi-view instance of one patient as input and outputs the predicted class label in the end. From Fig. 4, there are two basic procedures in the proposed CDRFR method: Deep Region-based Feature Representation (center) and Coding (right), please refer to Fig. 4. In terms of a description, each multi-view instance contains images captured from the face, tongue, and sublingual vein, respectively. In the first part, the Deep Region-based Feature Representation is utilized to generate effective features according to a given multi-view instance (left). In the second part (Coding), multiple features from each view will be transformed into a coding vector via ProCRC. Since the coding vector integrates information from the three views, it is considered as a collective representation, which fuses their information by concatenating the highest probability class label from the three views. Finally, the coding vector is fed into the Decision Making unit (right hand side of Fig. 4) using the proposed classification algorithm in order to determine the predicted class label. Finally, the coding vector is fed into the Decision Making unit (right hand side of Fig. 4) using the proposed classification algorithm. Lastly, the results from each view are combined altogether to determine the predicted class label.

Specifically speaking, a Deep Region-based Feature Representation is composed of three sub-components, which are View Classification (left), Deep Region Segmentation (center), and Feature Representation (right). The View Classification sub-

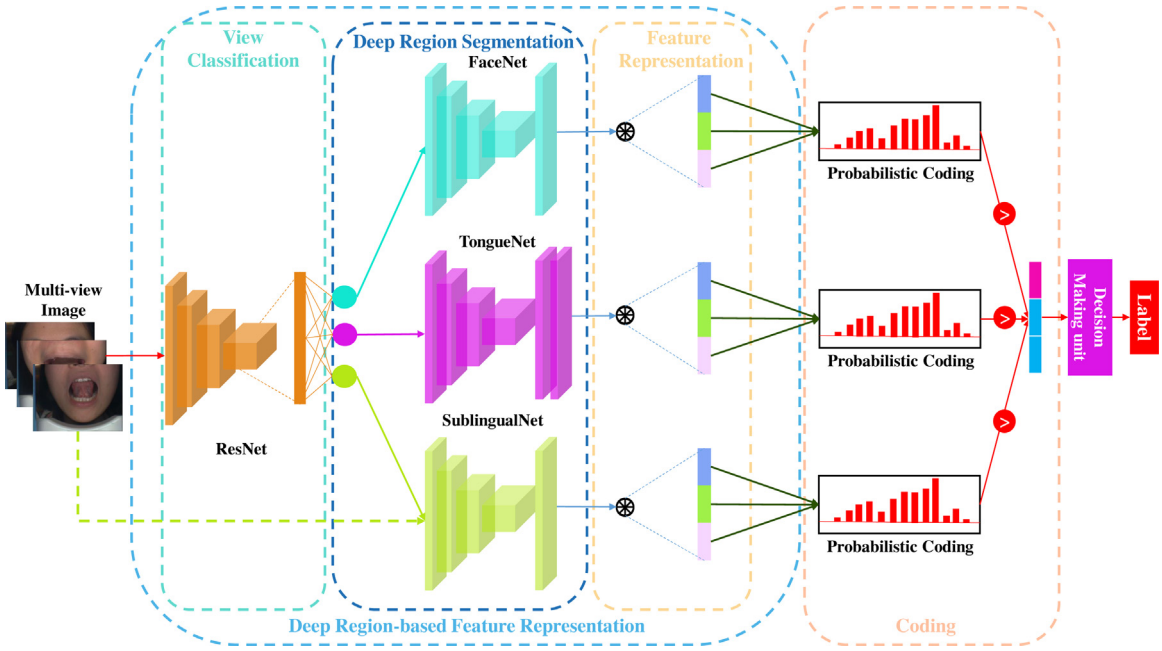


Fig. 4. An overview of the proposed collective deep region-based feature representation method. The left part is the deep region-based feature representation procedure, the right part is the coding procedure, and the right hand side illustrates the decision making unit.

component first classifies each image of the given image tuple and feeds it to the corresponding network for segmentation. By using this component, without the prior knowledge of view information, the system can automatically process an arbitrary image in the tuple, which reduces human errors to a large extent. Next, each segmentation network will generate a 'deep region' of corresponding views. After that, multiple features are extracted from this 'deep region' in the Feature Representation sub-component. At last, the Deep Region-based Feature Representation component will output a vector of features. In contrast, the Coding component is relative simple: three probabilistic coders which output 0 or 1 respectively are set in parallel waiting to be fed with features from the corresponding view. Then, a Decision Making unit for classification is placed after the probabilistic coders. More details about this unit are discussed in Section 3.3.

3.2. Collective Deep Region-based Feature Representation (CDRFR)

As mentioned above, there are two main components of CDRFR: (1) Deep Region-based Feature Representation and (2) Coding. In Deep Region-based Feature Representation, the architecture receiving images from each view will segment them via its corresponding networks to generate a new image only containing the deep region pixels in the first step. Next, multiple features will be extracted from the image of a 'deep region'. Here, three groups of features corresponding to the three views will be the final output, which are regarded as 'multi-view features'.

3.2.1. Deep region-based feature representation

As shown in Fig. 4, the Deep Region-based Feature Representation (center) is composed of three sub-components: (1) View Classification, (2) Deep Region Segmentation, and (3) Feature Representation. The details of the View Classification and Deep Region Segmentation are shown in Fig. 5 (along with a sample input) with all parameters marked beside its corresponding components. We can observe from this figure a 768×576 tongue image is the input (left), where the final output from this architecture is a well-segmented image preserving only the tongue region with a black background. The workflow shows the input tongue image is first fed to a ResNet for View Classification. To ensure the classification performance, here we utilize ResNet-18 [21], which is a powerful convolutional neural network in image classification. In addition, ResNet-18 is a lightweight deep residual neural network with a faster convergence.

To remove the influences of irrelevant noise and the background when extracting features, we performed the Deep Region Segmentation of each view. Here we term a set of key regions as the 'deep region', since these regions are obtained by its corresponding deep neural networks. After View Classification, a signal will be sent to the corresponding segmentation network according to the classification result. For example, the raw tongue image (shown in the left of Fig. 5) is transmitted to the TongueNet for deep region extraction (the green line) since it is classified as 'tongue' according to the result of the View Classification network. Afterwards, the TongueNet is activated in Fig. 5. Other networks in parallel such as FaceNet and SublingualNet remain dormant. In this proposed layout we used FCN-8s [24] as the main architecture of FaceNet, TongueNet, and SublingualNet since it has shown a satisfactory performance in the segmentation with key regions. We will show this later in Section 4.2. Each network is fine-tuned with its corresponding single-view images from the dataset, which can be considered as transfer learning. Noticeably, there is one extra layer at the end of TongueNet, which is termed the Morphological Processing layer to refine the tongue edge extraction. The effectiveness of the Morphological Processing layer is shown and discussed in [39]. The segmentation networks (FaceNet, TongueNet, and SublingualNet) with our pre-defined key region for each view is shown in Fig. 6. There are four key regions of the face: forehead (FH), nose bridge (NB), right cheek (RC), and left cheek (LC). According to TCM [32], the conditions of the different internal organs will be reflected on the corresponding regions of a human face. The partitioned regions we defined in Fig. 6(a) covers most of these areas. The tongue region includes the main body of the tongue [2] in the image as shown in Fig. 6(b). The sublingual vein is a pair of veins (colored as purple–blue) between the Plica Fimbriata on the lower surface of human tongue. Since the sublingual vein reflects the status of blood circulation through the appearance of the blood vessels [5],

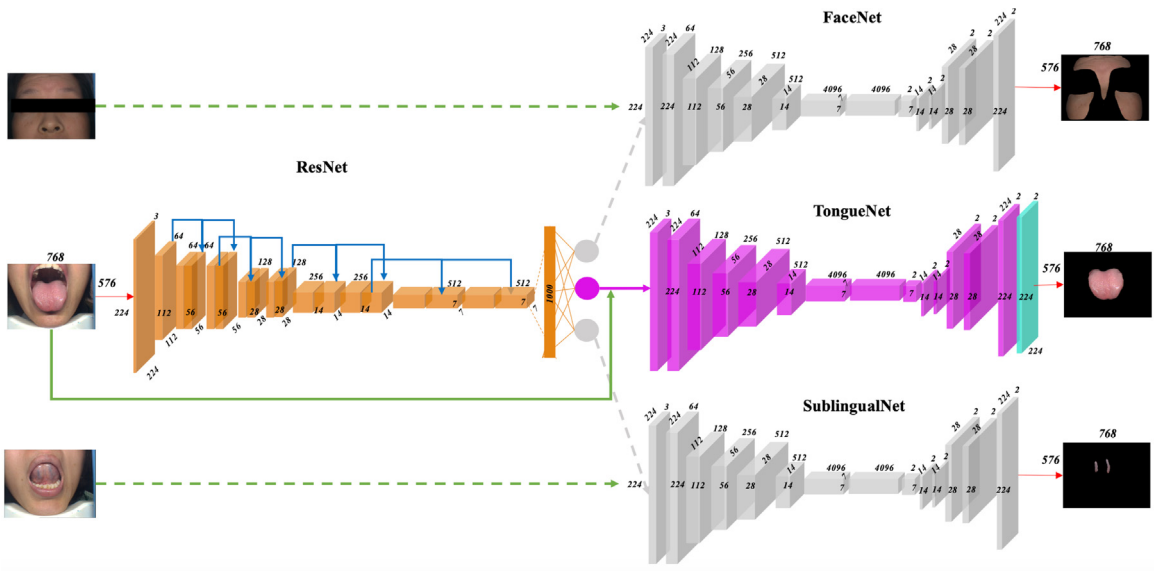


Fig. 5. The deep region-based feature representation. The solid green line represents the transmission of a raw input image to the corresponding network. The dotted green line indicates sending a single view image to the corresponding segmentation network, and the result (right hand side) is shown at the end of each segmentation network. Each block in FaceNet and SublingualNet contains three layers, while TongueNet has four layers: a Maxpooling layer, a Convolutional/Deconvolutional layer, and a ReLU layer. The last layer of TongueNet (aqua blue) is the Morphological Processing layer.

