A Learning Based Microultrasound System for the Detection of Inflammation of the Gastrointestinal Tract

Shufan Yang, Christina Lemke, Benjamin F. Cox, Ian P. Newton, Inke Näthke and Sandy Cochran

***Abstract*—Inflammation of the gastrointestinal (GI) tract accompanies several diseases, including Crohn's disease. Currently, video capsule endoscopy and deep bowel enteroscopy are the main means for direct visualisation of the bowel surface. However, the use of optical imaging limits visualisation to the luminal surface only, which makes early-stage diagnosis difficult. In this study, we propose a learning enabled microultrasound (µUS) system that aims to classify inflamed and non-inflamed bowel tissues. µUS images of the caecum, small bowel and colon were obtained from mice treated with agents to induce inflammation. Those images were then used to train three deep learning networks and to provide a ground truth of inflammation status. The classification accuracy was evaluated using 10-fold evaluation and additional B-scan images. Our deep learning approach allowed robust differentiation between healthy tissue and tissue with early signs of inflammation that is not detectable by current endoscopic methods or by human inspection of the µUS images. The methods may be a foundation for future early GI disease diagnosis and enhanced management with computer-aided imaging.**

***Index Terms*— Ultrasound, Gastrointestinal tract, Neural Network, Computer-aided detection and diagnosis**

I. Introduction

CROHN'S disease (CD) is a form of the Inflammatory Bowel Disease (IBD) group which includes ulcerative colitis (UC). The disease may affect any part of the gastrointestinal (GI) tract from mouth to anus, but most commonly manifests in the terminal ileum of the small bowel [1]. Disease pathogenesis is incompletely understood but is considered to be multifactorial and involve environment (e.g. diet and lifestyle), resident bowel microbes and genetics [2]. It is a chronic and progressive incurable disease marked by intermittent periods of quiescence (i.e. remission) and relapse (i.e. flare). During periods of relapse, an acute inflammatory response is superimposed upon a chronic inflammatory state. The acute inflammatory process is characterised by a rapid accumulation of immune cells, including neutrophils and monocytes, in the subsurface effector sites such as the mucosa of the bowel wall (Fig. 1).

Clinicians apply several methods to image the bowel wall to aid disease management. These include external methods such as MRI, CT, ultrasound (US) and internal endoscopic methods to assess disease activity and treatment response. Pill-sized ingestible capsule endoscopes (CE) can transit the entire GI tract and directly visualise the luminal surface. However, capsule endoscopy cameras are limited to imaging the luminal surface and US endoscopy is limited in its reach to the upper GI tract. Although an operator can obtain tissue samples (i.e. biopsy), the reach of even a standard endoscope is limited by insertion tube length [3]. This means remote areas are accessible only with deep bowel enteroscopy, providing only visual information, and cannot be assessed routinely.

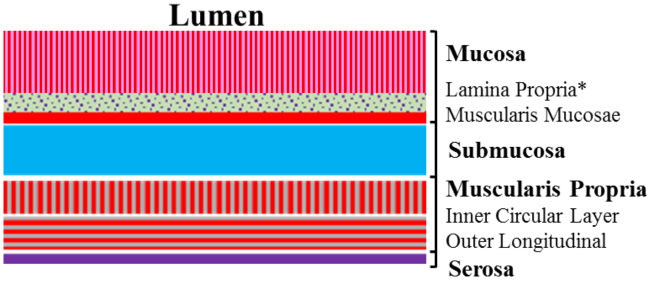


Fig.1. Histological organisation of the bowel wall. Principal layers (i.e. superficial Mucosa, Submucosa, Muscularis Propria, deep Serosa) are indicated in bold. The lamina propria (*), a sublayer of the mucosa, is the site of immune cell activity during inflammation.

The mouse bowel data collection was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Sonopill Programme EP/K034537/2.

S. Yang is with the Centre for Medical and Industrial Ultrasonics, James Watt School of Engineering, G12 8QQ, University of Glasgow, UK. (e-mail: shufan.yang@glasgow.ac.uk).

C. Lemke is with the Centre for Medical and Industrial Ultrasonics, James Watt School of Engineering, G12 8QQ, University of Glasgow, UK. (e-mail: christina.lemke@glasgow.ac.uk).

B.F. Cox is with the School of Life Sciences, University of Dundee, Dow Street, DD1 5EH Dundee, UK (e-mail:b.cox@dundee.ac.uk).

I.P. Newton is with the School of Life Sciences, University of Dundee, Dow Street, DD1 5EH Dundee, UK (i.z.newton@dundee.ac.uk)

I. Näthke is with the School of Life Sciences, University of Dundee, Dow Street, DD1 5EH Dundee, UK (e-mail: i.s.nathke@dundee.ac.uk).

S. Cochran is with the Centre for Medical and Industrial Ultrasonics, James Watt School of Engineering, G12 8QQ, University of Glasgow, UK. (e-mail:sandy.cochran@glasgow.ac.uk).

US images are generated by US reflecting (i.e. echoing) from structures within the tissue. Echogenic sources include the interfaces between tissue subtypes and internal layer compositions comprising cells and structural proteins [4]. The standard clinical frequency range is approximately 4 - 12 MHz, providing axial resolution in the range up to 195 µm in tissue. Hence, standard US can detect and display the four principal layers of tissue in the gut: superficial mucosa, submucosa, muscularis propria, deep serosa.

Microultrasound (µUS) is a term applied to ultrasound at frequencies above 30 MHz: the shorter pulse length leads to improved axial resolution, improving from 195 µm to 75 µm, which can be more sensitive to changes in spatial distribution of cells in a given specimen. Further information can be gathered with µUS implementing a higher range of frequencies, to detect additional ultrasonic layers [5]. Two independent groups have suggested that µUS has the potential to detect pathological changes in the bowel wall [6][7].

The inflammatory process, including the accumulation of immune cells, disrupts the normal architecture and composition of the bowel wall and, in turn, may be expected to alter the ultrasonic properties and imaging of the tissue. However, this approach will generate large volumes of data that will require review and interpretation. To help reduce the review time, a means of computer-aided diagnosis (CADx) will be necessary. In this work, we propose a learning-enabled computer-aided system to classify inflammation of the murine GI tract using µUS imaging methods.

Several attempts to use machine learning and deep learning for US-based disease diagnosis and characterisation have been reported [8]. For instance, super-resolution in US localisation microscopy and beamforming uses deep learning techniques for the removal of artefacts in element-wise complex in-phase and quadrature data [9]. Computer-aided US diagnosis methods for breast cancer imaging have been researched, incorporating features relating to shape, margin, orientation, echo patterns and acoustic shadowing [10]. A more recent development has been reported to use deep learning techniques for extracting nonlinear features from US images at frame-level and video-level using convex and linear probes with central frequency below 5 MHz [11]. Although this study demonstrated the feasibility of using deep learning to enable the characterisation of the state of high permeability and advanced disease, the issues of how to incorporate deep learning techniques into the US imaging chain from data acquisition to post image processing was ignored. Furthermore, the use of deep learning methods to find unique US features has not been fully explored in detail.

Taking advantage of recent developments in deep learning, we developed a learning-based system for US scans to distinguish inflamed and non-inflamed tissues in the murine GI tract, as a model for human disease. The potential for faster diagnoses at an earlier stage of disease and the classification of IBD in the GI tract is illustrated in this paper.

(1) We created a workflow to generate training and evaluation datasets from a limited number of µUS images.

(2) We introduced a simple and efficient approach of training with colored US images instead of traditional grey-scale images.

