Case-based Reasoning of a Deep Learning Network for Prediction of Early Stage of Oesophageal Cancer*

Xiaohong Gao[1], Barbara Braden[2], Leishi Zhang[1], Stephen Taylor[3], Wei Pang[4], and Miltos Petridis[1]

[1] Middlesex University, London, UK
[2] John Radcliffe Hospital Oxford, UK
[3] MRC Weatherall Institute of Molecular Medicine Oxford, UK
[4] University of Aberdeen, Old Aberdeen, UK

Abstract. Case-Based Reasoning (CBR) is a form of analogical reasoning in which the information for a (new) query case is determined based on the known cases in a database with established information. While deep machine learning tech-niques of AI have demonstrated state of the art results in many fields, their transparency status of those hidden layers have cast double in many applica-tions, especially in the medical field, where clinicians need to know the rea-sons of decision making delegated by a computer system. This study aims to provide a visual explanation while performing classification of endoscopic oesophageal videos. Towards this end, this work integrates the interpretation and decision-making together by producing a set of profiles that in appearance resemble the training samples and hence explain the outcome of classification, in an attempt to allay the concerns that using a different model to explain the predictions while employing varying priors from the original network. Fur-thermore, different from many explainable networks that highlight key regions or points of the input that activate the network, this work is based on whole training images, i.e. case-based, where each training image belongs to one of the classes. Preliminary results have demonstrated the classification accuracy of 95% for training and 75% for testing while applying 500 training data (with 10% for testing split randomly) for each of three classes of 'cancer', 'high grade' and 'suspicious' of oesophageal squamous cancer from endoscopy videos. When training with 2000 samples for the two classes of 'high grade' and 'suspicious', the testing result delivers an accuracy of 77%, implying the impact of sample sizes. Future work includes collection of large annotated dataset and improving classification accuracy.

Keywords: deep learning · visual recognition · classification.

* This project is financially funded by the Cancer Research UK (CRUK). Their financial support is gratefully acknowledged.

1 Introduction

While machine learning has turned into an integral and indispensable technique in assisting people to process big data in the current digital era, its transparency and interpretability become increasingly important. For example, in radiology, lack of transparency has caused challenges to Food and Drug Administration (FDA) approval of deep learning-based software products [1]. This is because artificial neural networks are consisted of high dimensional nonlinear functions that do not naturally lend an explanation to human beings. Consequently, making the black box transparent has gain more interests in both research and application communities.

1.1 Explainable neural network

Neural networks are designed mainly for achieving state of the art accuracy results, whereas interpretability is only analysed after the training, aiming to explain the trained model or the learned high-level features. As a result, this kind of interpretability analysis requires a separate model to decipher the achieved results, which leads to the questions that whether these explanations are creditable as they derive from a separate modelling process with priors that are not part of the training from the original networks [2]. To ensure the interpretation of the network is meaningful, understandable, and creditable, many researches have since been conducted with a focus on the visualisation of parts of images that most strongly activate a given feature map [3, 4]. More recently, progress has been made to allow case-based interpretation through prototyping [5, 6]. Rather than enforcing a particular structure on feature maps, prototype-based approach introduces a special prototype layer for explanation of decision making. While prototype classification constitutes a classical form of case-based reasoning [7], within a neural network, the analysis task takes place in a latent space, i.e. the distance between prototype and observation is measured in a latent space, which is flexible and adaptive and hence is able to realise high quality performance.

Inspired by the work in [6] and autoencoder [7] architecture, this study builds an enhanced network to classify precancerous stages for early diagnosis of oesophagus cancers. This network models a profile layer comprising a list of profiles whereby each profile resembles observations in one of classes in visual appearance. Hence this set of profiles learns toward being a representative of the whole training set. In addition, this network utilises case-based reasoning instead of extractive reasoning by explaining its predictions based on similarity between observations and profile cases, rather than highlighting the most relevant parts of the input. In this work, we use the term of 'profile' instead of 'prototype'. This is because the term 'prototype' has been applied in multiple contexts with varying meanings. For example, in few-shot [8] and zero-shot [9] learning, prototypes are points in the feature space used to represent a single class as well as the distance to the prototype which determines how an observation is classified.

