DOI:
10.1016/j.neucom.2021.02.049

Document Version
Peer reviewed version

A Case Study on Computer-Aided Diagnosis of Nonerosive Reflux Disease Using Deep Learning Techniques

Junkai Liao[a], Hak-Keung Lam[a,*], Guangyu Jia[a], Shraddha Gulati[b], Julius Bernth[c], Dmytro Poliyivets[a], Yujia Xu[a], Hongbin Liu[c], Bu Hayee[b]

[a]*Department of Engineering, King's College London, London, United Kingdom*

[b]*King's Institute of Therapeutic Endoscopy, King's College Hospital NHS Foundation Trust, London, United Kingdom*

[c]*School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom*

Abstract

This paper aims to develop deep-learning-based algorithms to automatically diagnose the nonerosive reflux disease (NERD) using the near focus narrow band imaging (NF-NBI) images, which are collected by clinicians of King's College Hospital. To diagnose this disease, we propose a deep learning classification system to distinguish the NF-NBI images captured in the esophagus of healthy people and the NERD patients, which is a binary classification of two classes: non-NERD and NERD. To achieve an effective and accurate classification, we first propose an algorithm to automatically extract the region of interest (ROI) from the NF-NBI images and then generate image patches through a patch-generating algorithm. After that, we train six representative state-of-the-art deep convolutional neural network (CNN) models (ResNet18, ResNet50, ResNet101, DenseNet201, InceptionV3, and Inception-ResNetV2) to extract robust hierarchical features from these patches and classify them based on the hierarchical features. Finally, to determine the classification results of each subject, majority voting is employed to the corresponding generated NF-NBI image patches. We verify our classification system by ten-fold cross-validation using the clinical dataset. We perform subject-dependent and subject-independent experiments. In both experiments, we compare the classification performance of the ROI-based CNN models (the CNN models with our proposed ROI-based algorithms) with the CNN models. Meanwhile, we compare the classification performance of the ROI-based CNN models with the local binary pattern (LBP)-based support vector machine (SVM) classifier, the histograms of oriented gradients (HOG)-based SVM classifier, and the scale-invariant feature transform (SIFT)-based SVM classifier. The results show that the ROI-based CNN models are able to obtain higher average mean of ten-fold test accuracy on image level than the CNN models in the subject-dependent experiment (29.0% improvement) and the subject-independent experiment (10.5% improvement), which demonstrate the effectiveness of our proposed ROI-based algorithms. Meanwhile, the ROI-based CNN models are able to obtain higher average mean of ten-fold test accuracy on image level than the SVM classifiers in the subject-dependent experiment (20.5% improvement) and the subject-independent experiment (14.0% improvement), which demonstrate the ROI-based CNN models have better classification performance than the SVM classifiers. Among the ROI-based CNN models, the ROI-based InceptionV3 model achieves the best classification performance in the subject-dependent experiment, while the ROI-based Inception-ResNetV2 model achieves the best classification performance in the subject-independent experiment, which suggests the ROI-based Inception-ResNetV2 model has better generalization ability than the ROI-based InceptionV3 model. Moreover, the highest mean of ten-fold test accuracy (77.8%) on subject level obtained by using the ROI-based InceptionV3 model or the ROI-based Inception-ResNetV2 model demonstrates the practicality of our proposed classification system for assisting clinical diagnosis of the NERD.

Keywords: deep learning, transfer learning, convolutional neural network (CNN), region of interest (ROI), endoscopic image, nonerosive reflux disease (NERD)

*Corresponding author

Email addresses: junkai.liao@kcl.ac.uk (Junkai Liao), hak-keung.lam@kcl.ac.uk (Hak-Keung Lam), guangyu.jia@kcl.ac.uk (Guangyu Jia), shraddha.gulati@nhs.net (Shraddha Gulati), julius.bernth@kcl.ac.uk (Julius Bernth), dpoliyivets@icloud.com (Dmytro Poliyivets), yujia.xu@kcl.ac.uk (Yujia Xu), hongbin.liu@kcl.ac.uk (Hongbin Liu), b.hayee@nhs.net (Bu Hayee)

1. Introduction

Gastroesophageal reflux disease (GERD) is a common condition defined by the pathological reflux of gastric contents into the esophagus.

Typical symptoms of GERD include heartburn, a burning sensation of stomach, regurgitation, cough, asthma and nocturnal dyspnea all resulting in a significant negative impact on quality of life [1]. GERD is common [2], with a prevalence of 20% among adults in the western world, of whom up to 40% will experience symptoms every month [1]. Additionally, GERD is associated with a significant economic burden, accounting for an annual direct cost of $9.3 billion in the USA [3]. Costs are attributed to the diagnostic pathway, medical treatment, surgical treatment in cases of medically refractory GERD and the complications of GERD including the development of pre-cancerous conditions such as Barrett's esophagus which require further surveillance procedures and esophageal cancer. Indirect costs also relate to loss of productivity secondary to absenteeism.

The GERD includes two main categories: nonerosive reflux disease (NERD) and erosive esophagitis. The NERD is defined as the presence of typical symptoms of GERD caused by intraesophageal acid in the absence of visible esophageal mucosal injury at conventional light endoscopy [4, 5, 6].

The current conventional diagnosis of NERD is complex and includes endoscopic examination, biopsy of the lining of the esophagus (mucosa) and the gold-standard test to measure the amount of acid using an ambulatory pH testing over 24-96 hours [7, 8, 5, 6, 9, 10]. All investigation methods including ambulatory pH testing require careful consideration of the limitations associated with the investigation and usually require multiple patient visits to conduct these investigations leading to a prolonged time to achieve a diagnosis. Computer-aided diagnosis has the potential to solve many of the current issues as clinicians potentially only need to input the endoscopic images into the computer and then obtain a near instant diagnostic result, which is more convenient and efficient than the conventional diagnosis.

To develop the computer-aided technology, we require endoscopic images captured in healthy people (non-NERD) and the NERD patients diagnosed by the current gold-standard of ambulatory pH testing. The endoscopic images are used to train a designed classification system that classifies the endoscopic images. After the training, when a new captured endoscopic image is input to the classification system, the system can judge whether the endoscopic image is from healthy people (non-NERD) or the NERD patients.

1.1. Classification of Endoscopic Images

The classification of endoscopic images consists of three sequential steps: pre-processing, feature extraction and classification. We will analyze the characteristics of the endoscopic images first and develop dedicated pre-processing algorithms. Then, we will extract the features from the endoscopic images via feature extraction algorithms. Lastly, we will use the extracted features to train a classifier.

1.1.1. Pre-Processing

The endoscopic image is one of medical images. For medical images, the issues influencing the classification include enhancing the image quality and locating the region of interest (ROI). Figure 1 shows the categories of issues, influencing factors and corresponding methods.

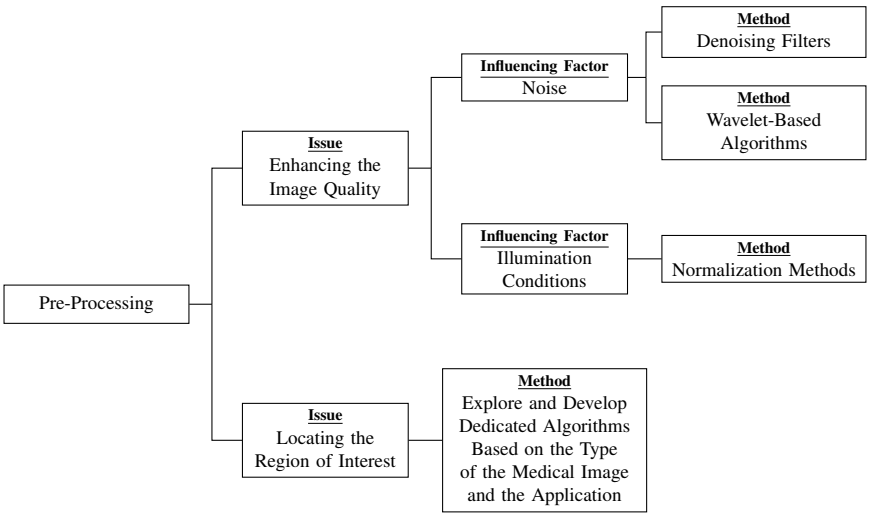


Figure 1: The Categories of Issues and Corresponding Methods for Pre-Processing

The influencing factors of image quality include noise and illumination conditions [11]. To address the issue of noise, various denoising filters including mean filters, median filters, Gaussian filters and bilateral filters were designed in literature and demonstrated good performance for denoising the additive noises, multiplicative noises and quantization noises. However, the images could be blurred during the above denoising process, which means that some useful information of the images may be lost [12, 13, 14, 15]. To deal with this problem, various wavelet-based algorithms [16, 17, 18, 19] were proposed, in which the researchers designed and implemented specific wavelet transforms to the images, and then extracted the features and employed the low-pass filtering in wavelet space. After that, the corresponding inverse wavelet transforms were implemented in [16, 17, 18, 19] to reconstruct the images. The wavelet-based algorithms separate the image signals from the noise signals, which not only achieve good performance for denoising, but also retain the image details. For the impact of illumination conditions, various normalization methods are employed for the medical images and demonstrated to be effective for the specific applications [20, 21, 11, 22]. As we can see, the methods for the issues of image quality have been intensively explored and well developed in the image processing field. Therefore, we will not focus on this aspect in this paper.

To deal with the problems of image quality, locating the ROI is another aspect of pre-processing the image that is more important. The ROI is an area that contains the key information for distinguishing the images of different classes. Literature shows that the related research generally explore and develop dedicated algorithms for locating the ROI based on the type of the medical image and the application. For example, [23] developed a principal component analysis (PCA) based dedicated algorithm for locating the optic disk in color retinal images. [24] performed the boundary detection using morphological operations for locating the regions of exudates in fundus images. [25] employed the Berkeley wavelet transform [26, 27] for locating the tumor regions. However, different applications have different ROIs, a dedicated pre-processing algorithm for locating the ROI have to be developed for the specific application.

1.1.2. Feature Extraction

After the pre-processing, the ROI of the endoscopic images have been located. Then, the features for classification need to be extracted from the ROI. The feature extraction includes supervised feature

extraction and unsupervised feature extraction [28, 29, 30, 31]. The supervised feature extraction algorithms require the priori knowledge including domain knowledge and professional experience from the researchers, while the unsupervised feature extraction algorithms automatically extract the features from a great quantity of raw data without any priori knowledge. Figure 2 shows the categories, representative algorithms and their development of feature extraction. In the figure, the arrows denote the development of the representative algorithms.

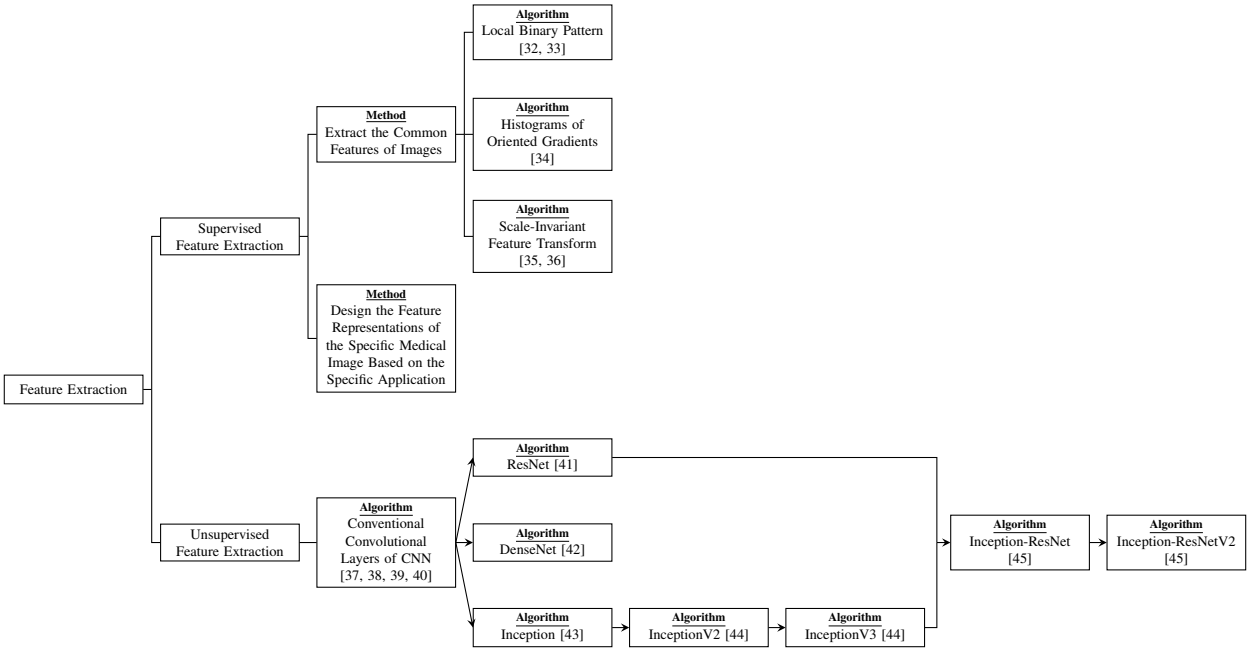


Figure 2: The Categories, Representative Algorithms and their Development of Feature Extraction

