Article

AI-based detection of aspiration for video-endoscopy with visual aids in meaningful frames to interpret the model outcome.

Jürgen Konradi[1*], Milla Zajber[2], Ulrich Betz[1], Philipp Drees[3], Annika Gerken[4], Hans Meine[4]

[1] Institute of Physical Therapy, Prevention and Rehabilitation, University Medical Center of the Johannes Gutenberg-University Mainz, Germany; juergen.konradi@unimedizin-mainz.de (J.K.); ulrich.betz@unimedizin-mainz.de (U.B.)
[2] Department for Health Care & Nursing, Catholic University of Applied Sciences, Mainz, Germany; millazajber@gmail.com (M.Z.)
[3] Department of Orthopedics and Trauma Surgery, University Medical Centre, Johannes Gutenberg University Mainz, 55122 Mainz, Germany; philipp.drees@unimedizin-mainz.de (P.D.)
[4] Fraunhofer-Institute for Digital Medicine MEVIS, Bremen, Germany; annika.gerken@mevis.fraunhofer.de (A.G.); hans.meine@mevis.fraunhofer.de (H.M.)

*Correspondence: juergen.konradi@unimedizin-mainz.de; Tel.: +49-6131-17-2362

Abstract: Disorders of swallowing often lead to pneumonia when material enters the airways (aspiration). Flexible Endoscopic Evaluation of Swallowing (FEES) plays a key role in the diagnostics of aspiration but is prone to human errors. An AI-based tool could facilitate this process. Recent non-endoscopic/non-radiologic attempts to detect aspiration using machine-learning approaches have led to unsatisfying accuracy and show black box characteristics. Hence, for clinical users it is hard to trust in these model decisions. Our aim is to introduce an explainable artificial intelligence (XAI) approach to detect aspiration in FEES. Our approach is to teach the AI about the relevant anatomical structures like the vocal cords and the glottis based on 92 annotated FEES videos. Simultaneously, it is trained to detect bolus that passes the glottis and becomes aspirated. During testing, the AI successfully recognized glottis and vocal cords, but could not yet achieve satisfying aspiration detection quality. Albeit detection performance has to be optimized, our architecture results in a final model that explains its assessment by locating meaningful frames with relevant aspiration events and by highlighting the suspected bolus. In contrast to comparable AI tools, our framework is verifiable, interpretable and therefor accountable for clinical users.

Keywords: XAI; segmentation; detection; aspiration; glottis; vocal cords; endoscopy; FEES; interpretability; meaningful sequences; key frames

1. Introduction

Machine Learning has a huge impact on biomedical applications and will play a continuously increasing role in diagnostics and patient care [1]. The underlying AI models can be divided into two classes: White-box and black-box models. White-box models, e.g., decision trees based on comprehensible input variables, allow the basic understanding of their algorithmic relationships; they are thus self-explanatory with regard to their mechanisms of action and the decisions they make. With black-box models, such as deep neural networks that have recently redefined the state of the art in many applications, it is generally no longer possible to understand their inner workings [2]. Instead, there are methods for the explanation of single decisions (local explainability) or attempts at deriving descriptions of specific input patterns that a trained model looks out for. Depending on

 doi:10.20944/preprints202212.0051.v1

the specific requirements, it is possible to apply established explanation tools, e.g., LIME, SHAP, Integrated Gradients, LRP, DeepLift or GradCAM [3]. But even these tools require expert knowledge for the interpretation of their output, and only a few of them provide intuitively understandable decision explanations (e.g., saliency maps, prototypes or surrogate models or contrastive and counterfactual explanations) [4]. This means that the importance of explanatory strategies will continue to increase in the future, while they are already an essential component of many AI applications today. The importance of explainability varies greatly depending on the field [5], with the healthcare sector being one of the most demanding ones. To serve this need, technical and non-technical challenges have to be overcome. This can lead to new and further development of suitable "hybrid" approaches that combine data- and knowledge-driven concepts and/or white- and black-box modeling attempts [6]. Additionally, behavioral or cognitive science aspects for explainable AI should be considered, such as transparency and measurability of the explanation as well as automated explanation adaptations for users. Employing this human-computer-interaction (HCI) provides transparency to the user, allowing them to trust the machine [7]. For instance, regarding digital applications that are based on video recordings, the identification of meaningful frames or key frames [8,9] in video sequences is one saliency map approach that can be very helpful to interpret algorithmic decisions. As an example for such a perceptive human-based interpretation approach [3], we will introduce a concept that can be used to facilitate the clinical diagnosis of swallowing disorders based on video-endoscopic swallowing examinations.

Disorders of swallowing are a relevant problem across various etiologies and all sectors of healthcare provision. Each year, approximately one in 25 adults will experience a swallowing problem in the United States. Dysphagia cuts across so many diseases and age groups that its true prevalence in adult populations is not fully known and is often underestimated [10]. A recent systematic review demonstrated that the presence of oropharyngeal dysphagia significantly increases healthcare utilization and cost, highlighting the need to recognize oropharyngeal dysphagia as an important contributor to pressure on healthcare systems [11]. The leading cause for the complications of dysphagia is the aspiration of boluses and saliva (i.e., when material passes the vocal cords and enters the airways). A comprehensive review summarizes that 43-54% of all acute stroke patients suffer from dysphagia, about 37% of those develop aspiration pneumonia, of which 3.8% die if no dysphagia diagnosis and therapy takes place. The aspiration pneumonia rate in the first 14 days can be lowered from 8.2% to 1.3% (relative risk reduction 84%) by early screening, instrumental diagnostics, and subsequent dysphagia therapy [12].

At present there are two instrumental diagnostics that can be regarded as gold standards: Videofluoroscopic Swallowing Study (VFSS) and Flexible Endoscopic Evaluation of Swallowing (FEES). In contrast to VFSS, FEES is appropriate for bedside administration, is radiation-free, can be administered by speech and language pathologists and is therefore not relying on rare medical personnel, which altogether leads to far lower costs for FEES [13,14]. All these aspects limit the clinical use of VFSS. Consequently, FEES is currently the most commonly used tool for instrumental dysphagia diagnostics. With the goal to improve and systematize training, a multi-level training curriculum was developed [15] that is now also implemented within the European Society for Swallowing Disorders (ESSD). Hence, FEES it is in widespread use across Europe. In 2010 in Germany, FEES was incorporated in the German Version of the International Classification of Procedures in Medicine with a cost estimation of 200€ [16], which is lower than retrospectively calculated mean reimbursements of $321.23 in the US [14]. At present, and to be in accordance with the reimbursement procedure (OPS), FEES has to be performed by two persons. According to literature the duration of FEES administration varies between 30-40 minutes [17] but can easily reach 90-120 minutes (own experience of > 1500 FEES). Furthermore, in a very time-consuming process, data has to be stored and inspected again for the diagnostic report to establish better reliability in detection of aspiration (Krippendorff's alpha ~.78 vs. second video inspection frame by frame ~.87) [18]. Beyond binary

