

Deep-learning-based Detection of Food Hypersensitivity from Confocal Laser Endomicroscopy Images of the Gastrointestinal Tract

Md Abid Hasan
German Research Center for AI (DFKI)
abid.hasan@dfki.de

Frédéric Li
University of Lübeck
fr.li@uni-luebeck.de

Vivian Tetzlaff-Lelleck
University of Lübeck
Vivian.tetzlafflelleck@student.uni-luebeck.de

Franziska Schmelter
University of Lübeck
Franziska.Schmelter@uksh.de

Greta Marie Ahlemann
University of Lübeck
GretaMarie.Ahlemann@uksh.de

Lennart Jablonski
German Research Center for AI (DFKI)
lennart.jablonski@dfki.de

Xinyu Huang
University of Lübeck
x.huang@uni-luebeck.de

Christian Sina
University of Lübeck
christian.sina@uksh.de

Marcin Grzegorzek
German Research Center for AI (DFKI)
marcin.grzegorzek@dfki.de

Abstract

Food hypersensitivity (FHS) is a relatively common pathological condition characterised by adverse reactions to specific foods or food components with currently limited diagnostic methods. Confocal laser endomicroscopy (CLE) has recently been proposed for the assessment of FHS, but their interpretation can be challenging even for trained physicians. We propose to alleviate this problem by training machine learning models on CLE images of the duodenal mucosa in the gastrointestinal tract for the binary classification problem of recognising images that show an adverse reaction after intestinal food challenge. More specifically, the performances of four state-of-the-art image classification models (VGG16, Inception-v3, Xception, MobileNet-v2) are compared on one dataset acquired from 38 patients with proven FHS. Additionally, the models decisions are interpreted using the Grad-CAM technique. Our study shows that although all four models achieve satisfying classification performances, they learn very different features in terms of interpretability from the clinical perspective.

Keywords: Food Hypersensitivity, Confocal Laser Endomicroscopy, Deep Learning, Grad-CAM

1. Introduction

Food hypersensitivity (FHS) refers to adverse reactions to food intake and encompasses various pathological conditions ranging from non-immune mediated food intolerances to immune-mediated food allergies (Lam et al., 2023; Turnbull et al., 2015). FHS is characterized by symptoms that can vary in their impact on daily life, such as watery eyes, runny nose, rash, abdominal pain, or diarrhea. In severe cases of food allergies, life-threatening reactions (anaphylaxis) can occur (Tedner et al., 2021). It is estimated that around 19% of the US population and between 2-37% of the European population are affected by FHS (Lam et al., 2023).

Predicting severe food reactions is challenging due to the variety of food-related factors (e.g., specific allergens) and host-related factors (such as individual sensitivity to allergens and the severity of previous reactions) (Lam et al., 2023). Current diagnostic methods for FHS, especially non-immunoglobulin E (IgE) mediated ones, are limited and highly variable. Food challenges or exclusion diets are time-consuming, while the clinical reaction to the ingested food occurs several hours after ingestion, and a negative response may not be known for days. There is a need for primary care physicians, gastroenterologists, nutritionists, and allergologists to have a diagnostic method for differentiating and identifying FHS in

patients who have tested positive or negative with current food allergy tests, especially those with negative IgE findings (Kiesslich et al., 2022).

An emerging new diagnostic tool is confocal laser endomicroscopy (CLE) with functional imaging (Bojarski et al., 2022). The magnification enables the identification of cells and vessels of the mucous membrane lining in the gastrointestinal tract. Additionally, CLE enables the imaging of dynamic processes, such as intestinal barrier dysfunction and intestinal barrier cell shedding, which constitute positive markers for adverse reaction to food (Fritscher-Ravens et al., 2014). Recent studies suggested that CLE may be useful for the detection of FHS or atypical allergies to food. Nevertheless, CLE is invasive, time-consuming and the interpretation by trained physicians is challenging. Consequently, the current status is inconclusive on whether CLE should be used as a diagnostic tool with sufficient accuracy (Bojarski et al., 2022). While image recognition through machine learning is a commonly applied tool, the application of machine learning for FHS detection from CLE images has, to the best of our knowledge, not yet been implemented.

We, therefore, investigate this problem by training machine learning models based on deep feature learning on a labeled dataset of 38 patients whose data were anonymised after their collection at the University Hospital Schleswig-Holstein (UKSH) Lübeck, Germany. The problem of FHS detection in intestinal CLE images is translated into a binary classification problem between the classes "affected" and "non-affected", that include images showing an adverse reaction to food intake or not, respectively. We compare the performances of several standard convolutional architectures pre-trained for image classification, including *VGG16* (Simonyan and Zisserman, 2014), *Inception-v3* (Szegedy et al., 2016), *Xception* (Chollet, 2016) and *MobileNet-v2* (Howard et al., 2017) for the downstream task of classifying CLE images between the affected and non-affected categories. We additionally provide explanations of how the models make their decision by analysing the *Grad-CAM* (Selvaraju et al., 2016) maps obtained on the input CLE images for all four tested models.