1.2 Challenges of detecting oesophageal squamous cancer

Oesophagus cancer (OC), or cancer of the gullet, is the 8th most common cancer worldwide [10] and the 6th leading cause of cancer-related death [11]. Two main histological types represent the most majority of all oesophageal cancers, which are adenocarcinoma and squamous cell carcinoma cancer (SCC). Worldwide, about 87% of all oesophageal cancers are SCC with the highest incidence rates occurring in Asia, the Middle East and Africa [12, 13].

While the five-year survival rate of oesophagus cancer is less than 20% [14], the rate can be improved significantly to more than 90% if the cancer is detected in its early stages when it still can be treated endoscopically [15]. Hence there is a clinical urgency to improve the detection of oesophageal pre-cancerous stages, e.g. dysplasia, to allow endoscopic treatment and monitoring of affected patients.

Precancerous stages (dysplasia in the oesophageal squamous epithelium) and early stages of SCC are easily missed at the time of conventional White Light Endoscopy (WLE) as these lesions grow usually flat with only subtle changes in colour and in microvasculature as demonstrated in Figure 1 for those suspicious regions ('S' and 'H'), where 'C' refers to 'cancer', 'H' for 'High grade' of possible cancer and 'S' for 'suspicious'. To overcome this shortcoming while viewing WLE images, Narrow-Band Imaging (NBI) can be turned on to display only two wavelengths ($415nm$ (blue) and $540nm$ (green)) (Figure 1(b)) to improve the visibility of those suspected lesions by filtering out the rest of colour bands. Another approach is dye-based chromoendoscopy, i.e. Lugol's staining technique, which highlights dysplastic abnormalities by spraying iodine [16] (Figure 1 (c)).

2 Methodology

2.1 Datasets and data augmentation

In this collection, 600 annotations are provided by a clinician from 350 frames extracted from 15 oesophagus videos with suspected SCC together with four normal subject videos. These data are collected from Oxford NHS University Hospital, UK. These videos last from 10 to 30 minutes with 50 frames per second (FPS). The resolution of these videos is of 1920×1080 pixels whereas still images have varying sizes between 256×256 and 1920×1080 after cropping out personal information.

Three categories of annotations are given, which are SCC 'cancer', 'high grade' of possibility of SCC and 'suspicious' of SCC as illustrated in Figure 1. Since each frame may contain multiple annotations, each annotation is then segmented, augmented, and resized into $128 \times 128 \times 3$ voxels to generate three groups where each group shares only one label. Figure 2 demonstrates the process of data augmentation applied in this work, including clipping, rotating, colouring and blurring. As a result, 500 images are selected from each class (cancer, high grade, suspicious, normal) with 90% of them utilised for training and 10% for testing with a random split.

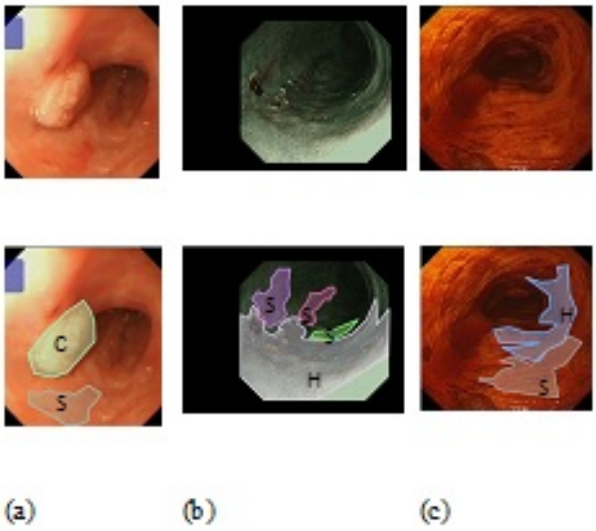

Fig. 1. Examples of SCC where C=cancer, S=suspicious, H=High grade. Top row: original images; bottom: with masks. (a) WLE; (b) NBI; (c) Lugol's.