 doi:10.20944/preprints202212.0051.v1

diagnosis (aspiration / no aspiration), more detailed scales like the Penetration-Aspiration Scale (PAS) [19] can be used to describe the results. The PAS classifies from Score 1 (Material does not enter the airway), via Score 2-5 (Penetration of material into the larynx of different depths and ability of clearance) to Score 6-8 (Aspiration with different ability for clearance; whilst 8 means no attempt for clearance at all). In these cases, the overall inter-rater reliability (IRR) across clinicians is stated to be between .35 (PAS score 5), .56 (PAS score 7) and .73. (PAS score 8) [20]; especially PAS 7 and 8 are highly relevant since they indicate aspiration. Most striking are the differences between intra-RR (.60) and inter-RR (.29) before specific trainings [21], but overall inter-RR score irrespective of clinical experience can also reach .85 [22]. Hence, the more differentiated the diagnostics should be and the less the staff is trained, the less reproducible human decisions become. The low intra- and inter-RR values especially for PAS score 7 and 8 clearly show the existence of relevant missing rates.

Taken together, there is room for improvements in FEES in the areas of validity, reliability, and duration in the context of aspiration detection, the reduction of staff needed for administration and report of findings, as well as general costs. One recent approach that addresses the problem of human misses for penetration and aspiration detection is narrow band imaging. It is implemented in certain types of endoscopes, can be used to sharpen the optical contrasts, and has proven to increase the IRR [23,24]. However, all described aspects for improvement could be addressed, when combining the administration of FEES with the help of an Artificial Intelligence (AI) tool that is capable of giving reproducible, quantitative output based on a frame-by-frame analysis without concentration errors.

Whilst up to date no one has developed an AI to detect aspiration for FEES videos, various other attempts using machine-learning approaches to detect aspiration or signs for unsafe swallowing have been performed. The only high potential application is a CNN for Aspiration detection of VFSS videos with an accuracy of AUC of 1.00 [25], but as described above, VFSS is limited in its clinical use. Further studies investigated the possibility of identifying dysphagia by means of localization of the hyoid bone or hyoid bone movements by an AI: On the one hand, the detection of auscultations, swallowing sounds and vibrations is used [26-29] and on the other hand, video material (VFSS or ultrasound) [30-32]. Both approaches yield good results. In general, detection of (impaired) swallowing based on auscultation is the subject of research in many studies [33-35]. Furthermore, there are studies investigating the combination of pressure build-up (lingual/palatal/pharyngeal) in combination with AI [36,37], sometimes additionally with the combination of VFSS data [38]; again, promising results can be obtained [39]. Combinations of various biometric data are also used for AI-based dysphagia diagnosis [40]. Studies looking at aspiration detection using image data (VFSS) [25] or swallow onset detection [41] also yield promising results. A different approach is an image analysis of the external neck appearance for the detection of sarcopenic dysphagia [42]. Finally, also speech recordings have been investigated for the presence of dysphagia [43]. Whilst some of these approaches already show sufficient accuracy, these existing machine-learning models all show black box characteristics and lack transparency regarding their classification results: After the model outcome there is no gold standard (ground truth) which can be used by the examiner to validate the models' assumptions, as they do not provide the examiner with any explanatory insight of the airways. Hence, these models only classify between healthy and at risk for aspiration. Only a subsequent FEES or VFSS could validate the model outcome. Hence, for clinical users it is hard to trust in these models and their decisions. Furthermore, such lack of transparency might not comply with requirements of the European General Data Protection Regulation (GDPR), as it prohibits decision solely based on automated processing [44] and therefor limits practical applications in the clinical context [45].

 doi:10.20944/preprints202212.0051.v1

Thus, our aim is to introduce an explainable artificial intelligence (XAI) approach to detect aspiration (i.e., liquids, jelly, saliva) during FEES for patients suffering from dysphagia. The automatic detection should improve IRR but also be interpretable, increasing its trustworthiness and transparency. To facilitate the detection of bolus aspiration, while at the same time achieving explainability goals, the AI should also learn the segmentation of relevant anatomical structures like the vocal cords and the glottis, a task that has been shown to be feasible before [46-48]. Simultaneously, the AI will be trained to detect bolus that passes the glottis and becomes aspirated into the airways. This interpretable architecture results in a final model that explains its assessment by locating specific video frames with relevant aspiration events and by highlighting the glottis, vocal cords, and suspected bolus in situ as visual aids in meaningful frames.

2. Materials and Methods

2.1 Video Data and Annotation

92 patient videos (50 showing aspiration, 32 videos showing penetration and 10 without aspiration) based on established PAS-Scores [22], 8-6 for aspiration, 5-2 for penetration (no aspiration), and 1 for healthy from an already existing data set of ~1500 FEES recordings were retrospectively analyzed by 2 experts for FEES as a basis for annotation. All recordings were made by the same type of endoscope (Orlvision, Video Rhino Laryngoscope RS1, 3.9mm diameter, 130°/130° probe control, 90° viewing angle, 291.000 px resolution, Orlvision GmbH) and recorded on an rpSzene system (Rehder/PartnerGmbH). The study was approved by the responsible ethics committee of the State Chamber of Physicians of Rhineland-Palatinate (No.: 2021-16141-retrospektiv) and is registered with WHO (INT: DRKS00026822). We split the videos into disjunct sets for training, validation during training, and final testing. The videos were graphically annotated using the highly customizable annotation tool for data curation and quality control SATORI [49]. To ensure a human-in-the-loop approach, the two domain experts performed annotation. The structures of the vocal cords, the glottis (open, closed, obscured) as the region of interest (ROI), and cases of aspiration (saliva, liquid, slurry) as well as no aspiration were drawn into certain frames and serve as the gold standard (ground truth) for the AI segmentation. In addition, frames not showing any of the structures or cases of aspiration were labeled as such using a frame-labeling tool, to reduce the number of false positive segmentations when processing a full video.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 December 2022 doi:10.20944/preprints202212.0051.v1