More specifically, the contributions of our work are as follows:

- We investigate the application of machine learning techniques to the problem of detecting reaction to FHS in CLE images of the gastrointestinal tract. This problem has not been explored in the literature yet to our best knowledge.

- We train and compare the performances of various standard convolutional-based neural architectures fine-tuned for the downstream task of classifying CLE images of the intestinal villi depending on whether they show an adverse reaction to food intake or not.
- We apply the *Grad-CAM* approach on the fine-tuned models to understand which parts of the input contribute the most to the classification decision. We show in particular that different models that yield similar classification performances may learn features that are quite different in terms of how meaningful they are, and of how easy they are to interpret.

The paper is organised as follows: Section 2 presents the related work from the literature. Section 3 describes the data and methods applied in the frame of this study. Section 4 shows the obtained experimental results, while Section 5 includes their discussion. Finally, a conclusion and outlook on future work is provided in Section 6.

2. Related Work

CLE in humans has revealed epithelial cell shedding and barrier defects, indicated by fluorescein plumes. Mouse studies showed inward flow through leakage events, increasing with lower luminal osmolarity. Inflammatory bowel disease (IBD) patients in remission with increased cell shedding and fluorescein leakage had higher relapse rates within 12 months (Kiesslich et al., 2012). These findings suggest confocal endomicroscopy can predict IBD relapse, making it a valuable diagnostic tool. It also aids in the rapid diagnosis of acute intestinal graft-versus-host disease (GvHD) during endoscopy, potentially reducing platelet transfusions and unnecessary biopsies (Bojarski et al., 2012). Further, CLE has revolutionised colorectal cancer diagnosis and treatment, allowing *in vivo* histology at subcellular resolution. CLE significantly impacts the diagnosis and management of patients during screening or surveillance colonoscopy for colorectal cancer, enabling immediate diagnosis and targeted interventions. Thus, CLE has become a crucial technique for *in vivo* diagnosis of colorectal cancer (Kiesslich et al., 2007). However, the usefulness and accuracy of CLE as a diagnostic tool are also subjects of ongoing debate. (Bojarski et al., 2022) criticised that the diagnostic accuracy for irritable bowel syndrome (IBS) is too low to recommend its widespread use, especially given the high response rate to a gluten-free diet (GFD) among IBS patients. A significant challenge is the interpretation of the images, which can be difficult even

for trained physicians.

Several attempts at leveraging machine learning techniques to address this issue by automating the CLE image analysis have been made in the past, although none of them are related to the detection of FHS to the best of our knowledge. (Aubreville et al., 2017) for instance carried out a study comparing the performances of Local Binary Pattern (LBP) features and features learnt by an *Inception-v3* convolutional architecture (Szegedy et al., 2016) to classify CLE images of the oral cavity for the detection of oral squamous cell carcinoma. Their analysis showed the superiority of the deep-learning-based method for this specific problem. (Izadyazdanabadi et al., 2018) checked the performances of various deep feature learning approaches for the detection of brain tumor in CLE images. Two convolutional-based models - *AlexNet* (Krizhevsky et al., 2012) and *Inception-v3* - were tested in different training configurations (trained from scratch, shallow or deep fine-tuning). The analysis showed that deep fine-tuning and creating an ensemble of the best models led to the best classification performances. (Aubreville et al., 2019) investigated the problem of the automated removal of motion artefacts in oral cavity and vocal folds probe-based CLE (p-CLE) images. A solution based on hand-crafted features using Histograms of Oriented Gradients (HOG), LBP, and angular offset features was compared to a deep learning model based on *Inception-v3*. The latter showed to be the most efficient at detecting motion artefacts. (Udriștoiu et al., 2021) applied machine learning techniques on colon CLE images to assist with the diagnosis of Crohn's disease. Two deep learning models either based on convolutional layers, or a combination of convolutional and Long-Short-term Memory (LSTM) layers were trained to solve the binary classification problem of detecting normal or inflamed colonic mucosa. The experimental results showed the superiority of the architecture involving LSTM layers. (van der Laan et al., 2021) reviewed the usages of CLE images for the diagnosis of IBD, in particular listing past work using convolutional architectures for the assessment of images for ulcerative colitis following the Mayo classification, or the assessment of the severity of the inflammation. They also highlighted the main challenges of machine-learning-based CLE analysis that consist mainly in the difficulty of obtaining good quality datasets (e.g. in terms of size, image quality, labels) and the lack of clear interpretability of the decisions made by deep learning models. (Lee et al., 2023) investigated the detection of pancreatic cystic lesions for the prevention of pancreatic cancer in CLE images. A *VGG-19* model (Simonyan and Zisserman,