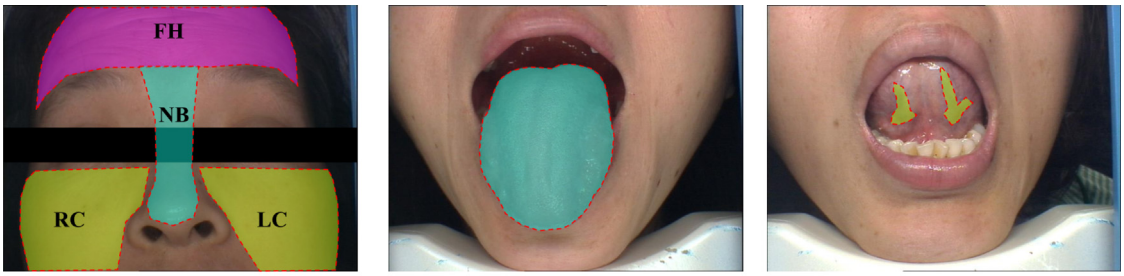


Fig. 6. The key region definition of three views: (a) face, (b) tongue, and (c) sublingual vein. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we defined this pair of blood vessels as the region to focus on (shown in Fig. 6(c))

After Deep Region Segmentation, we perform the Feature Representation. Specifically, we extract color and texture from the segmented image of each view. In addition, the geometry features are extracted for the face and tongue images. From these images, we focus on the patterns (e.g., luster, size, shape, roughness, etc.) of the different views to make a diagnosis. For the view images, there are three types of features that can be analyzed as the patterns to represent the appearance of the segmented regions: color, texture, and geometry. Researchers have examined these features for many years [2,3,34,38]. For example, color is a critical feature which carries plenty of valuable pathological information for disease detection [34]. Based on this idea, researchers made successful implementations that detected diseases through the color features [2,3,5]. For the texture features, Shu et al. [7] proved that it is helpful in disease detection. Finally, in regards to the geometry features, [36,40] showed its importance in detecting diabetes mellitus and other illnesses. Given an image I, we extract the following color features in each channel:

$$
\begin{aligned}
H_c &= \left\{\delta(x) \,|\, 1 \le x \le 255 \text{ and } x \in Z^+\right\} \\
\lambda_c &= \frac{1}{N}\sum_{i=1}^{N} v_{c,i} \\
\varphi_c &= \left[\frac{1}{N}\sum_{j=1}^{N}\left(v_{c,j}\right)^2\right]^{\frac{1}{2}} \\
s_c &= \left[\frac{1}{N}\sum_{j=1}^{N}\left(v_{c,j}\right)^3\right]^{\frac{1}{3}}
\end{aligned}
\tag{1}
$$

where H_c is a color histogram vector of channel c and $\delta(\cdot)$ calculates the frequency of pixel x in channel c. λ_c, φ_c, and s_c are the Color Moment [41], which represents the mean, variance, and skewness of color in channel c. Next, we define the color descriptor of a given image I as:

$$
A = \{H_R, H_G, H_B, \lambda_R, \lambda_G, \lambda_B, \varphi_R, \varphi_G, \varphi_B, s_R, s_G, s_B\} \tag{2}
$$

where R, G, B represents the red channel, green channel, and blue channel of image I, respectively.

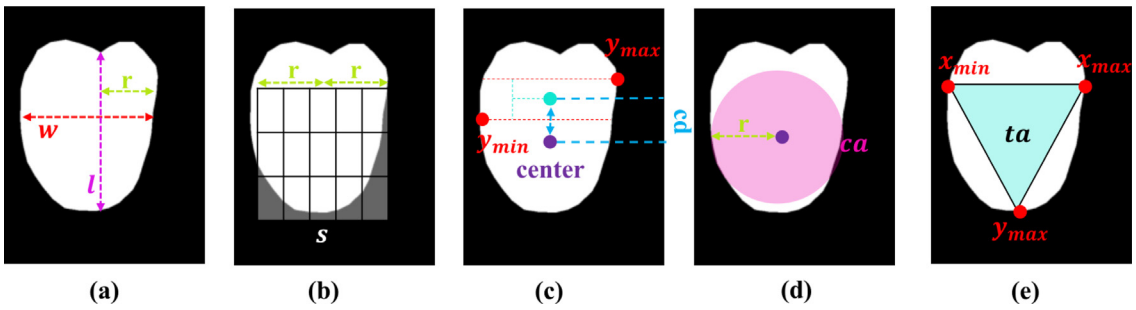


Fig. 7. The geometry features of a tongue image. (a) length (l) and width (w), (b) square area (s), (c) center distance (cd), (d) circle area (ca), and (e) triangle area (ta).

To extract the texture feature, we use an array of 2-D Gabor filters W as the descriptor. The 2-D Gabor filter [42] can be described as follows:

$$G(x, y) = \gamma exp\left\{-\frac{x^2 + y^2}{2\delta^2}\right\} cos\left(2\pi f\left(xcos\theta + ysin\theta\right)\right) \tag{3}$$

where γ is the normalizing factor, θ is the orientation of the observation, and f is the frequency. To extract the texture feature comprehensively, we constructed a Gabor filter array consisting of filters with varying wavelengths and frequencies and created the texture descriptor as follows:

$$\Upsilon = \left\{\frac{1}{N}\sum G_{\theta_i, f_i} | \theta \in \Theta, f \in \psi\right\} \tag{4}$$

where G_{θ_i, f_i} is the response of using different orientations and frequencies, and θ_i and f_i are the ith element in the orientation set Θ and frequency set ψ.

Besides the color and texture features, we also extract the geometry features from the face and tongue. Specifically speaking, 12 features in total including width (w), length (l), length–width ratio (r), smaller half-distance (shd), center distance (cd), center distance ratio (cdr), area (a), circle area (ca), square area ratio (car), square area (S), square area ratio (sar), and triangle area (ta) were calculated from the face and tongue. Fig. 7 visualizes some of these geometry features on a tongue image. Similarly, we extracted these features on the face images as well. More details can be found in [3]. Below we define the geometry features descriptor as follows:

$$\Gamma = \{w, l, r, shd, cd, cdr, a, ca, car, s, sar, ta\} \tag{5}$$

Then, all the features from each view are concatenated together and normalized. The new overall descriptor of each view can be described as follows:

$$\tilde{f_k} = \{A, \Upsilon, \Gamma\}$$
$$f_k = \frac{\tilde{f_k} - \frac{1}{N}\sum f_{k,j}}{max\left(\tilde{f_k}\right) - min\left(\tilde{f_k}\right)} \tag{6}$$

where f_k is the normalized feature vector of the kth view, A, Υ, and Γ are the color, texture, and geometry features in Eqs. (2), (4) and (5), respectively.

3.2.2. Coding

In order to perform comprehensive diagnosis based on the features from different views, we use a coding strategy to represent features from multiple views in a collective way. Here we define Ψ as the function mapping a feature vector from all views to a coding vector, which can be described in the following Eq. (7):

$$\Psi(\xi) : \xi \in \mathbb{R}^{n\times 3} \rightarrow \{c_1, c_2, c_3 | c_i \in \{0, 1\}\} \tag{7}$$

where c_i is the predicted class label of the ith view, $\xi \in \mathbb{R}^{n\times 3}$ is a matrix consisting of features extracted from the three views, whose columns correspond to one view. The output of function $\Psi(\cdot)$ is a tuple only containing two values: 0 or 1. To encode each column, we applied techniques termed as Probabilistic Coding to get the highest probability class label (shown as the right part of Fig. 4), inspired from a classifier named ProCRC [43]. For a given sample ξ, the probabilistic encoder E_i encodes the ith view in ξ with the following equation:

$$c_i = argmax_k \prod P\left(s\left(\xi_i\right) = k\right) \propto argmax_k\ exp\left(-\left(\|\xi_i - D\alpha_i\|_2^2 + \lambda\|\alpha_i\|_2^2 + \frac{\eta}{2}\sum\left\|D\alpha - D_j\alpha_i^j\right\|_2^2\right)\right) \tag{8}$$

where k is the class label, and $s(\cdot)$ represents the function mapping ξ_i to a class label. D is a dictionary consisting of all training samples. η is a scaling factor. α is the probabilistic collaborative coefficient, which can be obtained in Eq. (9) according to [43]:

$$\alpha_i = \left(D^T D + \frac{\eta}{2}\sum_{k=1}^{2}\left(\left(D_k'\right)^T D_k' + \lambda I\right)^{-1} D^T\right)\xi_i \tag{9}$$

where $D' = D - D_k$ represents the difference between dictionary D and the dictionary of subspace k, D_k.

3.3. Decision making unit

Given the features of a sample ξ, the classification result is determined as shown in the Eq. (10):

$$identity(\xi) = argmax_i \rho\left(\Psi(\xi)\right) \tag{10}$$

where $\rho(\cdot)$ calculates the frequencies of each element in the vector $\Psi(\xi)$, where $\Psi(\xi)$ represents the coding strategy shown in Section 3.2.2. The final decision is made according to the max-voting strategy.

Overall, we summarize the CDRFR-based classification for disease diagnosis in Algorithm 1.

4. Experimental results and discussion

Experiments were performed to evaluate the effectiveness and efficiency of the proposed diagnostic system based on the proposed method (CDRFR). First, we describe the experimental settings, including data preparation, experimental environment, as well as the evaluation metrics utilized. Next, we show the performance of Deep Region Segmentation in each view. Afterwards, comparisons between the CDRFR and five classifiers (KNN [44], LDA [45], SRC [46], CRC [47], and Random Forest [48]) fed with features based on the state-of-the-art feature extraction method for disease detection [7] are demonstrated on four diseases in total. Lastly, we compared the computation time of the proposed CDRFR and other methods.