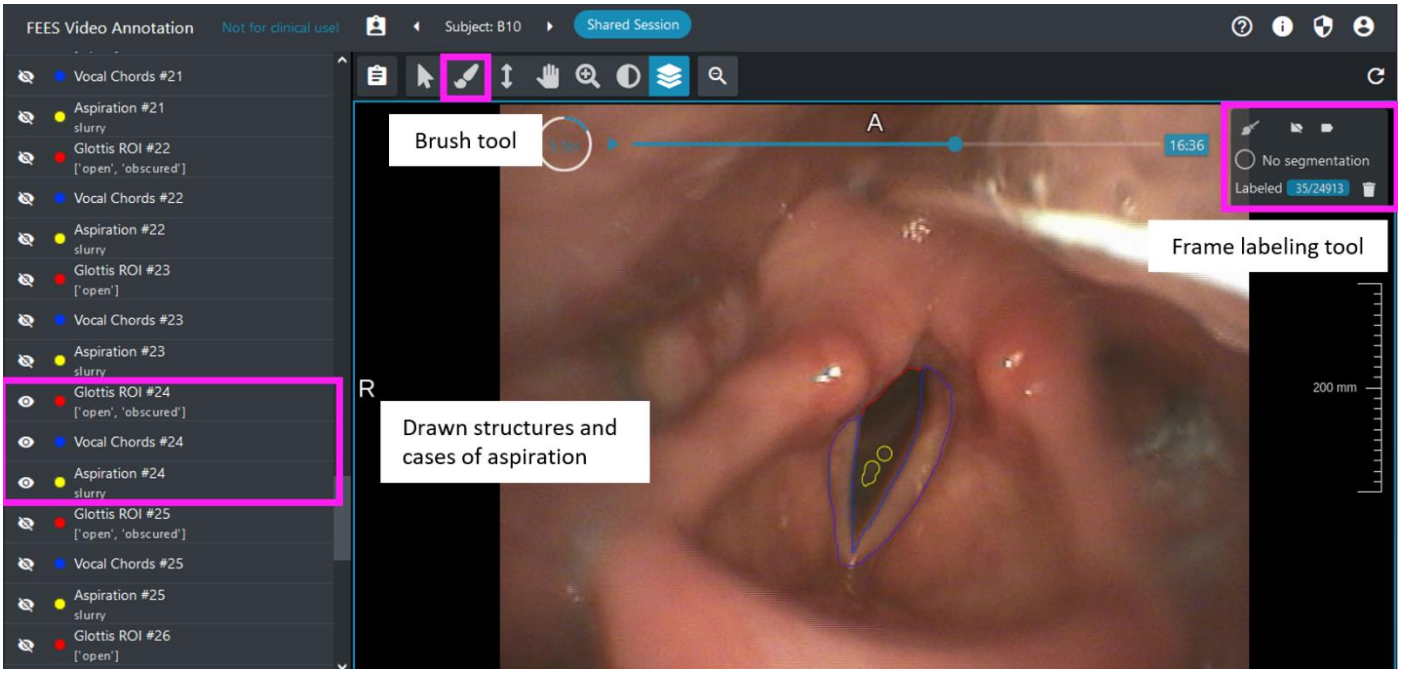


Figure 1. Annotation example with drawn structures of vocal cords (red), glottis (blue) and cases of aspiration (yellow) using the frame-labeling tool (screenshot).

These labelled pixels create the data basis for a subsequent convolutional neural network (CNN; U-Net) specifically designed for segmentation tasks. In addition to aspiration, the AI was also trained to segment the ROI and vocal cords, because they are easier to detect and the aspiration always appears within this region of interest.

Furthermore, the training data was augmented using geometric transformations and color modifications (rotations, zooming up to *1.5, mirroring left-right, change of contrast as well as picture brightness based on frame mean ± 25%) to add as many variants of the annotated frames as possible. This was done in order not to make the detection of aspiration dependent on incidental features, such as sharpness or contrast, since the detection of findings becomes more robust to such distortions and shape changes when sufficiently trained with appropriate data. Such augmentation techniques are commonly used to teach modern AI models that different positions, lighting conditions and camera angles, partial occlusions or horizontal and vertical shifts do not represent anomalies [50]. This approach increases robustness, reducing the expected performance drop when applying the resulting model to external test data. In summary, the goal is to train the model on a sample of videos that reasonably cover the expected variability occurring during practical application. Therefore, since the quality of the FEES videos also varies in reality, videos were also selected on which the structures are rather poorly visible.

2.2 Deep Neural Networks for Segmentation

A 2D U-Net was chosen as neural network architecture to segment glottis ROI, vocal cords and aspirated bolus. U-Nets were developed specifically for the segmentation of biomedical images [51] and variants of this idea have become the most commonly used and most successful architecture to date [52]. The architecture is based on a performant fully-convolutional design and consists of an encoder and a decoder that produce a result at the same resolution as the input image, with skip connections that facilitate information flow and feature re-use for a detailed result. Unlike the original U-Net architecture, convolutions with zero-padding and only 32 base filters in the first convolutional layer were used. For regularization, dropout [53]) and batch normalization [54] were added and PReLU [55]was chosen as activation function. The training was run on videos downscaled by factor two to remove comb artifacts from interlaced recording, on patches of size

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 December 2022 doi:10.20944/preprints202212.0051.v1

352x288 and batch size 16. The training error was optimized using the Adam optimizer [56] and Dice loss function [57] with an initial learning rate of 10^{-4}. The patches were sampled so that 80% included the glottis ROI and 25% of these showed aspirations, and the remainder were frames labeled as not containing the ROI. Every 500 iterations, the U-Net was evaluated on the validation data and the Jaccard score to the reference segmentation was computed. After 15 validation steps without an improvement in the Jaccard score, the training was stopped. The network state with the highest validation Jaccard score was retained and selected as output model ("early stopping"), which helps to prevent overfitting that this task on such a relatively small dataset would otherwise be prone to (in particular when considering frames from the same video to be correlated) [47]. The model output was post-processed by selecting the largest connected component for the glottis structure and restricting the vocal cords and aspiration segmentation to this ROI.

2.3 Evaluation

Comparison of the overlap of surface area (pixels) between human assignment during annotation and AI based segmentation of the vocal cords, the glottis ROI, and aspirated bolus was used to calculate the model's segmentation performance (Dice score). Given two binary masks $X = (x_{ij})$ and $Y = (y_{ij})$, the Dice score is defined as

$$Dice(X,Y) = \frac{2\sum_{i,j} x_{ij} y_{ij}}{\sum_{i,j} x_{ij} + \sum_{i,j} y_{ij}}$$

To assess the model's capability of correctly identifying frames where the glottis is not visible, the number of pixels falsely segmented as glottis ROI were calculated on all frames labeled as not containing the glottis. A confusion matrix was calculated to make the aspiration detection capabilities of the AI assessable. The detection performance is represented by its precision (positive predictive value: how many findings are actually aspirations)and recall (sensitivity: how many of the aspirations were found) metrics [58]. Based on them, the F1 score as the the harmonic mean of recall and precision was also calculated, to rate the AI performance between 0 and 1 in a single metric [59]. Given the number of true positive (TP), false positive (FP) and false negative (FN) predictions of aspirations, the metrics are defined as:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2}{Recall^{-1} + Precision^{-1}} = \frac{2TP}{2TP + FP + FN}$$

Spearman-Rho correlations were calculated for the overlap of the AI-segmented bolus with the size of the reference segmentation in order to investigate if larger entities / more pixels can be detected more easily. The calculation of metrics and correlation analysis was performed using the Python packages scikit-learn v0.24.2 [60] and SciPy v1.5.2 [61].