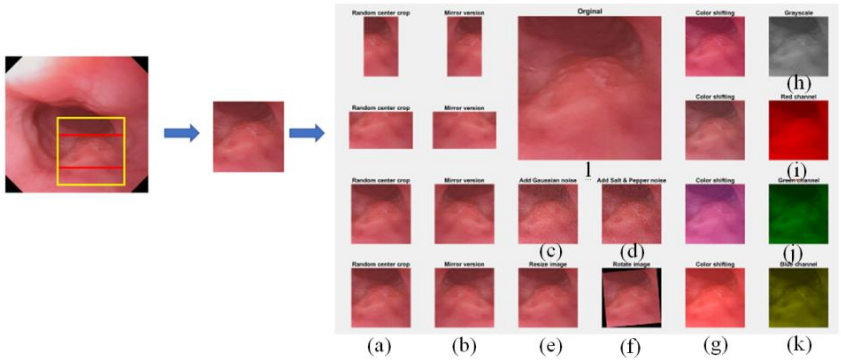

Fig. 2. Illustration of the process of data augmentation. Initial lesion (left most) (red box) is detected from the ground truth mask, then segmented (yellow box) into a segment (middle graph), which is augmented (right most). (a): Random centre crop; (b): Mirror conversion; (c): Add Gaussian noise; (d): Add salt pepper noise; (e): Resize image; (f): Rotation image; (g): Color shifting; (h): Greyscale; (i): Red channel; (j): Green channel; (k): Blue channel; (l): Original image.

2.2 Case-based reasoning of classification of cancerous stages using deep learning network

As illustrated in Figure 3, the proposed case-based reasoning architecture that comprises four components, encoder, decoder, classifier and the reasoning profiles. The network is analogous of an autoencoder architecture, where the profiles, $(p_1, p_2, ...p_m)$ as well as the classifier are in the latent space. These profiles are expected to give the explanation of the decision making towards classification by producing similar images in appearance to one of classes. Hence, when a test image is inputted to the trained model, the model calculates the overall distance between this test image and each of the profile images and delivers the final classification result

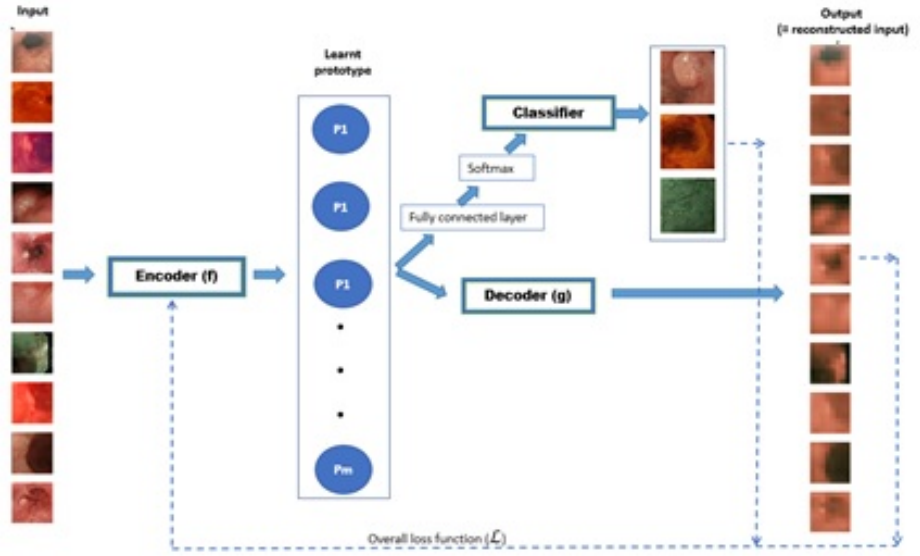

Fig. 3. The proposed profile network that explains the classification.

The function of encoder aims to reduce the dimensionality of the input (as well as noise) and to learn the weights (W) of transformation from input, leading to the final prediction of classes using Eq. (1), whereas the profiles layers (P) in between generates the profile units that resemble in appearance one of the K classes (K=3 in this study, i.e. 'cancer', 'high grade', and 'suspicious') to be studied.

$$p = f(x) = f'(WX + B) \tag{1}$$

Where the input $X = (x_1, x_2, ...x_n)^T$ of n samples with each image $(x_i, i = 1, 2, ...n)$ having a size of $128 \times 128 \times 3$ and produces an set of profiles $p = (p_1, p_2, ..., p_m)^T$.

In Eq. (1), B represents the bias that is generated randomly during the training. The profile number (m) is pre-defined and can have the size of class numbers (K) or more. W refers to the weight matrices in the encoder that are to be determined in the training.

Specifically, f' refers to the calculations from a range convolution layers from an encoder as illustrated in Figure 3.

In this study, the sizes of m varying from 3 to 30 are investigated. It has found that more profiles do not necessary lead to more accurate results as some profiles appear to be redundant by presenting near blank features.

This profile layer computes the squared distance between encoded input z (Eq. (2)) and each of the profile vectors as formatted in Eq. (3).

$$z = \left[f\left(x_1\right), f\left(x_2\right), ...f\left(x_n\right)\right] \tag{2}$$

$$P\left(z\right) = \left[\sum\left(z - p_1\right)^2, \sum\left(z - p_2\right)^2, ... \sum\left(z - p_m\right)^2\right]^T \tag{3}$$

After the profile layer, a fully connected layer and a classification layer follow to compute weighted sums of these distances $W_p\left(P\left(z\right)\right)$, where W_p is the $K \times m$ weight matrix and will be learnt through the training as shown in Figure 3. These weighted sums are then normalized by the *Softmax* layer to output a probability distribution over the K classes. In our case, $K = 4$, which refers to 'cancer', 'high-grade', 'suspicious' and 'normal'.

Hence, the distribution of probability of a test image that belongs to each class will be calculated in the *Softmax* layer that in a form of a vector with K elements, where the k_th ($k = 1, 2, ...K$) component of the output of the *Softmax* layer is defined by

$$S_{\text{Softmax}}\left(V_k\right) = \frac{\exp\left(V_k\right)}{\sum_{i=1}^{K}\exp\left(V_i\right)} \tag{4}$$

where V_k is the k_th component of the vector $V = W_p\left(P\left(z\right)\right) = \left(v_1, ..., v_k\right)$.

During the prediction, the neural network architecture depicted in Figure 3 delivers the class label that has the highest probability among the S vector.

In Figure 3, the Decoder is to reconstruct back the input $x \in X$, based on the profiles, i.e. from $m \times 1$ profile units to construct $128 \times 128 \times 3$ image using a function g as given in Eq. (5), which decodes the encoded feature vectors in $x, x \in X$.

$$x_{-} = g\left(x\right) \tag{5}$$

Hence, the multi-task loss function L for the network of Figure 3 is formulated in Eq. (6) by combining the loss of classification, decoding and two interpretability regularisation measures.

$$L = \lambda_1 L_{\text{classification}} + \lambda_2 L_{\text{decoder}} + \lambda_3 L_{\text{interpreter}-1} + \lambda_4 L_{\text{interpreter}-2} \tag{6}$$

Where λ_1 to λ_4 are the real valued hyper-parameters and applied to adjust the ratios between those four terms.

The classification loss applies the standard cross-entropy function as calculated in Eq. (7).

$$L_{\text{classification}} = \frac{1}{n}\sum_{i}^{n} log\,(\widehat{y_i}) \tag{7}$$

Where n is the total number of data samples, y_i refers to the i_{th} sample label and $\widehat{y_i}$ the predicted label.

The loss function for the reconstruction of decoding can be calculated using mean squared errors *(MSE)* by using Eq. (8).

$$L_{\text{decoder}} = \frac{1}{n}\sum (X - X_{-})^2 \tag{8}$$

Similar to [7], the two interpretability measures are calculated using Eqs.(9) and (10), which are established to safeguard respectively the distances of each profile to be as close as possible to at least one the training samples in the latent space, and the distances of each encoded training sample to be as close to one of the profiles as possible.

$$L_{\text{interpreter}-1} = \frac{1}{m}\sum_{j=1}^{m} \min((p_j - f(x_1^2)), ..., (p_j - f(x_n))^2) \tag{9}$$

$$L_{\text{interpreter}-2} = \frac{1}{n}\sum_{i=1}^{n} \min((p_1 - f(x_i))^2, ..., (p_M - f(x_i))^2) \tag{10}$$

In this way, $L_{\text{interpreter}-1}$ will propel the profile vectors to have meaningful decoding in the pixel space, whereas $L_{\text{interpreter}-2}$ will cluster the training samples closely around profiles in the latent space. Therefore, these two measures will lead to the tight closeness between profiles and training samples in appearance.

3 Results

The implementation is carried out by applying Python with Tensorflow library. The values of λ_1 to λ_4 are set to be 0.85, 0.05, 0.05, and 0.05, to give highest weight to classification. Similar to conventional convolutional neural network, the encoding process is composed of 6 convolutional layers with each one having filter size of 3×3. While the few of other λ_1 to λ_4 values are also workable, this combination appears to deliver the best training model.

After training for 2000 epochs, for classification of three classes of 'cancer', 'high grade', and 'suspicious', the proposed model (Figure 3) achieved accuracy of 94.46% for training and 75% for testing based on 450 training samples and 50 for testing when the profile numbers are trained to be 15.

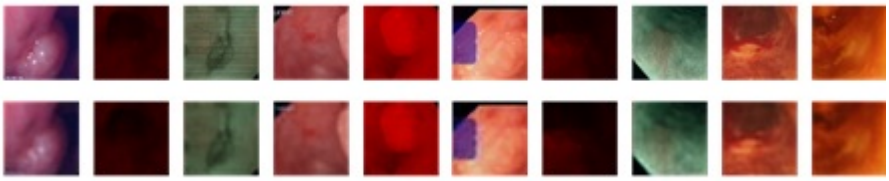

Fig. 4. The demonstration of re-generated samples (bottom) using the trained model of profiles from the original images (top row).

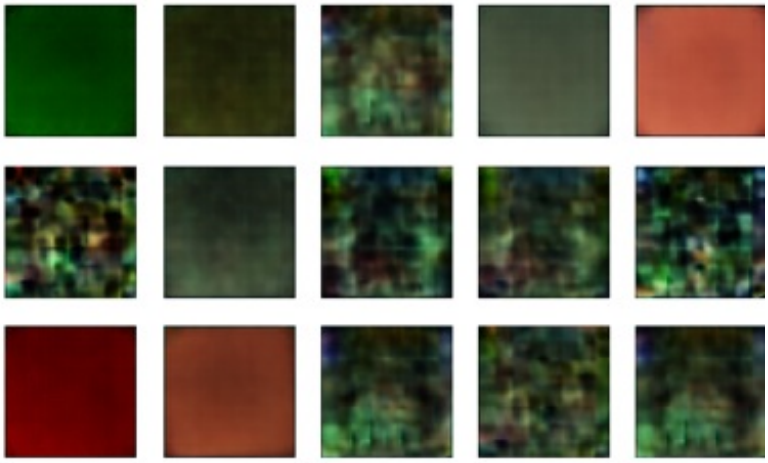

Fig. 5. The fifteen profiles that represent training samples of three classes, i.e., cancer, high grade, and suspicious.

Figure 4 demonstrates the results of decoding to reproduce ten training samples (top row) (randomly selected from training set) using trained profiles. Visually, the regenerated samples (bottom row) appear similar to the original images (top row), indicating the profiles tend to be representative of the training samples.

Figure 5 depicts the 15 profiles that are trained to be representative of 3 classes (i.e. cancer, high grade, suspicious) of training samples, whereas Figure 6 illustrates the 4 profiles for 4 classes (i.e. cancer, high grade, suspicious, normal).

While Figure 6 of four profiles tend to depict each class using only one profile that visually resembles one of the four classes, the accurate of classification is much worse with only 60% accuracy for testing after 2000 training epochs. This is because the training samples have varying forms, for instance, WLE, NBI. Using one profile to represent each class is apparently not sufficient. Another

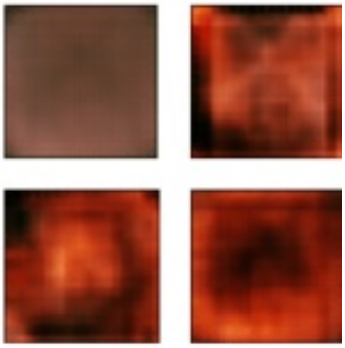

Fig. 6. Four profiles trained for 4 classes, i.e. cancer, high grade, suspicious and normal.

Fig. 7. Ten profiles that are trained with two classes, i.e. 'suspicious' and 'high grade'.

reason could the small training sample size with 500 images for each class after data augmentation.

Since the main purpose of this project is to detect SCC at its early stages, the investigate is also given to classify two classes, which are 'suspicious' and 'high grade', the two classes attracting the largest number of training samples with 2000 each. The ten profiles and reproduction of ten samples are demonstrated respectively in Figures 7 and 8, whereas the classification accuracy of test images is 77%, the highest among classification of 4, 3 and 2 classes. While the class of 'cancer' appears to be more obvious visually in comparison with the other two classes, it has the smallest sample size of 500. It is more challenge to distinguish between 'high grade' and 'suspicious' categories.

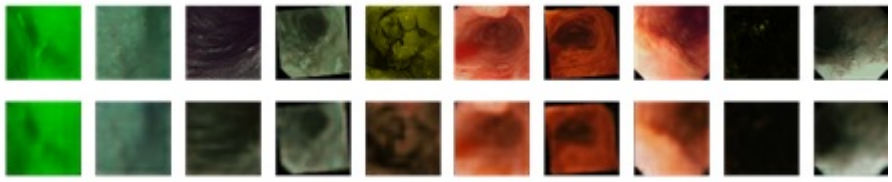

Fig. 8. The reproduction of two class images using the profiles in Figure 7

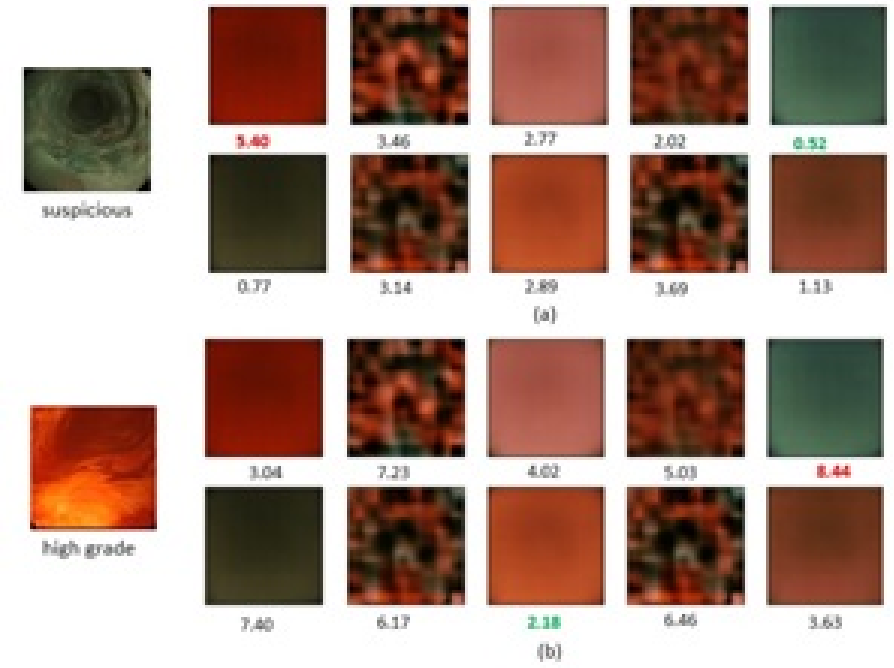

Fig. 9. The illustration of distance between each sample of 'suspicious' (a) and of 'high-grade' (b) (left most column) and ten profiles. The number in red refers to the largest distance (most dissimilar) and in green the shortest distance (most similar) between the test sample and the profile.

Figure 9 demonstrates the distances between test samples (left most) and profiles, where the number in red refers to the biggest distance (very dissimilar) and green the shortest (very similar).

4 Discussion and Future work

While more profiles may cover the variations of training samples, too many of them does not necessarily produce better accuracy. For example, in this study,