(3) We addressed the importance of using factorised inception blocks for µUS scans [12].

(4) Convolutional neural networks (CNNs) tend to overfit; in another words, those networks rarely generalise well from training data to unseen data [13]. We developed training strategies to demonstrate successful transfer learning for US scans in *ex vivo* tissues.

(5) We showed that the features learned by the deep learning network perform well to identify the discriminative image regions used for US image categorisation.

(6) Finally, we evaluated three deep learning networks for studying GI IBD and identified the most suitable deep learning architectures and dataset characteristics for improved diagnostic accuracy in the classification of µUS images.

II. Method

A. Inflammation Study Data

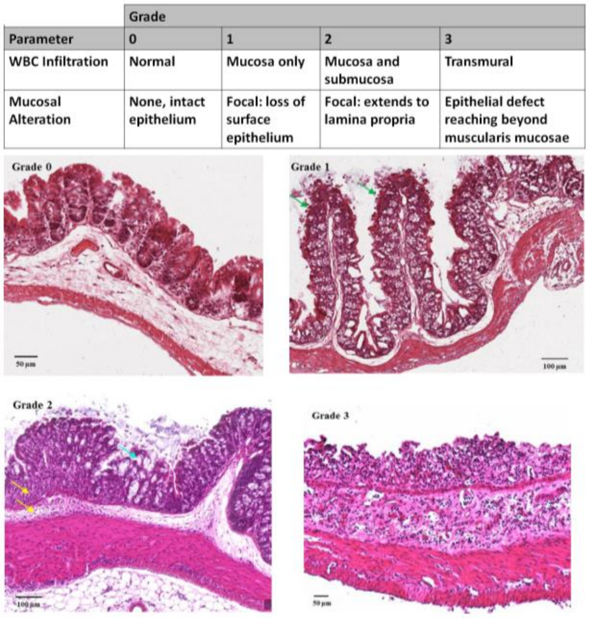
A mouse model of bowel inflammation was used to study the feasibility of machine learning in conjunction with µUS. Bowel inflammation was induced in mice using dextran sodium sulphate (DSS) [19]. Following inflammation, bowel tissue was scanned *ex vivo*. Ground truths, such as confirmation of inflammation and severity grading, were established with histology. Wild-type C57BL/6 mice were used for all stages of acute inflammation except for Stage 2B, although for Stage 2B (see below), two female mice heterozygous for mutation in the adenomatous polyposis coli gene (ApcMin/+) were included. The ages of the mice were in the range 67 - 88 days with a median age of 74.3 days, and animals were grouped by sex in each experimental stage (Table I).

The mice were housed in the University of Dundee Wellcome Building Resource Unit (WBRU) and maintained in accordance with Home Office (UK) guidelines for the care and use of laboratory animals. This study was conducted under a Home Office (UK) Procedure Project Licence: P3800598E, in accordance with the Animal (Scientific Procedures) Act 1986. Humane endpoints were predefined in cooperation with the Named Veterinary Surgeon (NVS), along with means to assess status. All mice underwent daily observations which included health assessment, weigh-in and faecal examination

The experimental endpoint was to determine if µUS could detect visually obvious inflammation. The study was divided into three stages based on primary endpoints. Stages included 1A/B, 2/B and 3. Stages 1A and 1B were pilot studies designed to induce overt signs and symptoms of GI inflammation. Stages 2 and 2B were designed to deduce the lowest grade of inflammation detectable by µUS. Stage 3, a blinded randomised control trial (RCT), was also designed to deduce the lowest grade of inflammation detectable by µUS. By conducting the experiment as a blinded RCT, potential sources of bias when interpreting the results were controlled for.

Stage 1A (N = 4) was an all-male group and stage 1B (N = 4) was an all-female group. It was necessary to conduct two experiments to control for different responses by sex [15]. Each stage began on Day 0 with introduction of DSS. Mice were either dosed for 7 days (Day 6) or culled when a humane endpoint had been reached. Stages 2 (N = 12) and 2B (N = 7) were evenly divided between sexes. As the aim of Stage 2 was to determine the lowest grade of inflammation detectable by µUS, one mouse of each sex was culled daily. Stage 3 (N = 16) mice were randomly assigned to either a control or treatment group. The treatment group was further assigned to length of treatment randomly. This was done by the lab's Scientific Officer using a list randomiser [16].

Animals were culled by cervical dislocation and confirmation of death was by exsanguination by femoral incision. Post-mortem dissection of the distal small bowel, caecum and colon was performed after confirmation of death. This was followed by preparation of each anatomical section for scanning by cleaning and transecting along the long axis of the bowel to allow exposure of the mucosa to the µUS probe. Tissue was cleaned and mounted for scanning (See Section 2B).



	Grade			
Parameter	0	1	2	3
WBC Infiltration	Normal	Mucosa only	Mucosa and submucosa	Transmural
Mucosal Alteration	None, intact epithelium	Focal: loss of surface epithelium	Focal: extends to lamina propria	Epithelial defect reaching beyond muscularis mucosae

Fig. 2. Inflammation severity using histology information. The top table shows mouse demographics for acute inflammation studies of immune cell activity during inflammation. Key: WBC-White blood cells.

After scanning, the tissue was fixed in 4% Paraformaldehyde (PFA) then stained with Haematoxylin and Eosin (H&E) and slide mounted. Stage 3 tissue slides were randomised and coded by the lab's scientific officer [17]. Slides were then batch scanned and digitised by the Tayside Tissue Bank (Dundee, UK) on a digital pathology slide scanner (Aperio Scanscope XT, Leica, Germany) at 40x or by the TMA (tissue microarray) and Image Analysis Unit at the University of Glasgow (Glasgow, UK) on a NanoZoomer NDP (Hamamatsu, Japan) at 40x. Tissue was assessed and graded for severity of changes in morphology, and presence of inflammatory cells using QPath [18].

An ordinal grading scheme was used to assess bowel inflammation. H&E stained tissue was assessed for histomorphologic alterations, which included white blood cell infiltration and mucosal alterations. This experiment focused on acute inflammation, where white blood cells are predominately neutrophils and monocytes. The table in the top part of Fig. 2 summarises the inflammatory grading scheme based on the aforementioned criteria, adapted from Elsheikh *et al.* and Erben *et al.* [19][20], respectively.

As shown in Fig. 2, Grade 0 corresponded to lack of signs of inflammatory cell infiltration and the presence of a continuous, intact epithelium. Grade 1 corresponded to cases with signs of mucosal inflammatory cell infiltration (red arrows) but no mucosal disruption. Grade 2 corresponded to cases with infiltration at the mucosa and submucosa (green arrows) and focal epithelial disruption (blue arrow). Grade 3 corresponded to cases such as the one shown, with transmural inflammation and infiltration at all histologic levels and confluent disruption of the epithelium.

B. Microultrasound Scanning System Setup

Microultrasound scanning was performed with focused, single-element, high frequency transducers using a bespoke mechanical scanning system (Fig. 3a) [21]. Scanning was controlled via a graphical user interface (GUI) based on a LabVIEW program (National Instruments, Austin, TX, USA). US was generated and echoes were received by a remote pulser-receiver unit (DPR500, JSR Ultrasonics, NY, USA), buffered, AC/DC-converted and sampled with a NI PCI-e Express Card in a PXIe-1071 chassis (National Instruments, Austin, Texas, USA).

Ex-vivo tissue was mounted and secured using a bespoke Lego clamp (Fig. 3b). The clamp rested on an acoustic absorber (Aptflex F28, Precision Acoustics, UK) which mitigated stray echoes and acted as a pinning board. The scaffold and absorber were placed in a plastic container that was either back-filled with 1% (w/v) agar to the top level of the struts (Stages 1A, 1B 2) or omitted (Stages 2B and 3). For Stage 2, a 3 mm space between the tissue and agar was maintained with acoustic gel to prevent the tissue from contacting the agar and giving a separate acoustic signal from each of them. Tissue was arranged with the mucosa from the small bowel (SB), Caecum (Cae) and Colon (Co) positioned towards the transducer. In addition to anchoring the tissue, the clamps also reduced excessive *post-mortem* tissue curling. To ensure consistent and identifiable tissue orientation during optical imaging and µUS scanning, visual markers were used to indicate the correct surface and proximal tissue portion. The scanning tray was placed in a second, slightly larger container and filled with Krebs Henseleit solution titrated to pH 7.4. The solution acted as nutrient fluid to maintain tissue viability and coupling fluid for the µUS.

Two focused single element µUS transducers were used: a 37.5 MHz transducer made with a composite comprising lead zirconate titanate (PZT) and polymer and a 62 MHz lithium niobate (LNO) transducer. The amplitude scan (A-scan) data measured at each point were averaged 32 times per scan point. The scan parameters were: travel distance in x-direction (X mm); travel speed in x-direction (X mm/s); travel distance in y-direction (Y mm); travel step size in y-direction (Y mm/step); sampling frequency (800 MHz); and number of samples per A-scan (8000).

Stages 1A and 1B had visible haemorrhagic lesions. Stages 2B and 3 lacked visual clues as inflammation was visually occult and prevented targeted scanning. This necessitated a comprehensive scan of all tissue contained in the Lego clamp. Scan parameters were set to 32 mm x 18 mm, X and Y directions, respectively. This overlapped the 30 mm x 15 mm window between struts to include a Lego signal on all 4 sides. This was done to include reflections from the Lego as fiducial markers, ensure full tissue coverage, and reduce errors due to imprecise scanner motor movements. The X direction speed was set to 0.2 mm/s and the Y direction step size was set to 0.2

TABLE I
MOUSE DEMOGRAPHICS FOR ACUTE INFLAMMATION STUDIES

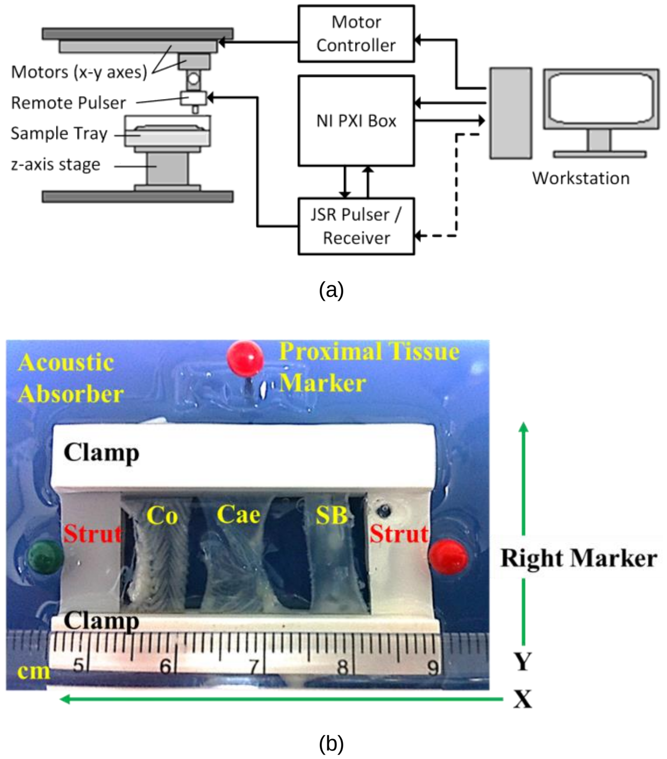
Stage	Treatment Regimen	Number	Sex	Genotype	Weight Day 0	Transducer Centre Frequency	No. of Images	Deep Learning
1A	5% DSS max. 7 days	4 Treated 1 Control	5 Male	WT C57BL/6	24.2-25.2g	37.5 MHz	470	Training
1B	5% DSS max. 7 days	4 Treated 3 Control	7 Female	WT C57BL/6	19.6-21.6g	37.5 MHz	614	Training
2	5% DSS max. 5 days	10 Treated 2 Control	6 Female 6 Male	WT C57BL/6	18.4-24.3g	37.5 MHz	2087	Training
2B	5% DSS max. 5 days	6 Treated 1 Control	3 Female 4 Male	5x WT C57BL/6 2x ApcMin/+	18.4-26.3g	62 MHz	699	Unseen Testset
3	5% DSS max. 5 days	11 Treated 5 Control	11 Female 5 Male	WT C57BL/6	18.1-24.8g	37.5 MHz	1092	Unseen Testset

Key: DSS - Dextran Sodium Sulphate, WT - Wild Type, Apc$^{Min/+}$ - adenomatous polyposis coli (heterozygous)

mm/step. These parameters ensured complete tissue coverage in a reasonable overall time of ≈ 4 hours for each scan.

C. B-Scan Generation

More than one tissue sample was present in most scans in the experimental US data measurement platform we used (Fig. 3(a)) to collect US data, corresponding to the example with three tissue samples shown in Fig. 3(b). In this study, we developed a workflow (Fig. 4) to generate B-scan images from the raw ultrasound data captured by the µUS scanning system. Thus, some of the original raw data per tissue sample were extracted, with boundaries defined after visually inspecting each original scan, in order to create B-scan images containing data from only one single tissue sample. Subsequently, bandpass filtering and image segmentation were applied and each image was log-compressed and normalized. Bandpass filtering was achieved through a combination of low-pass and high-pass Butterworth filters defined by the parameters shown in Table II.



(a)

(b)

Fig. 3. Scanning Setup: (a). Schematic overview of the mechanical scanning system used to capture micro-ultrasound data. (b) Example of a tissue sample scanning tray embedded in agar with sample of small bowel (SB), Caecum (Cae) and Colon (Co) of a single mouse.

TABLE II
LOW-PASS AND HIGH-PASS FILTER COMBINATION PARAMETERS

Parameter	Low-pass Filter	High-Pass Filter
Passband frequency	55 MHz	15 MHz
Stopband frequency	105 MHz	1 MHz
Attenuation	60 dB	60 dB
Ripple	3 dB	3 dB

Segmentation allowed separation of the tissue echoes from other image content such as bubbles and the agar layer [21]. We created log-compressed, normalized B-scan images with a dynamic range of 40 dB first and applied an adaptive threshold approach. Then a binary image was constructed from the filtered B-scan, where pixel values were set to either 1 or 0, depending on whether their logarithmic intensity level were above or below a set threshold respectively. The threshold was adaptively determined for each B-scan by modelling the histogram of the intensity levels as a combination of two Gaussian distributions:

$$g(x) = a_1 e^{-\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + a_2 e^{-\left(\frac{x-\mu_2}{\sigma_2}\right)^2} \tag{1}$$

where a_1, a_2 are the amplitudes, μ_1, μ_2 are the means, and σ_1, σ_2 are the scaling parameters. The threshold τ was calculated from the b and σ values of the distribution with the highest amplitude, which represents the noise floor values of the B-scan:

$$\tau = \mu_n + 2 \cdot \sigma_n, n = \begin{cases} 1, a_1 > a_2 \\ 2, a_1 \leq a_2 \end{cases} \tag{2}$$

Through a combination of closing and filling operations, the binary image was cleaned and unconnected pixels were eliminated. The largest continuous cohort of pixels was selected and applied to the filtered B-scan data as a binary mask to separate tissue values from background and the noise floor. Finally, each image was log-compressed and normalized. Export of the processed images to 300 dpi tif-files was achieved using the “export_fig” function package [22] in MATLAB (The

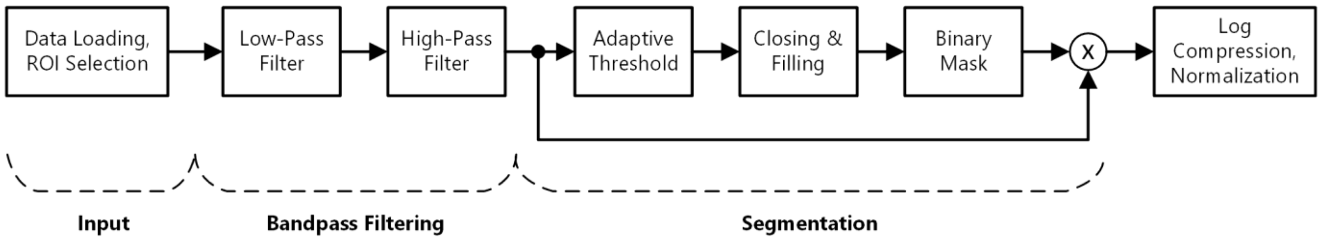
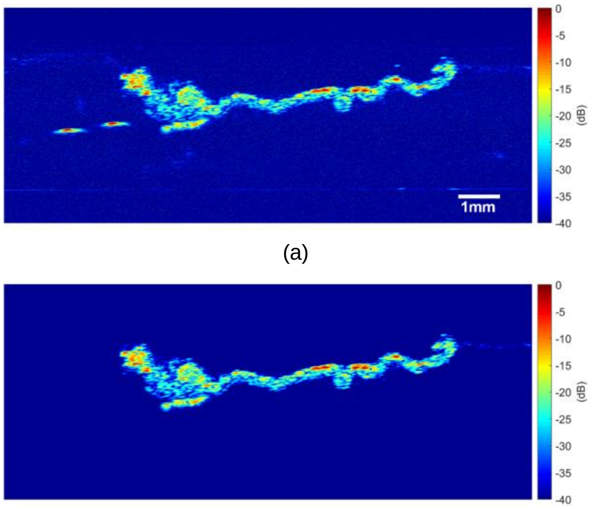


Fig. 4. Flow diagram of B-scan image reconstruction, including bandpass filtering (by a combination of low- and high pass filters) , segmentation and log compression for display purposes. Not shown: exporting images with a resolution of 300 dpi for further processing.

Mathworks, Cambridge, UK).



(a)

(b)

Fig. 5. Normalized B-scan images showing a caecum tissue sample from a female mouse, 69 days old, weight 19.8 g: a) original data after ROI selection: depth 2 mm to 8 mm, x-axis: 9 mm to 23 mm; b) post-processing scans after adaptive filter the in RGB format with signal strength. Red represents maximum signal intensity, normalized to 0 dB. In b), the dark blue colour in the background no longer contains any ultrasonic data.

D. Pre-processing B-Scan Images for Deep Learning Networks

Following the methods outlined above, the original B-scan was processed to remove noise and present clear echoes from the tissue (Fig. 5(b)). In order to increase the number of training samples, adaptive filtering was applied to divide the B-scan into four patches. However, the B-scan cannot simply be cut into small patches, as edge information used to detect the thickness of the tissue would be lost. Instead, the following method was applied to generate patches so that they provided maximal echo information and minimal background information. A linear shift-invariant filter was also used to achieve corresponding noise resistance with 2nd order of statistics of all the B-scan images [23]. The observed tissue echo signal, ß, was modelled at position i, j, z in a block of dimensions X * Y * Z by the sum of the ideal echo signals, γ, and a noise term, æ:

$$\text{ß}\,(\mathrm{i},\mathrm{j},\mathrm{z}) = \gamma(\mathrm{i},\mathrm{j},\mathrm{z}) + \text{æ}\,(\mathrm{i},\mathrm{j},\mathrm{z}) \qquad (3)$$

where the ß, γ, æ are vectors.

The vector pixel set ß can be written as a scalar product using a filter coefficient vector η [25]. Eq. (4) provides the actual filter output, η̂.

$$\hat{\eta}=\mathrm{x}^{T}\text{ß} \qquad (4)$$

where x is the filter coefficient vector. Finally,

$$\widehat{\text{ß}}= \mathrm{x}^{T}\gamma+ \mathrm{x}^{T}\text{æ} \qquad (5)$$

Eq. 5 shows the shift-invariant system to produce output $\widehat{\text{ß}}$, which has the desired minimum background information. Finally the output scan $\widehat{\text{ß}}$ divided into four patches (with size of 48 × 48 pixels for each patch).

E. Deep Learning Networks

In this study our learning-based system is constructed as a combination of computational modules from deep learning networks. This type of network has stacked modules, such as a convolutional layer, pooling layer, and a fully connected layer, one on top of the other. Fig. 6(a) shows the basic architecture of three different deep learning networks, including convolution, pooling and fully connected layers, which are all followed by the softmax layer for classification. Those stacked layers perform abstraction on representations of training data, in which higher level abstract features are defined by combining them with lower level features.

In this work, InceptionV3[28], InceptionResnetV2 [32][28] and NASnet(mobile) [30] were modified to evaluate their applicability to the inflammation study. A simplified illustration of the architectures is shown in Fig. 6(a). All three networks share the same feedforward convolution layout, except that the InceptionResnetV2 framework uses residual mapping and the Inception blocks are used in both InceptionV3 and InceptionResnetV2 networks (Fig. 6(b) - (c)). NASnet (mobile) network uses its own architectural building blocks (Fig. 6(d)). InceptionResnetV2 uses slightly different blocks and includes a feedback channel; NASnet(mobile) and InceptionV3 do not have this type of internal feedback channel in their networks. The parameters of the trained InceptionV3, InceptionResnetV2 and NASnet(mobile) models are 21M, 54M and 4M, respectively. The large size of the networks will create problems of implementation into a portable platform. It is therefore interesting to investigate which size of network can provide optimal performance whilst remaining relatively small.