2.4 Timeline for interpretation of the model outcome

The XAI concept of our approach is based on a human-centered design of the model output. In order to enable full perceptive interpretability of the model outcome by a post-hoc analysis that relies on the expert knowledge of the diagnostician, we implemented the concept of identifying meaningful or key frames in sequences [8,9]. This became possible, since we have automated the video analysis and can apply the CNN to an entire video to generate a new video in which all AI-based segmentations and detections of aspirations

 doi:10.20944/preprints202212.0051.v1

are drawn into all frames of the video sequence. Furthermore, on a separate screen window a timeline gets generated that plots a curve displaying the number of pixels for the segmentation tasks and the detected aspiration candidates. This provides the examiner with an overview across the complete video capture at one glance resulting in a human-in-the-loop process. In other words, one could then look by scrolling at the at time points where aspiration is detected on several consecutive frames to decide about the correctness of the AI detection. This human computer interaction guaranties for the demands of the EU GDPR [44] and provides transparency to the user.

3. Results

In order to provide transparency for the development and explainability for the system process, the general distribution of the videos in the different datasets will be shown, and then the number of annotated frames is presented. After that, first general results on the AI performance are outlined before going into more detail on the segmentation and detection results. Finally, the XAI approach will be demonstrated.

3.1. Video distribution across data sets and annotated frames

92 videos were included and were split into disjunct sets for training (77.2%; 71), validation during training (6.5%, 6), and final testing (16.3%, 15). Amongst the 50 videos with aspirations, the distribution of bolus types was slurry (21), saliva (18), and liquids (11). During preparation, 1330 frames were segmented and 2895 frames were labeled as not showing the glottis. Table 1 shows their distribution across the three data subsets for the development and evaluation of the AI.

Table 1. Distribution of annotated frames across the different data sets.

AI data subset	Segmented frames	Frames with aspiration	Frames not showing glottis
training	1029	424	2220
validation	103	17	186
test	199	63	489

3.2. AI training

Figure 2 shows the learning process of the AI. The curves of the learning progress for glottis and vocal cord segmentation (Jaccard scores 1 and 2) rise steeply from the beginning, unlike the aspiration detection task (Jaccard score 4), where the AI does not learn until about 19,000 iterations. After the rise of Jaccard score 4 (aspiration detection) only the training loss curve but not the validation loss curve progresses to decline. The best model performance based on the mean Jaccard score is reached at 32,000 iterations, building the basis for the test run.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 December 2022 doi:10.20944/preprints202212.0051.v1

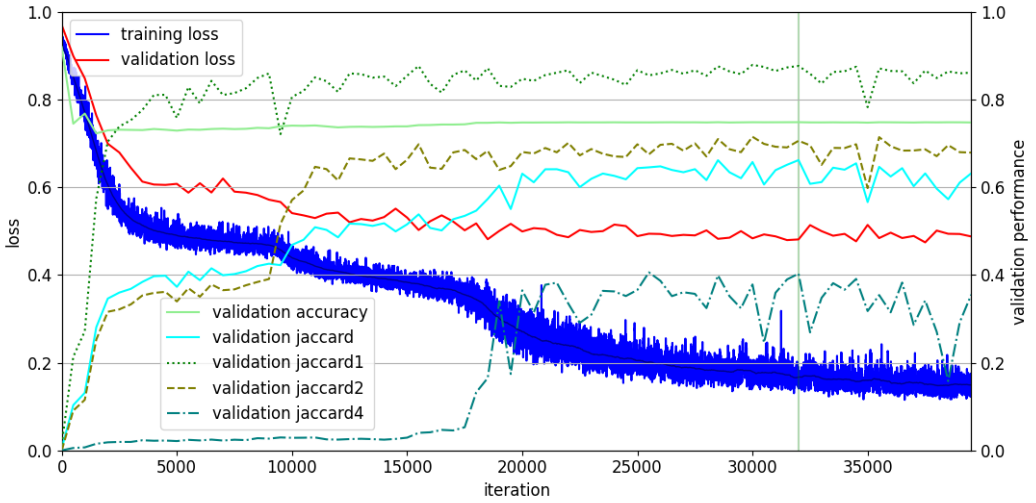
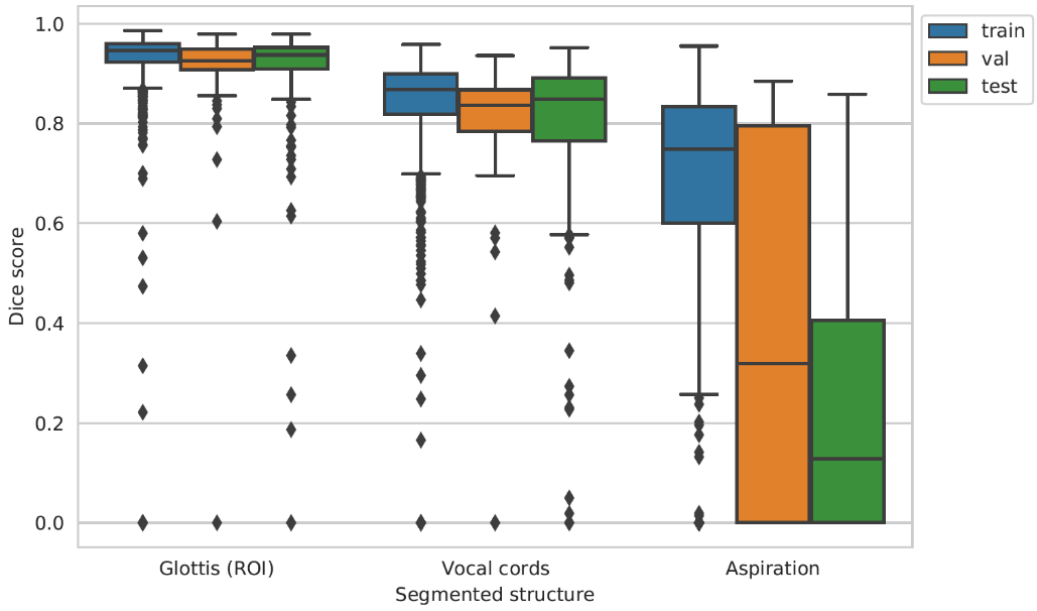


Figure 2. Loss curves of training (blue) and validation (red) as well as validation Jaccard scores (turquoise = mean of all; 1/dotted = segmentation of glottis; 2/dashed = segmentation of vocal cords; 4/dashed and dotted = detection of aspiration) show overlaps with the references. The vertical line shows the moment of the optimally working model.