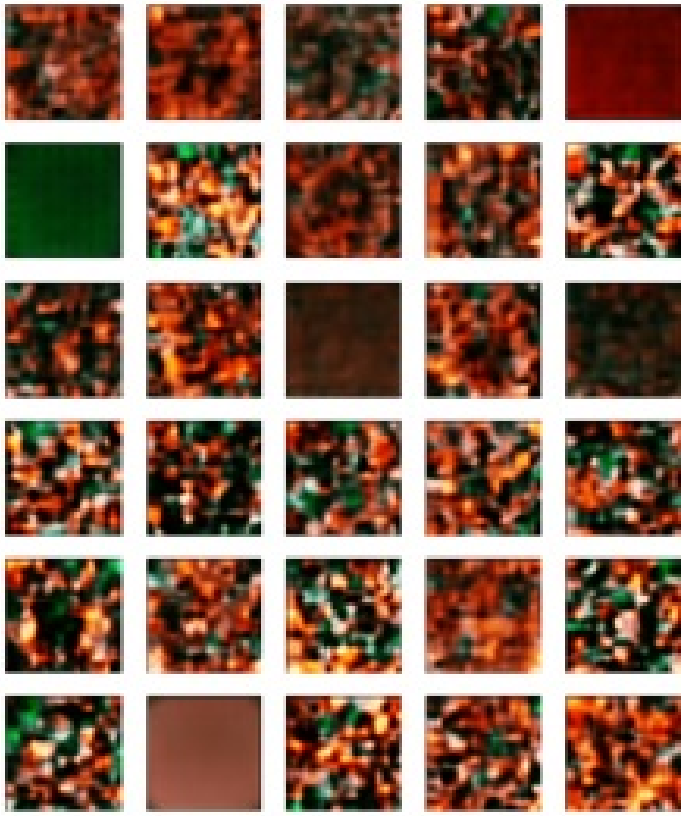

Fig. 10. Thirty profiles trained for three classes after 2000 epochs.

for training 3 classes (i.e. without 'normal' dataset), increasing profile numbers from 15 (Figure 5) to 30 (Figure 10) appear to reduce classification accuracy to 66% from 75%. This again could be due to the limitation of the size of training samples (500 for each class) and will be further investigated in the future. The two classes of 'suspicious' and 'high grade' have the largest number of datasets with 2000 each. Hence the training results appear to be improved from 75% for 3 classes to 77% for two classes, even these two categories appear to be similarly visually.

In addition, when 'normal' class is added, the classification accuracy is decreased, partially due to the similarity between 'suspicious' and 'normal' patterns. Another reason could be again the small size of training samples. Since training takes place using the conventional 6-layer CNN structure (plus one fully connection layer) without transfer learning, small sample size will make considerable impact to the training process. In the future, more data sets will be annotated and applied, in addition to data augmentation.

References

1. Wexler R. : When a computer program keeps you in jail: How computers are harming criminal justice. New York Times. http://www.springer.com/lncs. Last accessed 10 Oct 2019 (2017)
2. Montavon G., Samek W., and Müller K. : Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)
3. Zeiler MD., Fergus R. : Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision 2014 (ECCV) pp. 818–833. (2014)
4. Pinheiro PO., Collobert R. : From image level to pixel-level labelling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015 (CVPR), pp. 1713 –1721. (2015)
5. Kolodner J. : An introduction to case-based reasoning. Artificial Intelligence Review **6**, 3–34 (1992)
6. Li O., Liu H., Chen C., and Rudin C. : Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), (2018)
7. Hinton G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
8. Snell J., Swersky K., and Zemel R. S. : Prototypical networks for few-shot learning. CoRR abs/1703.05175 (2017)
9. Li Y., Wang D. : Zero-shot learning with generative latent prototype model. CoRR abs/1705.09474 (2017)
10. Bray F., Ferlay J., Soerjomataram I., Siegel RL., Torre LA, and Jemal A. : Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. **68**(6), 394–424 (2018)
11. Pennathur A., Gibson MK., Jobe BA., and Luketich JD.: Oesophageal carcinoma. The Lancet **381**(9864), 400–12 (2013)
12. Arnold M., Laversanne M., Brown LM., Devesa SS., and Bray F.: Predicting the Future Burden of Esophageal Cancer by Histological Subtype: International Trends in Incidence up to 2030. Am J Gastroenterol **112**(8), 1247–55 (2017)
13. Arnold M., Soerjomataram I., and J., Forman D. : Global incidence of oesophageal cancer by histological subtype in 2012. Gut. BMJ Publishing Group **64**(3), 381–7 (2015)
14. Siegel R., Ma J., Zou Z., and Jemal A.: Cancer statistics, 2014. CA Cancer J Clin. 3rd ed. American Cancer Society **64**(1), 9–29 (2014)
15. Shimizu Y., Tsukagoshi H., Fujita M., Hosokawa M., Kato M., and Asaka M.: Long-term outcome after endoscopic mucosal resection in patients with esophageal squamous cell carcinoma invading the muscularis mucosae or deeper. Gastrointest Endosc **56**(3), 387–90 (2002)
16. Trivedi PJ., Braden B. :Indications, stains and techniques in chromoendoscopy. QJM **106**(2), 117–31 (2013)