2014) was trained for the classification of five different lesion types, which yielded high specificity but low sensitivity. (Angelina et al., 2024) also proposed an approach for the very same problem, i.e. classification of five types of pancreatic cystic lesions in CLE images. Two models following a *ResNet* architecture (He et al., 2015) were tested and yielded promising results, although they failed to outperform the *VGG* model of (Lee et al., 2023). (Sievrt et al., 2024) investigated the performances of the foundation model *GPT4.0* (OpenAI., 2024) for the classification of CLE images regarding whether they contained squamous cell carcinoma or not. The CLE images were provided as input of a not fine-tuned *GPT4.0* with a prompt asking the model to identify the image as healthy or malignant, and provide a justification of its decision. The performances of the model were compared to the assessment of three medical experts. *GPT4.0* yielded promising sensitivity, but significantly worse specificity than the human evaluation, thus highlighting the need for further refinement of such a model to properly assist humans in complex medical decision processes.

The analysis of the literature indicates a lack of investigations regarding the application of machine learning techniques for the assessment of FHS. We therefore propose to investigate this topic by training several deep feature learning models to recognise intestinal villi CLE images depending on whether they display an adverse reaction to food intake or not. We additionally perform an analysis of the *Grad-CAM* attention maps of the trained models to further understand how their decisions were made.

3. Material and Methods

3.1. Dataset description

To conduct this study, a dataset of CLE images of human duodenal mucosa was collected from 38 subjects in an anonymous manner to assess FHS. During the CLE procedure, the CLE Food Allergy Sensitivity Test (CLE FAST) was performed. CLE was performed by the endoscopy department of the Medical Clinic I (UKSH, Lübeck, Germany). For the investigation a Cellvizio device (Mauna Kea Technologies, Paris, France) and a gastroscope probe (Mauna Kea Technologies, Paris, France) was used. Subjects were recruited by the physicians of gastroenterology outpatient clinic at University Hospital Schleswig-Holstein in Lübeck. The subjects reported gastrointestinal symptoms associated with food intake without a clear diagnosis of food allergy or food intolerance. During the CLE FAST, various food solutions were applied to the mucosa of the

duodenum in randomized order. These foods included wheat flour (Stelzermühle, Bad Wurzach, Germany), dry yeast (Dr. August Oetker Nahrungsmittel KG, Bielefeld, Germany), soy flour (Alnatura Produktions- und Handels GmbH, Darmstadt, Germany), milk powder (gb-foods GmbH, Schillingsfürst, Germany) and egg white powder (gb-foods GmbH, Schillingsfürst, Germany). Fresh dilutions of 3g wheat and soy flour and 1.5g of the other foods were prepared with 30ml sodium chloride in a Falcon tube. After performing a standard esophagogastroduodenoscopy (EGD) to detect any structural defects, 2.5ml fluorescein 10% was injected intravenously. Endomicroscopy was performed on at least four sites to examine the baseline condition and CLE images were recorded. If there were no signs of a positive reaction (leakage of fluorescein into the lumen and cell detachment), the intestinal food sample was performed. For this, 30ml of a food solution was applied to the duodenal mucosa through the working channel of the endoscope. After two minutes, imaging was performed with the endomicroscope. If no reaction was visible, the duodenum and the working channel were flushed with saline and the test continued with the next food solution (Kiesslich et al., 2022). To ensure comprehensive coverage of the duodenal mucosa, images were then extracted from the CLE videos at a rate of two frames per second. Ethical approval was granted by the Ethics Committee of the University of Lübeck, Germany (approval number: AZ 19-233).

To annotate the dataset, two nutrition experts conducted a thorough manual review of the acquired images post-capture to identify frames that show an adverse reaction to food intake, i.e. belonging to the class "affected", and those that show no particular reaction, i.e. belonging to the class "not affected". The two annotators were in particular instructed to check the presence of either of the two following features that are markers of a reaction to food: (1) cell shedding, and (2) lumen fluorescence leakage between the villi (Kiesslich et al., 2022). The annotations were conducted independently, with each expert labelling one half of the dataset.

From this process, a labelled dataset consisting of a total of 9,095 images was obtained, which is comparable to the number of CLE images used in similar studies (Aubreville et al., 2017, 2019; Izadyazdanabadi et al., 2018). Out of these, only 890 images were classified as affected, while the remaining 8,205 images were classified as non-affected. This imbalance in class distribution reflects the natural occurrence of hypersensitivity reactions within the study population. The image order was then randomised. We

used stratification to build a balanced representation of the two classes in both the training and testing sets, i.e. 80% of both the "affected" and "unaffected" images were selected for the training set, with the remaining 20% of each class allocated to the test set.