3.3. AI performance

Boxplots for the Dice Scores in Figure 3 show high values for the segmentation of the glottis and little lower values for the vocal cords across all data subsets, with median values of .94 and .85 respectively on the test set.

The plot clearly shows some overfitting despite early stopping, as the performance is overall higher on the training set than validation and test sets, in particular for aspiration segmentation. During training, the segmentation of aspiration achieves a median Dice score of .75 but drops to a median of .32 during validation and .13 during testing accompanied by a large increase of the inter quartile range. On frames that were labeled as not containing the glottis, the model detects false positive pixels in 5.6 %, 4.3 %, and 9.4 % of frames for training, validation, and test sets.



 doi:10.20944/preprints202212.0051.v1

Figure 3. Boxplots of the Dice Scores for all tasks (segmentation of the glottis, vocal cords, and detection of aspiration) for training, validation, and testing.

For aspiration detection a confusion matrix was also calculated (Figure 4) to determine the performance of the AI. Whenever aspiration was segmented and the Dice overlap with the reference was greater than 0, the frame was counted as true positive detection. For the training data, a very good result can be obtained for true negative and a good result for true positive outcomes. The achieved value for false positives is in the lower range, the range for false negatives as well.

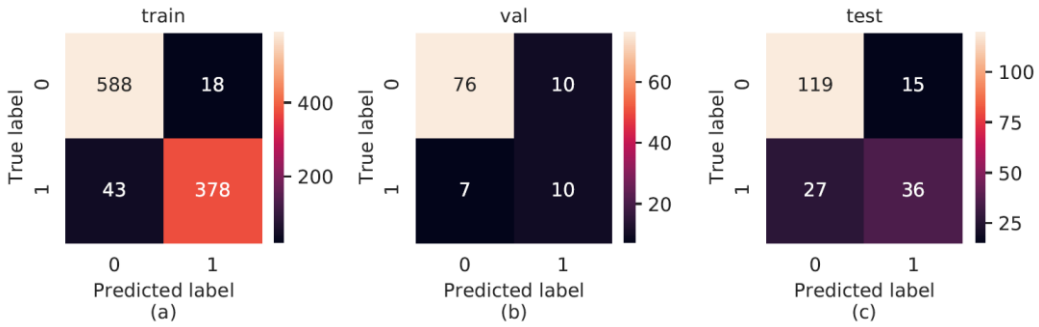


Figure 4. Confusion matrices with predicted and true labels (0 = negative / 1 = positive) for aspiration detection for training (a), validation (b), and testing (c) with heat-map scales for result interpretation (right in each case).

Additionally, we calculated the resulting values for precision, recall and F1 score (Table 2). On the training set, the precision is very high (.955), meaning that most detections were indeed aspirations. This drops to .5 during validation and .706 during testing. Amongst all annotated aspirations, the AI has detected 90% during training but only 59% during validation and 57% during testing. The harmonic mean of both metrics (F1-Score) also drops from .925 to .541 and .632 for training versus validation and test respectively. As for the segmentation, this also shows the overfitting on the training set.

Table 2. Metrics for aspiration detection for all data sets.

Metrics	Training	Validation	Test
Precision	.955	.500	.706
Recall	.898	.588	.571
F1-Score	.925	.541	.632

Selected video frames in Figure 5 demonstrate this heterogeneity of results. When the glottis is well visible, the segmentation of glottis and vocal cords is very precise (a-c), but may be less robust when the glottis is only partially visible or near the image edge (d). Aspirations can be detected in the correct location (d-e), but can also be overlooked (f) or falsely detected for example due to light reflections (g). In addition, a ROI segmentation can appear even though the relevant anatomical structure is not visible/present within the respective frame (h, piriform recess). Despite a correct detection of the aspiration itself as in (d-e), the segmentation itself may be imprecise leading to low Dice score values.

 doi:10.20944/preprints202212.0051.v1

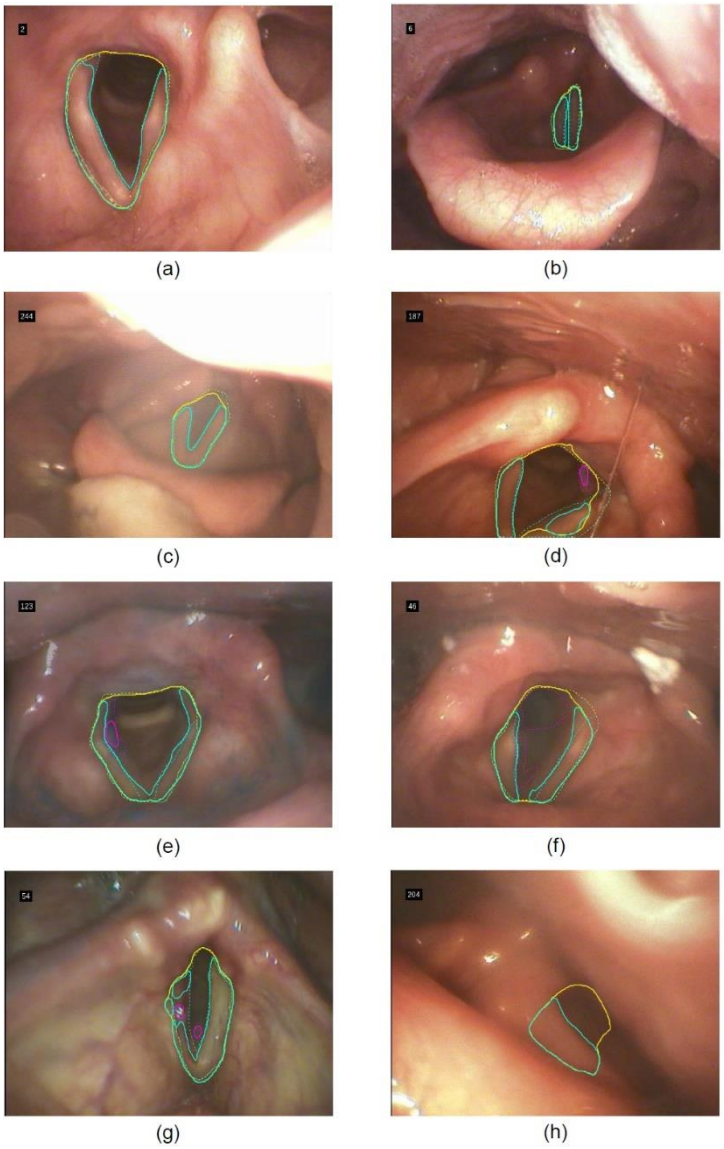


Figure 5. Examples of segmentation results in the test set across different videos: (a-c) high overlap between references (dotted) and AI-based (drawn through) segmentation in different states (open, closed) and light conditions, (d) segmentation errors of partially visible glottis close to the image edge, (e) correct detection of aspiration, (f) missed detection of aspiration, (g) false positive detection of aspiration, (h) false positive segmentation of glottis and vocal cords on frame without visible glottis.