For medical images, there are two ways for supervised feature extraction. One of which is to extract the common features of images, like the textures, edges, angular points, and marginal points of images. In 1994, local binary pattern (LBP) [32] was purposed to extract the local texture features of images. The LBP has the property of gray-scale invariance, which is insensitive to illumination changes, but it is sensitive to non-uniform illumination changes. Then, [33] polished the LBP from rectangle to circle, and purposed the rotational invariant LBP, which is insensitive to image rotation, but it is still sensitive to non-uniform illumination changes. In 2005, histograms of oriented gradients (HOG) [34] was purposed to extract the edge features of images by calculating and statistically counting the oriented gradients' histogram of the image's local area. The HOG is insensitive to non-uniform illumination changes and image rotation, but it cannot deal with the issue of image occlusion. Also, due to the property of oriented gradients, the HOG is very sensitive to noises. Over the same period, [35, 36] proposed the scale-invariant feature transform (SIFT) to extract the representative feature points of images, including angular points, marginal points, dark points in the light area and light points in the dark area. The SIFT is insensitive to non-uniform illumination changes, image rotation and noises. Meanwhile, it is robust to image occlusion. The weakness of this algorithm is that it needs considerably large amounts of computation due to the continuous downsampling and interpolation operations. Besides, it cannot accurately extract the feature points of the objects with smooth edges.

Another way for supervised feature extraction is to design the feature representations of the specific medical image based on the specific application using the priori knowledge. In 1996, [46] designed

a three-dimensional feature space representation for magnetic resonance images, where normal tissues are clustered around prespecified positions and abnormal tissues are clustered elsewhere. In 2004, [23] designed the feature representations based on the shape, color and depth of optic disks in color retinal images for diagnosis of various eye diseases. In 2009, [24] designed the feature representations including the optic disks, blood vessels, exudates, microaneurysms and hemorrhages for diagnosis of the diabetic retinopathy. In 2017, 11 textural feature representations were designed in [25] that are composed of mean, standard deviation, entropy, skewness, kurtosis, energy, contrast, inverse difference moment, directional moment, correlation and coarseness in brain magnetic resonance images for classification of brain tumor. The experiments of the above research showed satisfactory results for the specific applications, but all of the above supervised feature extraction algorithms need the priori knowledge.

Comparing to the supervised feature extraction algorithms, the unsupervised feature extraction algorithms automatically extract the features from a great quantity of raw data without any priori knowledge. The convolutional layers of the convolutional neural network (CNN) [37, 38, 39, 40, 47, 48] is a representative unsupervised feature extraction algorithm, which automatically extracts the features of input image by convolutional operations.

However, with the increasing number of convolutional layers, the issue of degradation arises, where the information of the data may be lost during the transmission between the convolutional layers. In 2016, [41] proposed a residual function and embedded it into the convolutional layers to resolve this issue. The residual function can be understood as adding a shortcut connection between the previous layer and the following layer. The connection skips some layers and transmit the original data directly to the "Following Layer" in Figure 7 (b). In 2017, [42] gave full play to this idea and proposed a novel structure of CNN, where the inputs of each layer are from the outputs of all the previous layers. Each layer has the direct access to the gradients from the loss function and the original input signal, which leads an implicit deep supervision. On the other hand, [43] proposed a new module called Inception, which performs multiple convolutional or pooling operations in parallel to the feature maps from the "Previous Layer" in Figure 9 and then concatenates all the outputs to be a larger feature map. The experiments showed that performing these operations in parallel and concatenating all the outputs can obtain better feature representations of the input samples. However, the Inception needs remarkably high computational costs and easily overfits the dataset. To deal with these issues, [44] proposed InceptionV2 in which researchers employed convolutional decomposition to replace the original large-size convolutional operations in the Inception. Specifically, a 3×3 convolutional operation with n filters over a grid with m filters has 2.78 times less computational cost than a 5×5 convolutional operation with the same number of filters. Hence the researchers employed two 3×3 convolutional operations to replace a 5×5 convolutional operation in the InceptionV2. Meanwhile, a $1 \times n$ convolutional operation followed by an $n \times 1$ convolutional operation is equivalent to an $n \times n$ convolutional operation while the former dramatically saves the computational cost with the increasing of n. In experiments, the researchers found that a 1×7 convolutional operation followed by a 7×1 convolutional operation can obtain very good results. Hence the researchers further proposed InceptionV3 employing the factorization of the 7×7 convolutional operation. Moreover, [44] also proposed a model regularization approach via label smoothing to prevent the model from overly biasing to one category, which is an approach avoiding overfitting. In 2017, [45] combined the advantages of the residual function and the InceptionV3, and proposed the Inception-ResNet. The researchers found that the Inception-ResNet considerably saves the training time and slightly increases the classification accuracy than the plain Inception. Meanwhile, the researchers further proposed the Inception-ResNetV2 by modifying the combination order of the residual function and the InceptionV3. The results of the experiments showed that the Inception-ResNetV2 further slightly increases the classification accuracy

than the previous version.

The CNN is a representative deep structured algorithms in the deep learning field [49, 50], which have hierarchical structures to extract the features from low-dimensional to high-dimensional space to obtain the better representations than the supervised feature extraction algorithms.

1.1.3. Classification

After the feature extraction, the extracted features will be input to a classifier to do the classification. For example, a conventional CNN could be regarded as a series of convolutional layers followed by a neural network. The convolutional layers are for the feature extraction as discussed in Section 1.1.2, and the neural network is for the classification [49, 50]. In the following paragraphs, we will introduce two representative classifiers: support vector machine (SVM) [51, 52, 53, 54, 55, 56, 57] and neural network [58, 59, 60, 61, 62, 63, 64, 65, 66, 67].

The SVM [51, 52] is a discriminative classifier for binary classification. Training the SVM is to employ a kernel function to find an optimal boundary to separate the extracted features of two classes in the feature hyperplane. Then the separated feature hyperplane can be used to classify new samples. The SVM is suitable to deal with small dataset, which generally contains no more than 10000 samples [68, 69, 70, 71]. Meanwhile, the SVM can be used to explain which extracted feature contributes to classify the samples. However, calculating the vectors of SVM needs to implement the quadratic programming, which requires a great amount of memory in the case of dealing with large dataset. Also, the SVM cannot be employed for multi-class classification directly. To deal with the multi-class classification using the SVM, researchers generally need to train multiple SVMs for every pair of two classes, and then classify the new samples by majority voting of all the trained SVMs.

The neural network [58, 59] is a biologically inspired classifier composed of a series of nodes (neurons) and connection weights. The connection weights connect the nodes between two layers, which determines the weights of values from the nodes of the previous layer. Then, the weighted values will be output to the following layer via a nonlinear activation function. During the training, the connection weights are modifiable by the back-propagation algorithm [72], which is inspired from the memory of human brain for cognition. It was proven that a neural network with one hidden layer containing a finite number of nodes can approximate any continuous nonlinear function arbitrarily well in a compact domain [73]. Hence, the neural network can learn the complex nonlinear mapping between the extracted features and the labels.

1.2. Problem Formulation and Proposed Method

In 2015, [74] proposed a hierarchical heterogeneous feature representation fusion SVM framework for diagnosis of GERD. In [74], the researchers designed a hierarchical heterogeneous feature representation using priori knowledge for endoscopic images based on six-color models including RGB, normalized RGB, opponent, C-invariant, transformed RGB and index of hemoglobin. Then, they separated the endoscopic image into 4×4 equal-sized rectangle regions, and extracted the features of each rectangle region. And then, they used the extracted features to train an SVM classifier. After the training, they used the trained SVM classifier to classify new endoscopic images to judge whether the endoscopic images are from a person suffering from GERD or not.

However, the method in [74] did not locate the ROI. Instead, it simply separated the endoscopic image into 4×4 equal-sized rectangle regions, some of which may have redundant information, so the classifiers cannot focus on the regions that contain important information. Therefore, we propose an ROI-extraction algorithm to locate the ROI. With the ROI-extraction algorithm, the classifiers can focus on the regions that contain the key information for distinguishing the images of different classes. Hence, the classifiers can have better classification performance. We will perform comparative experiments (to be presented in Section 4) where the classifiers with and without the ROI-extraction algorithm to verify the effectiveness of the ROI-extraction algorithm. In addition, the method in [74] designed the feature representations of the endoscopic image based on the six-color models, which needs the priori knowledge. While our unsupervised feature extraction algorithms automatically extract the features from a great quantity of raw data without any priori knowledge.

To the best of our knowledge, the work in [74] is the first and the only computer-aided diagnosis method based on endoscopic images for GERD. The GERD includes two main categories: nonerosive reflux disease (NERD) and erosive esophagitis. The NERD is defined as the presence of typical symptoms of GERD caused by intraesophageal acid in the absence of visible esophageal mucosal injury at conventional light endoscopy [4, 5, 6]. Because the visible esophageal mucosal injuries shown in the endoscopic images are apparent for diagnosing the case as erosive esophagitis [75], we focus on the challenging part of the diagnosis, which is to diagnose the NERD.

To diagnose the NERD, we propose a deep learning classification system to classify endoscopic images captured in the esophagus of health people and the NERD patients, which is a binary classification problem of two classes: non-NERD and NERD. To achieve an effective and accurate classification, we propose an algorithm to automatically extract the ROI from the endoscopic images and then generate image patches through a patch-generating algorithm. After that, we train six representative state-of-the-art deep CNN models (ResNet18 [41], ResNet50 [41], ResNet101 [41], DenseNet201 [42], InceptionV3 [43, 44], and Inception-ResNetV2 [45]) to extract robust hierarchical features from these patches and classify them based on the hierarchical features. Finally, to determine the classification results of each subject, majority voting is employed to the corresponding generated endoscopic image patches.

We will perform subject-dependent and subject-independent experiments. In the subject-dependent experiment, the samples of training and test datasets are from the same groups of subjects. Hence, the test dataset is highly correlated to the training dataset, so the classifiers may not generalize to the unseen samples from new subjects. Therefore, we also designed the subject-independent experiment. In the subject-independent experiment, the samples of training and test dataset are from different groups of subjects. We used the trained classifiers from the subject-dependent experiment to classify the images from additional 18 subjects. We divided all the images by subject. Hence, through the subject-independent experiment we can verify whether the classifiers can generalize to the unseen samples from new subjects.

In both the subject-dependent and the subject-independent experiments, we will compare the classification performance of the ROI-based CNN models (the CNN models with our proposed ROI-based algorithms) with the CNN models using ten-fold cross-validation. After that, we will compare the classification performance of the ROI-based CNN models with the LBP-based SVM classifier, the HOG-based SVM classifier, and the SIFT-based SVM classifier [32, 33, 34, 35, 36, 51, 52] using ten-fold cross-validation.

The results (to be presented in Section 4) will show that the ROI-based CNN models are able to obtain better classification performance than the CNN models in both the subject-dependent experiment and the

subject-independent experiment, which demonstrate the effectiveness of our proposed ROI-based algorithms. Meanwhile, the ROI-based CNN models are able to obtain better classification performance than the SVM classifiers in both the subject-dependent experiment and the subject-independent experiment, which demonstrate the ROI-based CNN models have better classification performance than the SVM classifiers. Among the ROI-based CNN models, the ROI-based InceptionV3 model is able to achieve the best classification performance in the subject-dependent experiment, while the ROI-based Inception-ResNetV2 model is able to achieve the best classification performance in the subject-independent experiment, which suggests the ROI-based Inception-ResNetV2 model has better generalization ability than the ROI-based InceptionV3 model. Moreover, the best classification performance obtained by using the ROI-based InceptionV3 model or the ROI-based Inception-ResNetV2 model demonstrates the practicality of our proposed classification system for assisting clinical diagnosis of the NERD. The main contributions of this paper are summarized as follows:

1) We propose the first deep learning classification system that only uses the endoscopic images for fully automatic diagnosis of the NERD.

2) We propose an ROI-extraction algorithm to locate the ROI. The ROI is an area that contains the key information for distinguishing the images of different classes.