3.2. CLE image classification

The classification performances of several pre-trained image classification approaches were evaluated for the binary classification of CLE images regarding whether they display an adverse reaction to food intake or not. More specifically, we selected the *VGG-16* (Simonyan and Zisserman, 2014), *Inception-v3* (Szegedy et al., 2016), *Xception* (Chollet, 2016), and *MobileNet-v2* (Howard et al., 2017) architectures. Both *VGG16* and *Inception-v3* are architectures that were commonly used in the CLE processing literature. *Xception* was introduced as an improvement over *Inception-v3* where *Inception* modules are replaced by depthwise separable convolutions (Chollet, 2016), while *MobileNet-v2* is a relatively lightweight model that has shown to yield competitive performances with other state-of-the-art models for various image processing tasks (Howard et al., 2017).

After replacing their softmax classification layer, we fully fine-tuned each model following the findings of (Izadyazdanabadi et al., 2018), which showed that the deep fine-tuning of models processing CLE images outperforms shallow fine-tuning in both classification performances and superior semantic interpretability of the obtained features.

To evaluate the performances of our models, we compute the accuracy, sensitivity, specificity, and average F1-score (AF1) on the testing set. Their formulas in terms of true positive (tp), true negative (tn), false positive (fp), and false negative (fn) are provided in Equations 1 to 6:

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (1)$$

$$Sensitivity = Recall_+ = \frac{t_p}{t_p + f_n} \quad (2)$$

$$Specificity = Recall_- = \frac{t_n}{t_n + f_p} \quad (3)$$

To address possible bias from class imbalance, we additionally calculated the average F1 score (AF1), also

known as the macro-averaged F1 score. The AF1 score is the average of the F1 scores for all classes, with each class F1 score being the harmonic mean of that class precision and recall.

$$F1_{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{Precision} = \frac{t_p}{t_p + f_p} \quad (5)$$

$$AF1_{score} = \frac{1}{c} \sum_{i=1}^c \text{F1 score}_i \quad (6)$$

4. Experiments and Results

Our experiments were implemented using the *Tensorflow* [2.12.0] Python library on a machine with an AMD Ryzen 9 7950X 16-Core Processor CPU and NVIDIA GeForce RTX 4090 GPU. All models were trained using the ADAM optimiser with a learning rate of 0.0001. The training process spanned over 100 epochs, and a batch size of 16 was used.

The classification metrics obtained by each model on the testing set are shown in Table 1. It can be seen that all four tested models yield comparable classification metrics. *Inception-v3* performs the best with an accuracy / sensitivity / specificity / AF1 of 98.14 / 85.96 / 99.43 / 94.44% respectively. *VGG16* on the other hand returns slightly worse performances evaluation metrics than the other tested models, with an accuracy / sensitivity / specificity / AF1 of 99.07 / 80.11 / 98.87 / 91.21% respectively. The confusion matrices of all models are also provided in Figure 1. It can be observed from them that all models obtain satisfying classification performances for both classes, despite the class unbalance observed on this dataset.

5. Discussion

The obtained classification performances have revealed that all of the four tested convolutional architectures are able to obtain promising performances for the binary classification problem of recognising whether an adverse reaction to food was detected in CLE images or not. Slight differences in performances between the models could still be observed, which raised the question of understanding by what they could be caused.

To understand the reasons behind the decisions of our tested models, we use the *Grad-CAM* approach (Selvaraju et al., 2016) that is a state-of-the-art interpretability method for convolutional architectures. *Grad-CAM* is a technique for visualising and interpreting the decisions of Convolutional Neural Networks (CNNs) by highlighting important regions in the input image. It is based on the principle of flowing back the gradients of a target concept (i.e. class) with respect to the feature map activations of a layer. More specifically, *Grad-CAM* works by computing the gradients of the target class score y^c with respect to the feature maps A^k of the last convolutional layer. These gradients are globally averaged to obtain importance weights α_k^c , which are then used to create a weighted combination of the feature maps as shown in Equation 7.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (7)$$

where Z is the number of pixels in the feature map A^K .

The resulting heatmap is passed through a ReLU function to focus on the positive influences and then upsampled to the size of the input image ($L_{\text{Grad-CAM}}^c$), highlighting areas that significantly impact the prediction of the model as shown in Equation 8.

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (8)$$

In accordance with the common recommendation of the literature (Fu et al., 2020; Selvaraju et al., 2016), we compute the *Grad-CAM* maps of the last convolutional layer of each model. This means, more specifically, the layer "block5_conv3" for *VGG16*, "mixed10" for *Inception-v3*, "block14_sepconv2" for *Xception* and "Conv_1" for *MobileNet-v2*. The generated *Grad-CAM* map are then reshaped to the input image size and superimposed to the input CLE image for further interpretation.