Algorithm 1: CDRFR-based classification

Input: Feature matrix ξ_i, Dictionary D, scaling factor η, λ, number of views V

Output: predicted label L

$i = 1$;

$C = \{\emptyset\}$;

do

$\alpha_i = \left(D^T D + \frac{\eta}{2}\sum_{k=1}^{2}\left(\left(D_k'\right)^T D_k' + \lambda I\right)^{-1} D^T\right)\xi_i$;

$c_i = argmax_k exp\left(-\left(\|\xi_i - D\alpha_i\|_2^2 + \lambda\|\alpha_i\|_2^2 + \frac{1}{2}\left\|D\alpha_i - D_k\alpha_i^k\right\|\right)\right)$;

$C = C \cup c_i$;

$i = i + 1$;

while $i \leq V$

$L = argmax_i \rho(C)$

Table 2
The description of data for the four diseases and healthy control group.

Class	Number of samples	Number of training samples	Number of testing samples
Breast	103	50	53
Heart	86	43	43
Liver	200	100	100
Lung	100	50	50
Healthy	180	90	90

4.1. Experimental settings

4.1.1. Data preparation and experimental environment

In this paper, we perform binary disease detection among five groups of multi-view images, which are malignant breast tumor (Breast), heart disease (Heart), fatty liver (Liver), malignant lung tumor (Lung), and a healthy control group (Healthy). We placed instances of different classes altogether and performed binary classification to assign the predicted class label as the diagnostic result. The number of samples in each class are listed in Table 2. We will first take experiments on healthy control group versus each disease respectively, and them perform experiments on each disease versus other one. Each instance (refer to Section 2.2) is composed of one facial image, one tongue image, and one sublingual vein image from a patient. We placed instances of different classes altogether and performed binary classification to assign the predicted class label as the diagnostic result. The number of samples in each class are listed in Table 2:

For the training and testing in View Classification and Deep Region Segmentation, we randomly selected the training samples and testing samples with a ratio of 7 (training):3 (testing) from a subset of the medical dataset. The size of the subset is 480 images in total, evenly distributed in three views.

For disease detection, there are 489 samples from all diseases and 180 samples of healthy control group in the experiments, which equates to 2007 images (669 samples $\times$ 3 views) in total. Also, we divided the dataset into training and testing sets with a ratio of 1 (training):1 (testing) for classification while keeping the original size of the image (768 $\times$ 576) resolution as mentioned in Section 2.2.

All experiments were conducted on a PC with a 3.40 GHz Intel Core CPU and 16.0 GB RAM. In addition, we used the GEFORCE GTX 1070 Ti graphic card to train ResNet and FCNs in the architecture.

4.1.2. Evaluation metrics

To sufficiently evaluate the performance, we utilized different metrics for the different stages in the system. For the Deep Region Segmentation, the pixel accuracy (PA), weighted IOU (WIOU), and mean BF score (MBF) [49] were adopted. As for disease detection, Accuracy, Precision, Recall, and F-score were adopted. Metrics mentioned above are described as follows:

Given a predicted mask I from a segmentation output and a ground truth mask GT, we denote I_{TP} as the correctly predicted pixels. The pixel accuracy (PA) measures the correctly predicted pixels over several pixels of I, which provides a comprehensive insight into the Deep Region Segmentation performance. The PA is described as follows:

$$PA = \frac{|I_{TP}|}{|I|} \tag{11}$$

where $|I_{TP}|$ and $|I|$ represent the number of pixels in I_{TP} and the number of pixels in I, respectively.

The weighted IOU (WIOU) is an average IOU (Intersection over Union) metric weighted by the number of pixels from each class:

$$WIOU = \sum_{k=1}^{K}\frac{|I_k|}{|I|}\frac{\left|GT \bigcap IP\right|}{|GT| + |IP| - \left|GT \bigcap IP\right|} \tag{12}$$

where K is the number of classes and IP is the number of positive pixels in I. The WIOU reflects more on the pixels of the majority class.

The mean BF score (MBF) calculates the F-score of the predicted pixels of the boundary. Given B_I as the boundary map of I, B_{gt} as the boundary map of GT, MBF can be described as follows:

$$MBF = \frac{1}{K}\sum_{k=1}^{K}\frac{2 \cdot \frac{|I \bigcap B_{gt}|}{|B_I|} \cdot \frac{|I \bigcap B_I|}{|B_I|}}{\frac{|I \bigcap B_{gt}|}{|B_I|} + t\frac{|I \bigcap B_I|}{|B_I|}} \tag{13}$$

where K is the number of classes, Compared to PA, MBF emphasizes the performance of the boundary prediction.

Given a vector of prediction results V, we denote TP, FP, TN, and FN as the true positive, false positive, true negative, and false negative, respectively. Based on four indicators above, the Accuracy (Acc), Precision (P), Recall (R), and F-score (F) are defined as follows:

The Accuracy measures the number of correctly predicted samples over the number of all samples in the testing set.

$$Acc = \frac{TP + TN}{|TP + FP + FN + TN|} \tag{14}$$

The Precision measures the number of samples correctly predicted as positive over the number of all samples labeled as positive.

$$P = \frac{TP}{TP + FP} \tag{15}$$

The Recall measures the number of samples correctly predicted as positive over the number of all samples predicted correctly.

$$R = \frac{TP}{TP + TN} \tag{16}$$

The F-score is a weighted harmonic average between Precision and Recall, which comprehensively measures the performance in class-specific and general perspectives.

$$F = 2 \cdot \frac{P \cdot R}{P + R} = \frac{2TP}{2TP + FN + FP} \tag{17}$$

4.2. Results of region based feature extraction

In order to ensure the quality of Deep Region Segmentation and feature representation, the accuracy of View Classification should be close to 100% as much as possible. Table 3 shows the

Table 3
Comparison of ResNet and other classifiers in view classification.

Classifier	Accuracy (%)	Computation time (s)
KNN	77.71	7.08
SRC	66.67	313.13
CRC	64.62	123.7
VGG16	99.54	10.42
VGG19	**100**	10.8
AlexNet	99.09	14.25
ResNet	**100**	9.7

Table 4
Deep Region Segmentation performance of face, tongue, and sublingual vein.

View	PA (%)	WIOU (%)	MBF
Face	92.59	85.97	68
Tongue	99.04	98.11	92.51
Sublingual vein	99.18	98.60	79.21

comparison between ResNet and six other classifiers, including KNN, SRC, CRC, VGG16 [20], VGG19 [20], and Alexnet [19]. The highest value in Accuracy is marked in bold font. From Table 3, ResNet had a relatively lower computation time (9.701 s) and achieved an accuracy of 100%, which shows the superiority both on effectiveness and efficiency over other classifiers.

Next, we evaluated the Deep Region Segmentation performance using three metrics introduced in Section 4.1.2, which are pixel accuracy (PA), weighted intersection over union (WIOU), and mean BF-score (MBF). The performance of face, tongue and sublingual vein images are shown in Table 4. Besides the evaluation in a quantitative way, we present the qualitative evaluation in Fig. 8. First, from Table 4, we can see that all PA values of the three views are over 90%, indicating that the performance of pixel prediction is promising. The WIOU of the tongue and sublingual vein are 98.11% and 98.60%, representing the outstanding performance of pixel prediction in the class with a majority number of pixels. The highest MBF of three views is tongue (92.51%), which proves that the Morphological Processing layer added at the end of TongueNet is effective. Although the MBF value of sublingual vein is relatively low, its high PA and WIOU values reflect the superiority in other aspects. Noticeably, the WIOU and MBF of the face are low in contrast to tongue, and sublingual vein since the prediction area of the face is much larger than the other two. What is more, the boundary of the face region is unclear sometimes, unlike the samples of the tongue and sublingual vein images. In other words, we put more emphasis on the 'region' of the face images rather than the 'boundary' since we intend to extract features only from the pixels on a large area of the face instead of extracting features like shape or space. The Deep Region Segmentation results of face, tongue, and sublingual vein are shown in Fig. 8(a)–(l).

From Fig. 8(a)–(d) (the Deep Region Segmentation results of face), it is clear that the FaceNet is able to segment four basic face regions (FH, NB, RC, and LC introduced in Fig. 6(a)) precisely. No matter how the surroundings change, the predicted pixels did not contain any irrelevant areas (such as the eyes, eyebrows, etc.). For example, in Fig. 8(c), the boundary of the FH region is a skin region between the hair and the eyebrows. The Pearson correlation indexes of Fig. 8(a) and (b) indicate an extremely high correlation between the segmentation result and ground truth. For the Deep Region Segmentation results of the tongue depicted in Fig. 8(e)–(h), the boundary of the segmented area (aqua blue) fits well along the verge of the tongue. In spite of the varying size, shape, and surroundings, TongueNet still predicted precisely, especially for Figs. 8(g)–(h). All the results of Figs. 8(e)–(h) have high Pearson correlation coefficients (>0.8), implying an exceptionally extremely high correlation between the segmentation results and ground truths. As for the Deep Region Segmentation results of sublingual vein illustrated in Figs. 8(i)–(l), the majority of pixels belong to blood vessels, indicating a good performance in sublingual vein Deep Region Segmentation. The MSE of this view is also relatively low (0.008), meaning the small difference between the segmentation result and ground truths.

4.3. Results of disease detection

We first show experimental results on healthy control group versus one disease listed in Table 2, and then show results on one disease versus other one disease. We performed experiments on disease detection using four diseases listed in Table 2. To ensure the results are convincing and statistically reliable, we ran each experiment 30 times with random samplings and calculated the average value as the final result. Besides this, the confidence interval was also computed for each result, where Accuracy, Precision, Recall, and F1-score introduced in Section 4.1.2 were employed. We compared our proposed CDRFR method with five other classifiers fed with facial and tongue block features based on the feature extraction method of Shu et al. [1,7]. Furthermore, we appended the sublingual vein features (refer to Section 3.2) to the facial and tongue block features. This allows us to compare with three views. A 6-layers CNN architecture was applied in the comparisons. We show the results under the optimal parameter setting of each method in Tables 5–8 and Figs. 9–11. In the following subsections we conducted two groups of experiments. The first group is a healthy control group versus a single disease. The other group is a disease versus another disease.