In order to investigate if the bolus segmentation might depend on the amount of aspirated bolus (e.g., the more aspirate, the easier to be detected), as visualized in Figure 6, we calculated Spearman's Rho correlations for the overlap of reference and AI segmentation (Dice score) with the size of aspiration (number of pixels segmented in reference). The correlation decreases from training (r=.62, p=0) to validation (r=.43, p=.08) and testing (r=.37, p=.003).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 December 2022 doi:10.20944/preprints202212.0051.v1

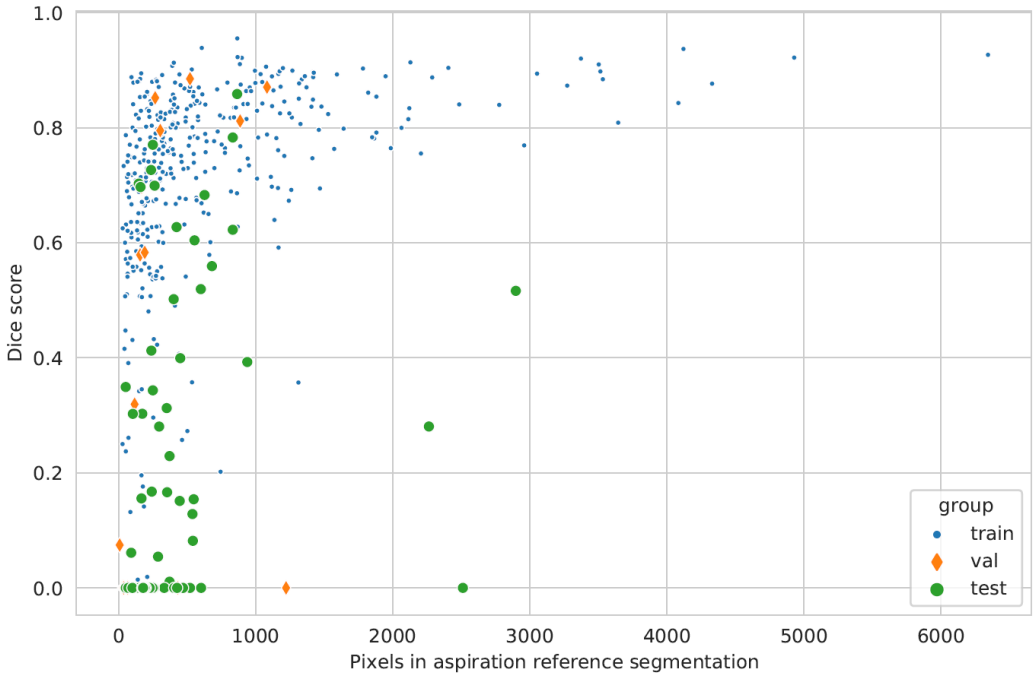


Figure 6. Scatterplot of the overlap between the number of pixels for the aspiration in the reference annotation (x-axis) and the Dice Score for the overlap of AI detection (y-axis) and reference during training, validation, and testing.

3.4. Interpretability by identifying meaningful frames

As a means for post-hoc interpretation of the model outcome by the examiner, we implemented a concept of identifying meaningful frames in sequences. Therefore, an automated video analysis applies the CNN to an entire video to create a new video in which all AI-based segmentations and detections of aspirations are drawn into all frames of the video sequence (Figure 7), serving as a first visual aid for key frames. The unmarked video can be seen in parallel. As a second visual aid, on a separate screen window a timeline gets generated that plots a curve displaying the number of pixels for the segmentation tasks and the detected aspiration candidates. It also features a further zoom window. Hence, the examiner is provided with an overview across the complete video capture at one glance and can scroll to meaningful frames for diagnostic purposes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 December 2022 doi:10.20944/preprints202212.0051.v1

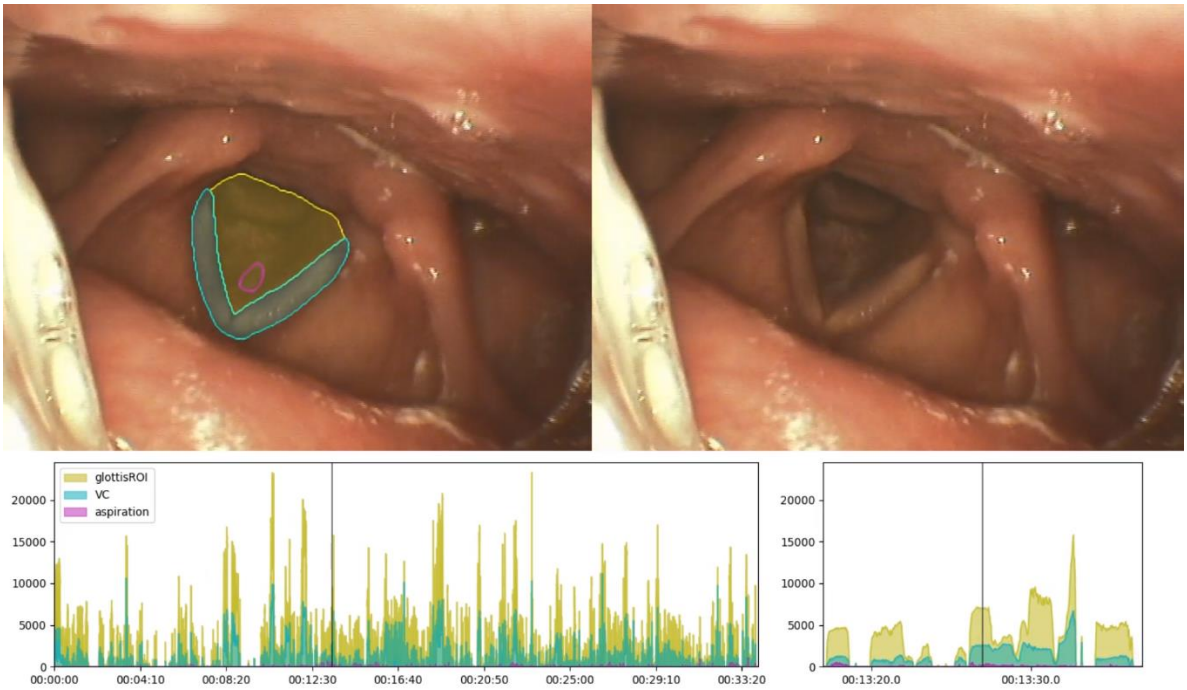


Figure 7. Visual aids to find meaningful frames for interpretation of model output. Screenshot of AI-based segmentation and detection of aspiration results (upper left) respective normal view (upper right). Timeline and timeline zoom with a curve displaying the number of pixels for the segmentation tasks and the detected aspiration (below). Vertical line indicates the point in time.