To perform an analysis of the models decision, we computed the *Grad-CAM* maps from 10 random images randomly selected from the testing set, that were unambiguous true positives, i.e. CLE images depicting an adverse reaction to food classified as such by all four tested models. The analysis was performed on unambiguous true positive examples only to facilitate the interpretation and comparison across models. True negative examples in particular were not analysed in depth since the negative class is mostly characterised

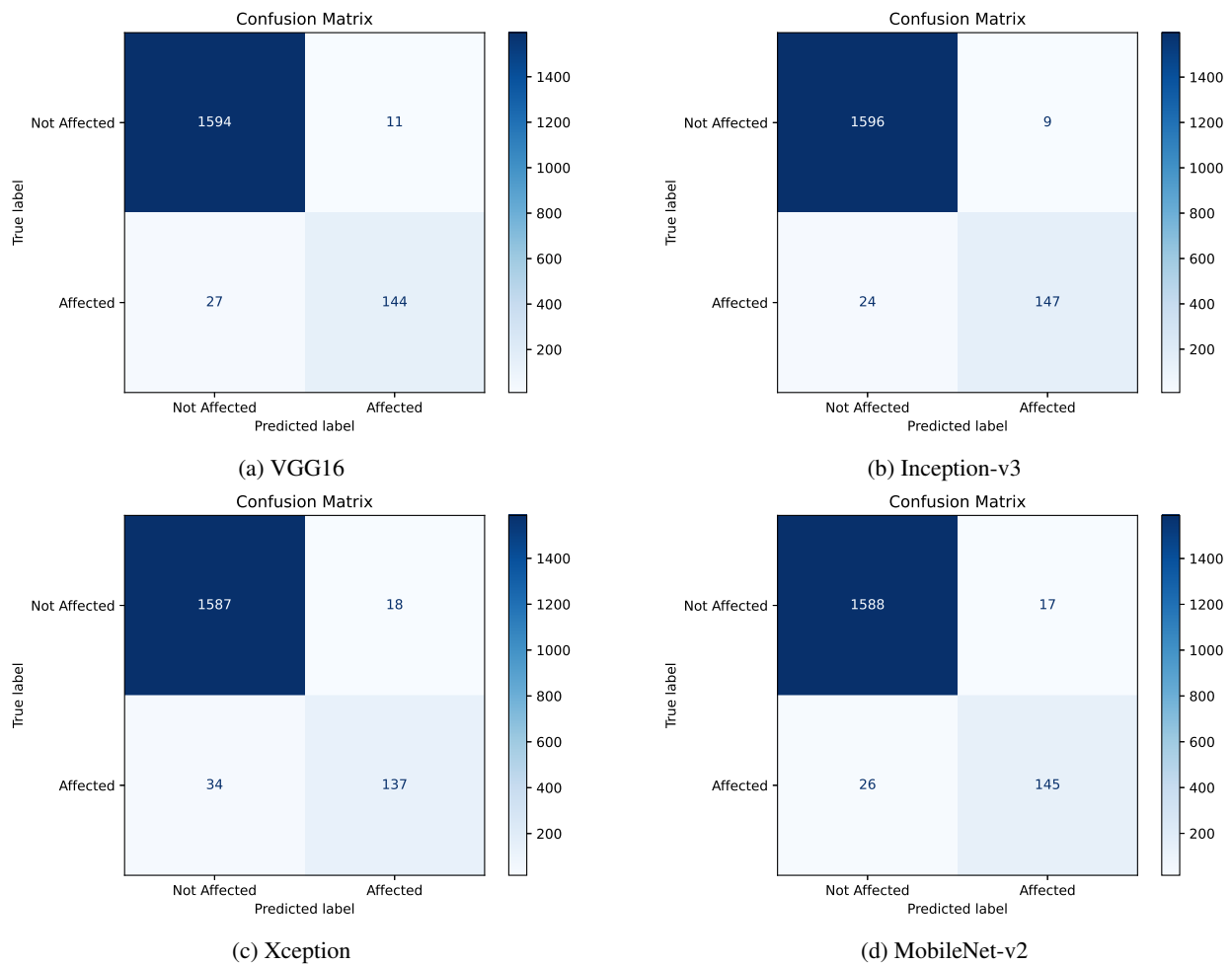


Figure 1: Confusion matrices of the four tested models.

Table 1: Classification performances of the four tested models obtained on the testing set.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Average F1-score (%)
VGG16	97.07	80.11	98.87	91.21
Inception-v3	98.14	85.96	99.43	94.44
Xception	97.57	84.79	98.94	92.87
MobileNet-v2	97.86	84.21	99.31	93.58

by the absence of cell shedding or lumen fluorescence, which is harder to interpret from *Grad-CAM* maps. The latter were reviewed by one clinical expert who indicated how meaningful from the human perspective the highlighted parts of the image are for each model. The *Grad-CAM* maps for all four models are displayed in Figure 2.