4.3.1. Results of healthy versus a disease

The detection results of healthy control group (Healthy) versus four diseases, including breast tumor (Breast), heart disease (Heart), fatty liver (Liver), and malignant lung tumor (Lung) are shown in Tables 5–8. The highest value in each metric is marked in bold font. From Tables 5–8, it is clear that the proposed CDRFR outperforms the other six classifiers including the CNN in all metrics and detection tasks. With the varying of diseases ranging from breast tumor (Breast) to lung tumor (Lung) shown in the four tables, our proposed method showed good generalization. Specifically, the proposed method achieved over 95% on both Accuracy and F-score. In addition, the confidence interval of the proposed method is relatively small, which indicates strong robustness.

Fig. 9 shows boxplots of the error rates on the four diseases. We can observe intuitively that our proposed method (cyan) has clearly the lowest error rate compared with the other six classifiers. Besides this, its standard deviations are relatively low (e.g., in the Healthy vs. Heart group), indicating that CDRFR is stable.

To further prove the superiority of the proposed method, the ROC curves of the four detection groups are shown in Fig. 10. Generally speaking, the four sub-figures in Fig. 10. shows that the proposed method surpasses the other classifiers under most circumstances. In particular, CDRFR covers all other classifiers in Fig. 10(b), (c). Although Random Forest shows a competitive performance in Fig. 10(a), (d), the overall area of CDRFR is larger.

4.3.2. Results of disease versus a disease

Next, we tested our proposed method on detection between diseases, to further evaluate its performance. There are 6 groups of comparisons in this experiment, which covers all combination of the four diseases. The results and comparisons are shown in Fig. 11(a)–11(d).

According to Fig. 11(a)–11(d), the proposed method achieved the best results compared with other classifiers fed with the

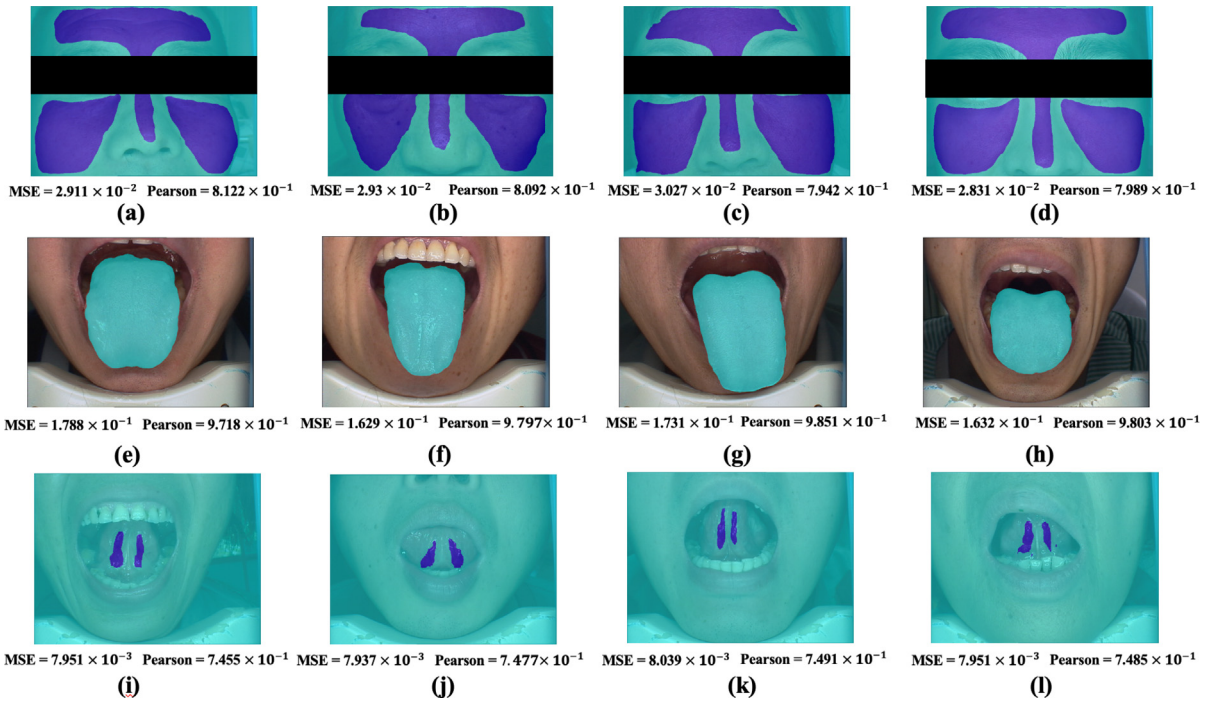


Fig. 8. Deep Region Segmentation results of the face, tongue, and sublingual images. (a)–(d) Deep Region Segmentation results of the face. (e)–(h) Deep Region Segmentation results of the tongue. and (i)–(l) Deep Region Segmentation results of the sublingual vein. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Results and comparisons of Healthy vs. Breast.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-score
KNN	64.83 ± 9.80	64.82 ± 8.85	65.20 ± 9.16	65.01 ± 9.01
LDA	84.38 ± 3.77	84.38 ± 8.85	84.73 ± 3.50	84.56 ± 3.63
Random Forest	82.59 ± 4.65	82.59 ± 4.65	82.96 ± 4.69	82.78 ± 4.66
CRC	82.25 ± 3.38	75.65 ± 4.65	76.38 ± 5.69	76.01 ± 5.88
SRC	74.95 ± 5.03	74.33 ± 3.90	73.39 ± 3.94	73.34 ± 3.92
CNN	80.69 ± 3.23	81.22 ± 4.56	81.49 ± 2.77	81.36 ± 3.87
CDRFR	**95.8 ± 0.76**	**96.34 ± 1.35**	**95.56 ± 2.09**	**95.93 ± 1.29**

Table 6
Results and comparisons of Healthy vs. Heart.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-score
KNN	66.61 ± 8.35	66.61 ± 8.35	66.97 ± 8.65	66.79 ± 8.49
LDA	78.23 ± 2.75	78.23 ± 2.61	78.34 ± 2.77	78.28 ± 2.76
Random Forest	82.59 ± 2.87	75.04 ± 2.72	75.16 ± 2.88	75.10 ± 2.87
CRC	73.92 ± 6.44	73.92 ± 4.17	74.60 ± 2.88	74.26 ± 2.87
SRC	74.33 ± 2.90	75.33 ± 1.89	76.40 ± 0.93	73.34 ± 2.91
CNN	83.78 ± 2.58	84.13 ± 3.79	84.09 ± 2.86	84.11 ± 3.13
CDRFR	**96.49 ± 0.12**	**94.49 ± 1.33**	**95.00 ± 1.44**	**95.70 ± 1.79**

Table 7
Results and comparisons of Healthy vs. Liver.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-score
KNN	69.77 ± 3.99	72.10 ± 2.50	70.17 ± 4.29	69.97 ± 4.12
LDA	83.21 ± 2.11	83.21 ± 2.61	83.28 ± 2.11	83.25 ± 2.11
Random Forest	78.84 ± 2.02	78.84 ± 2.02	78.89 ± 2.03	78.87 ± 2.02
CRC	79.56 ± 2.46	75.17 ± 5.53	75.89 ± 5.13	75.53 ± 5.32
SRC	73.80 ± 2.77	75.33 ± 1.90	74.40 ± 1.01	73.53 ± 3.72
CNN	85.69 ± 1.23	85.78 ± 2.56	86.32 ± 1.77	86.05 ± 1.87
CDRFR	**96.00 ± 1.76**	**96.34 ± 2.35**	**94.06 ± 1.59**	**96.00 ± 2.29**

state-of-the-art feature extraction method. Furthermore, values in all metrics are over 90%, showing its good quality in disease detection. We can observe that our proposed method (light blue) has outperformed CNN classifier (dark blue) on four metrics. The

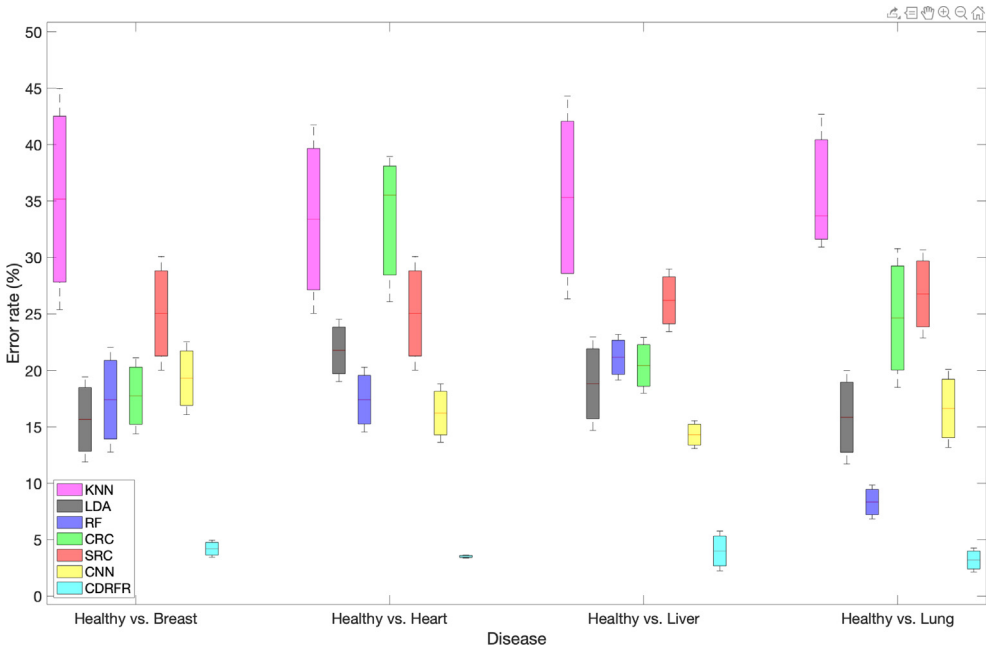


Fig. 9. Error rates of the different classifiers for the four diseases detections.

Table 8
Results and comparisons of Healthy vs. Lung.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-score
KNN	66.31 ± 8.99	72.11 ± 8.85	65.20 ± 2.75	72.37 ± 2.55
LDA	84.15 ± 4.14	83.15 ± 8.85	84.73 ± 1.57	84.15 ± 3.57
Random Forest	91.66 ± 1.50	82.59 ± 4.65	82.96 ± 3.05	90.66 ± 2.95
CRC	75.36 ± 6.14	71.95 ± 4.65	72.94 ± 3.07	82.38 ± 3.47
SRC	73.23 ± 3.90	73.63 ± 3.60	72.93 ± 3.96	76.48 ± 4.27
CNN	83.37 ± 3.46	84.27 ± 3.66	84.98 ± 2.84	84.36 ± 3.13
CDRFR	**96.8 ± 1.06**	**96.55 ± 2.35**	**95.56 ± 2.65**	**95.71 ± 1.40**

Table 9
Computation time comparisons.

Classifiers	Healthy vs. Breast	Healthy vs. Heart	Healthy vs. Liver	Healthy vs. Lung
KNN	0.0038	0.0065	0.0113	0.0073
LDA	0.0195	0.0151	0.0188	0.0162
Random Forest	1.7941	1.7695	1.8592	1.7952
CRC	0.0315	0.0279	0.0528	0.0307
SRC	0.1541	0.1419	0.3099	0.1451
CDRFR	**0.0031**	**0.003**	**0.0046**	**0.0033**

large gaps between our proposed method (light blue) and other classifiers highlights the superiority of CDRFR.

4.4. Efficiency evaluation

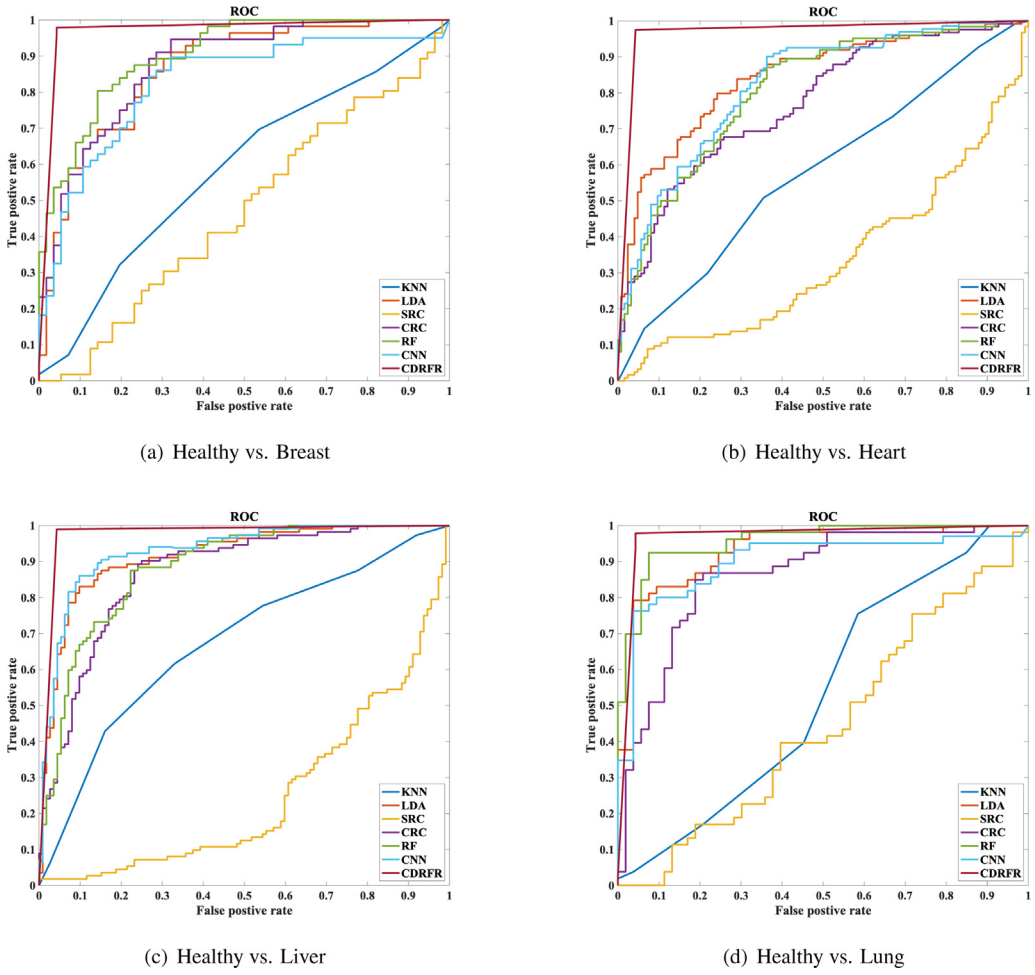
In addition to the quality of disease detection, the computational complexity is critical to a disease detection system. We recorded the computation time during each diagnosis and compared our proposed method with the other five classifiers mentioned above. The comparison results are shown in the Table 9. The least computation time consumed from each column is marked in bold font.

From Table 9, it can be observed that our proposed CDRFR consumes less time in different detection tasks. The fastest speed of detection is 0.003 s for Healthy vs. Heart, which is faster than other methods by about 0.39 s on average. Among the detection tasks, the longest computation time is Heart vs. Liver, which took 0.0046 s. In this case, the proposed system consumes 0.45 s less on average compared to other methods. Overall, Table 9 shows the high efficiency of the system.

4.5. Ablation study

We performed various ablation studies to demonstrate the contribution(s) of each operation. Besides this, we added VGG16 and AlexNet as deep learning-based classification model in comparison. Here, we designed four experiments to show: (1) the effectiveness of applying multi-view disease detection, (2) the effectiveness of applying the combined features (color, texture, and geometry), (3) the effectiveness of applying late fusion as the multi-view learning strategy, (4) the effectiveness of our method compared with deep learning models. Correspondingly, we defined the following 14 configurations of our proposed CDRFR method:

① CDRFRv1: CDRFR using only a single view of the face with multiple features.
② CDRFRv2: CDRFR using only a single view of the tongue with multiple features.
③ CDRFRv3: CDRFR using only a single view of the sublingual vein with multiple features.
④ CDRFRv4: CDRFR using the face and tongue views with multiple features.
⑤ CDRFRv5: CDRFR using the face and sublingual vein views with multiple features.
⑥ CDRFRv6: CDRFR using the tongue and sublingual vein views with multiple features.
⑦ CDRFRv7: CDRFR using multiple views with only the color feature.



(a) Healthy vs. Breast

(b) Healthy vs. Heart

(c) Healthy vs. Liver

(d) Healthy vs. Lung

Fig. 10. ROC curves of Healthy vs. a Disease.

⑧ CDRFRv8: CDRFR using multiple views with only the texture feature.
⑨ CDRFRv9: CDRFR using multiple views with only the geometry feature.
⑩ CDRFRv10: CDRFR using multiple views with the color and texture features.
⑪ CDRFRv11: CDRFR using multiple views with the color and geometry features.
⑫ CDRFRv12: CDRFR using multiple views with the texture and geometry features.
⑬ CDRFRv13: CDRFR using multiple views fused with early fusion [31] applying multiple features.
⑭ CDRFRv14: CDRFR using multiple views fused with late fusion [30] applying multiple features. Please note this is the configuration of the proposed method.

To investigate these configurations more distinctly, we have summarized them in the Table 10:

The results of four experiments are shown as follows:

(1) The effectiveness of applying multi-view disease detection

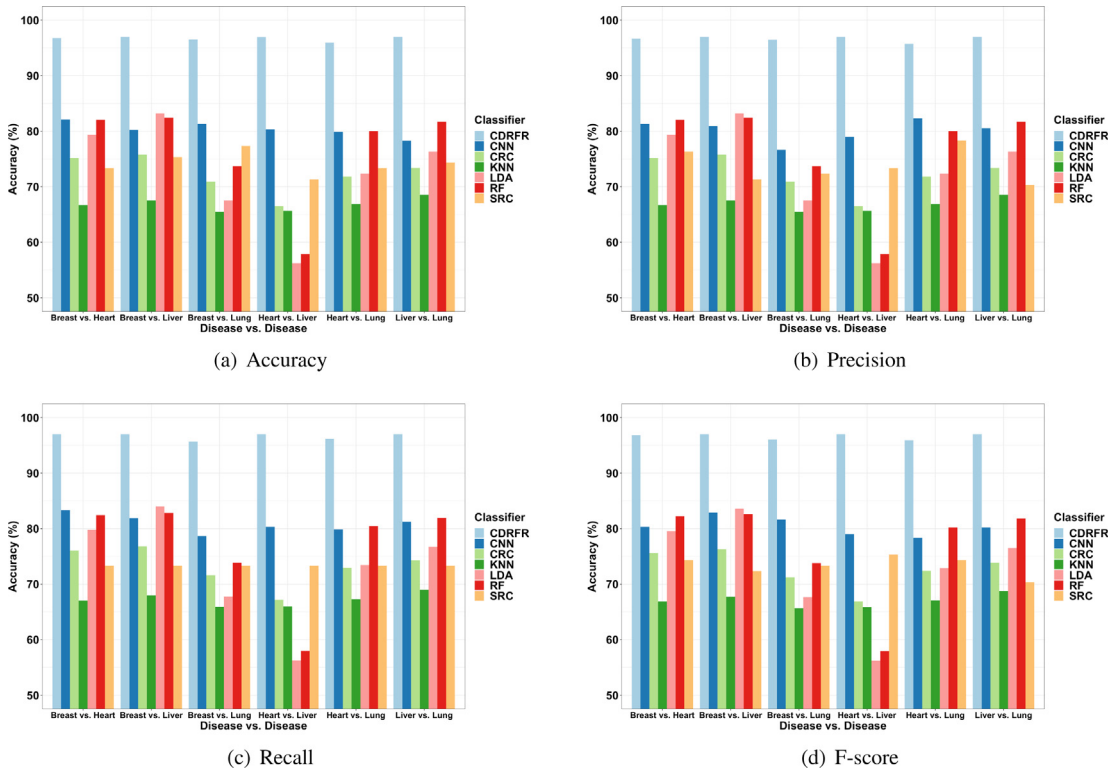
To prove the effectiveness of applying multi-view disease detection, we conducted an ablation study by changing the number of views used and the combination of views in the proposed CDRFR method. Table 3 shows the accuracy comparison among the proposed method using a single view (CDRFRv1, CDRFRv2, and CDRFRv3), CDRFR using two views (CDRFRv4, CDRFRv5, and CDRFRv6), and CDRFR (CDRFRv14 using multiple views).

Observing Table 11, CDRFR using multiple views (CDRFR14) outperformed the other configurations on all four disease detection tasks. Compared to a single view (CDRFRv1, CDRFRv2, and CDRFRv3), CDRFR is 2.21%, 3.39%, 2.63%, 3.65% higher than the highest accuracy on breast, heart, fatty liver, and lung disease detection among these three configurations. Furthermore, compared to using two views (CDRFRv4, CDRFRv5, CDRFRv6), CDRFRv14 is 1.03%, 1.2%, 1.12%, 2.84% higher than the highest accuracy for these three formations. We can find that the highest accuracy increases when the number of views increase as well. This is typically due to the fact that different views can disambiguate mistakes made by the other view(s). The comparison in Table 11 indicates the advantage of multi-view disease detection.

(2) The effectiveness of applying the combined features (color, texture, and geometry)

To prove the effectiveness of applying the combined features in our method, we performed an ablation study by changing the number of features used and the combination of features in the proposed CDRFR method. Table 12 shows the accuracy comparison with the proposed method CDRFR using a single feature (CDRFRv7, CFRFRv8, and CDRFRv9), CDRFR using two features (CDRFRv10, CDRFRv11, and CDRFRv12), and CDRFR using multiple features (CDRFRv14).

As can be seen from Table 12, CDRFRv14 using the multiple features achieved a greater accuracy than the other configurations on all four disease detection tasks. When compared to a single view (CDRFR7, CDRFR8, and CDRFR9), CDRFRv14 was 2.03%, 1%, 0.28%, and 1.99% higher than the highest accuracy on breast,



(a) Accuracy (b) Precision

(c) Recall (d) F-score

Fig. 11. Performance and comparisons of Disease vs. a Disease. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 10
Configurations of the ablation studies.

Configuration	Component							
	Face	Tongue	Sublingual vein	Color	Texture	Geometry	Early fusion	Late fusion
CDRFRv1	√			√	√	√		√
CDRFRv2		√		√	√	√		√
CDRFRv3			√	√	√	√		√
CDRFRv4	√	√		√	√	√		√
CDRFRv5	√		√	√	√	√		√
CDRFRv6		√		√	√	√		√
CDRFRv7	√	√	√	√				√
CDRFRv8	√	√	√		√			√
CDRFRv9	√	√	√			√		√
CDRFRv10	√	√	√	√	√			√
CDRFRv11	√	√	√	√		√		√
CDRFRv12	√	√	√		√	√		√
CDRFRv13	√	√	√	√	√	√	√	
CDRFRv14	√	√	√	√	√	√		√

Table 11
The accuracy comparison among CDRFR using single view, two views, and multiple views for disease detection.

Classifiers	Healthy vs. Breast	Healthy vs. Heart	Healthy vs. Liver	Healthy vs. Lung.
CDRFRv1	93.59 ± 0.77	93.1 ± 1.05	93.23 ± 1.22	93.15 ± 0.84
CDRFRv2	92.16 ± 0.77	92.50 ± 1.04	92.61 ± 1.32	76.10 ± 0.73
CDRFRv3	83.41 ± 0.76	88.55 ± 1.05	93.37 ± 0.98	85.24 ± 0.71
CDRFRv4	94.77 ± 0.77	95.29 ± 1.05	94.88 ± 0.79	87.35 ± 1.21
CDRFRv5	89.44 ± 0.82	90.21 ± 1.05	94.53 ± 1.22	93.96 ± 0.84
CDRFRv6	90.58 ± 0.76	93.4 ± 0.96	94.31 ± 1.51	84.52 ± 1.32
CDRFRv14	**95.8 ± 0.76**	**96.49 ± 0.12**	**96.00 ± 1.06**	**96.8 ± 1.06**

heart, fatty liver, and lung disease detections, respectively. Moreover, compared to using two views (CDRFR10, CDRFR11, and CDRFR12), CDRFRv14 is 1.9%, 0.34%, 0.12%, and 1.08% higher than the highest accuracy among these three configurations, correspondingly. The results in Table 12 denote the benefits of applying multiple features in disease detection.

(3) The effectiveness of applying late fusion as multi-view learning strategy

To prove the effectiveness of applying the late fusion strategy in our method, we made a comparison with CDRFR using an early fusion strategy (CDRFRv13). Table 13 shows the accuracy comparison between CDRFR using early fusion strategy (CDRFRv13) and CDRFR using late fusion strategy (CDRFRv14).

Table 12
The accuracy comparison among CDRFR using single feature, two features, and multiple features for disease detection.

Classifiers	Healthy vs. Breast	Healthy vs. Heart	Healthy vs. Liver	Healthy vs. Lung.
CDRFRv7	93.77 ± 0.56	95.49 ± 1.22	95.72 ± 0.73	94.81 ± 1.06
CDRFRv8	87.8 ± 0.83	92.1 ± 0.75	93.74 ± 0.09	89.8 ± 0.08
CDRFRv9	78.14 ± 1.71	75.49 ± 1.23	73.01 ±	72.89 ± 1.08
CDRFRv10	93.89 ± 0.14	96.12 ± 0.93	95.88 ± 0.88	94.81 ± 0.37
CDRFRv11	91.24 ± 1.23	96.03 ± 0.35	95.72 ± 0.46	95.72 ± 0.77
CDRFRv12	84.32 ± 0.21	86.21 ± 0.23	82.3 ± 0.23	85.95 ± 0.34
CDRFRv14	**95.8 ± 0.76**	**96.49 ± 0.12**	**96.00 ± 1.06**	**96.8 ± 1.06**

Table 13
The accuracy comparison between CDRFR using early fusion strategy and CDRFR using late fusion strategy.

Classifiers	Healthy vs. Breast	Healthy vs. Heart	Healthy vs. Liver	Healthy vs. Lung.
CDRFRv13	94.98 ± 0.51	95.71 ± 0.86	94.88 ± 0.33	95.73 ± 0.29
CDRFRv14	**95.8 ± 0.76**	**96.49 ± 0.12**	**96.00 ± 1.06**	**96.8 ± 1.06**

Clearly as can be seen in Table 13, CDRFRv14 outperformed CDRFRv13 on all four disease detection tasks. Specifically, CDRFRv14 achieved improvements of 0.82%, 0.78%, 1.12%, and 1.07% on breast, heart, fatty liver, and lung disease detections, respectively. This proved that the late fusion strategy we applied in CDRFRv14 has an advantage over the early fusion strategy.

(4) The effectiveness of our method comparing with deep learning models

We compared our method with the deep learning-based classification method (VGG19 and AlexNet) in Table 14. In this experiment, we took the images of the different regions from the three views respectively as the training and testing images for the deep learning architectures. Before performing classification, each network was trained on the images based on the training and testing ratio introduced in Section 4.1.1. In Table 14, the view name (in parentheses) represents the input view of this deep learning architecture.

As shown in Table 14, CDRFR (CDRFRv14) had a better performance compared with VGG19 and AlexNet, which belongs to the deep learning-based classification method. Specifically speaking, CDRFRv14 is 11.08%, 15%, 11.58%, and 14.48% higher than the highest accuracy achieved by VGG19 on breast tumor, heart disease, fatty liver, malignant lung tumor detections, respectively. Furthermore, the accuracy of CDRFRv14 is 17.98%, 19.76%, 20.76%, and 16.36% higher than AlexNet for these disease detections, correspondingly. Here, the deep learning models achieved poorer performances when facing the relatively insufficient training data.

4.6. Robustness evaluation

we have made the experiments on image with different conditions to show the robustness of the proposed method. We designed 3 experiments: ① Disease detection with random noise, ② Disease detection with rotations, ③ Disease detection with block occlusion. The results are shown in the Fig. 12.

(1) Disease detection with random noise

To show the robustness of our proposed method against noise, we added salt and pepper noise to each image on three views respectively. The percentage of noisy pixels ranges from 0% to 80%. The first row of Fig. 12 corresponds to the experiment of disease detection with noisy images. Images to the left of the graphs illustrate the noisy face, tongue, and sublingual vein images. Three graphs (Fig. 12(a), (b), and (c)) show the results in detail on the four diseases. Obviously, with the increasing percentage of noisy pixels (0%–80%), the accuracy of detection decreased steadily. For the noisy face image, the accuracy dropped from 95.4% (breast), 96.49% (heart), 96% (liver), and 96.8% (lung) with 0% noise added to 90.02% (breast), 90.38% (heart), 88.3% (liver), and 91.73% (lung) with 80% noise added. The largest accuracy loss was liver disease, at 7.7%. For tongue, the accuracy decreased from 95.4% (breast), 96.49% (heart), 96% (liver), and 96.8% (lung) with 0% noise added to 90.13% (breast), 90.03% (heart), 91.83% (liver), and 91.21% (lung) with 80% noise added. In this case, 6.43% was the largest difference from heart disease. For the sublingual vein, the accuracy decreased from 95.4% (breast), 96.49% (heart), 96% (liver), and 96.8% (lung) with no noise added to 93.41% (breast), 94.89% (heart), 91.89% (liver), and 92.89% (lung) with 80% noise added. Here, the greatest accuracy drop was again liver disease, at 4.11%. We can observe that the noise on the sublingual vein brings less loss than the face and tongue. Generally speaking, except for the noisy face from liver disease, the proposed method ensured an accuracy above 90% even with 80% of added noise, indicating a strong robustness.