III. Results

A. Evaluations on Training set

At the neural network training stage, we used two types of

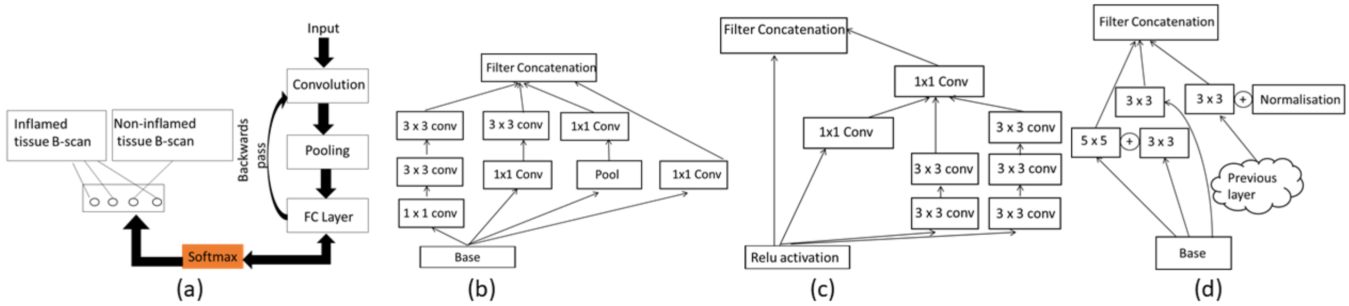
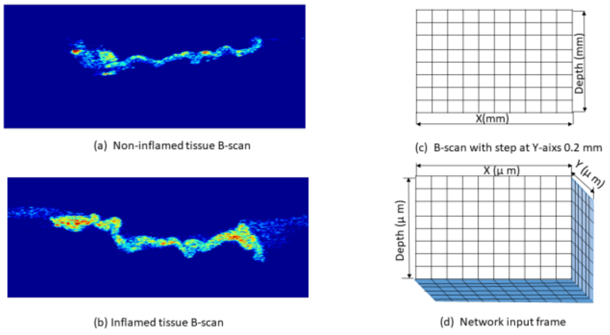


Fig. 6. Architecture of a deep learning framework (a) general architecture of a deep learning network, (b) a basic block used in InceptionV3 [27], (c) a basic block used in InceptionResnetV2 [28] , (d) a basic block used in NASnet(mobile) network [29].

labelled B-scans: non-inflamed and inflamed. Fig. 7 (a) - (b) shows one B-scan each, collected using our µUS system with the ground-truth label based on histological examination. A series of scans (Fig. 7 (d)) were grouped into mini-batch size of 32 for training.

We split our dataset into training and test sets from the five experimental stages (Table I). In total 3,171 scans from Stage 1A, 1B and 2 form the training set, including 1,270 scans without inflammation and 1,901 scans with inflammation.



(a) Non-inflamed tissue B-scan

(c) B-scan with step at Y-aixs 0.2 mm

(b) Inflamed tissue B-scan

(d) Network input frame

Fig. 7. (a) Input image training samples for non-inflamed class and(b) inflamed class. (c)The step size of grid. (d) the training batch (for demonstration purpose).

A further 1,791 scans (699 from Stage 2B and 1,092 from Stage 3) were used as an ‘unseen’ test set (categorised based on histological observations) to evaluate our models. A plot showing the distribution of the scores of inflammation grade for each experimental stage is given in Fig. 8. The inflammation severity is labelled using a histology grading method (Table in Fig. 2). In practice, Grade 1 inflammation was difficult to distinguish from healthy tissue since only white blood cell infiltration appears in the mucosa; therefore, for our study, Grades 0 and 1 of inflammation were classified as non-inflamed, whereas Grade 2 and Grade 3 were classified as inflamed. By mixing scans in each experimental stage, we formed a relatively balanced dataset.

Data augmentation [30] was introduced to facilitate training of the neural network. We used random image shifts vertically and random channel shifts as our data augmentation methods. Other data augmentations including rotation, horizontally shift and flip were not applied, because these can potentially change the pattern of power intensity from received echoes in the images.

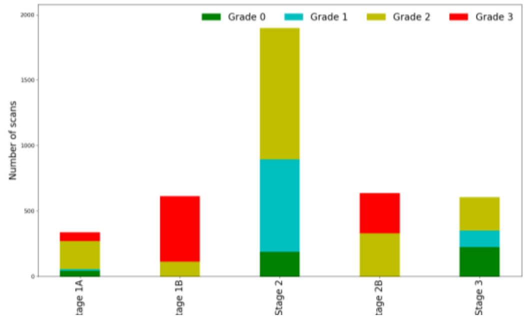


Fig. 8. Distribution of post-processed B-scans with inflammation severity for each experimental stage

Although transfer learning has been used since 2015, the main difference between this study and others [31] is that we used more aggressive extraction of various filter sizes. At the training stage, the stochastic gradient descent optimizer was employed to train the network on one Nvidia GTX1080 card. The learning rates were set at 0.1, 0.01,0.001 and 0.00005 at epochs 80, 120, 160 and 180. The categorical cross-entropy was used in the training. The code was developed using Keras application interface with TensorFlow as backend. (source code can be accessed from the link:

https://github.com/WOLVS/Ultrasound_Inflammation_IEEE_TME).

We conducted two experiments to evaluate the classification performance of the networks. In the first experiment, we used patches that were generated using the pre-processing method (see details in Section II). In the second experiment, we used the same patches with subtracted dark blue backgrounds as those backgrounds do not contain any echo information.

Training curves from the first experiment for 200 epochs and 400 epochs are shown in Figs. 9 - 11. In those figures, the blue lines indicate the training results and the orange lines indicate validation results. The average training accuracies of InceptionV3, InceptionResnetV2 and NASnet(mobile) were 78%, 79% and 76%, respectively. For 400 epochs of training, the training losses in all three networks decreased to values near the evaluation losses, which means that all networks finally converged (Figs. 9 – 11). The training loss significantly decreased as the number of training epochs increased, while the validation loss notably decreased. The validation accuracy reached a plateau rapidly, with the exception of the NASnet(mobile) network where the training loss significantly decreased but reached another peak, then settled into a plateau in the epoch range 350 - 400 (Fig. 11(b)). The training stage

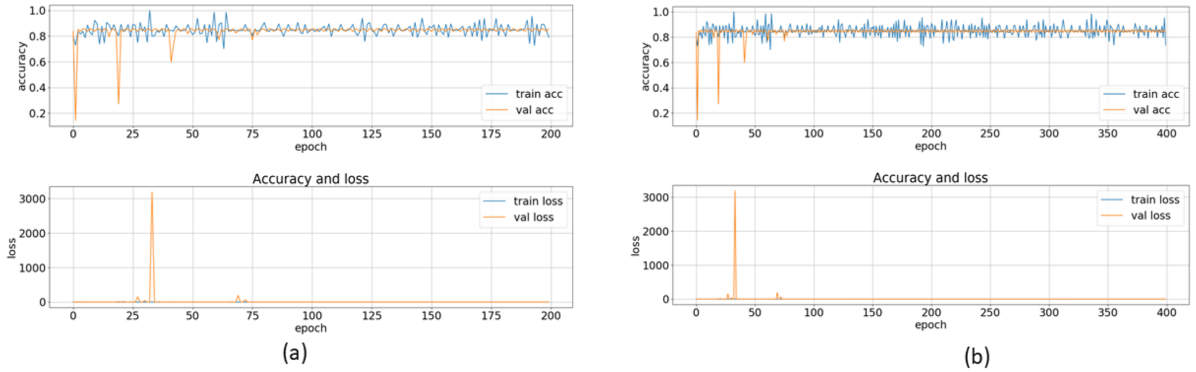


Fig. 9. Training result of InceptionV3 network using input images with background information. The blue line is for training set and the orange line is the validation set. (a) shows the results in the range of 200 epochs. (b) shows the results in the range of 400 epochs.

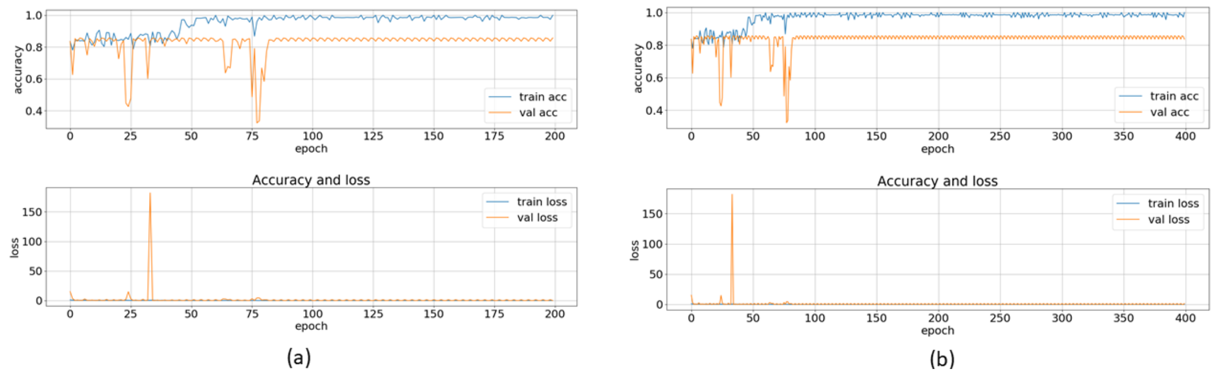


Fig. 10. Training result of InceptionResnetV2 network using input images with background information. The blue line is for training set and the orange line is the validation set. (a) shows the results in the range of 200 epochs. (b) shows the results in the range of 400 epochs.

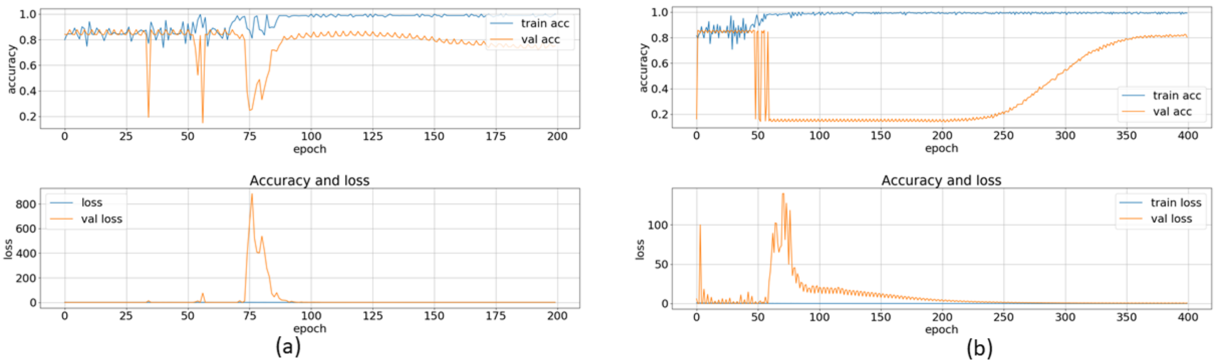


Fig. 11. Training result of NASnet(mobile) network using input images with background information. The blue line is for training set and the orange line is the validation set. a) shows the results in the range of 200 epochs. (b) shows the results in the range of 400 epochs.

around epochs 25, 50 and 75 show repetitive behaviour since the input training data have a high degree of similarity. The motor moves at 0.2 mm/step in the y direction. With such small steps, the similarity of those multiple scans was high due to tiny hypoechoic targets changing in tissue samples.

In the first experiment, the training images included blue background information (Fig. 7). It is highly likely that neural networks used background information as a part of image features during the neural network learning processes. To make sure the network "learned" information solely from echoes from tissue, the second experiment of training was conducted using B-scan images without the blue background. In this way, each B-scan represents only the tissue itself. After retraining the networks using this type of image, the training loss significantly decreased as the number of training epochs increased, while the validation loss notably increased, and the validation accuracy did not improve much (Figs. 12 - 14), which allowed the network to reach the convergence stage at around 100 epochs.

Importantly, InceptionV3 and InceptionResnetV2 showed better and more consistent performance in terms of training loss, validation loss and validation accuracy than NASnet(mobile). The average classification accuracies for InceptionV3, InceptionResnetV2 and NASnet(mobile) were 85%, 83% and 76% respectively. With careful selection of less similar images or "rigorous reshuffling" it may be possible to remove those fluctuations in the training curves. However, to keep the original experimental results, we chose to use those figures to present the real data features.

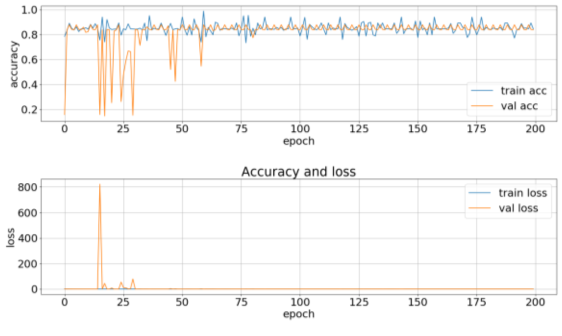


Fig. 12. InceptionV3 network training without blue background information

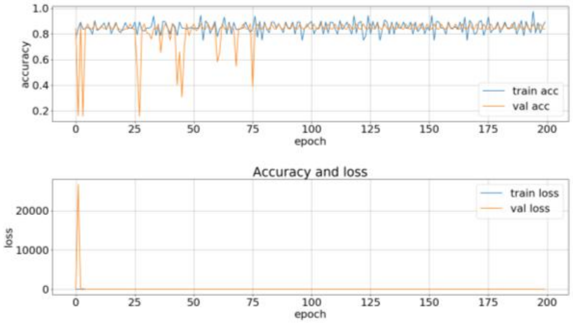


Fig. 13. InceptionResnetV2 neural network training without blue background information

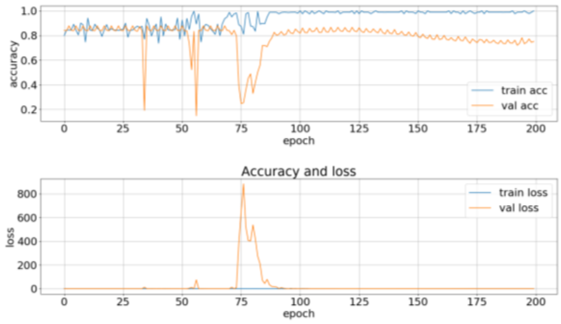


Fig. 14. NASnet(mobile) network training without blue background information

Fig. 15 shows average confusion matrices in a 10-fold cross validation using the weights trained from the second experiments. All the experiments were trained using the same hyper-parameters and implementation. Results from InceptionV3 show that the aggregation was more balanced than for the other two networks. The average accuracy of Inception V3 was 90%. The average accuracies of InceptionResNetV2 and NASnet(mobile) were 88% and 81% respectively. The higher average accuracy than training accuracy is attribute to the 10 fold calculation.

We suspect that the InceptionV3 and InceptionResNetV2 networks responded well to the white blood cell infiltration, a typical histomorphological feature of inflammation. Thanks to the inclusion of inception blocks in both networks, the network weights can be reformulated to learn the desired unknown mapping between inputs and ground-truth.

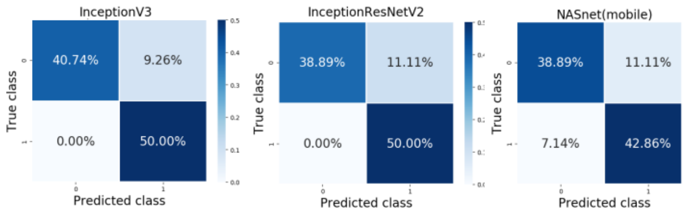


Fig. 15. Normalised Confusion matrices

B. *Evaluation on Test set*

After the neural networks were trained using the training set from the second experiment, we used 1,791 unseen scans for evaluation, including Stage 2B and Stage 3 scans. Notably, the 699 scans from Stage 2B were collected with the higher frequency transducer than the training data (62 MHz versus 37.5 MHz) producing higher axial resolution. All Stage 2B samples were predicted correctly as the inflamed class. Stage 3 scans (Table I) were produced with a transducer of centre frequency 37.5 MHz and the evaluation results are reported in Fig. 16.

The area-under-the-ROC-curve (AUC) and true positive rate (TPR, recall or sensitivity) and false negative rates (FN) are traditionally used as performance metrics. The traditional ROC (Receiver Operating Characteristic) curve provides a single quantitative index of network prediction accuracy with the assumption that the underlying distributions for normal and inflamed groups are Gaussian distributed. In real -life data this

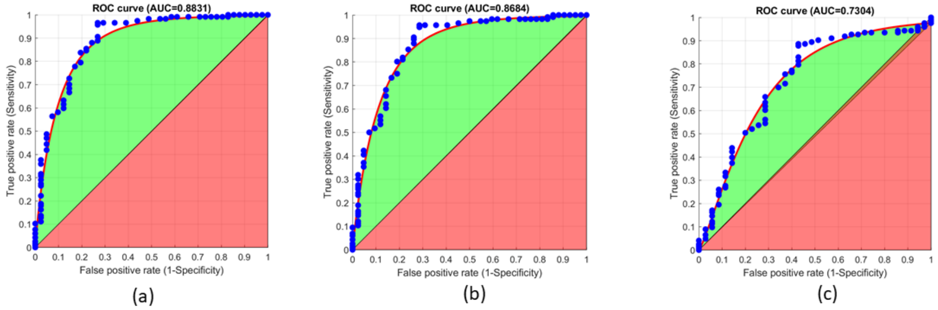


Fig. 16. ROC curves averaged for inflammation using deep learning networks. (a) ROC results from prediction generated from InceptionV3.(b) ROC results from predictions generated from InceptionResnetV2. (c) ROC results from prediction generated from NASnet (mobile).

assumption is not always true. In this study we used a permutation test (10,000 permutations) with a 95% confidence interval and P-value to access statistical significance, by randomly reallocating all of the scans into two groups and re-computing the AUC and coefficient of determination. The 95th percentile points of the empirical distributions were used as critical values to estimate P values (0.05) [33] [34].

Fig. 16 shows that the ROC results are strongly significant for InceptionV3 and InceptionResnetV2, appearing to corroborate the findings of the validation processes with ROC-AUC 0.8831 and 0.8631. The red line is the ROC curve that was calculated using iterative numerical methods from the determination coefficient derived from the logistic regression analysis. The ROC analysis did not yield a significant result for the NASnet(mobile) method due to bigger P values (3.24).

All our deep learning models achieved notably good ROC-AUC results. InceptionV3 consistently outperformed InceptionResnetV2, though by small margins (0.02). NASnet(mobile) performed poorly and was susceptible to over-fitting; we did not have enough data to train this network due to its high complexity and 771-layer depth. Encouragingly, the Inception blocks deployed in both the InceptionV3 and InceptionResnetV2 network demonstrated their suitability for classification when trained on thousands of medical images.

C. Discriminative Features for Inflamed Tissues

Although deep learning models are far from being able to reproduce the full chain of reasoning required for medical interpretation, deep learning networks can learn from higher, middle to lower level abstract features from inputs, providing a means to differentiate many more features in images.

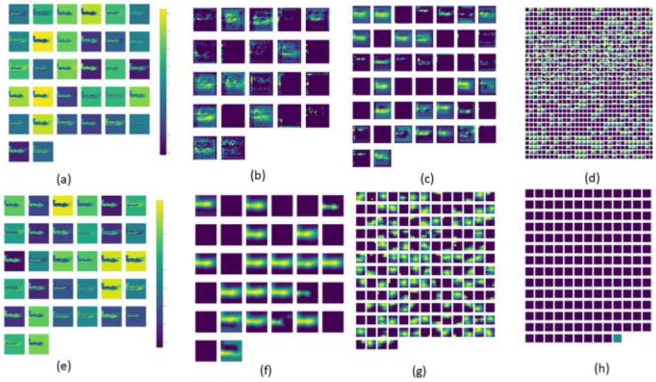


Fig. 17. Visualisation of learned weight of neural networks for inferencing on B-scan images for inflamed tissue. (a) visualisation of first layer convolution filters of NASnet(mobile) network. (b) middle activation layer of NASnet(mobile); (c) lower middle activation layer NASnet(mobile); (d) final layer of InceptionV3. (e) visualisation of first layer convolution filters of InceptionV3 network; (f) middle activation layer of InceptionV3; (g) lower middle activation layer of InceptionV3; (h)the last activation layer before average pooling layer of InceptionV3.

To understand how a deep learning network can improve the classification of bowl tissue scans, the first layer convolution filters from two deep learning networks are visualized in Fig. 17. Due to the memory limitations of our computer, we only used InceptionV3 and NASnet(mobile) to demonstrate the visualisation of neural network weights. We expect InceptionResnetV2 will show very similar results to those from InceptionV3 since they both share very similar architectural structures. We noticed that the visualisation of weights from InceptionV3 (Fig. 17(e) – (h)) showed a clearer pattern of recognition than NASnet(mobile) (Fig. 17(a)- (b)). Many higher orders of contrast were evident that the learning in the neural networks gradually processed into two classes. We believe the greater layer depth that provided the pattern significantly to activate a certain feature may respond to higher excitation of strong echo intensity in the inflamed tissues.

According to [35], by modifying the global average pooling layer with class activation mapping, we believe the visualisation of activation could potentially present white blood cells infiltration in US scans, although this needs further experiments to confirm.

IV. Discussion & Conclusion

The main goal of this work was to investigate the effects of two factors, CNN architectures and dataset characteristics, on performance in a domain-specific medical image analysis problem. The proposed deep learning training method was effective and efficient for the automatic classification of µUS B-scans with inflamed and non-inflamed tissue. The investigated networks (InceptionV3, InceptionResnetV2 and NASnet(mobile)) demonstrated applicability and a powerful classification capacity. Removal of the blue background information, representing anechogenic regions, led to faster convergence.

A. Training on μUS Images of ex vivo Tissues

We believe that the difference in the µUS signals produced by inflamed and non-inflamed tissue was caused by the immune response and corresponding influx of white blood cells, but further work is required to confirm this.

Deep learning networks, especially those including inception blocks, are suitable for the classification of inflamed tissue with a minimum 85% training accuracy and 83% evaluation accuracy. Findings based on these measurements could ultimately lead to the integration of deep learning in µUS CADx. To advance our understanding of how networks complete feature abstraction, future investigations will be required using tissue exhibiting more subtly different grades of inflammation, confirmed by histological data.

B. Limitation of Dataset

In this work, data collection processes were extended because of the use of single element transducers. The variety of inflammation severity in the test dataset was limited. The test dataset (Stages 2B and 3) has a high percentage of inflamed tissue scans which makes it susceptible to bias. Also, the similarity of scans results in fluctuation, even with the shuffling of the validation dataset.

C. Possible Applications

High resolution µUS imaging has the potential to substitute the need for physical biopsy and with virtual biopsy, providing highly detailed subsurface information along the entire length of the bowel, including anatomically remote regions of the small bowel [36]. However, realizing this goal would result in enormous numbers of data requiring physician review. Our ultimate objective is thus to deliver clinically relevant tools using deep learning methods. Those tools may then be used effectively to monitor treatment response, reducing workload and importantly, identifying pathologies that would otherwise be missed during the processes of diagnosis and management of GI tract disorders[37][38].

Acknowledgments

We thank Prof Kwok-Ho Lam of the Hong Kong Polytechnic University for the kind provision of the 62 MHz high-frequency ultrasound transducer.

References

[1] D. C. Baumgart and W. J. Sandborn, “Crohn’s disease,” *Lancet*, vol. 380, no. 9853, pp. 1590–1605, 2012.

[2] R. J. Xavier and D. K. Podolsky, “Unravelling the pathogenesis of inflammatory bowel disease,” *Nature*, vol. 448, no. 7152, pp. 427–34, Jul. 2007.

[3] R. V Bryant, S. Winer, S. P. L. Travis, and R. H. Riddell, “Systematic review: histological remission in inflammatory bowel disease. Is ‘complete’ remission the new treatment paradigm? An IOIBD initiative,” *J. Crohns. Colitis*, vol. 8, no. 12, pp. 1582–97, Dec. 2014.

[4] M. B. Kimmey, *et al*, “Histologic correlates of gastrointestinal ultrasound images,”. *Gastroenterology*, vol. 96, no. 2, pp. 433-441, 1989.

[5] S. Ødegaard, L. B. Nesje, O. D. Lærum, and M. B. Kimmey, “High-frequency ultrasonographic imaging of the gastrointestinal wall.,” *Expert Rev. Med. Devices*, vol. 9, no. 3, pp. 263–73, May 2012.

[6] A. Fatehullah, S. Sharma, I. P. Newton, A. J. Langlands, H. Lay, and S. A. Nelso, “Increased variability in ApcMin/+ intestinal tissue can be measured with microultrasound,” *Sci. Rep.*, vol. 6, 29570, Jul. 2016.

[7] A. B. Le Roux, L. A. Granger, N. Wakamatsu, M. T. Kearney and L. Gaschen, “Ex Vivo Correlation of Ultrasonographic Small Intestinal Wall Layering With Histology in Dogs,” *Vet. Radiol. Ultrasound*, vol. 57, no. 5, pp. 534–545, 2016.

[8] Van Sloun, Ruud JG, Regev Cohen, and Yonina C. Eldar. "Deep learning in ultrasound imaging." Proceedings of the IEEE 108.1 (2019): 11-29.

[9] Luchies, Adam C., and Brett C. Byram. "Deep neural networks for ultrasound beamforming." IEEE transactions on medical imaging 37, no. 9 (2018): 2010-2021.

[10] J. Shan, S. K. Alam, B. Garra, Y. Zhang and T. Ahmed, “Computer-Aided Diagnosis for Breast Ultrasound Using Computerized BI-RADS Features and Machine Learning Methods”, *Ultrasound Med Biol.*, vol. 42, no. 4, pp. 980–988, 2016.

[11] Roy, Subhankar, *et al.* "Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound." IEEE Transactions on Medical Imaging (2020).

[12] L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, “Machine Learning for Medical Ultrasound: Status, Methods, and Future Opportunities,” *Abdom Radiol*, vol. 43, pp. 786–799, 2018.

[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.

[14] B. Chassaing, *et al*, "Dextran sulfate sodium (DSS)‐induced colitis in mice," *Current protocols in immunology*, no. 104.1, pp.15-25, 2014.

[15] S. Wirtz*et al.*, “Chemically induced mouse models of acute and chronic intestinal inflammation,” *Nat. Protoc.*, vol. 12, no. 7, pp. 1295–1309, 2017.

[16] M. Haahr and S. Haahr, “List Randomizer,” RANDOM.ORG. [Online]. Available: https://www.random.org/lists/. Accessed: Nov. 29,2018.

[17] A. Hedges, “Random Number Generator / Picker.” [Online]. Available: https://andrew.hedges.name/experiments/random/. Accessed on November 29, 2018.

[18] P. Bankhead *et al* “QuPath: Open source software for digital pathology image analysis,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, 2017.

[19] W. Elsheikh, K. L. Flannigan, W. McKnight, J. G. P. Ferraz, and J. L. Wallace, “Dextran sulfate sodium induces pan-gastroenteritis in rodents: implications for studies of colitis.,” *J. Physiol. Pharmacol.*, vol. 63, no. 5, pp. 463–469, Oct. 2012.

[20] U. Erben *et al.*, “A guide to histomorphological evaluation of intestinal inflammation in mouse models,” *Int. J. Clin. Exp. Pathol.*, vol. 7, no. 8, pp. 4557–4576, 2014.

[21] T. Anbarasan *et al*, “High Resolution Micro-ultrasound (μUS) Investigation of the Gastrointestinal (GI) Tract”, *in: Biosensors and Biodetection, Methods in Molecular Biology*, Prickril B., Rasooly A. (eds) , vol 1572, pp. 541-561, Springer, New York, NY, 2017

[22] Y. Altman, “export_fig: A Matlab Toolbox for exporting quality publication figures”, [Online]. Available: https://www.github.com/altmany/export_fig, . Accessed on: Dec. 8, 2019.

[23] Z. Dokur and T. Ölmez, “Segmentation of ultrasound images by using a hybrid neural network,” *Pattern Recognition Letters*, vol. 23, no. 14, pp. 1825-1836, 2002.

[24] S. Sharma, “Micro-Ultrasound Imaging of Tissue Dysplasia,” Doctoral dissertation, University of Dundee, 2015.

[25] M. Manchanda and R. Sharma, “A novel method of multimodal medical image fusion using fuzzy transform,” *Journal of Visual Communication and Image Representation*, vol. 40, pp. 197-217, 2016.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking the inception architecture for computer vision,” *In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2818-2826, Dec. 2015, DOI: 10.1109/CVPR.2016.308.

[27] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity mappings in deep residual networks." In European conference on computer vision, pp. 630-645. Springer, Cham, 2016.

[28] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition”, *In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[29] Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning." In Thirty-first AAAI conference on artificial intelligence. 2017.

[30] A. Gulli and S. Pal, “Deep learning with Keras”, Packt Publishing Ltd., 2017.

[31] H. C. Shin*et al.*, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning”. *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285-1298, 2016,

[32] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” *In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8697-8710, 2018. arXiv preprint arXiv:1707.07012, 2017.

[33] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. *Radiology*, vol. 143, no. 1, pp. 29-36, 1982

[34] Cardillo G: ROC curve: compute a Receiver Operating Characteristics curve.[http://www.mathworks.com/matlabcentral/fileexchange/19950-roc-curve]

[35] Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921-2929. 2016.

[36] J. Layland, A. M. Wilson, I. Lim and R. J. Whitbourn, “Virtual Histology: A Window to the Heart of Atherosclerosis,” Heart Lung Circ vol. 20, no. 10, pp. 615–621, 2011.

[37] B. F. Cox, F. Stewart, H. Lay, G. Cummins, I. P. Newton, M. P. Desmulliz, *et. al*, “Ultrasound capsule endoscopy: sounding out the future,” Ann. Transl. Med., vol. 5, no. 9, pp. 201–201, May 2017.

[38] G. Cummins., B. F. Cox, G. Ciuti, T. Anbarrasan, M. P. Desmuiliez, S. Cochran, *et. al* “Gastrointestinal diagnosis using non-white light imaging capsule endoscopy.,” Nat. Rev. Gastroenterol. Hepatol., Apr. 2019.