In the given example (Figure 7), slurry parts of yoghurt and saliva become aspirated and reside above the first cartilage of the trachea (membrana cricothyroidea). The AI detects the part next to the vocal cords as an aspiration.

4. Discussion

The discussion will firstly focus on the XAI aspects of our approach; afterwards the model accuracy will be mooted.

Due to the human-in-the-loop process and the HCI, our XAI can be considered a "hybrid" concept that combines data- and knowledge-driven as well as white- and black-box modeling approaches. Our attempt provides full post-hoc human-based perceptive interpretability of the model outcome by the examiner. Hence, our concept of identifying meaningful frames by adding visual aids adds a further example to the notion of key frame identification as XAI approaches [8,9]. The user can decide, if the AI explanation is suitable, and based on that the further course for the patient can be planned (e.g., oral feeding is possible). Therefore, the final explanation provided by the system is effective and acceptable. This goes beyond most existing approaches for this task (except the VFSS approach [25]), since they only provide predictions or classifications without providing proper interpretable information for the diagnostician [28,35,37,42,43,62,63]. Since this lack of transparency conflicts with EU GDPR, as it prohibits decision solely based on automated processing [44,64], a subsequent FEES or VFSS would become necessary anyway before critical decisions like abstinence from food, insertion of a nasogastric tube, or even re-intubation and tracheotomy could be made. Furthermore, regarding the interpretability, our concept does not only enable interpreting the model output for diagnostic purposes, this concept of meaningful frames also facilitates the ongoing quality and performance assessment of the model compared to a patient-level black-box prediction, helping to further develop our decision support system. Additionally, since in current FEES prac-

 doi:10.20944/preprints202212.0051.v1

tice a retrospective video analysis may already be preferable [18], but is very time-consuming, our interpretable model output is an appropriate tool to focus on relevant meaningful frames instead of viewing the whole video again.

Regarding the accuracy of the segmentation of anatomical structures, we obtained very satisfying results, similar to comparable work in the field [46-48]. Despite using only selected frames for training and not full videos, we achieved a false positive rate for glottis segmentations of only 5% on frames labeled as not containing the ROI, allowing for the processing of full videos. We expect to be able to further reduce this rate by labeling more negative example frames, which is a relatively fast annotation operation as no segmentation is required. Taking into account information from consecutive frames (e.g., using a recurrent network architecture) would likely help to further reduce spatiotemporal noise in the predictions.

Hence, we conclude that the general requirement for the second step, the AI-based detection of aspiration, was fulfilled. To be of use for the clinical workflow, both high recall (i.e., identification of true aspiration events) and high precision (i.e., not too many false positives) are desirable. In our preliminary study, we achieved satisfying precision during training (.955) and testing (.706), but slightly lower recall during training (.925) and testing (.571), meaning that a large amount still is overlooked. A trade-off between precision and recall is typical for detection algorithms, therefore we might increase recall at the cost of lower precision. Newly emerging false positive detections might be eliminated in a post-processing step. The segmentation accuracy of the detected aspiration was satisfactory during training (Dice score .75) but still to be improved during validation (.32) and testing (.13), and was accompanied by a large increase in the inter-quartile range. Overall, the decline of detection and segmentation performance from training to validation and testing is unsatisfying. In a qualitative analysis, we looked at samples of mispredictions to evaluate if the type of bolus (slurry, saliva, liquid) played a role, especially since we had no equal distribution for them in the training data, but we were not able to identify such a contributing factor. When considering the size of the aspiration as a potential explanation for its detectability within a frame with known aspiration (i.e., a segmentation), we saw a strong correlation of the true bolus size with the Dice score in the training data, but not as high during validation and testing. While the Dice score itself is known to correlate with the area to contour ratio of a 2D object, this still indicates that other factors next to bolus size may impact the segmentation accuracy during testing, for example changed lighting conditions. This limited performance can already be seen in the loss and validation plot of the training process (Figure 2), where the AI shows signs of overfitting to training data.

Hence, the currently trained AI lacks sufficient generalization for aspiration detection but not for segmentation of vocal cords and glottis ROI. Regarding the aspiration detection task, the current model performance might be comparable to an untrained human examiner [21]. Hence, at the present state our model does not lead to better results than comparable non-endoscopic / non-radiologic approaches [28,33,35,62,63,65]; but in clear contrast to them, our model outcome, also the false positives and negatives, is fully interpretable and can therefore be corrected by an experienced examiner. This becomes particularly easy since the examiner can perform the correct assignment by jumping at the respective point in the timeline of the video sequence.

Since explainability forms a crucial aspect of XAI, but has to come along with a profound model performance, as only this concurrence will lead to the acceptance of our developed system, we are at present in a re-evaluation process to understand the limitations of our model for aspiration detection. As explanations for this particular unsatisfying model performance at its present state, we have already identified various limiting aspects that can be addressed specifically. First, with 4225 annotated (thereof 1330 segmented) frames we have only achieved a basic sample for the training. Furthermore, we did not achieve a homogeneous distribution of annotated frames regarding the different subtypes of aspiration (i.e., slurry, saliva, liquids); this was only given for the samples with and

 doi:10.20944/preprints202212.0051.v1

without aspiration (50 vs. 42). Therefore, regarding the training data there are far more frames in which the ROI appears than in which aspirations occur. We did apply a sampling strategy to account for part of the imbalance; however, we did not fully optimize the ratio of frames with and without aspiration or not showing the ROI at all in a hyperparameter tuning step. In the future, we will therefore include more patient videos, annotate significantly more frames, especially more frames with aspirations, and in addition apply more data augmentation techniques, to strengthen the robustness. Moreover, we have currently processed the videos in a pure 2D approach, analyzing the video frame by frame. The 2D approach for training and prediction was chosen based on our sparsely labeled training set, in which only a few frames per video have been manually annotated and can be directly used for supervised training. To further strengthen the aspiration detection, we will consider a 2D+T approach, for example using recurrent neural networks, which take a temporal sequence of frames into account. To achieve this, we need to explore strategies for combining labeled and unlabeled frames into the training. Additionally, we want to implement an online augmented reality approach to highlight moments of potential aspiration as detected by the AI in a separate small window whilst the endoscopic procedure can continue. This would enable real-time verification, possibly with an adjustment of the FEES procedure (e.g., retesting a certain type of bolus).