3) We verify the practicality of our proposed classification system by conducting ten-fold cross-validation using the clinical dataset.

4) We conduct both the subject-dependent and the subject-independent experiments to compare the ROI-based CNN models with the CNN models. Meanwhile, we compare the ROI-based CNN models with the SVM classifiers. The results demonstrate the superiority of the ROI-based CNN models.

5) We select the best ROI-based CNN model by comparing the classification performance between the six representative state-of-the-art deep CNN models.

The rest of the paper is organized as follows: Section II describes the clinical data collection. The proposed methods in this paper are introduced in Section III. Section IV demonstrates the experiments and results. Section V draws a conclusion.

2. Data Collection

The dataset for training and test the classifiers in this paper was collected by clinicians of King's College Hospital, which has obtained an ethical approval [76]. The clinicians applied near focus narrow band imaging (NF-NBI) to collect the data. The narrow band imaging technology employs filters to filter out the broadband spectrum of red, blue and green light waves from the endoscope light source, leaving only narrowband spectrum, which can enhance the contrast of the lesion, hence it can take the place of endoscopic staining process [77].

The dataset contains 1394 clinical NF-NBI images from 50 subjects, of which 554 images from 21 subjects were labeled as positive samples (NERD patients), and 840 images from 29 subjects were labeled as negative samples (healthy people). After that, the clinicians collected and provided extra 556 clinical

NF-NBI images from additional 18 subjects. We used them to further test the classifiers on both image level and subject level, and then compare the results with the clinical diagnosis by the clinicians.

3. Methods

We aim to develop algorithms to automatically diagnose the NERD using the NF-NBI images. To diagnose this disease, a deep learning classification system is proposed to distinguish the NF-NBI images captured in the esophagus of healthy people and the NERD patients, which is a binary classification of two classes: non-NERD and NERD.

Figure 3 illustrates the workflow diagram of the proposed system. During the training phase, the full-size training images are input to the system. The ROIs of the images are extracted first, and then image patches are generated from the ROIs by a patch-generating algorithm. After that, the image patches are used to train the CNN model by transfer learning [78]. During the test phase, the full-size test images are input to the system and processed by the same procedures before the CNN model, and then the image patches are input to the CNN model. After that, the CNN model gives classification result for each image patch. The classification results are recorded for final diagnosis by majority voting. The methods are described in detail in the following subsections.

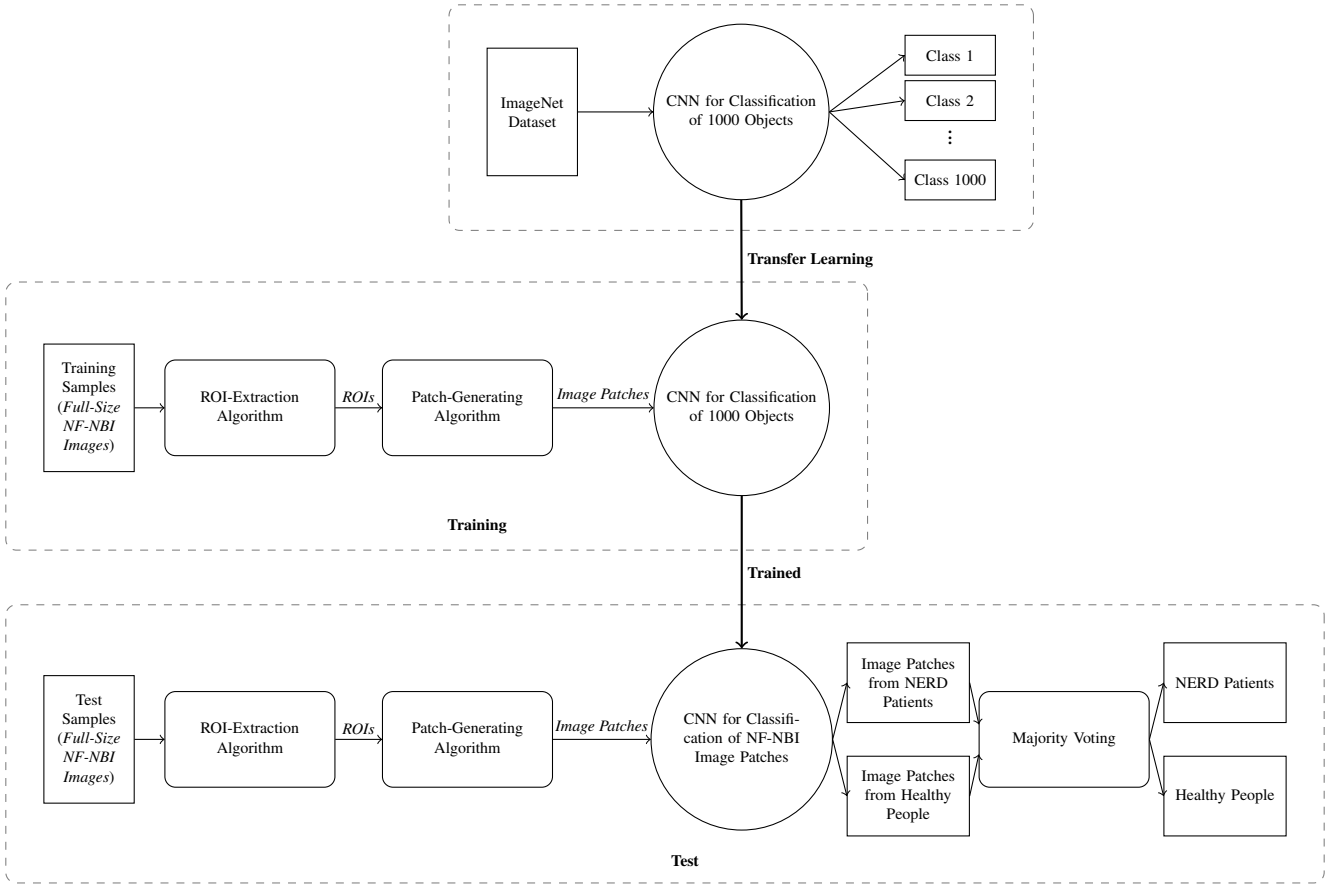


Figure 3: Workflow Diagram of the Proposed System

3.1. ROI-Extraction Algorithm

The ROI is an area that contains the key information for distinguishing the images of different classes. The work in [79] summarized that the endoscopic images show the intra-epi-thelial papillary capillary loops (IPCLs) at the lower side of the esophagus in the NERD patients dilated and elongated with the regular arrangement, which is a potential feature for diagnosis of the NERD. Hence, we consider the region of IPCLs as the ROI of our classification system.

We ever considered manually marking the region of IPCLs. However, it is a laborious and impractical task for thousands of images. Meanwhile, correctly marking the region of IPCLs requires considerable medical knowledge. In practice, new clinically captured full-size NF-NBI images may have complicated shooting conditions, which could be very different from the training dataset. Hence, if the regions of the training data were manually marked, the trained classification system is not easy to be generalized to the new data.

Therefore, we propose an ROI-extraction algorithm that automatically marks the region of IPCLs. Figure 4 (a) shows a full-size NF-NBI image, which has edges of camera lens and some areas of glare that seriously affect the judgement of classifiers. Hence, the first step of the ROI-extraction algorithm is to remove the edges of camera lens and the areas of glare. After that, we observed that the areas of IPCLs have relatively high intensity than other areas after we converted the full-size NF-NBI image to be a greyscale image. Hence, in the second step, we utilize this feature to segment the full-size NF-NBI image by adaptive image thresholding. In the last step, we employ morphological closing and morphological opening to obtain the complete region of IPCLs.

The ROI-extraction algorithm is shown in Algorithm 1. In Algorithm 1, the block of lines 1 to 13 is for searching the edges of camera lens, and the input of this block is the full-size NF-NBI image shown in Figure 4 (a). Firstly, we summarize the range in HSV (hue, saturation and value) color space of black color that covers the edges of camera lens, and then perform the thresholding on the full-size NF-NBI image. Secondly, we employ the morphological closing to remove holes inside the areas, and then we perform dilation to increase the areas. After that, the output of this block is the mask of camera lens' edges shown in Figure 4 (b).

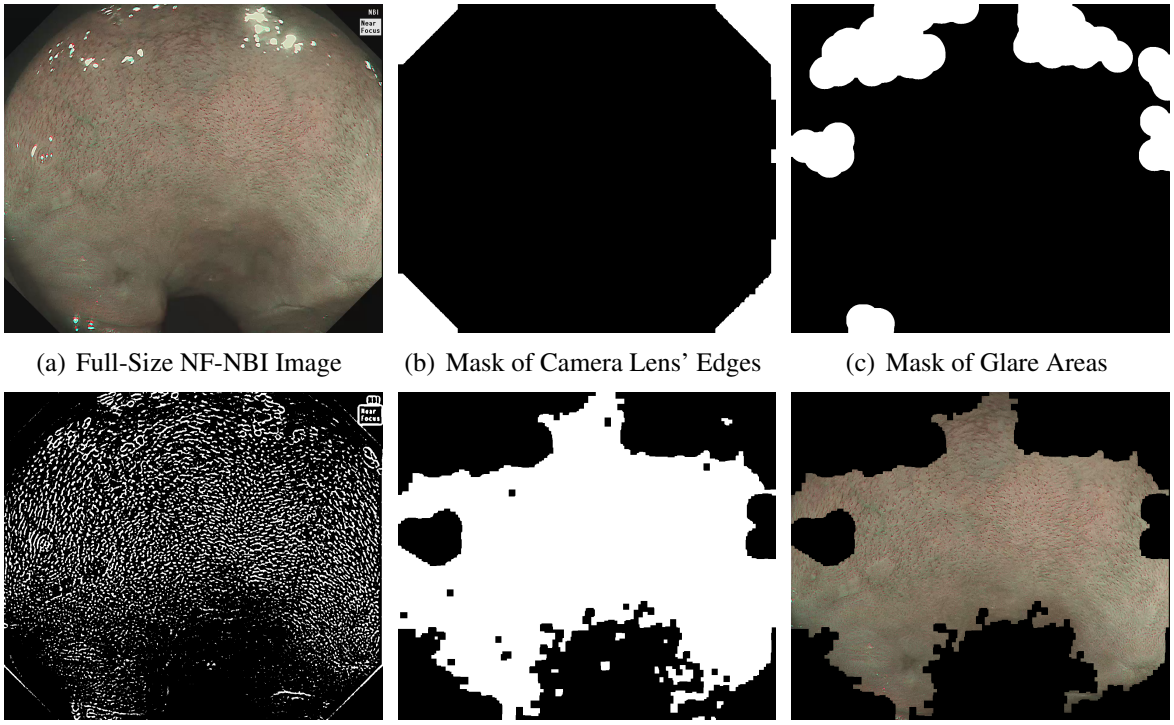
The block of lines 14 to 25 is for searching the areas of glare. The input of this block is the full-size NF-NBI image shown in Figure 4 (a). Similarly, we summarize the range in HSV color space of white color that covers the areas of glare, and then perform the thresholding on the full-size NF-NBI image. Secondly, we employ the dilation to increase the areas, and then we perform the morphological closing to remove holes inside the areas. After that, the output of this block is the mask of glare are as shown in Figure 4 (c).

The two steps above define a fixed range of color to perform the thresholding. However, different areas of the full-size NF-NBI image have different illumination conditions such as different brightness, contrast and white balance. The influence of the different illumination conditions is especially strong on IPCLs. In order to locate the region of IPCLs, we perform the adaptive thresholding on the full-size NF-NBI image. The threshold for a pixel is a Gaussian-weighted sum of its surrounding pixels' values minus a pre-defined constant. The pre-defined constant was selected by experiments. The block of lines 26 to 39 is for segmenting the full-size NF-NBI image by the adaptive thresholding. The input of this block is the full-size NF-NBI image shown in Figure 4 (a). Firstly, we convert the full-size NF-NBI image to be a grayscale image. Secondly, we define the pre-defined constant as 2 based on experiments. After that, we

perform the adaptive thresholding algorithm on the full-size NF-NBI image, and the output of this block is the mask of IPCLs shown in Figure 4 (d).

The block of lines 40 to 50 is for combining the mask. The inputs of this block are the mask of camera lens' edges, the mask of glare areas and the mask of IPCLs shown in Figure 4 (b), (c) and (d). Firstly, we remove the areas of camera lens' edges and glare in the mask of IPCLs by subtraction. Secondly, we employ morphological closing to cluster the scattered areas. Thirdly, we employ the morphological opening to remove redundant scattered areas. After that, the output of this block is the combined mask shown in Figure 4 (e).

The line 51 is for the extraction of the ROI by the combined mask. The inputs are the full-size NF-NBI image and the combined mask shown in Figure 4 (a) and (e), and the output is the ROI shown in Figure 4 (f).



(a) Full-Size NF-NBI Image (b) Mask of Camera Lens' Edges (c) Mask of Glare Areas

(d) Mask of IPCLs (e) Combined Mask (f) ROI

Figure 4: ROI-Extraction Algorithm

Algorithm 1 ROI-Extraction Algorithm

Input: Full-Size NF-NBI Image $\mathbb{I}$
Output: ROI $\mathbb{R}$

Convert $\mathbb{I}$ from RGB (red, green, blue) color space to HSV (hue, saturation and value) color space
for $h_1 = 0 : 1 : \mathbb{I}_{height}$ **do**
  for $w_1 = 0 : 1 : \mathbb{I}_{width}$ **do**
    if $(0, 0, 0) \leq \mathbb{I}[h_1, w_1] \leq (165, 105, 20)$ **then**
      $\mathbb{B}[h_1, w_1] = 1$
    else
      $\mathbb{B}[h_1, w_1] = 0$
    end if
  end for
end for
Employ morphological closing to $\mathbb{B}$
Employ dilation to $\mathbb{B}$
Obtained the mask of camera lens' edges $\mathbb{B}$
for $h_2 = 0 : 1 : \mathbb{I}_{height}$ **do**
  for $w_2 = 0 : 1 : \mathbb{I}_{width}$ **do**
    if $(4, 8, 237) \leq \mathbb{I}[h_2, w_2] \leq (150, 102, 255)$ **then**
      $\mathbb{C}[h_2, w_2] = 1$
    else
      $\mathbb{C}[h_2, w_2] = 0$
    end if
  end for
end for
Employ dilation to $\mathbb{C}$
Employ morphological closing to $\mathbb{C}$
Obtained the mask of glare areas $\mathbb{C}$
Convert $\mathbb{I}$ to be a grayscale image
Define constant $c = 2$ based on experiments
for $h_3 = 0 : 1 : \mathbb{I}_{height}$ **do**
  for $w_3 = 0 : 1 : \mathbb{I}_{width}$ **do**
    $\mathbb{N} = \mathbb{I}[h_3 - 1 : h_3 + 1, w_3 - 1 : w_3 + 1]$
    $S =$ Gaussian-weighted sum of $\mathbb{N}$
    if $\mathbb{I}[h_3, w_3] \geq S - c$ **then**
      $\mathbb{D}[h_3, w_3] = 1$
    else
      $\mathbb{D}[h_3, w_3] = 0$
    end if
  end for
end for
Obtained the mask of IPCLs $\mathbb{D}$
for $h_4 = 0 : 1 : \mathbb{D}_{height}$ **do**
  for $w_4 = 0 : 1 : \mathbb{D}_{width}$ **do**
    if $\mathbb{B}[h_4, w_4] = 1$ **then**
      $\mathbb{D}[h_4, w_4] = 0$

44: **end if**
45: **end for**
46: **end for**
47: $\mathbb{E} = \mathbb{D} - \mathbb{C}$
48: Employ morphological closing to $\mathbb{E}$
49: Employ morphological opening to $\mathbb{E}$
50: Obtained the combined mask $\mathbb{E}$
51: $\mathbb{R} = \mathbb{I} \cdot \mathbb{E}$

3.2. Patch-Generating Algorithm

As mentioned in Section 2, we have 1394 full-size NF-NBI images from 50 subjects for training and testing the classifiers, which is far from enough to train a deep CNN model. Hence, we designed a patch-generating algorithm to generate the image patches as input samples to the deep CNN model, which increases the quantity of training and test samples.

The idea of patch-generating algorithm is illustrated in Figure 5. We apply a sliding window of 224×224 pixels to generate the image patches from the ROI. The step length of sliding is designed as 100 pixels based on experiments.

The patch-generating algorithm is shown in Algorithm 2. The input of Algorithm 2 is the ROI shown in Figure 4 (f). In Algorithm 2, lines 1 to 3 and lines 19 to 20 are for creating a sliding window of 224×224 pixels, and the step length of sliding is 100 pixels. Line 4 is for generating the image patches. Lines 5 to 18 are for determining which generated image patches should be discarded, i.e., the generated image patches that contain the areas outside the ROI should be discarded. After that, the outputs of Algorithm 2 are the generated image patches. Figure 6 shows a part of the generated image patches from the ROI shown in Figure 4 (f). For the dataset in this paper, totally 6484 image patches were generated from 1394 ROIs.

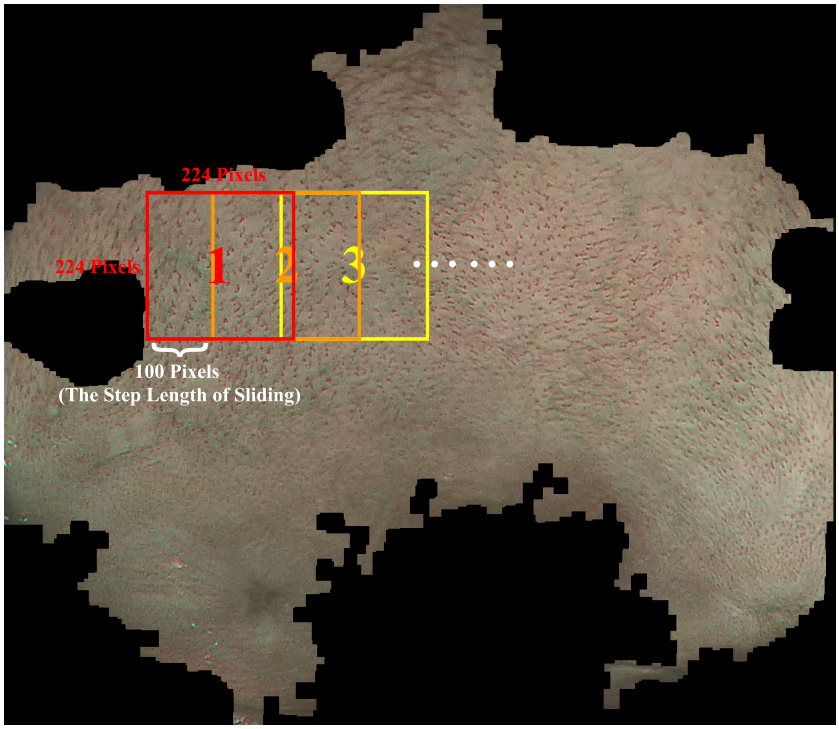


Figure 5: Illustration of Patch-Generating Algorithm

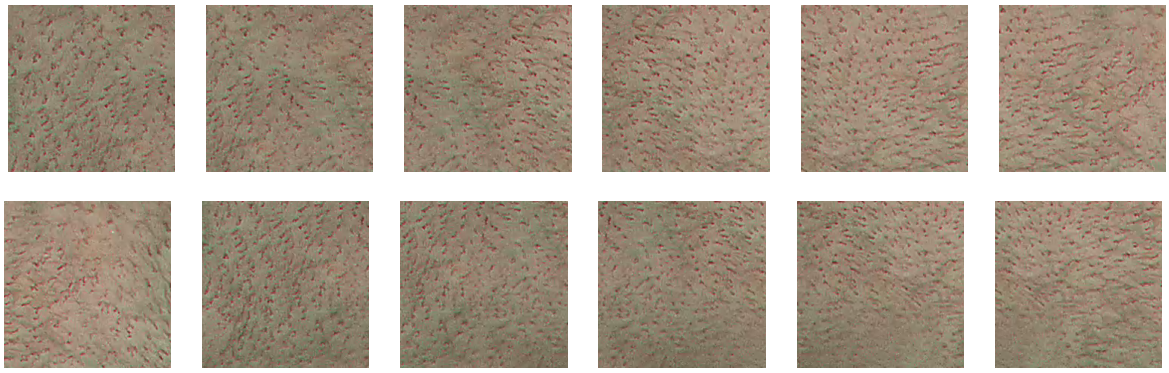


Figure 6: Generated Image Patches

Algorithm 2 Patch-Generating Algorithm

Input: ROI $\mathbb{R}$
Output: Generated Image Patches $\mathbb{P}_1, \mathbb{P}_2, ...$

$n = 1$
for $h_1 = 0 : 100 : (\mathbb{R}_{height} - 224)$ **do**
  for $w_1 = 0 : 100 : (\mathbb{R}_{width} - 224)$ **do**
    $\mathbb{T} = \mathbb{R}[h : h + 224, w : w + 224]$
    for $h_2 = 0 : 1 : \mathbb{T}_{height}$ **do**
      for $w_2 = 0 : 1 : \mathbb{T}_{width}$ **do**
        if $\mathbb{T}[h_2, w_2] = (0, 0, 0)$ **then**
          $Discard = 1$
          break
        else
          $Discard = 0$
        end if
      end for
    end for
    if $Discard = 0$ **then**
      $\mathbb{P}_n = \mathbb{T}$
      $n = n + 1$
    end if
  end for
end for

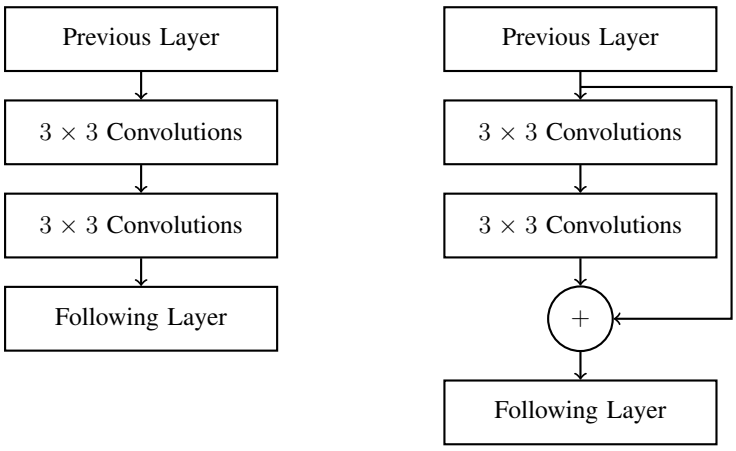
3.3. CNN Models

For selecting the best ROI-based CNN model, we implemented and compared six representative state-of-the-art deep CNN models including ResNet18 [41], ResNet50 [41], ResNet101 [41], DenseNet201 [42], InceptionV3 [43, 44] and Inception-ResNetV2 [45]. The CNN models will be introduced in the following subsections.

3.3.1. ResNet Model

With increasing the number of parameter layers, some prominent training problems are appearing. The most significant problem is vanishing or exploding gradient, which affects the network converging in the very beginning of the training. The problem of convergence can be partially resolved by renormalization. On the premise that the deep network can converge, with increasing the number of parameter layers, the classification accuracy begins to saturate or even decrease, which is known as the degradation problem of the network, where the information of the data may be lost during the transmission between the convolutional layers [41].

In 2016, [41] proposed a residual function and embedded it into the convolutional layers to resolve this issue. Figure 7 shows the structural difference between a conventional CNN and a CNN with a residual function.



(a) In a Conventional CNN (b) In a CNN with a Residual Function

Figure 7: The Structural Difference

The residual function can be understood as adding a shortcut connection between the previous layer and the following layer. The connection skips some layers and transmit the original data directly to the following layer. The added shortcut connection neither increases the parameters nor complexity of the model.

For the ImageNet dataset [80], [41] demonstrated that the classification accuracy of the ResNet CNN is steadily rising with increasing the number of parameter layers. In this paper, we implemented and compared the ResNet18, the ResNet50 and the ResNet101 that have 18, 50 and 101 parameter layers respectively.

3.3.2. DenseNet Model

As mentioned in last subsection, on the premise that the deep network can converge, with increasing the number of parameter layers, the classification accuracy begins to saturate or even decrease, which is known as the degradation problem of the network [41]. While adding a shortcut connection between the previous layer and the following layer can partially resolved this problem. In 2017, [42] gave full play to this idea and proposed a novel structure of CNN named DenseNet, where the inputs of each layer are from the outputs of all the previous layers. Figure 8 shows a dense module in DenseNet CNN, each layer takes all previous feature maps as input.

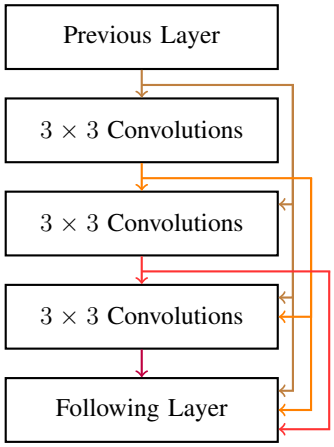


Figure 8: A Dense Module in DenseNet CNN

Each layer has the direct access to the gradients from the loss function and the original input signal, which leads an implicit deep supervision [42]. Owing to adding these shortcut connections between parameter layers, the network could be designed as very deep, which means the network could have a large number of parameter layers without degradation.

3.3.3. Inception Model

In 2015, [43] proposed a CNN structure with a new module named Inception, which performs multiple convolutional or pooling operations in parallel to the feature maps from the previous layer and then concatenates all the outputs to be a larger feature map. Figure 9 shows an Inception module in a CNN.

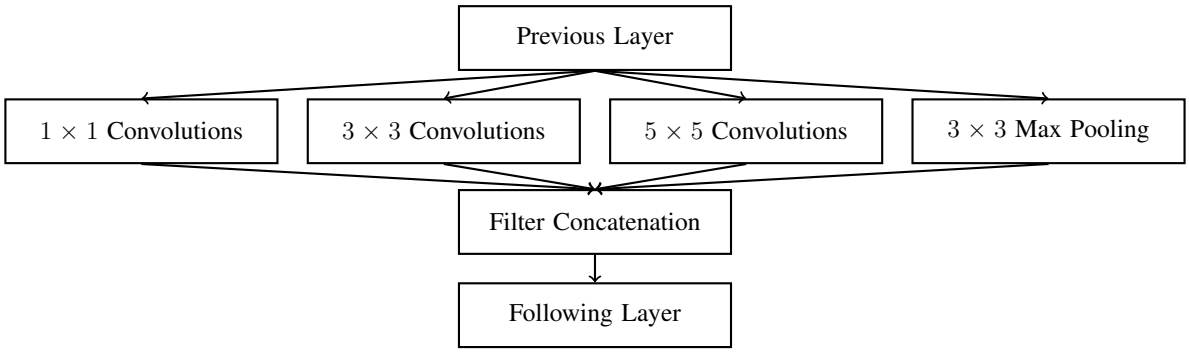


Figure 9: An Inception Module in a CNN

Because different sizes of convolutional and pooling operations such as 1×1, 3×3 and 5×5 can extract different information of the input data, performing these operations in parallel and concatenating all the outputs can obtain better feature representations of the input data.

However, the Inception needs remarkably high computational cost and easily overfits the dataset. To deal with these issues, [44] proposed InceptionV2 where researchers employed convolutional decomposition to replace the original large-size convolutional operations in the Inception. Specifically, a 3×3 convolutional operation with n filters over a grid with m filters has 2.78 times less computational cost than a 5×5 convolutional operation with the same number of filters. Hence the researchers employed two

3×3 convolutional operations to replace a 5×5 convolutional operation in the InceptionV2. Meanwhile, a $1 \times n$ convolutional operation followed by an $n \times 1$ convolutional operation is equivalent to an $n \times n$ convolutional operation while the former dramatically saves the computational cost with the increasing of n. In experiments, the researchers found that a 1×7 convolutional operation followed by a 7×1 convolutional operation can obtain very good results. Hence the researchers further proposed InceptionV3 employing the factorization of the 7×7 convolutional operation. Moreover, [44] also proposed a model regularization approach via label smoothing to prevent the model from overly biasing to one category, which is an approach avoiding overfitting.

In 2017, [45] combined the advantages of the residual function and the InceptionV3, and proposed the Inception-ResNet. The researchers found that the Inception-ResNet considerably saves the training time and slightly increases the classification accuracy than the plain Inception. Meanwhile, the researchers further proposed the Inception-ResNetV2 by modifying the combination order of the residual function and the InceptionV3. The results of the experiments showed that the Inception-ResNetV2 further slightly increases the classification accuracy than the previous version.

3.4. Transfer Learning

Transfer learning [78] is a common learning method for deep learning classifiers. It transfers the knowledge that has learned from a classification task to improve the learning performance for a new classification task. Specifically, we first train the classifier for a classification task A using a corresponding dataset α. After the training, the classifier has learned the features from the dataset α to deal with the classification task A. The learned features include common image features (e.g. textures, edges, angular points, and marginal points) and task-specific features. Then, we keep the learned common image features in the classifier and retrain the part of task-specific features using another dataset β for a corresponding classification task B. This learning method can improve the learning performance for the classification task B, especially when we have insufficient number of labeled images in the dataset β [78, 81].

We use the transfer learning to train the CNN models introduced in Section 3.3. We first train the CNN models using the ImageNet dataset [80] as shown in Figure 3. The ImageNet dataset [80] contains 1000 classes of objects. Each class contains 1000 labeled images on average. After the training, the CNN models learned the features for classifying the 1000 classes of objects. Then, we keep the early layers of the CNN models that learned the common image features (e.g. textures, edges, angular points, and marginal points) and discard the last three layers of the CNN models that learned the task-specific features, which are the fully connected layer, the softmax layer and the output layer. We replace a new fully connected layer, a new softmax layer and a new output layer to the last three layers of the CNN models. After that, we use the image patches generated from the patch-generating algorithm (refer to Section 3.2) to train the CNN models. After the training, the CNN models learned the new task-specific features and can classify the image patches.

3.5. Majority Voting

After training the CNN models, we use the CNN models to classify the image patches and obtain the classification results on image level. In order to obtain the classification results on subject level, we employ the majority voting. That is, if more than 50% of the image patches corresponding to a subject were classified as one class by the CNN models, we classify this subject as this class.

4. Experiments and Results

We performed subject-dependent and subject-independent experiments (refer to Section 1.2). In both the subject-dependent and the subject-independent experiments, we compared the classification performance of the six representative state-of-the-art deep CNN models (ResNet18 [41], ResNet50 [41], ResNet101 [41], DenseNet201 [42], InceptionV3 [43, 44], and Inception-ResNetV2 [45]) with and without our proposed ROI-based algorithms using ten-fold cross-validation.

After that, we compared the classification performance of the six CNN models with our proposed ROI-based algorithms (referred to as "ROI-based CNN models") and the SVM classifiers [51, 52] using ten-fold cross-validation. To extract the features for classification using the SVM classifiers, we selected three representative supervised feature extraction algorithms to extract the common features of images, which are the LBP [32, 33], the HOG [34], and the SIFT [35, 36] (introduced in Section 1.1.2). Hence, the SVM classifiers are the LBP-based SVM classifier, the HOG-based SVM classifier, and the SIFT-based SVM classifier. During the comparison, the ROI-based CNN models are replaced by the SVM classifiers in Figure 3, which means the inputs of the SVM classifiers are the generated image patches.

The classification performance is measured by accuracy [82], which is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP and FN denote the number of true-positive, true-negative, false-positive and false-negative classification results, respectively.

In the following subsections, the subject-dependent and the subject-independent experiments are described in detail.

4.1. Subject-Dependent Experiment

In this experiment, we use the dataset containing 1394 clinical full-size NF-NBI images from 50 subjects, of which 554 images from 21 subjects were labeled as positive samples (NERD patients), and 840 images from 29 subjects were labeled as negative samples (healthy people). We divide the dataset into ten folds with the same quantity of images. For the CNN models, the images are the full-size NF-NBI images; for the ROI-based CNN models and the SVM classifiers, the images are the generated image patches. The scheme of the division is illustrated in Figure 10. Each fold contains 10% images of each subject.

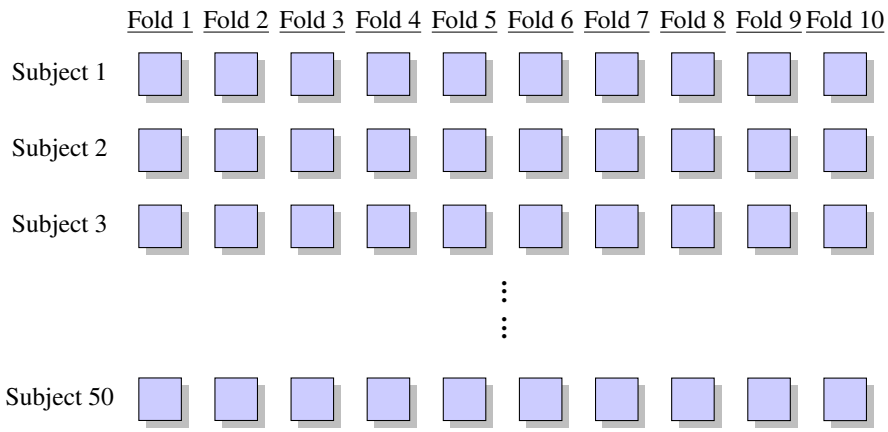


Figure 10: Scheme of the Division

Then, we implement the ten-fold cross-validation for training and testing the proposed system, where nine folds are used for training, and the remaining one fold is used for test. After that, the training and test process are repeated ten times with each of the ten folds used once for test. The advantage of the ten-fold cross-validation is that all the data are used for both training and test, and each sample of data is used for test exactly once.

As the result of this experiment configuration, each fold contains a portion of images from all the subjects, hence it is a subject-dependent classification. Therefore, we only obtained the classification results on image level in the subject-dependent experiment.

Tables 1, 2, 3, 4, 5 and 6 show the training accuracy and test accuracy of CNN models, ROI-based CNN models and SVM classifiers on image level in the subject-dependent experiment. Note that the labels 18, 50 and 101 in ResNet18, ResNet50 and ResNet101 denote the quantity of parameter layers in the ResNet CNNs [41]. The label 201 in DenseNet201 denotes the quantity of parameter layers in the DenseNet CNN [42]. The label V3 in InceptionV3 denotes the third version of the Inception CNN [43, 44]. The label V2 in Inception-ResNetV2 denotes the second version of the combination of the Inception CNN and the ResNet CNN [45].

Accuracy / Fold	Training Accuracy on Image Level					
	ResNet18	ResNet50	ResNet101	DenseNet201	InceptionV3	Inception-ResNetV2
1	60.0%	56.2%	59.2%	59.5%	58.7%	55.3%
2	59.8%	58.4%	56.2%	41.6%	59.8%	52.7%
3	59.0%	59.9%	59.9%	44.2%	59.4%	49.9%
4	57.9%	60.0%	54.5%	54.4%	41.8%	40.2%
5	59.8%	59.9%	57.3%	49.3%	46.2%	50.7%
6	59.8%	60.6%	43.7%	40.2%	41.7%	58.5%
7	59.8%	63.5%	45.1%	62.0%	59.8%	67.7%
8	59.8%	58.0%	41.7%	41.6%	58.3%	75.9%
9	40.2%	59.8%	59.8%	58.1%	60.0%	61.2%
10	55.2%	72.1%	59.9%	63.3%	59.8%	59.8%
Mean	**57.1%**	**60.9%**	**53.7%**	**51.4%**	**54.6%**	**57.2%**
SD	5.8%	4.2%	6.9%	8.7%	7.5%	9.4%
Best	60.0%	72.1%	59.9%	63.3%	60.0%	75.9%
Worst	40.2%	56.2%	41.7%	40.2%	41.7%	40.2%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 1: Training Accuracy of CNN Models on Image Level in Subject-Dependent Experiment

Accuracy / Fold	Training Accuracy on Image Level					
	ROI-Based ResNet18	ROI-Based ResNet50	ROI-Based ResNet101	ROI-Based DenseNet201	ROI-Based InceptionV3	ROI-Based InceptionResNetV2
1	94.1%	100.0%	100.0%	95.6%	99.7%	99.1%
2	82.1%	100.0%	99.4%	100.0%	99.8%	99.2%
3	76.9%	100.0%	99.8%	95.8%	99.9%	98.2%
4	94.3%	100.0%	99.7%	97.2%	99.6%	98.4%
5	97.4%	99.9%	100.0%	100.0%	99.9%	99.3%
6	61.2%	99.7%	100.0%	95.3%	99.9%	99.7%
7	61.7%	99.9%	99.9%	94.3%	100.0%	98.5%
8	94.8%	99.9%	99.7%	98.2%	99.4%	99.5%
9	82.2%	99.9%	100.0%	100.0%	99.9%	99.9%
10	54.2%	99.6%	100.0%	99.9%	100.0%	99.9%
Mean	**79.9%**	**99.9%**	**99.8%**	**97.6%**	**99.8%**	**99.2%**
SD	15.2%	0.1%	0.2%	2.1%	0.2%	0.6%
Best	97.4%	100.0%	100.0%	100.0%	100.0%	99.9%
Worst	54.2%	99.6%	99.4%	94.3%	99.4%	98.2%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 2: Training Accuracy of ROI-Based CNN Models on Image Level in Subject-Dependent Experiment

Accuracy / Fold	Training Accuracy on Image Level		
	LBP-Based SVM	HOG-Based SVM	SIFT-Based SVM
1	79.7%	76.3%	62.1%
2	79.6%	76.4%	62.0%
3	79.6%	76.8%	62.0%
4	80.0%	76.3%	62.2%
5	80.0%	76.3%	61.9%
6	79.5%	76.3%	61.9%
7	80.0%	76.2%	61.9%
8	79.6%	76.1%	61.8%
9	80.3%	75.8%	62.2%
10	79.5%	74.6%	60.2%
Mean	**79.8%**	**76.1%**	**61.8%**
SD	0.3%	0.6%	0.6%
Best	80.3%	76.8%	62.2%
Worst	79.5%	74.6%	60.2%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation
** SVM Kernel: Gaussian Kernel

Table 3: Training Accuracy of SVM Classifiers on Image Level in Subject-Dependent Experiment

Accuracy / Fold	Test Accuracy on Image Level					
	ResNet18	ResNet50	ResNet101	DenseNet201	InceptionV3	Inception-ResNetV2
1	60.6%	59.8%	61.4%	60.6%	62.2%	55.9%
2	60.6%	57.5%	55.1%	40.2%	59.8%	55.9%
3	54.3%	59.8%	59.8%	40.9%	60.6%	53.5%
4	59.8%	60.6%	49.6%	56.7%	40.9%	40.9%
5	59.8%	59.8%	63.0%	47.2%	49.6%	49.6%
6	59.8%	55.1%	46.5%	40.2%	40.9%	51.2%
7	59.8%	58.3%	48.0%	60.6%	59.8%	62.2%
8	59.8%	63.0%	40.9%	43.3%	59.1%	70.1%
9	40.2%	59.8%	59.8%	55.9%	59.8%	57.5%
10	56.7%	64.6%	58.3%	59.1%	59.1%	59.8%
Mean	**57.2%**	**59.8%**	**54.3%**	**50.5%**	**55.2%**	**55.7%**
SD	6.0%	2.5%	7.1%	8.4%	7.8%	7.4%
Best	60.6%	64.6%	63.0%	60.6%	62.2%	70.1%
Worst	40.2%	55.1%	40.9%	40.2%	40.9%	40.9%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 4: Test Accuracy of CNN Models on Image Level in Subject-Dependent Experiment

Accuracy / Fold	Test Accuracy on Image Level					
	ROI-Based ResNet18	ROI-Based ResNet50	ROI-Based ResNet101	ROI-Based DenseNet201	ROI-Based InceptionV3	ROI-Based InceptionResNetV2
1	77.5%	86.3%	84.9%	84.0%	84.3%	82.7%
2	77.2%	84.2%	84.7%	85.2%	86.1%	85.0%
3	74.1%	84.2%	83.7%	77.8%	85.6%	83.5%
4	85.0%	87.8%	87.8%	82.0%	86.1%	84.1%
5	80.8%	82.8%	87.9%	90.0%	85.5%	86.6%
6	62.0%	85.5%	91.3%	84.6%	89.1%	91.3%
7	61.1%	89.3%	87.7%	84.0%	87.8%	86.5%
8	86.6%	88.0%	85.3%	81.7%	89.3%	87.5%
9	76.9%	87.1%	86.2%	87.6%	86.0%	85.7%
10	52.8%	91.8%	90.7%	88.7%	94.5%	97.4%
Mean	**73.4%**	**86.7%**	**87.0%**	**84.6%**	**87.4%**	**87.0%**
SD	10.6%	2.5%	2.4%	3.4%	2.8%	4.1%
Best	86.6%	91.8%	91.3%	90.0%	94.5%	97.4%
Worst	52.8%	82.8%	83.7%	77.8%	84.3%	82.7%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 5: Test Accuracy of ROI-Based CNN Models on Image Level in Subject-Dependent Experiment

Accuracy / Fold	Test Accuracy on Image Level		
	LBP-Based SVM	HOG-Based SVM	SIFT-Based SVM
1	75.5%	54.9%	61.0%
2	74.8%	52.8%	60.6%
3	75.5%	52.4%	62.0%
4	71.7%	55.9%	61.1%
5	75.4%	54.9%	63.1%
6	77.1%	53.7%	61.7%
7	71.0%	57.5%	62.7%
8	75.3%	58.1%	62.9%
9	71.1%	58.2%	61.0%
10	74.7%	61.1%	58.3%
Mean	**74.2%**	**56.0%**	**61.4%**
SD	2.0%	2.6%	1.3%
Best	77.1%	61.1%	63.1%
Worst	71.0%	52.4%	58.3%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation
** SVM Kernel: Gaussian Kernel

Table 6: Test Accuracy of SVM Classifiers on Image Level in Subject-Dependent Experiment

Tables 1, 2, 4 and 5 show the comparison of CNN models and ROI-based CNN models in the subject-dependent experiment.

It can be seen that for the training accuracy on image level, all the CNN models with our proposed ROI-based algorithms obtain higher mean of ten-fold accuracy than the CNN models without our algorithms. As shown in Tables 1 and 2, the maximal improvement is for the DenseNet201 that is 46.2% (i.e. $97.6\% - 51.4\%$). The minimal improvement is for the ResNet18 that is 22.8% (i.e. $79.9\% - 57.1\%$).

Similarly, for the test accuracy on image level, all the CNN models with our proposed ROI-based algorithms obtain higher mean of ten-fold accuracy than the CNN models without our algorithms. As shown in Tables 4 and 5, the maximal improvement is for the DenseNet201 that is 34.1% (i.e. $84.6\% - 50.5\%$). The minimal improvement is for the ResNet18 that is 16.2% (i.e. $73.4\% - 57.2\%$).

The average mean of ten-fold test accuracy of the CNN models with our proposed ROI-based algorithms on image level is 84.4% (i.e. (73.4% + 86.7% + 87.0% + 84.6% + 87.4% + 87.0%) / 6), while the average mean of ten-fold test accuracy of the CNN models without our proposed ROI-based algorithms on image level is 55.4% (i.e. (57.2% + 59.8% + 54.3% + 50.5% + 55.2% + 55.7%) / 6). Hence, the

improvement is 84.4% − 55.4% = 29.0%. The paired t-test [83] for the CNN models with our proposed ROI-based algorithms against the CNN models without our proposed ROI-based algorithms shows that the p-value (1-tailed) is 0.000065 $<$ 0.01, which demonstrates the effectiveness of our proposed ROI-based algorithms.

Tables 2, 3, 5 and 6 show the comparison of ROI-based CNN models and SVM classifiers in the subject-dependent experiment.

It can be seen that the average mean of ten-fold test accuracy of the CNN models with our proposed ROI-based algorithms on image level is 84.4% (i.e. (73.4% + 86.7% + 87.0% + 84.6% + 87.4% + 87.0%) / 6), while the average mean of ten-fold test accuracy of the SVM classifiers on image level is 63.9% (i.e. (74.2% + 56.0% + 61.4%) / 3). Hence, the improvement is 84.4% − 63.9% = 20.5%, which demonstrates the SVM classifiers only extract a small amount of information from the data and considerable information may be lost, while the ROI-based CNN models take all information into account, and extract the features from low-dimensional to high-dimensional space by its hierarchical structure, so that they have better classification performance than the SVM classifiers.

On the other hand, it can be seen that the means of ten-fold test accuracy of the ROI-based ResNet50 (86.7%), the ROI-based ResNet101 (87.0%), the ROI-based InceptionV3 (87.4%) and the ROI-based Inception-ResNetV2 (87.0%) are very close, while the mean of ten-fold test accuracy of the ROI-based DenseNet201 (84.6%) is distinctly lower than others, which indicates that this dataset may not be suitable for the DenseNet models that make full use of each layer's features. It implies this dataset may not has enough features that could be directly used, hence we need to do more non-linear transformation from these few features, like what the ResNet models and the Inception models do.

Moreover, the means of ten-fold test accuracy of the ROI-based ResNet50, the ROI-based ResNet101, the ROI-based InceptionV3 and the ROI-based Inception-ResNetV2 reach 86.7%, 87.0%, 87.4% and 87.0%, respectively, which demonstrates the practicality of our selected CNN models for the proposed classification system.

4.2. Subject-Independent Experiment

In this experiment, we use the new collected 556 clinical full-size NF-NBI images from additional 18 subjects to test the trained classifiers from the previous experiment on both image level and subject level. We retain all the components of the proposed system shown in Figure 3 except the classifier part, which means the inputs of the CNN models are the full-size NF-NBI images, while the inputs of the ROI-based CNN models and the SVM classifiers are the generated image patches.

Specifically, we divide all the full-size NF-NBI images (for the CNN models) and the generated image patches (for the ROI-based CNN models and the SVM classifiers) by subject. Then we test all the trained classifiers on image level and record the classification result of each image. After that, we employ the majority voting to obtain the classification results on subject level. That is, if more than 50% of the images corresponding to a subject were classified as one class, we classify this subject as this class. Hence, it is a subject-independent classification.

Tables 7, 8, 9, 10, 11 and 12 show the test accuracy of CNN models, ROI-based CNN models and SVM classifiers on image level and subject level in the subject-independent experiment.

It can be seen that the test accuracy of all the classifiers in the subject-independent experiment are remarkably lower than the test accuracy in the subject-dependent experiment, which demonstrates the intra-class variety including the complicated shooting conditions in the esophagus of human and the difference of esophagus' physiological characteristics between different human significantly affect the performance of the classifiers.

Accuracy / Fold	Test Accuracy on Image Level					
	ResNet18	ResNet50	ResNet101	DenseNet201	InceptionV3	Inception-ResNetV2
1	60.1%	49.8%	59.5%	58.8%	58.0%	46.1%
2	59.2%	57.2%	54.6%	41.3%	59.0%	51.1%
3	57.7%	59.0%	59.0%	46.1%	59.0%	44.3%
4	60.8%	59.8%	53.8%	54.9%	40.8%	41.0%
5	53.8%	59.0%	58.2%	49.5%	50.8%	52.1%
6	51.6%	46.1%	44.1%	41.0%	42.8%	55.2%
7	59.0%	57.0%	48.4%	58.3%	59.0%	63.7%
8	53.8%	57.8%	41.0%	41.7%	61.1%	67.2%
9	41.0%	59.0%	58.8%	59.3%	58.8%	59.3%
10	51.6%	59.5%	59.2%	52.6%	59.3%	59.0%
Mean	**54.9%**	**56.4%**	**53.6%**	**50.4%**	**54.9%**	**53.9%**
SD	5.7%	4.4%	6.5%	7.1%	7.0%	8.1%
Best	60.8%	59.8%	59.5%	59.3%	61.1%	67.2%
Worst	41.0%	46.1%	41.0%	41.0%	40.8%	41.0%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 7: Test Accuracy of CNN Models on Image Level in Subject-Independent Experiment

Accuracy / Fold	Test Accuracy on Image Level					
	ROI-Based ResNet18	ROI-Based ResNet50	ROI-Based ResNet101	ROI-Based DenseNet201	ROI-Based InceptionV3	ROI-Based InceptionResNetV2
1	55.6%	60.9%	67.8%	62.0%	67.7%	70.4%
2	60.9%	69.4%	63.1%	70.0%	67.7%	72.1%
3	49.4%	62.1%	67.3%	60.4%	67.3%	73.6%
4	64.0%	65.8%	67.4%	58.0%	69.2%	66.7%
5	65.0%	65.0%	61.9%	66.7%	69.4%	67.6%
6	44.0%	62.5%	66.8%	62.1%	69.5%	69.3%
7	44.5%	66.2%	64.7%	54.4%	70.0%	70.8%
8	60.6%	63.5%	64.0%	58.2%	71.0%	70.7%
9	57.1%	67.5%	66.1%	68.5%	69.3%	71.5%
10	55.9%	65.5%	67.6%	58.2%	69.1%	67.4%
Mean	**55.7%**	**64.9%**	**65.7%**	**61.8%**	**69.0%**	**70.0%**
SD	7.1%	2.5%	2.0%	4.8%	1.1%	2.1%
Best	65.0%	69.4%	67.8%	70.0%	71.0%	73.6%
Worst	44.0%	60.9%	61.9%	54.4%	67.3%	66.7%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 8: Test Accuracy of ROI-Based CNN Models on Image Level in Subject-Independent Experiment

Accuracy / Fold	Test Accuracy on Image Level		
	LBP-Based SVM	HOG-Based SVM	SIFT-Based SVM
1	60.3%	46.6%	44.1%
2	60.1%	45.3%	43.6%
3	60.3%	46.9%	44.1%
4	60.0%	45.0%	44.1%
5	61.5%	45.7%	43.8%
6	60.4%	49.2%	43.9%
7	60.4%	49.0%	43.7%
8	60.6%	47.2%	43.8%
9	61.2%	48.5%	43.9%
10	62.5%	47.5%	42.7%
Mean	**60.7%**	**47.1%**	**43.8%**
SD	0.7%	1.4%	0.4%
Best	62.5%	49.2%	44.1%
Worst	60.0%	45.0%	42.7%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation
** SVM Kernel: Gaussian Kernel

Table 9: Test Accuracy of SVM Classifiers on Image Level in Subject-Independent Experiment

Accuracy / Fold	Test Accuracy on Subject Level					
	ResNet18	ResNet50	ResNet101	DenseNet201	InceptionV3	Inception-ResNetV2
1	66.7%	61.1%	66.7%	66.7%	66.7%	50.0%
2	66.7%	66.7%	66.7%	33.3%	66.7%	44.4%
3	66.7%	66.7%	66.7%	33.3%	66.7%	27.8%
4	66.7%	66.7%	66.7%	66.7%	33.3%	33.3%
5	66.7%	66.7%	66.7%	38.9%	38.9%	55.6%
6	66.7%	33.3%	33.3%	33.3%	33.3%	72.2%
7	66.7%	66.7%	33.3%	66.7%	66.7%	66.7%
8	66.7%	66.7%	33.3%	33.3%	66.7%	72.2%
9	33.3%	66.7%	66.7%	66.7%	66.7%	66.7%
10	66.7%	66.7%	66.7%	61.1%	66.7%	66.7%
Mean	**63.3%**	**62.8%**	**56.7%**	**50.0%**	**57.2%**	**55.6%**
SD	10.0%	10.0%	15.3%	15.7%	14.5%	15.3%
Best	66.7%	66.7%	66.7%	66.7%	66.7%	72.2%
Worst	33.3%	33.3%	33.3%	33.3%	33.3%	27.8%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 10: Test Accuracy of CNN Models on Subject Level in Subject-independent Experiment

Accuracy / Fold	Test Accuracy on Subject Level					
	ROI-Based ResNet18	ROI-Based ResNet50	ROI-Based ResNet101	ROI-Based DenseNet201	ROI-Based InceptionV3	ROI-Based InceptionResNetV2
1	66.7%	55.6%	83.3%	66.7%	72.2%	88.9%
2	72.2%	77.8%	61.1%	77.8%	72.2%	77.8%
3	66.7%	55.6%	72.2%	55.6%	66.7%	83.3%
4	77.8%	66.7%	77.8%	66.7%	72.2%	77.8%
5	72.2%	61.1%	55.6%	83.3%	83.3%	72.2%
6	66.7%	55.6%	83.3%	83.3%	83.3%	77.8%
7	66.7%	72.2%	72.2%	61.1%	83.3%	72.2%
8	72.2%	55.6%	55.6%	66.7%	83.3%	77.8%
9	72.2%	72.2%	72.2%	77.8%	83.3%	77.8%
10	33.3%	72.2%	61.1%	50.0%	77.8%	72.2%
Mean	**66.7%**	**64.4%**	**69.4%**	**68.9%**	**77.8%**	**77.8%**
SD	11.7%	8.3%	10.0%	10.9%	6.1%	5.0%
Best	77.8%	77.8%	83.3%	83.3%	83.3%	88.9%
Worst	33.3%	55.6%	55.6%	50.0%	66.7%	72.2%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation

Table 11: Test Accuracy of ROI-Based CNN Models on Subject Level in Subject-Independent Experiment

Fold \ Accuracy	Test Accuracy on Subject Level		
	LBP-Based SVM	HOG-Based SVM	SIFT-Based SVM
1	61.1%	55.6%	61.1%
2	66.7%	55.6%	61.1%
3	66.7%	50.0%	61.1%
4	66.7%	61.1%	61.1%
5	66.7%	50.0%	61.1%
6	66.7%	50.0%	61.1%
7	66.7%	61.1%	61.1%
8	66.7%	55.6%	61.1%
9	66.7%	61.1%	61.1%
10	66.7%	55.6%	61.1%
Mean	**66.1%**	**55.6%**	**61.1%**
SD	1.7%	4.3%	0.0%
Best	66.7%	61.1%	61.1%
Worst	61.1%	50.0%	61.1%

* Mean, SD, Best and Worst denote mean, standard deviation, the best and the worst of results from ten-fold cross-validation
** SVM Kernel: Gaussian Kernel

Table 12: Test Accuracy of SVM Classifiers on Subject Level in Subject-Independent Experiment

Tables 7, 8, 10 and 11 show the comparison of CNN models and ROI-based CNN models in subject-independent experiment.

It can be seen that for the test accuracy on image level, all the CNN models with our proposed ROI-based algorithms obtain higher mean of ten-fold accuracy than the CNN models without our algorithms. As shown in Tables 7 and 8, the maximal improvement is for the InceptionResNetV2 that is 16.1% (i.e. $70.0\% - 53.9\%$). The minimal improvement is for the ResNet18 that is 0.8% (i.e. $55.7\% - 54.9\%$).

Similarly, for the test accuracy on subject level, all the CNN models with our proposed ROI-based algorithms obtain higher mean of ten-fold accuracy than the CNN models without our algorithms. As shown in Tables 10 and 11, the maximal improvement is for the InceptionResNetV2 that is 22.2% (i.e. $77.8\% - 55.6\%$). The minimal improvement is for the ResNet50 that is 1.6% (i.e. $64.4\% - 62.8\%$).

The average mean of ten-fold test accuracy of the CNN models with our proposed ROI-based algorithms on image level is 64.5% (i.e. (55.7% + 64.9% + 65.7% + 61.8% + 69.0% + 70.0%) / 6), while the average mean of ten-fold test accuracy of the CNN models without our proposed ROI-based algorithms on image level is 54.0% (i.e. (54.9% + 56.4% + 53.6% + 50.4% + 54.9% + 53.9%) / 6). Hence, the improvement is $64.5\% - 54.0\% = 10.5\%$. The paired t-test [83] for the CNN models with our proposed ROI-based algorithms against the CNN models without our proposed ROI-based algorithms shows that the p-value (1-tailed) is $0.002488 < 0.01$, which demonstrates the effectiveness of our proposed ROI-based algorithms.

Tables 8, 9, 11 and 12 show the comparison of ROI-based CNN models and SVM classifiers in subject-independent experiment.

It can be seen that the average mean of ten-fold test accuracy of the CNN models with our proposed ROI-based algorithms on image level is 64.5% (i.e. (55.7% + 64.9% + 65.7% + 61.8% + 69.0% + 70.0%) / 6), while the average mean of ten-fold test accuracy of the SVM classifiers on image level is 50.5% (i.e. (60.7% + 47.1% + 43.8%) / 3). Hence, the improvement is $64.5\% - 50.5\% = 14.0\%$, which demonstrates the ROI-based CNN models have better classification performance than the SVM classifiers.

On the other hand, it can be seen that the ROI-based InceptionV3 obtains considerably higher means of ten-fold test accuracy (69.0% and 77.8%) than all the ROI-based ResNet models on both image level

and subject level, and the ROI-based Inception-ResNetV2 obtains a slightly higher mean of ten-fold test accuracy (70.0%) on image level and the same mean of ten-fold test accuracy (77.8%) on subject level. An important difference between the ResNet models and the Inception models is that the Inception models have different sizes of filters in each Inception module, which is shown in Figure 9. It implies the features of this dataset need different sizes of filters in a CNN model to be extracted.

Moreover, it can be seen that the mean of ten-fold test accuracy of the ROI-based InceptionV3 (87.4%) is higher than the ROI-based Inception-ResNetV2 (87.0%) on image level in the subject-dependent experiment, but the mean of ten-fold test accuracy of the ROI-based InceptionV3 (69.0%) is lower than the ROI-based Inception-ResNetV2 (70.0%) on image level in the subject-independent experiment, which suggests the ROI-based Inception-ResNetV2 has better generalization ability than the ROI-based InceptionV3.

Furthermore, the highest mean of ten-fold test accuracy (77.8%) on subject level obtained by using the ROI-based InceptionV3 or the ROI-based Inception-ResNetV2 demonstrates the practicality of our proposed classification system for assisting clinical diagnosis of the NERD.

5. Conclusion

In this paper, we proposed a deep learning classification system using the NF-NBI images to diagnose the NERD. In the classification system, we proposed an algorithm to automatically extract the ROI from the NF-NBI images and then generated image patches through a patch-generating algorithm. After that, we trained six representative state-of-the-art deep CNN models (ResNet18, ResNet50, ResNet101, DenseNet201, InceptionV3, and Inception-ResNetV2) to extract robust hierarchical features from these patches and classify them based on the hierarchical features. Finally, to determine the classification results of each subject, majority voting was employed to the corresponding generated NF-NBI image patches. We verified our classification system by ten-fold cross-validation using the clinical dataset. We performed the subject-dependent and the subject-independent experiments. In both experiments, we compared the classification performance of the ROI-based CNN models (the CNN models with our proposed ROI-based algorithms) with the CNN models. Meanwhile, we compared the classification performance of the ROI-based CNN models with the SVM classifiers (LBP-based SVM, HOG-based SVM, and SIFT-based SVM).

The results from the experiments demonstrated that the ROI-based CNN models obtained higher average mean of ten-fold test accuracy on image level than the CNN models in the subject-dependent experiment (29.0% improvement) and the subject-independent experiment (10.5% improvement), which demonstrated the effectiveness of our proposed ROI-based algorithms. Meanwhile, the ROI-based CNN models obtained higher average mean of ten-fold test accuracy on image level than the SVM classifiers in the subject-dependent experiment (20.5% improvement) and the subject-independent experiment (14.0% improvement), which demonstrated the ROI-based CNN models had better classification performance than the SVM classifiers.

Among the ROI-based CNN models, the ROI-based InceptionV3 model achieved the best classification performance in the subject-dependent experiment, while the ROI-based Inception-ResNetV2 model achieved the best classification performance in the subject-independent experiment, which suggested the ROI-based Inception-ResNetV2 model had better generalization ability than the ROI-based InceptionV3

model. Moreover, the highest mean of ten-fold test accuracy (77.8%) on subject level obtained by using the ROI-based InceptionV3 model or the ROI-based Inception-ResNetV2 model demonstrated the practicality of our proposed classification system for assisting clinical diagnosis of the NERD.

In our proposed classification system, we used the majority voting (refer to Section 3.5) to determine the class of a subject, which could be replaced by weighted voting in the future. Because different image patches corresponding to a subject may have different classification accuracy corresponding to the classification on subject level. For example, some image patches may have fewer task-specific features and more common image features (e.g. textures, edges, angular points, and marginal points) than others, while the classification accuracy based on the task-specific features may be better than the classification accuracy based on the common image features. In such cases, if a higher voting weight is given to the image patches that have more task-specific features, the classification performance on subject level could be improved. In future work, we plan to design an algorithm to distinguish which image patches have more task-specific features and which image patches have more common image features. Then, assigning a higher voting weight to the image patches that have more task-specific features may improve the classification performance on subject level.

Disclosure statement

No potential conflict of interest was reported by the authors.

Acknowledgement

This work was partly supported by King's College London and China Scholarship Council.

References

[1] C. Antunes, S. A. Curtis, Gastroesophageal Reflux Disease, https://www.ncbi.nlm.nih.gov/books/NBK441938/, accessed May 15, 2019 (2019).

[2] A. F. Peery, E. S. Dellon, J. Lund, S. D. Crockett, C. E. McGowan, W. J. Bulsiewicz, L. M. Gangarosa, M. T. Thiny, K. Stizenberg, D. R. Morgan, et al., Burden of gastrointestinal disease in the United States: 2012 update, Gastroenterology 143 (5) (2012) 1179–1187.

[3] R. S. Sandler, J. E. Everhart, M. Donowitz, E. Adams, K. Cronin, C. Goodman, E. Gemmen, S. Shah, A. Avdic, R. Rubin, The burden of selected digestive diseases in the United States, Gastroenterology 122 (5) (2002) 1500–1511.

[4] R. Fass, M. B. Fennerty, N. Vakil, Nonerosive reflux disease - Current concepts and dilemmas, The American Journal of Gastroenterology 96 (2) (2001) 303–314.

[5] I. M. Modlin, R. H. Hunt, P. Malfertheiner, P. Moayyedi, E. M. Quigley, G. N. J. Tytgat, J. Tack, R. C. Heading, G. Holtman, S. F. Moss, Diagnosis and management of non-erosive reflux disease - the Vevey NERD Consensus Group, Digestion 80 (2) (2009) 74–88.