(2) Disease detection with rotations

In this experiment, we rotated the images of the face, tongue, and sublingual veins to show the robustness of the proposed method. The second row of Fig. 12 corresponds to the rotated images. Here, we rotated the images with four angles: 0°, 90°, 180°, and 270° on the face, tongue, and sublingual veins, respectively (refer to the left of the graphs on the second row). According to Fig. 12(d), (e), and (g), the accuracies of each view only have minor fluctuations, which indicates the performances are immune to the varying rotation conditions.

(3) Disease detection with block occlusions

To show the robustness (of the proposed method) against occlusions, we placed image blocks on the face, tongue, and sublingual vein images. We set five relative positions: upper-left ($I(0,0)$), upper-right($I(0,W)$), lower-left($I(L-h,0)$), lower-right($I(L-h,W-h)$), and center ($I((L-h)/2,(W-h)/2)$). The size of each block was $h = 128 \times 128$ pixels. The third row in Fig. 12 corresponds to the experiments of disease detection with block occlusions. The blocks are placed on the face, tongue, and sublingual vein images, respectively. The results in detail of the face, tongue, and sublingual veins are shown in Fig. 12(g), (h), (i), along with samples of its occlusion to the left. Clearly, the accuracies in each view change ever so slightly (if at all) with the varying positions of the blocks. Taking breast disease as an example, the accuracy of the face, tongue, and sublingual vein images with occlusions ranges from 95.25% to 95.4%, from 95.32% to 95.4%, and from 95.4% to 95.4% (no change), correspondingly. These results demonstrate that our proposed method has strong robustness against block occlusions.

4.7. Discussion

According to the results of the experiments in the previous subsections, we can discuss this work in the following 3 aspects:

Table 14
The effectiveness of our method comparing with deep learning models.

Classifiers	Healthy vs. Breast	Healthy vs. Heart	Healthy vs. Liver	Healthy vs. Lung.
VGG19 (face)	84.72 ± 0.84	79.33 ± 0.85	81.93 ± 0.33	82.32 ± 0.29
AlexNet (face)	77.82v0.51	76.73 ± 0.51	75.22 ± 0.52	80.44 ± 0.56
VGG19 (tongue)	80.36 ± 0.21	81.49 ± 0.88	84.42 ± 0.34	81.32 ± 0.29
AlexNet (tongue)	78.22 ± 0.39	79.41 ± 0.51	75.24 ± 0.45	78.05 ± 0.86
VGG19 (sublingual vein)	75.86 ± 0.24	74.23 ± 0.86	71.89 ± 0.33	70.21 ± 0.7
AlexNet (sublingual vein)	73.82 ± 0.84	72.58 ± 0.2	71.25 ± 0.42	69.44 ± 0.04
CDRFRv14	**95.8 ± 0.76**	**96.49 ± 0.12**	**96.00 ± 1.06**	**96.8 ± 1.06**

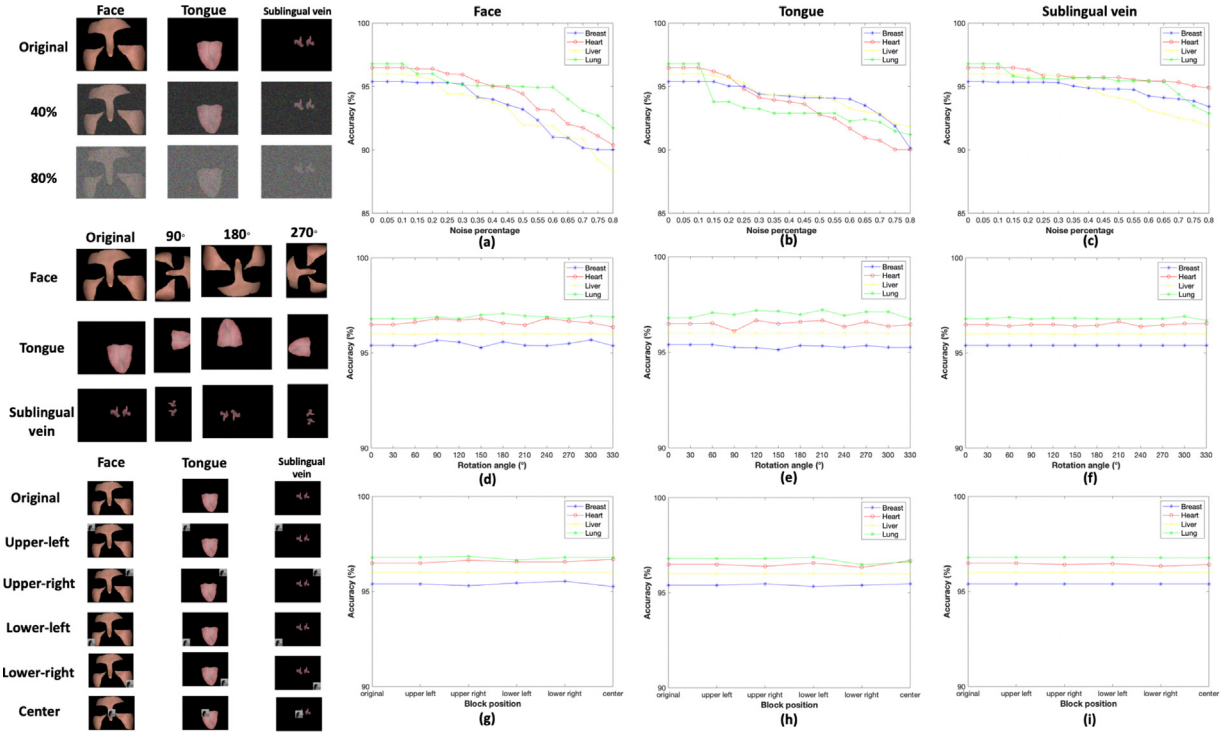


Fig. 12. Robustness evaluation on different image parameters. The first row shows the results of images with random noise. The second row shows the results of images with rotations. The third row shows the results of images with block occlusions.

(1) It is noticeable that our presented system shows its effectiveness as well as its efficiency in disease diagnosis according to Sections 4.3 and 4.4. Tables 5–8 shows our system achieved the best performance in all four groups of Healthy vs. a Disease under four metrics. What is more, the results of Disease vs. Disease in Fig. 11 also reflects the large gap with other classifiers. Meanwhile, our system takes little computation time ranging from 0.003 s (Healthy vs. Breast) to 0.0046 s (Healthy vs. Liver), which is 1.85 s less compared with the most time-consuming classifier. The best detection result we achieved is lung tumor detection, which had a 96.8%-Accuracy (96.55%-Precision, 95.55%-Recall, and 95.71%-F-score) with a 0.0033 s detection time.

(2) The proposed CDRFR provides a novel framework to fuse features extracted from different views for a given instance. It is the first time to use region-based features of different organs on the human body. Compared with the methods using local information (e.g., facial blocks) [7] and [3], the region-based features provide more information in feature extraction. Fig. 10 shows the superiority of the proposed method, whose ROC curves cover almost all classifiers fed with the features based on blocks in our experiments. Noticeably, compared with the ROC curves of the other classifiers, the proposed method holds better robustness as the slope of the true positive rate over the false positive rate is almost constant. Therefore, this indicates the high confidence of the proposed method when performing disease detection. The same conclusion can be reached in Tables 5–8 since it is observed that the confidence interval of the proposed CDRFR method (between 0 to 2) is smaller than other classifiers (between 0 to 9). In addition, both results from Tables 5–8 and Fig. 8 maintain the high values in four metrics, implying the good generalization of the proposed method.

(3) Transfer learning has proven to be suitable for our disease detection task. We implemented transfer learning to the View Classification and Deep Region Segmentation by using only 489 images from each view, which is a small subset of the whole dataset. The promising performances of the networks shown in Table 4 and Fig. 8 indicate that it has been a success using transfer learning in this application. The ResNet we utilized in the View Classification obtained a high accuracy, while maintaining a high efficiency (9.7s). For the Deep Region Segmentation, the FCNs achieved a high pixel accuracy on the face (92.59%), tongue (99.04%), and sublingual vein (99.18%), respectively. The Deep Region Segmentation quality shown in Fig. 8 proves that the networks of the three views segmented the regions successfully. On the other hand, information extraction was accomplished in a manual or semi-automatic way, such

as works presented by [1,13,50]. These methods usually require prior knowledge (e.g., experience from the medical professionals) from experts for each individual case, which cannot be integrated into an automatic diagnostic system. Furthermore, they are sensitive to the environment. If the medical profile is not captured in the same standard, it is hard to perform feature extraction. The proposed system solves these problems by using the Deep Region Segmentation (refer to Section 3.2.1), which ensures the pixels in key regions of corresponding views are considered in the feature extraction procedure.

(4) The multi-view fusion strategy and feature representation are effective to the disease diagnosis. The ablation studies in Section 4.5 shows both multiple views and features are beneficial to the disease detection. It also demonstrate the late fusion strategy is better than the early fusion strategy. This combination of components enables the proposed method to achieve the better disease detection performance. In addition, results in Section 4.6 indicate the proposed CDRFR with late fusion strategy and multiple hand-made features is robust to different image parameters. Therefore, the CDRFR is a powerful and robust method in disease detection.

5. Conclusion

In this paper, we presented an automatic disease detection system based on a multi-view instance captured from an individual. The system will output a class label (either a specific disease or healthy) via the analysis of a group of multi-view images from the face, tongue, and sublingual vein. To detect the correct output accurately and efficiently on the multi-view images, a **C**ollective **D**eep **R**egion-based **F**eature **R**epresentation method was proposed. CDRFR consists of Deep Region-based Feature Representation and Coding, where the output from the latter was fed to a unique classification algorithm. Extensive experiments on four diseases and one healthy control group demonstrated the effectiveness of the proposed system in terms of classification and computation with healthy control group versus a disease as well as a disease versus another disease.

As part of our future work, we will try to explore more effective representations from a novel perspective (e.g., graph networks) to perform disease detection. Meanwhile, we will try to improve our proposed system in real application scenarios.

CRediT authorship contribution statement

Jianhang Zhou: Conceptualization of this study, Methodology, Software, Writing - original draft. **Qi Zhang:** Conceptualization of this study, Software. **Bob Zhang:** Data curation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Ting Shu, Bob Zhang, Yuan Yan Tang, An improved noninvasive method to detect diabetes mellitus using the probabilistic collaborative representation based classifier, Inform. Sci. 467 (2018) 477–488.
[2] Bob Zhang, David Zhang, et al., Noninvasive diabetes mellitus detection using facial block color with a sparse representation classifier, IEEE Trans. Biomed. Eng. 61 (4) (2013) 1027–1033.
[3] Bob Zhang, B.V.K. Vijaya Kumar, David Zhang, Detecting diabetes mellitus and nonproliferative diabetic retinopathy using tongue color, texture, and geometry features, IEEE Trans. Biomed. Eng. 61 (2) (2013) 491–501.
[4] Jinxing Li, Bob Zhang, David Zhang, Joint discriminative and collaborative representation for fatty liver disease diagnosis, Expert Syst. Appl. 89 (2017) 31–40.
[5] Jinxing Li, Bob Zhang, Guangming Lu, Jane You, David Zhang, Body surface feature-based multi-modal learning for diabetes mellitus detection, Inform. Sci. 472 (2019) 1–14.
[6] Ting Shu, Bob Zhang, Yuan Tang, Novel noninvasive brain disease detection system using a facial image sensor, Sensors 17 (12) (2017) 2843.
[7] Ting Shu, Bob Zhang, Yuan Yan Tang, An extensive analysis of various texture feature extractors to detect diabetes mellitus using facial specific regions, Comput. Biol. Med. 83 (2017) 69–83.
[8] Jian Wu, Bob Zhang, Yong Xu, David Zhang, Tongue image alignment via conformal mapping for disease detection, IEEE Access (2019).
[9] Rafael C. Gonzalez, Paul Wintz, Digital image processing (Book), (13) Reading, Mass., Addison-Wesley Publishing Co., Inc. (Applied Mathematics and Computation, 1977, p. 451.
[10] David Y. Zhang, Medical Biometrics: First International Conference, ICMB 2008, Hong Kong, China, January 4-5, 2008, Proceedings, Vol. 4901, Springer, 2007.
[11] David Zhang, Dongmin Guo, Ke Yan, Breath Analysis for Medical Applications, Springer, 2017.
[12] Jianfeng Li, Jinhuan Shi, Hongzhi Zhang, Yanlai Li, Naimin Li, Changming Liu, Tongue image texture segmentation based on gabor filter plus normalized cut, in: International Conference on Medical Biometrics, Springer, 2010, pp. 115–125.
[13] Bo Pang, David Zhang, Kuanquan Wang, Tongue image analysis for appendicitis diagnosis, Inform. Sci. 175 (3) (2005) 160–176.
[14] Li Zhang, Bob Zhang, Non-invasive multi-disease classification via facial image analysis using a convolutional neural network, in: 2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), IEEE, 2018, pp. 66–71.
[15] Ting Shu, Bob Zhang, Yuan Yan Tang, Effective heart disease detection based on quantitative computerized traditional chinese medicine using representation based classifiers, Evid.-Based Complement. Altern. Med. 2017 (2017).
[16] Ke Yan, David Zhang, Darong Wu, Hua Wei, Guangming Lu, Design of a breath analysis system for diabetes screening and blood glucose level prediction, IEEE Trans. Biomed. Eng. 61 (11) (2014) 2787–2795.
[17] Mingjia Liu, Zhenhua Guo, Hepatitis diagnosis using facial color image, in: International Conference on Medical Biometrics, Springer, 2008, pp. 160–167.
[18] Xingzheng Wang, Bob Zhang, Zhimin Yang, Haoqian Wang, David Zhang, Statistical analysis of tongue images for feature extraction and diagnostics, IEEE Trans. Image Process. 22 (12) (2013) 5336–5347.
[19] Alex Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural, in: Neural Information Processing Systems, 2014, pp. 1–9.
[20] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[22] Forrest N. Iandola, Matthew W. Moskewicz, Sergey Karayev, Ross B. Girshick, Trevor Darrell, Kurt Keutzer, DenseNet: Implementing efficient convnet descriptor pyramids, 2014, ArXiv, abs/1404.1869.
[23] Lorien Y. Pratt, Jack Mostow, Candace A. Kamm, Ace A. Kamm, Direct transfer of learned information among neural networks, in: AAAI, Vol. 91, 1991, pp. 584–589.
[24] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[25] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-Net: Convolutional networks for biomedical image segmentation, 2015, ArXiv, abs/1505.04597.
[26] Ran D. Zilca, Yuval Bistritz, Feature concatenation for speaker identification, in: 2000 10th European Signal Processing Conference, 2000, pp. 1–4.
[27] Michael P. Perrone, Leon N. Cooper, When networks disagree: Ensemble methods for hybrid neural networks, 1992.
[28] Sokol Koço, Cécile Capponi, A boosting approach to multiview classification with cooperation, in: ECML/PKDD, 2011.
[29] Lynn Houthuys, Rocco Langone, Johan A.K. Suykens, Multi-view least squares support vector machines classification, Neurocomputing 282 (2018) 78–88.
[30] Weiran Wang, Raman Arora, Karen Livescu, Jeff A. Bilmes, On deep multi-view representation learning, in: ICML, 2015.
[31] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y. Ng, Multimodal deep learning, in: ICML, 2011.

[32] Lillian Bridges, Face Reading in Chinese Medicine, 2nd edn., 2012.
[33] B. Tomlinson, T.T.W. Chu, Traditional Chinese Medicine in the treatment of Diabetes, Springer, 2007.
[34] David Zhang, Hongzhi Zhang, Bob Zhang, Tongue image analysis, 2017.
[35] Megan McAuliffe, Elizabeth C. Ward, Bruce E. Murdoch, Speech production in parkinson's disease: I. an electropalatographic investigation of tonguepalate contact patterns, Clin. Linguist. Phonetics 20 (2006) 1–18.
[36] Brian V. Reamy, Richard Derby, Christopher W. Bunt, Common tongue conditions in primary care., Amer. Fam. Phys. 81 5 (2010) 627–634.
[37] Xingzheng Wang, David Zhang, An optimized tongue image color correction scheme, IEEE Trans. Inf. Technol. Biomed. 14 (6) (2010) 1355–1364.
[38] Chuang-Chien Chiu, Chen-Yen Lan, Yung-Hsien Chang, Objective assessment of blood stasis using computerized inspection of sublingual veins, Comput. Methods Programs Biomed. 69 (1) (2002) 1–12.
[39] Jianhang Zhou, Qi Zhang, Bob Zhang, Xiaojiao Chen, Tonguenet: A precise and fast tongue segmentation system using u-net with a morphological processing layer, Appl. Sci. 9 (15) (2019) 3128.
[40] Bob Zhang, Han Zhang, Significant geometry features in tongue image analysis, evidence-based complementary and alternative medicine, Evid.-Based Complement. Altern. Med. (2015) 897580.
[41] Markus Andreas Stricker, Markus Orengo, Similarity of color images, in: Storage and Retrieval for Image and Video Databases III, Vol. 2420, International Society for Optics and Photonics, 1995, pp. 381–392.
[42] Itzhak Fogel, Dov Sagi, Gabor filters as texture discriminator, Biol. Cybernet. 61 (2) (1989) 103–113.
[43] Sijia Cai, Lei Zhang, Wangmeng Zuo, Xiangchu Feng, A probabilistic collaborative representation based approach for pattern classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2950–2959.
[44] Naomi S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Amer. Statist. 46 (3) (1992) 175–185.
[45] Ronald A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[46] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, Yi Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2008) 210–227.
[47] Lei Zhang, Meng Yang, Xiangchu Feng, Sparse representation or collaborative representation: Which helps face recognition? in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 471–478.
[48] Tin Kam Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1, IEEE, 1995, pp. 278–282.
[49] Gabriela Csurka, Diane Larlus, Florent Perronnin, France Meylan, What is a good evaluation measure for semantic segmentation? in: BMVC, Vol. 27, Citeseer, 2013, p. 2013.
[50] Jifeng Ning, David Zhang, Chengke Wu, Feng Yue, Automatic tongue image segmentation based on gradient vector flow and region merging, Neural Comput. Appl. 21 (8) (2012) 1819–1826.

Jianhang Zhou received the B.S. degree in Computer Science from Nanjing Forestry University in 2018, the M.S. degree in Computer Science from University of Macau in 2020. He is currently pursuing the Ph.D. degree in Computer Science in the Department of Computer and Information Science, Faculty of Science and Technology at the University of Macau. His research interest includes pattern recognition, biometrics, deep learning for medical image processing, and computer vision.

Qi Zhang received his B.S. in Automation from Central South University of Forestry and Technology (CSUFT) 2015, Changsha, China, in 2015, a M.S. in Electrical and Computer Engineering from University of Macau, Macau, China, in 2018. He received the M.S. degree in computer science in the Department of Computer and Information Science, Faculty of Science and Technology at the University of Macau in 2020. He is currently pursuing the Ph.D. degree in Computer Science in the Department of Computer and Information Science, Faculty of Science and Technology at the University of Macau.

Bob Zhang received his B.A. in Computer Science from York University in 2006, a M.A.Sc. in Information Systems Security from Concordia University in 2007, and a Ph.D. in Electrical and Computer University from the University of Waterloo in 2011.

After graduating from Waterloo he remained with the Center for Pattern Recognition and Machine Intelligence, and later worked as a Post-Doctoral Researcher in the Department of Electrical and Computer Engineering at Carnegie Mellon University. Currently, he is an Associate Professor in the Department of Computer and Information Science at the University of Macau. His research interests focus on biometrics, pattern recognition, and image processing.