As a further future goal as well as a general idea for other research groups that development of XAI systems in the field of dysphagia diagnostics, we want to propose to implement a feedback system, especially for corrections and negative feedback information that can be provided by the domain experts. In such an active learning scenario, the algorithm itself could suggest frames in which it is not sure whether aspiration was detected or not, and ask for feedback. This would enable continuous training or planned re-trainings in certain intervals to enhance the model performance, while at the same time reducing annotation effort when compared with an undirected approach. Moreover, when gaining research partners that are in possession of a reasonable amount of narrow band imaging videos showing aspirations [23,24] this could also be used to further facilitate the AI based detection. Additionally, we could also implement other XAI concepts like a combination of frame-wise classification for aspiration detection and XAI methods like GradCAM or Saliency maps [3]. We could compare the output of these methods to the proposed segmentation, to evaluate their usefulness as a visual aid. Taken together and despite the discussed limitations of the current model's state, our novel concept of AI-based detection of aspiration during video-endoscopy with visual aids in meaningful frames makes it possible to interpret the model outcome. With the proposed XAI approach, the AI segmentation and the pixel-wise classification as an aspiration can be verified thereby providing proper interpretable information for the diagnostician to understand why subjects were classified, and beyond that, it enables the identification of misclassifications. This substantially reduces the black box character of the machine-learning model. Therefore, our current attempt is an important step for making the identification of meaningful frames an XAI approach that will become more applicable in clinical contexts.

5. Conclusions

For the first time we have introduced an XAI that has been trained to detect aspiration in endoscopic swallowing videos. Albeit detection performance has to be optimized significantly in future studies, our architecture results in a final model that explains its assessment by locating specific video frames with relevant aspiration events and by highlighting the suspected bolus in situ as a meaningful sequence. Hence, in contrast to existing machine learning tools for aspiration detection, the AI decision in our framework is verifiable, interpretable, and thus accountable for clinical users. During the next development steps, the interaction with the dysphagia experts will continuously improve the outcome.

After implementation of this tool in a FEES software, it will aid endoscopists to improve accuracy (thereby potentially saving lives), shorten the duration of the administration, and altogether safe costs as positive contributions for healthcare.

Author Contributions: Conceptualization, J.K., H.M., and A.G.; methodology, A.G., H.M., and J.K.; software, A.G. and H.M.; validation, A.G., M.Z., H.M., and J.K.; formal analysis, A.G., M.Z., H.M., and J.K.; investigation, M.Z., J.K.; resources, U.B. and P.D.; data curation, M.Z., A.G., H.M., and J.K.; writing—original draft preparation, J.K., A.G., H.M.; writing—review and editing, J.K., A.G., H.M., and M.Z.; visualization, A.G., H.M., M.Z.; supervision, U.B. and P.D.; project administration, J.K., U.B., and P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, was approved by the responsible ethics committee of the State Chamber of Physicians of Rhineland-Palatinate (Nr.: 2021-16141-retrospektiv), and is registered with WHO (INT: DRKS00026822).

Informed Consent Statement: The approval of the ethics committee covers the retrospective analysis and data sharing (restricted to MEVIS) of video sequence from endoscopic evaluations of swallowing that were gathered as a part of the daily clinical routine. Chosen video sequences show no faces, only internal anatomical structures, the audio track was removed. In this case, no patient consent has to be obtained and cannot be obtained since we cannot trace back the patient the video was taken from. Publication of these anonymized data is accepted, especially when part of aggregated data.

Data Availability Statement: The data are not publicly available due to restrictions on the use of clinical patient data. Furthermore, a public accessibility is not covered by the given ethics vote. In this, data sharing is limited to the research partner (MEVIS).

Acknowledgments: Foremost, the authors would like to thank all of the anonymous patients for their video donations. All clinical colleagues are acknowledged for conducting the endoscopic examinations over the last decade.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Muller, H.; Mayrhofer, M.; Van Veen, E.; Holzinger, A. The Ten Commandments of Ethical Medical AI" in Computer. *Computer* **2021**, *54*, 119-123, doi:10.1109/MC.2021.3074263.
2. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138-52160, doi:10.1109/ACCESS.2018.2870052.
3. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *32*, 4793-4813, doi:10.1109/tnnls.2020.3027314.
4. Stepin, I.; Alonso, J.M.; Catala, A.; Pereira-Fariña, M. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* **2021**, *9*, 11974-12001, doi:10.1109/ACCESS.2021.3051315.
5. Li, X.H.; Cao, C.C.; Shi, Y.; Bai, W.; Gao, H.; Qiu, L.; Wang, C.; Gao, Y.; Zhang, S.; Xue, X.; et al. A Survey of Data-Driven and Knowledge-Aware eXplainable AI. *IEEE Transactions on Knowledge and Data Engineering* **2020**, *34*, 29-49, doi:10.1109/TKDE.2020.2983930.
6. Dağlarli, E. Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In *Advances and Applications in Deep Learning*, Aceves-Fernandez, M.A., Ed.; IntechOpen: 2020.
7. Nazar, M.; Alam, M.M.; Yafi, E.; Su'ud, M.M. A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE Access* **2021**, *9*, 153316-153348, doi:10.1109/ACCESS.2021.3127881.

8. Ujwalla, G.; Kamal, H.; Yogesh, G. Deep Learning Approach to Key Frame Detection in Human Action Videos. In *Recent Trends in Computational Intelligence*, Ali, S., Tilendra Shishir, S., Eds.; IntechOpen: Rijeka, 2020; p. Ch. 7.
9. Yan, X.; Gilani, S.Z.; Feng, M.; Zhang, L.; Qin, H.; Mian, A. Self-Supervised Learning to Detect Key Frames in Videos. *Sensors* **2020**, *20*, 6941.
10. Bhattacharyya, N. The prevalence of dysphagia among adults in the United States. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery* **2014**, *151*, 765-769, doi:10.1177/0194599814549156.
11. Attrill, S.; White, S.; Murray, J.; Hammond, S.; Doeltgen, S. Impact of oropharyngeal dysphagia on healthcare cost and length of stay in hospital: a systematic review. *BMC Health Serv Res* **2018**, *18*, 594-594, doi:10.1186/s12913-018-3376-3.
12. Doggett, D.L.; