ORIGINAL ARTICLE

Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning and machine learning

Tien-Yu Huang[a,b], Shan-Quan Zhan[c], Peng-Jen Chen[a], Chih-Wei Yang[a], Henry Horng-Shing Lu[d,e,*]

[a]Division of Gastroenterology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, ROC; [b]Taiwan Association for the Study of Small Intestinal Diseases, Taoyuan, Taiwan, ROC; [c]Institute of Computer Science and Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC; [d]Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC; [e]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan, ROC

Abstract

Background: In clinical applications, mucosal healing is a therapeutic goal in patients with ulcerative colitis (UC). Endoscopic remission is associated with lower rates of colectomy, relapse, hospitalization, and colorectal cancer. Differentiation of mucosal inflammatory status depends on the experience and subjective judgments of clinical physicians. We developed a computer-aided diagnostic system using deep learning and machine learning (DLML-CAD) to accurately diagnose mucosal healing in UC patients.

Methods: We selected 856 endoscopic colon images from 54 UC patients (643 images with endoscopic score 0-1 and 213 with score 2-3) from the endoscopic image database at Tri-Service General Hospital, Taiwan. Endoscopic grading using the Mayo endoscopic subscore (MES 0-3) was performed by two reviewers. A pretrained neural network extracted image features, which were used to train three different classifiers—deep neural network (DNN), support vector machine (SVM), and k-nearest neighbor (k-NN) network.

Results: DNN classified MES 0 to 1, representing mucosal healing, vs MES 2 to 3 images with 93.8% accuracy (sensitivity 84.6%, specificity 96.9%); SVM had 94.1% accuracy (sensitivity 89.2%, specificity 95.8%); and k-NN had 93.4% accuracy (sensitivity 86.2%, specificity 95.8%). Combined, ensemble learning achieved 94.5% accuracy (sensitivity 89.2%, specificity 96.3%). The system further differentiated between MES 0, representing complete mucosal healing, and MES 1 images with 89.1% accuracy (sensitivity 82.3%, specificity 92.2%).

Conclusion: Our DLML-CAD diagnosis achieved 94.5% accuracy for endoscopic mucosal healing and 89.0% accuracy for complete mucosal healing. This system can provide clinical physicians with an accurate auxiliary diagnosis in treating UC.

Keywords: Colectomy; Deep learning; Machine learning; Ulcerative colitis

1. INTRODUCTION

Ulcerative colitis (UC) is an idiopathic chronic inflammatory bowel disease (IBD) with a course of alternating relapse and remission. In clinical management of UC, achieving mucosal healing (MH) is an important treatment objective. Instead of clinical remission, endoscopic remission is associated with lower rates of colectomy, relapse, hospitalization, and colorectal cancer.[1,2] To evaluate endoscopic remission, we commonly use the Mayo endoscopic subscore (MES) to evaluate the mucosal status of inflammation, as illustrated in Fig. 1. MESs range from 0 to 3 (0, complete remission; 1, erythema, decreased vascular pattern, mild friability; 2, marked erythema, absence of vascular pattern, friability, erosions; 3, spontaneous bleeding, ulceration).[1] We defined MH according to the commonly used definition for clinical trials, an MES of 0 to 1. In recent UC therapeutic clinical trials, MH was defined by achieving a MES of 0 or 1.[3] One recent study showed that patients with an MES score of 0 have less risk of relapse than those with a score of 1.[4] This means that the outcomes of UC patients with complete MH (MES 0) are better than those with an MES of 1. However, the interpretation of Mayo endoscopic scores depends on the physician's observational experience of UC images. Accurate differentiation of colonic inflammation scores requires expertise and could be subjective in clinical situations. Endoscopic scoring of changing mucosal severity has substantial interobserver and intraobserver variability despite being performed by experienced physicians.[5] In addition, interpretation of whole colon evaluation and the extent of inflammation are still limited by the MES system. Therefore, if one method or tool could provide an accurate and convenient means of interpretation and scoring, it could be useful and have clinical impact for physicians and patients.

In recent years, medical image recognition and detection using artificial intelligence with machine learning has improved and been applied in many disease fields. Machine learning includes

Address correspondence: Prof. Henry Horng-Shing Lu, Institute of Statistics, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu 300, Taiwan, ROC. E-mail address hslu@stat.nctu.edu.tw (H.H.-S. Lu).

Author contributions: Dr. Tien-Yu Huang and Dr. Shan-Quan Zhan contributed equally to this study.

Conflicts of interest: The authors declare that they have no conflicts of interest related to the subject matter or materials discussed in this article.

Journal of Chinese Medical Association. (2021) 84: 678-681.

Received March 20, 2021; accepted May 2, 2021.

doi: 10.1097/JCMA.0000000000000559.

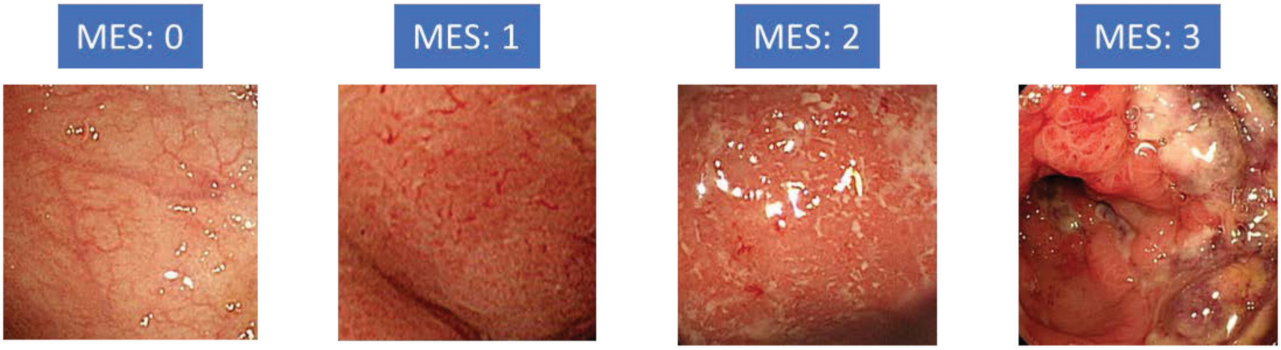

Fig. 1 Sample endoscopic images showing four levels of inflammation with corresponding Mayo endoscopic subscore (MES).

computer-aided diagnosis (CAD), radiomics, and medical image analysis.[6] Recently, deep learning, using convolutional neural networks (CNNs), has become popular in many areas of medicine. It has been applied in endoscopic imaging to distinguish early gastric cancers or neoplastic colorectal polyps from hyperplastic colorectal polyps.[7,8] However, the application of this system to mucosal inflammation status is still limited.[9–11]

In this study, we developed a computer-aided diagnosis system with deep learning and machine learning (DLML-CAD) and analyzed the accuracy of its diagnosis of colonic mucosal inflammation status in patients with UC.

2. METHODS

2.1. Resources

Colonoscopy images in patients with UC (856 endoscopic colonic images with varying MES scores from 54 patients) were obtained from the endoscopic image database at Tri-Service General Hospital, Taiwan. Endoscopic grading of the images using the MES 0-3 was performed by two experienced endoscopists (each with more than 15 years of experience in the field of diagnostic colonoscopy). Disagreements between the two reviewers were resolved by an independent third reviewer. This retrospective study was approved by the institutional review board at the Tri-Service General Hospital (TSGHIRB 2-108-05-109, the date of IRB approval: June 14, 2019) and conducted according to Helsinki Declaration principles. The DLML-CAD setup and image analysis were conducted at the Data Research Center and Institute of Statistics of National Chiao Tung University.

2.2. Database and splitting

In experiment 1, the 856 images described above were first classified as MES 0 to 1 or MES 2 to 3 to detect the presence or absence, respectively, of endoscopic MH. These images (643 images of MES 0 to 1 and 213 images of MES 2 to 3) were used to train and test the DLML-CAD. The training and test sets were stratified samples of the full dataset, split into the ratio 7:3. There were 600 images for training (452 with MES 0 to 1 and 148 with MES 2 to 3) and 256 for testing (191 with MES 0 to 1 and 65 with MES 2 to 3).

In experiment 2, the 643 MES 0 to 1 images were further classified as MES 0, showing complete MH, or MES 1, revealing incomplete healing. There were 411 images with MES 0 and 232 with MES 1. The training-to-test ratio was again 7:3; hence, there were 452 images for training (282 with MES 0 and 170 with MES 1) and 191 images for testing (129 with MES 0 and 62 with MES 1). Additionally, the images for testing were used to compare the performance of IBD, non-IBD, and trainee endoscopists in classifying MES.

2.3 Image preprocessing

Before model training, to increase the size of the training dataset, data augmentation was used. Rotated, flipped, and shifted versions of images in the training set were added to increase the training data. This type of data augmentation is consistent with our objective because the bowel is circular. In addition, the images were resized to 299 × 299 pixels, and the pixel values were rescaled to the range [−1, 1] to fit the pretrained CNN model (Inception-v3) used in this study.

2.4. Image classification through DLML-CAD

The DLML-CAD concept is derived from transfer learning, using a network which is pretrained using millions of nonmedical images to extract desired features. We used the feature to train our classifiers, a deep neural network (DNN), a support vector machine (SVM), and a k-nearest neighbor (k-NN) network. This technique is common in medical image classification, since the amount of data from the medical images is usually not large enough to train a whole DNN.

2.5. Pretrained model: feature extraction

We used an Inception-v3, a deep CNN, which had already been trained by ImageNet, using millions of nonmedical images. Inception-v3's feature extraction capability was utilized to extract the desired features from our colonoscopy images.

2.6. Training and calculation

The workflow is illustrated schematically in Fig. 2. We used the extracted features to train three different classifiers: DNN, SVM, and k-NN. The hyperparameters of each model were decided by hold-out validation. We used ensemble learning (voting) to obtain the final results by merging the results of the three classifiers (Fig. 2).

3. RESULTS

In experiment 1, we determined the accuracy of differentiation between endoscopic MH (MES 0-1) and nonmucosal healing (MES 2-3) by DLML-CAD, applying the three classifiers DNN, SVM, and k-NN. We evaluated the experimental results in terms of accuracy, area under the (receiver operating) curve (AUC), sensitivity, specificity, precision, and F1 score. The diagnostic performances of the DLML-CAD in experiment 1 are shown in Table 1. Among the three classifiers, DNN achieved the following: accuracy, 93.8%; AUC, 90.7%; sensitivity, 84.6%; specificity, 96.9%; precision, 90.2%; and F1 score, 87.3%. Moreover, the SVM classifier yielded the following: accuracy, 94.1%; AUC, 92.5%; sensitivity, 89.2%; specificity, 95.8%; precision, 87.9%; and F1 score, 88.5%. k-NN revealed the following: accuracy, 93.4%; AUC,

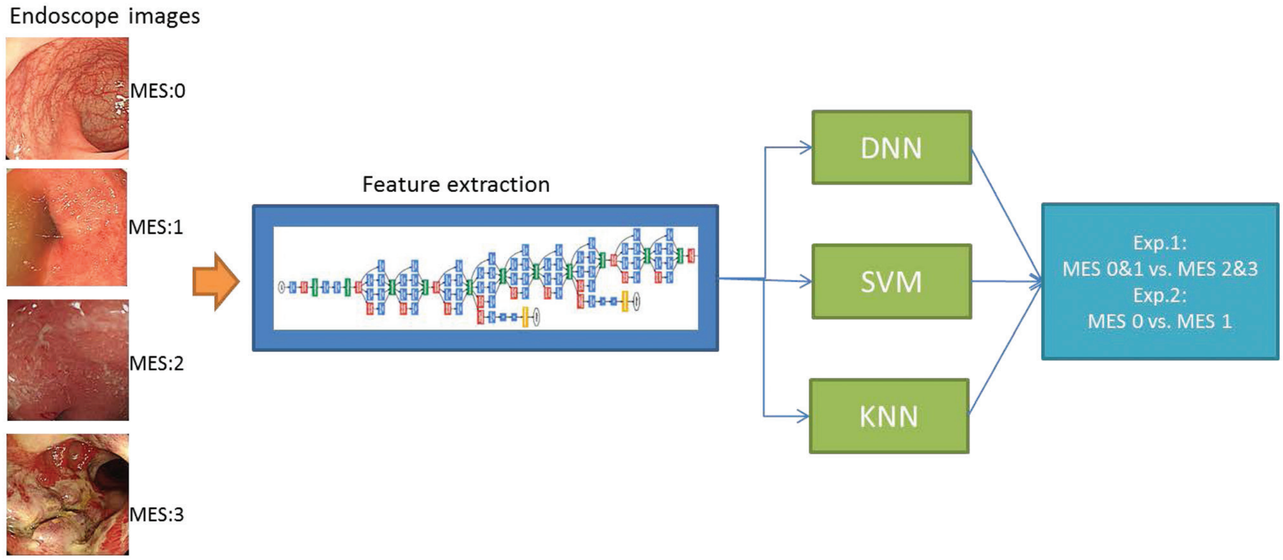

Fig. 2 Schematic illustration of the workflow, representing endoscopic image feature extraction, analysis by three neural network qualifiers, and ensemble merging. DNN = deep neural network; KNN = k-nearest neighbors; MES = Mayo endoscopic subscore; SVM = support vector machine.

90.7%; sensitivity, 86.2%; specificity, 95.8%; precision, 87.5%; and F1 score, 86.8%. Finally, using ensemble learning among the 256 endoscopic testing images, the DLML-CAD achieved accuracy, 94.5%; AUC, 92.8%; sensitivity, 89.2%; specificity, 96.3%; precision, 89.2%; and F1 score, 89.2% (Table 1).

In experiment 2, we determined the accuracy of differentiating endoscopic complete MH (MES 0) images from MES 1 images by DLML-CAD. The same three classifiers DNN, SVM, and k-NN were used, and the diagnostic performance of the networks in experiment 2 is shown in Table 2. DNN achieved accuracy, 86.4%; AUC, 82.4%; sensitivity, 71.0%; specificity, 93.8%; precision, 84.6%; and F1 score, 77.2%. SVM yielded accuracy, 88.8%; AUC, 85.6%; sensitivity, 79.0%; specificity, 92.2%; precision, 85.6%; and F1 score, 81.0%. k-NN showed the following: accuracy, 81.6%; AUC, 78.9%; sensitivity, 71.0%; specificity, 86.8%; precision, 72.1%; and F1 score, 71.5%. Finally, using ensemble learning on the 191 endoscopic images, DLML-CAD correctly classified 51 of the 62 MES 1 images (sensitivity, 82.3%) and 119 of the 129 MES 0 images (specificity, 92.2%). In addition, the ensemble yielded an accuracy of 89.0%; AUC, 87.3%; and F1 score, 82.9% (Table 2).

As compared with human endoscopists, the performance of DLML-CAD is similar to that of the IBD endoscopists and superior to that of the non-IBD and trainee endoscopists. Additionally, the performance of IBD endoscopists was greater than that of the non-IBD and trainee endoscopists. The performance of the senior IBD endoscopist was similar to that of the junior IBD endoscopist (Table 3).

4. DISCUSSION

Endoscopic scoring for inflammation severity is a crucial clinical parameter for treat-to-target management in UC patients and an important endpoint for therapeutic clinical trials. Adequate and accurate endoscopic scoring is challenging for local site IBD endoscopists. Previous studies showed that interobserver agreement on endoscopic severity from viewing colonoscopy video images was unsatisfactory.[5] In one double-blind UC study with central reading, the results showed an intraobserver agreement of 0.89 and an interobserver agreement of 0.79.[12]

Studies focusing on deep learning determination of endoscopic mucosal severity in UC patients are still limited. Recently, Stidham et al[9] demonstrated that the performance of CNN in distinguishing remission (MES 0-1) from moderate to severe disease (MES 2-3) was excellent (AUC, 96.6%; sensitivity, 83.0%; specificity, 96.0%). Ozawa et al[11] showed high-level performance to identify MES 0 and 0 to 1 using CNN-based CAD system. Our results showed similar performance using DLML-CAD (accuracy, 94.5%; sensitivity, 84.6%; specificity, 96.9%; Table 1). We further distinguished MES 0 from MES 1, also with excellent performance (accuracy, 89.1%; sensitivity, 82.3%; specificity, 92.2%; Table 2).

Table 1

The performances of DLML-CAD for classifying mucosal healing (MES 0-1 vs MES 2-3) from the endoscopic colonic images of UC patients

	Accuracy	AUC	Sensitivity	Specificity	Precision	F1 score
DNN	0.9375	0.9074	0.8461	0.9685	0.9016	0.8730
SVM	0.9414	0.9252	0.8923	0.9581	0.8787	0.8854
KNN	0.9335	0.9074	0.8615	0.9581	0.8750	0.8682
Ensemble	0.9453	0.9278	0.8923	0.9633	0.8923	0.8923

AUC = area under the curve; DLML-CAD = deep learning and machine learning computer-aided diagnosis; DNN = deep neural network; KNN = k-nearest neighbors; MES = Mayo endoscopic subscore; SVM = support vector machine; UC = ulcerative colitis.

Table 2

Performance of DLML-CAD in classifying complete mucosal healing (MES 0 vs MES 1) from endoscopic colonic images of UC patients

	Accuracy	AUC	Sensitivity	Specificity	Precision	F1 score
DNN	0.8639	0.8238	0.7096	0.9379	0.8461	0.7719
SVM	0.8879	0.8564	0.7903	0.9224	0.8305	0.8099
KNN	0.8160	0.7889	0.7096	0.8682	0.7213	0.7154
Ensemble	0.8901	0.8725	0.8225	0.9224	0.8360	0.8292

AUC = area under the curve; DLML-CAD = deep learning and machine learning computer-aided diagnosis; DNN = deep neural network; KNN = k-nearest neighbors; MES = Mayo endoscopic subscore; SVM = support vector machine; UC = ulcerative colitis.

Table 3

Performance comparison of DLML-CAD and human endoscopists in classifying mucosal healing (MES 0-1 vs MES 2-3) and complete mucosal healing (MES 0 vs MES 1) from the endoscopic colonic images of UC patients

	MES 0-1 vs MES 2-3			MES 0 vs MES 1		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
DLML-CAD	0.846	0.969	0.938	0.823	0.922	0.891
IBD endoscopists						
Senior	0.871	1.000	0.933	0.900	0.933	0.917
Junior	0.849	1.000	0.917	0.833	1.000	0.900
Non-IBD endoscopists						
Senior	0.750	0.950	0.817	0.732	1.000	0.817
Junior	0.833	1.000	0.900	0.714	0.944	0.783
Trainee endoscopist	0.714	0.944	0.783	0.769	0.952	0.833

DLML-CAD = deep learning and machine learning computer-aided diagnosis; IBD = inflammatory bowel disease; MES = Mayo endoscopic subscore; UC = ulcerative colitis.

In our study, DNN, SVM, and k-NN models were chosen as classifiers for DLML-CAD. Ensemble learning was performed to obtain the final results. Our system classified endoscopic MH for UC (MES 0-1) with 94.5% accuracy and complete MH (MES 0) with 89.0% accuracy. Although the number of training images was small, the results from combined deep learning and machine learning, with subsequent ensemble learning, are satisfying. Each classifier achieved >80% accuracy (Tables 1 and 2). The ensemble learning test performance was superior to that of all the individual classifiers. Ensemble learning is used to aggregate multiple classifiers,[13] and the result of ensemble learning is better than that of any one classifier. DNN was not the best-performing classifier in testing. Classifiers other than DNN available on Inception-v3 worked well and demonstrated better performance than DNN.

In comparing DLML-CAD and human endoscopists, the performance of DLML-CAD reached the level of the IBD endoscopists and was greater than that of the non-IBD and trainee endoscopists. Additionally, DLML-CAD appeared to be correlated with the experiences in the IBD field but not the experiences in the field of general endoscopy. However, more data are required to make this conclusion.

Some limitations of our study should be addressed. First, the number of images was small. To improve the performance and stability of DLML-CAD, more data are required. More data should be added to increase the training data amount and update the model. The DLML-CAD could then be updated and provide better performance. Second, this is a retrospective study conducted only in one medical center. There could be biases influencing the quality of the input images. Third, the result of this study was only shown as a (binary) classification between two groups (MES 0-1 vs MES 2-3, then MES 0 vs MES 1). Multiclass classification results using this system (for MES 0-3) were lacking. Further, the assessment of MES is subjective and needs more validation. We had two experienced endoscopists assess the training and testing images. We minimized the influence of subjective judgement as much as possible; however, this limitation remains. More collected images assessed by a broader pool of experienced IBD endoscopists are needed to decrease this confounding factor.

In conclusion, we developed a DLML-CAD system to accurately diagnose both MH and complete MH from endoscopic colon images. Combining deep learning and machine learning with transfer learning, using three qualifiers, ensemble learning achieved 94.5% and 89.0% accuracy, respectively, for MH and complete MH. In future, this DLML-CAD system could be extended to the multiclass classification of MES (0-3) and real-time scoring in different parts of the colon while performing colonoscopy on UC patients.

ACKNOWLEDGMENTS

The research funding for this study was supported by the Ministry of Science and Technology (MOST 108-2314-B-016-053, 109-2314-B-016-011, 107-2118-M-009-006-MY3), Taiwan.

REFERENCES

1. Narang V, Kaur R, Garg B, Mahajan R, Midha V, Sood N, et al. Association of endoscopic and histological remission with clinical course in patients of ulcerative colitis. *Intest Res* 2018;**16**:55–61.
2. Ponte A, Pinho R, Fernandes S, Rodrigues A, Alberto L, Silva JC, et al. Impact of histological and endoscopic remissions on clinical recurrence and recurrence-free time in ulcerative colitis. *Inflamm Bowel Dis* 2017;**23**:2238–44.
3. Marchal Bressenot A. Which evidence for a treat to target strategy in ulcerative colitis? *Best Pract Res Clin Gastroenterol* 2018;**32–33**:3–8.
4. Barreiro-de Acosta M, Vallejo N, de la Iglesia D, Uribarri L, Bastón I, Ferreiro-Iglesias R, et al. Evaluation of the risk of relapse in ulcerative colitis according to the degree of mucosal healing (Mayo 0 vs 1): a longitudinal cohort study. *J Crohns Colitis* 2016;**10**:13–9.
5. Travis SP, Schnell D, Krzeski P, Abreu MT, Altman DG, Colombel JF, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the ulcerative colitis endoscopic index of severity (UCEIS). *Gut* 2012;**61**:535–42.
6. Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol* 2017;**10**:257–73.
7. Wang Z, Meng Q, Wang S, Li Z, Bai Y, Wang D. Deep learning-based endoscopic image recognition for detection of early gastric cancer: a Chinese perspective. *Gastrointest Endosc* 2018;**88**:198–9.
8. Chen PJ, Lin MC, Lai MJ, Lin JC, Lu HH, Tseng VS. Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology* 2018;**154**:568–75.
9. Stidham RW, Liu W, Bishu S, Rice MD, Higgins PDR, Zhu J, et al. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Netw Open* 2019;**2**:e193963.
10. Maeda Y, Kudo SE, Mori Y, Misawa M, Ogata N, Sasanuma S, et al. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). *Gastrointest Endosc* 2019;**89**:408–15.
11. Ozawa T, Ishihara S, Fujishiro M, Saito H, Kumagai Y, Shichijo S, et al. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointest Endosc* 2019;**89**:416–21.e1.
12. Feagan BG, Sandborn WJ, D'Haens G, Pola S, McDonald JWD, Rutgeerts P, et al. The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. *Gastroenterology* 2013;**145**:149–57.e2.
13. Ju C, Bibaut A, van der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J Appl Stat* 2018;**45**:2800–18.