The analysis of the *Grad-CAM* attention maps revealed the following insights:

- Although they yield somewhat similar

classification performances, the four tested models make decisions based on quite different features.

- Despite returning the worst classification performances out of the four tested models, *VGG16* is the neural network that makes decisions the most aligned with the clinical expertise. It is in particular good at highlighting lumen fluorescence (as seen in examples #2, #3, #4, #5 and #9 in Figure 2), but tends to overlook

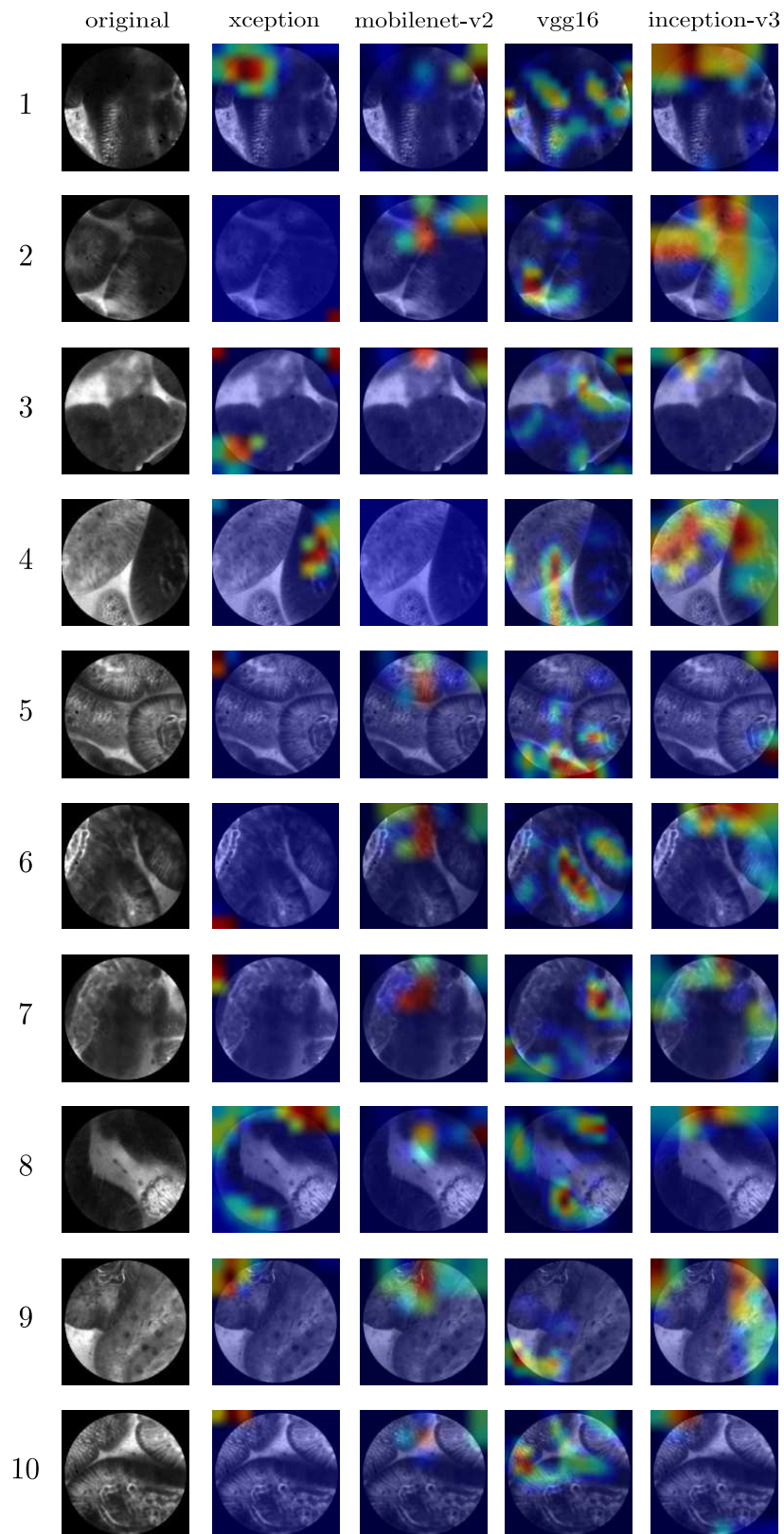


Figure 2: Grad-CAM attention maps obtained for ten random true positive examples of the testing set. Each row corresponds to one example while the columns represent in order: the original CLE image, and the attention maps of the *Xception*, *MobileNet-v2*, *VGG16*, and *Inception-v3* models, respectively. High and low activations are respectively indicated by the yellow/red and blue colouring.

cell shading.

- On the other hand, *Xception* seems to be mostly making decisions based on noise as it consistently highlights the dark borders of the CLE images, that do not contain any meaningful content from the semantic point of view. This behaviour could be an indication of overfitting on the acquired dataset.
- *Inception-v3* and *MobileNet-v2* are both between *VGG16* and *Xception* in terms of interpretability, as they sometimes highlight meaningful features of the CLE images, but also sometimes their border.
- For some examples, none of the models highlight the information traditionally used in clinical practice to assess the CLE images, i.e. cell shading or lumen fluorescence. This could be an indication that further information currently not considered by trained physicians can be used to recognise adverse reactions to food.

The findings of our study suggest that care should be taken when selecting the best performing deep-learning model, as commonly used classification evaluation metrics may not be indicative of the quality of the learnt features. A model with lower classification performances but better interpretability (such as *VGG16* in this study) is likely to generalise better than a model with better metrics but that seemingly makes its decisions based on noise (such as *Xception*). Similar findings were also found in additional experiments that involved freezing a variable number of layers in the transferred models (from one to all layers frozen), instead of fine-tuning all of them during the training process. These configurations yielded similar classification performances to the ones obtained after full fine-tuning that are reported in Table 1. However, the obtained features in them were evaluated as notably less interpretable after visual inspection of the *Grad-CAM* maps by one expert. These findings match the observations of (Izadyyazdanabadi et al., 2018) who showed that the deeply fine-tuned approach outperforms shallow fine-tuning in providing the best CLE features in terms of semantic content.

Despite the care we took in this study, its scope remains limited by a few aspects. Firstly, the dataset acquired for this study is relatively small which lowers the generalisation capacity of our results. Secondly, the annotation of the data was carried out by two nutrition experts independently, without cross-checking of the assigned labels. Finally, the analysis of the *Grad-CAM* maps was performed with a single interpretability

method, and only qualitatively on a small subset of the dataset, since the manual analysis of all images would have been too costly in terms of invested time. To address these issues, it is planned to acquire data from additional subjects with a third expert assessing the reliability of the assigned labels. Additionally, other state-of-the-art deep learning interpretability approaches such as *Grad-CAM++* (Chattopadhyay et al., 2018) or *Score-CAM* (Wang et al., 2020) will be tested to confirm the observations obtained with *Grad-CAM*.

6. Conclusion

The study presented in this paper investigates the usage of machine learning models to assess FHS from CLE images of the duodenal mucosa in the gastrointestinal track. Four state-of-the-art image classification deep neural networks (*VGG16*, *Inception-v3*, *Xception*, *MobileNet-v2*) were fine-tuned on a dataset of CLE images of FHS patients for the binary classification problem of detecting whether an image displays an adverse reaction to food or not. The experimental results show that all tested models yield promising classification performances, with *Inception-v3* slightly outperforming the other architectures in terms of accuracy, sensitivity, specificity, and AFI. Despite obtaining comparable classification metrics, an analysis of the *Grad-CAM* attention maps reveals that the tested models learn notably different features, with *VGG16* aligning with the clinical expertise, while *Xception* seemingly makes decisions based on noise. These findings suggest that considering classification evaluation metrics only may not be sufficient to assess how good a transferred deep-learning model is to solve a given target problem.

Future work will include the acquisition of a dataset containing more subjects as well as annotations that are cross-checked by a third nutrition expert. Furthermore, the analysis of additional interpretability techniques such as *GradCAM++* or *ScoreCAM* will be performed to check the generalisation capacity of our results. A larger dataset will also open up possibilities to investigate additional research questions, such as whether it is possible to recognise adverse reactions to specific food allergens (e.g. wheat, dry yeast, soy flour, milk powder, egg white powder) in the CLE images, or whether it would be possible to predict FHS from the CLE images acquired before the food intake using anatomical features.