[6] C. L. Chen, P. I. Hsu, Current advances in the diagnosis and treatment of nonerosive reflux disease, Gastroenterology Research and Practice 2013 (2013) 1–8.

[7] S. D. Martinez, I. B. Malagon, H. S. Garewal, H. Cui, R. Fass, Non-erosive reflux disease (NERD) - acid reflux and symptom patterns, Alimentary Pharmacology & Therapeutics 17 (4) (2003) 537–545.

[8] R. I. Narayani, M. P. Burton, G. S. Young, Utility of esophageal biopsy in the diagnosis of nonerosive reflux disease, Diseases of the Esophagus 16 (3) (2003) 187–192.

[9] M. Q. Khan, A. Alaraj, F. Alsohaibani, K. Al-Kahtani, S. Jbarah, H. Al-Ashgar, Diagnostic utility of impedance-pH monitoring in refractory non-erosive reflux disease, Journal of Neurogastroenterology and Motility 20 (4) (2014) 497–505.

[10] C. Barrett, Y. Choksi, M. F. Vaezi, Mucosal impedance: a new approach to diagnosing gastroesophageal reflux disease and eosinophilic esophagitis, Current Gastroenterology Reports 20 (33) (2018) 1–7.

[11] P. Vasuki, J. Kanimozhi, M. B. Devi, A survey on image preprocessing techniques for diverse fields of medical imagery, in: Proceedings of the 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE), IEEE, 2017, pp. 1–6.

[12] P. Deepa, M. Suganthi, Performance evaluation of various denoising filters for medical image, International Journal of Computer Science and Information Technologies 5 (3) (2014) 4205–4209.

[13] D. Bhonsle, V. Chandra, G. R. Sinha, Medical image denoising using bilateral filter, International Journal of Image, Graphics and Signal Processing 4 (6) (2012) 36–43.

[14] S. Saladi, N. Amutha Prabha, Analysis of denoising filters on MRI brain images, International Journal of Imaging Systems and Technology 27 (3) (2017) 201–208.

[15] M. Mafi, H. Martin, M. Cabrerizo, J. Andrian, A. Barreto, M. Adjouadi, A comprehensive survey on impulse and Gaussian denoising filters for digital images, Signal Processing 157 (2019) 236–260.

[16] Y. Wang, H. Zhou, Total variation wavelet-based medical image denoising, International Journal of Biomedical Imaging 2006 (2006) 1–6.

[17] S. Gupta, R. C. Chauhan, S. C. Sexana, Wavelet-based statistical approach for speckle reduction in medical ultrasound images, Medical and Biological Engineering and Computing 42 (2) (2004) 189–192.

[18] S. Sudha, G. R. Suresh, R. Sukanesh, Speckle noise reduction in ultrasound images by wavelet thresholding based on weighted variance, International Journal of Computer Theory and Engineering 1 (1) (2009) 7–12.

[19] Z. Gan, F. Zou, N. Zeng, B. Xiong, L. Liao, H. Li, X. Luo, M. Du, Wavelet denoising algorithm based on NDOA compressed sensing for fluorescence image of microarray, IEEE Access 7 (2019) 13338–13346.

[20] M. Foracchia, E. Grisan, A. Ruggeri, Luminosity and contrast normalization in retinal images, Medical Image Analysis 9 (3) (2005) 179–190.

[21] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, P. F. Sharp, Automated microaneurysm detection using local contrast normalization and local vessel detection, IEEE Transactions on Medical Imaging 25 (9) (2006) 1223–1232.

[22] M. Zhou, K. Jin, S. Wang, J. Ye, D. Qian, Color retinal image enhancement based on luminosity and contrast adjustment, IEEE Transactions on Biomedical Engineering 65 (3) (2017) 521–527.

[23] H. Li, O. Chutatape, Automated feature extraction in color retinal images by a model based approach, IEEE Transactions on Biomedical Engineering 51 (2) (2004) 246–254.

[24] S. Ravishankar, A. Jain, A. Mittal, Automated feature extraction for early detection of diabetic retinopathy in fundus images, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 210–217.

[25] N. B. Bahadure, A. K. Ray, H. P. Thethi, Image analysis for MRI based brain tumor detection and feature extraction using biologically inspired BWT and SVM, International Journal of Biomedical Imaging 2017 (2017).

[26] B. Willmore, R. J. Prenger, M. C. K. Wu, J. L. Gallant, The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform, Neural Computation 20 (6) (2008) 1537–1564.

[27] R. P. Remya, K. P. Soman, Berkeley wavelet transform based image watermarking, in: Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE, 2009, pp. 357–359.

[28] M. C. Chuang, J. N. Hwang, K. Williams, Supervised and unsupervised feature extraction methods for underwater fish species recognition, in: Proceedings of the 2014 ICPR Workshop on Computer Vision for Analysis of Underwater Imagery, IEEE, 2014, pp. 33–40.

[29] A. Datta, S. Ghosh, A. Ghosh, Supervised feature extraction of hyperspectral images using partitioned maximum margin criterion, IEEE Geoscience and Remote Sensing Letters 14 (1) (2016) 82–86.

[30] O. Irsoy, E. Alpaydın, Unsupervised feature extraction with autoencoder trees, Neurocomputing 258 (2017) 63–73.

[31] C. T. Sari, C. Gunduz Demir, Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images, IEEE Transactions on Medical Imaging 38 (5) (2018) 1139–1149.

[32] T. Ojala, M. Pietikainen, D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, in: Proceedings of the 12th International Conference on Pattern Recognition, Vol. 1, IEEE, 1994, pp. 582–585.

[33] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis & Machine Intelligence 24 (7) (2002) 971–987.

[34] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 International Conference on Computer Vision and Pattern Recognition, Vol. 1, 2005, pp. 886–893.

[35] D. G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 99, IEEE, 1999, pp. 1150–1157.

[36] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[37] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological Cybernetics 36 (4) (1980) 193–202.

[38] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Computation 1 (4) (1989) 541–551.

[39] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al., Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[40] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778.

[42] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 4700–4708.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 1–9.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 2818–2826.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 4278–4284.

[46] H. Soltanian Zadeh, J. P. Windham, D. J. Peck, Optimal linear transformation for MRI feature extraction, in: Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis, IEEE, 1996, pp. 64–73.

[47] F. Özyurt, E. Sert, E. Avci, E. Dogantekin, Brain tumor detection based on convolutional neural network with neutrosophic expert maximum fuzzy sure entropy, Measurement 147 (2019) 106830.

[48] G. Jia, H. K. Lam, J. Liao, R. Wang, Classification of electromyographic hand gesture signals using machine learning techniques, Neurocomputing 401 (2020) 236–248.

[49] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27–48.

[50] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F. E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26.

[51] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM, 1992, pp. 144–152.

[52] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[53] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, D. Liu, Evolving support vector machines using fruit fly optimization for medical data classification, Knowledge-Based Systems 96 (2016) 61–75.

[54] S. S. Mazlan, M. Z. Ayob, Z. A. K. Bakti, Anterior cruciate ligament (ACL) injury classification system using support vector machine (SVM), in: Proceedings of the 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), IEEE, 2017, pp. 1–5.

[55] H. Alquran, I. A. Qasmieh, A. M. Alqudah, S. Alhammouri, E. Alawneh, A. Abughazaleh, F. Hasayen, The melanoma skin cancer detection and classification using support vector machine, in: Proceedings of the 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), IEEE, 2017, pp. 1–5.

[56] M. Zurita, C. Montalba, T. Labbé, J. P. Cruz, J. D. da Rocha, C. Tejos, E. Ciampi, C. Cárcamo, R. Sitaram, S. Uribe, Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data, NeuroImage: Clinical 20 (2018) 724–730.

[57] A. García Floriano, Á. Ferreira Santiago, O. Camacho Nieto, C. Yáñez Márquez, A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images, Computers & Electrical Engineering 75 (2019) 218–229.

[58] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review 65 (6) (1958) 386–408.

[59] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[60] S. H. Ling, F. H. F. Leung, H. K. Lam, Y. S. Lee, P. K. S. Tam, A novel genetic-algorithm-based neural network for short-term load forecasting, IEEE Transactions on Industrial Electronics 50 (4) (2003) 793–799.

[61] S. H. Ling, H. K. Lam, Playing Tic-Tac-Toe using genetic neural network with double transfer functions, Journal of Intelligence Learning Systems and Application 3 (2011) 37–44.

[62] F. H. F. Leung, H. K. Lam, S. H. Ling, P. K. S. Tam, Tuning of the structure and parameters of a neural network using an improved genetic algorithm, IEEE Transactions on Neural Networks 14 (1) (2003) 79–88.

[63] S. H. Ling, F. H. F. Leung, H. K. Lam, An improved genetic algorithm based fuzzy-tuned neural network, International Journal of Neural Systems 15 (6) (2005) 457–474.

[64] H. K. Lam, U. Ekong, B. Xiao, G. Ouyang, H. Liu, K. Y. Chan, S. H. Ling, Variable weight neural networks and their applications on material surface and epilepsy seizure phase classifications, Neurocomputing 149 (2015) 1177–1187.

[65] B. Xiao, W. Xu, J. Guo, H. K. Lam, G. Jia, W. Hong, H. Ren, Depth estimation of hard inclusions in soft tissue by autonomous robotic palpation using deep recurrent neural network, IEEE Transactions on Automation Science and Engineering (2020).

[66] Y. Wang, Y. Chen, N. Yang, L. Zheng, N. Dey, A. S. Ashour, V. Rajinikanth, J. M. R. S. Tavares, F. Shi, Classification of mice hepatic granuloma microscopic images based on a deep convolutional neural network, Applied Soft Computing 74 (2019) 40–50.

[67] S. Pang, A. Du, M. A. Orgun, Z. Yu, A novel fused convolutional neural network for biomedical image classification, Medical & Biological Engineering & Computing 57 (1) (2019) 107–121.

[68] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, T. Wang, Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset, Neurocomputing 194 (2016) 87–94.

[69] J. Yang, Y. Xie, L. Liu, B. Xia, Z. Cao, C. Guo, Automated dental image analysis by deep learning on small dataset, in: Proceedings of the 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Vol. 1, IEEE, 2018, pp. 492–497.

[70] P. Shi, C. Wu, J. Zhong, H. Wang, Deep learning from small dataset for BI-RADS density classification of mammography images, in: Proceedings of the 2019 10th International Conference on Information Technology in Medicine and Education (ITME), IEEE, 2019, pp. 102–109.

[71] Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, M. Fujimoto, Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis, British Journal of Dermatology 180 (2) (2019) 373–381.

[72] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.

[73] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (5) (1989) 359–366.

[74] C. R. Huang, Y. T. Chen, W. Y. Chen, H. C. Cheng, B. S. Sheu, Gastroesophageal reflux disease diagnosis using hierarchical heterogeneous descriptor fusion support vector machine, IEEE Transactions on Biomedical Engineering 63 (3) (2015) 588–599.

[75] R. Fass, Erosive esophagitis and nonerosive reflux disease (NERD): comparison of epidemiologic, physiologic, and therapeutic characteristics, Journal of Clinical Gastroenterology 41 (2) (2007) 131–137.

[76] S. Gulati, J. Bernth, J. Liao, D. Poliyivets, S. Chatu, A. Emmanuel, A. Haji, H. Liu, B. Hayee, Tu1962 Near focus narrow band imaging driven artificial intelligence for the diagnosis of gastroesophageal reflux disease, Gastrointestinal Endoscopy 89 (6) (2019) AB633.

[77] Y. Sano, S. Tanaka, S.-e. Kudo, S. Saito, T. Matsuda, Y. Wada, T. Fujii, H. Ikematsu, T. Uraoka, N. Kobayashi, et al., Narrow-band imaging (NBI) magnifying endoscopic classification of colorectal tumors proposed by the Japan NBI Expert Team, Digestive Endoscopy 28 (5) (2016) 526–533.

[78] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2009) 1345–1359.

[79] M. Kato, J. Yamamoto, Y. Shimizu, H. Takeda, M. Asaka, Magnifying endoscopic findings of non-erosive reflux disease, Digestive Endoscopy 18 (2006) S33–S35.

[80] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, F. F. Li, Imagenet: A large-scale hierarchical image database, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[81] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, IGI Global, 2010, pp. 242–264.

[82] C. E. Metz, Basic principles of ROC analysis, Seminars in Nuclear Medicine 8 (4) (1978) 283–298.

[83] H. Hsu, P. A. Lachenbruch, Paired t test, Encyclopedia of Biostatistics 6 (2005).