Dear editor/reviewer of Applied Sciences,

Thank you for reviewing our submission and providing us with valuable comments. We have revised the article such that all concerns are addressed and the suggested alterations are incorporated. Below, we respond to these concerns in a point-wise fashion, where our responses are in *italic dark blue*. Quotes from the former version of the manuscript are in *italic red text*, whereas quotes from the new manuscript are in *italic green text*.

Yours sincerely,

Roger Fonolla and Thom Scheeve,

Also on behalf of the co-authors

Reviewer 2

The manuscript "Ensemble of Deep Convolutional Neural Networks for classification of early Barrett's neoplasia using volumetric laser endomicroscopy" by Fonollà et al. is devoted to development Neural Network-assisted classification of non-dysplastic and neoplastic BE patients based on VGG16 NNs. The manuscript is well-written in general and can be of interest for a broad scientific audience. It can be published if the following deficiencies were addressed:

We would like to thank the reviewer for his appreciation of our work, the presented manuscript and the overall positive review.

1. While the NN side was explained reasonably well, the preprocessing steps are a little bit unclear. I would recommend adding a diagram with all preprocessing steps

We would like to thank the reviewer for this suggestion. Accordingly, we have added the following diagram to help better understand the preprocessing steps at the end of the Section 2.4.2.

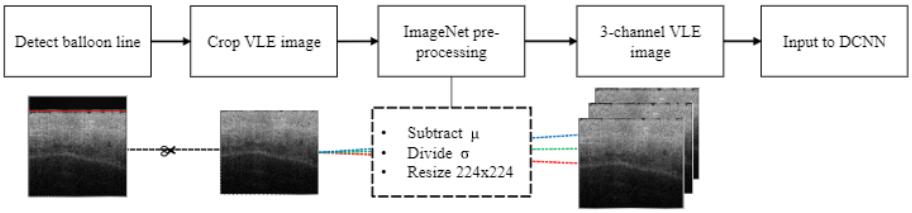

Figure 2. Automatic workflow of the preprocessing applied to each VLE frame. First the balloon line is detected in the VLE frame. Secondly, the image is cropped below the detected balloon line. The VLE image is then resized to 224 × 224. The mean (μ) is subtracted and the standard deviation (σ) divided according to the pre-trained ImageNet weights. Finally, the gray-scale channel is then triplicated to simulate the RGB requirements of the pre-trained network.

2. Line 141: It is unclear why authors moved from grayscale images to quasi-RGB images with triplicated grayscale channels. Did you use any pre-training?

We would like to thank the reviewer for pointing out the insufficient clarifications during the preprocessing steps. As the reviewer hints at, we indeed used a pre-trained VGG16 with ImageNet weights, which is further explained in the Training section. To avoid any confusion, we have changed the following paragraphs to help understand the reader the steps we followed in our work.

- *In Section 2.4.2. Preprocessing DCNN we have extended the following paragraph*

Old text:

"Each of the networks were initialized using ImageNet weights [29], hence each image was resized to 224 X 224 pixels to match the DCNN input shape. Furthermore, VLE images are originally embedded into the grayscale space, we triplicated the gray channels to represent the image in the RGB space. As final step, the dataset was

normalized by subtracting the mean and dividing by the standard deviation specified by the original ImageNet weights."

New text:

"Each of the networks were initialized using pre-trained ImageNet weights [29]. One limitation of using a pre-trained model is that the associated architecture cannot be changed, since the weights are originally trained for a specific input configuration. Hence, to match the requirements of the pre-trained ImageNet weights, each image was resized to 224 X 224 pixels. In addition, the dataset was normalized by subtracting the mean and dividing by the standard deviation, specified by the pre-trained ImageNet weights. As final step, the grayscale channel of each VLE image was triplicated to simulate the RGB input requirement of the pre-trained model (Figure 2)."

3. Authors should comment on how all images from 1 measurement were labeled. It is unclear whether all 51 images were labeled the same: NDBE or HGD?

We thank the reviewer for the suggestion. We agree with the reviewer that further explanation should be added into the setup of the presented experiments.