References

- Angelina, C. L., Pan, C.-M., Lee, T.-C., Han, M.-L., Kongkam, P., Wang, H.-P., Chang, C.-Y., & Chang, H.-T. (2024). Classification of pancreatic cystic lesions using resnet deep learning network in confocal laser endomicroscopy videos. *Proc. of ISICO 2023*, 357–363.
- Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., Bohr, C., Neumann, H., Stelzle, F., & Maier, A. (2017). Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. *Scientific Reports*, (11979).
- Aubreville, M., Stoeve, M., Oetter, N., Goncalves, M., Knipfer, C., Neumann, H., Bohr, C., Stelzle, F., & Maier, A. (2019). Deep learning-based detection of motion artifacts in probe-based confocal laser endomicroscopy images. *International Journal of Computer Assisted Radiology and Surgery*, 31–42.
- Bojarski, C., Günther, U., Rieger, K., Heller, F., Loddenkemper, C., Grünbaum, M., Uharek, L., Zeitz, M., & Hoffmann, J. (2012). In vivo diagnosis of acute intestinal graft-versus-host disease by confocal endomicroscopy. *Endoscopy*, (5).
- Bojarski, C., Tangermann, P., Barmeyer, C., Buchkremer, J., Kiesslich, R., Ellrichmann, M., Schreiber, S., Schmidt, C., Stallmach, A., Roehle, R., Loddenkemper, C., Daum, S., Siegmund, B., Schumann, M., & Ullrich, R. (2022). Prospective, double-blind diagnostic multicentre study of confocal laser endomicroscopy for wheat sensitivity in patients with irritable bowel syndrome. *Gut*.
- Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *Proc. of IEEE WACV 2018*.
- Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *Proc. of IEEE CVPR 2017*.
- Fritscher-Ravens, A., Schuppan, D., Ellrichmann, M., Schoch, S., Röcken, C., Brasch, J., Bethge, J., Böttner, M., J., Klose, & Milla, P. (2014). Confocal endomicroscopy shows food-associated changes in the intestinal mucosa of patients with irritable bowel syndrome. *Gastroenterology*, (147).
- Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., & Li, B. (2020). Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv:2008.02312*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *Proc. of CVPR 2015*.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*.
- Izadyyazdanabadi, M., Belykh, E., Mooney, M., Martirosyan, N., Eschbacher, J., Nakaji, P., Preul, M. C., & Yang, Y. (2018). Convolutional neural networks: Ensemble modeling, fine-tuning and unsupervised semantic localization for neurosurgical cle images. *Journal of Visual Communication and Image Representation*, 10–20.
- Kiesslich, R., Duckworth, C., Moussata, D., Gloeckner, A., Lim, L., Goetz, M., Pritchard, D., Galle, P., Neurath, M., & Watson, A. (2012). Confocal endomicroscopy shows food-associated changes in the intestinal mucosa of patients with irritable bowel syndrome. *Gut*, (8).
- Kiesslich, R., Goetz, M., Vieth, M., Galle, P., & Neurath, M. (2007). Technology insight: Confocal laser endoscopy for in vivo diagnosis of colorectal cancer. *Nature Reviews Clinical Oncology*, (4), 480–490.
- Kiesslich, R., Rusticeanu, M., Langhorst, J., Sina, C., Benamouzig, R., & Tack, J. (2022). *Endoscopic diagnosis of food-induced allergy-like reactions*. <https://www.maunakeatech.com/en/media/download/2532> (accessed: 14.06.2024).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.
- Lam, H. C. Y., Neukirch, C., Janson, C., Garcia-Aymerich, J., Clausen, M., Idrose, N. S., Demoly, P., Bertelsen, R. J., Ruiz, L. C., Raheison, C., & Jarvis, D. L. (2023). Food hypersensitivity: An examination of factors influencing symptoms and temporal changes in the prevalence of sensitization in an adult sample. *European Journal of Clinical Nutrition*, (77), 833–840.

- Lee, T.-C., Angelina, C. L., Kongkam, P., Wang, H.-P., Rerknimitr, R., Han, M.-L., & Chang, H.-T. (2023). Deep-learning-enabled computer-aided diagnosis in the classification of pancreatic cystic lesions on confocal laser endomicroscopy. *Diagnostics*, (13(7)).
- OpenAI. (2024). GPT-4 technical report.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proc. of IEEE ICCV 2017*.
- Sievert, M., Aubreville, M., Mueller, S. K., Eckstein, M., Breininger, K., Iro, H., & Goncalves, M. (2024). Diagnosis of malignancy in oropharyngeal confocal laser endomicroscopy using GPT 4.0 with vision. *European Archives of Oto-Rhino-Laryngology*, (281), 2115–2122.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proc. of IEEE CVPR 2016*, 2818–2826.
- Tedner, S., Asarnej, A., Thulin, H., Westman, M., Konradsen, J., & Nilsson, C. (2021). Food allergy and hypersensitivity reactions in children and adults—a review. *Journal of Internal Medicine*.
- Turnbull, J. L., Adams, H. N., & Gorard, D. A. (2015). Review article: The diagnosis and management of food allergy and food intolerances. *Alimentary pharmacology & therapeutics*.
- Udriștoiu, A. L., Ștefănescu, D., Gruionu, G., Gruionu, L. G., Iacob, A. V., Karstensen, J. G., Vilmann, P., & Săftoiu, A. (2021). Deep learning algorithm for the confirmation of mucosal healing in crohn's disease, based on confocal laser endomicroscopy images. *Journal of Gastrointestinal and Liver Diseases*, (30(1)), 59–65.
- van der Laan, J. J. H., van der Waaij, A. M., Gabriëls, R. Y., Festen, E. A. M., Dijkstra, G., & Nagengast, W. B. (2021). Deep learning algorithm for the confirmation of mucosal healing in crohn's disease, based on confocal laser endomicroscopy images. *Expert Review of Gastroenterology and Hepatology*, (15(2)), 115–126.
- Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020).

Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proc. of IEEE CVPR 2020*.