Therefore, we have interchanged Section 2.2 VLE Imaging System to Section 2.1 and Section 2.1 Data Collection and Description to Section 2.2.

We have replaced the following paragraph in the Section 2.2. Data Collection and Description, as well as added three new references that support the new explanations:

Old:

"For each patient, one or several regions of interest (ROIs) were extracted. VLE images in the ROIs were extracted from the VLE data, guided by the VLE-histology correlations. The histopathological correlation was assumed to apply over 1.25 mm (25 cross-sectional images) in both vertical directions (i.e, distal and proximal), resulting in 51 images per region."

New:

"For each patient, one or several regions of interest (ROIs) were extracted in the following manner. First, four-quadrant laser-mark pairs were placed at 2-cm intervals using the VLE system, according to the Seattle biopsy protocol [20,21]. Next, a full VLE scan was performed, after which the VLE balloon was retracted from the esophagus. Then, regular endoscopy was used to obtain biopsies in between the laser-mark pairs. Finally, ROIs were cropped from the full scan in between the same laser-mark pairs, and were labeled according to pathology outcome, ensuring histology-correlation [22] of the extracted ROIs. The histopathological correlation was assumed to apply over 1.25 mm, conform a small biopsy specimen, comprising 25 cross-sectional images in both vertical directions (i.e., distal and proximal), and thus resulting in 51 images per ROI."

4. Were all 3 DCNNs identical? with different weights seeded?

The reviewer addresses a very interesting issue.

1. *Yes, all the DCNNs were trained identically until convergence using the same optimizer hyperparameters. The only remarkable difference arises due to the cyclic learning rate, where each DCNN might converge at a different cyclic point.*
2. *By going back to the earlier point of the reviewer on pre-training in Line 141 (question 2 of the reviewer), we have already clarified whether we performed pre-training or not. Since in this work we take advantage of a pre-trained network, the weights are not differently seeded and are initialized exactly the same way. In the text we have covered this issue by stating:*

Section 2.4.2 Preprocessing DCNN, Line 154: "Each of the networks were initialized using pre-trained ImageNet weights [32]."

Minor deficiencies:

5. Line 44: Acronyms (e.g., NDBE, and HGD) should be defined on their first appearance

We have checked this and found this was done in Section 1: HGD, defined in Line 20 (old manuscript) / Line 22 (updated manuscript). NDBE, defined in Line 44 (old manuscript) / Line 48 (updated manuscript).

6. Line 93: In "In the previous works of Klomp et al. and Scheeve et al.," the explicit references should be included.

We have done this.

7. Line 204: Incorrect reference to Equation (3).

We would like to thank the reviewer for identifying and pointing out these errors in the manuscript and we have corrected them accordingly.

8. Authors need to explain what "single-frame" and "multi-frame" is in their context.

This is indeed a useful suggestion and we extended the paragraph in Section 2.3 with the aim to better clarify the work done by Scheeve et al. were we first refer to the terms if single-frame and multi-frame.

Old:

"In the previous works of Klomp et al. and Scheeve et al., several clinically-inspired quantitative image features were developed, the layer histogram (LH) and gland statistics (GS), to detect BE neoplasia in single frames. For a fair comparison with these studies, and per our request to the authors, we present our results by extending the single-frame analysis, further referred to as multi-frame analysis. The development of the clinically-inspired features has been described previously [12,17,18]. A summary of the methodology for the multi-frame analysis is given in the following sections."

New:

"In the previous works of Klomp et al. [23] and Scheeve et al. [15], several clinically-inspired quantitative image features were developed, the layer histogram (LH) and gland statistics (GS), to detect BE neoplasia in single frames. We refer to analysing one VLE image in a ROI to predict BE neoplasia as single-frame analysis. In Scheeve et al. [15], a single VLE image per ROI was used to compute the resulting prediction for each ROI. For a fair comparison with these studies, and per our request to the authors, we present our results by extending the single-frame analysis to 51 VLE images per ROI, further referred to as multi-frame analysis. The development of the clinically-inspired features has been described previously [15,23,24]. A summary of the methodology for the multi-frame analysis is given in the following sections."

General comment:

**We want to point out that we have as well corrected some minor grammatical errors throughout the manuscript.*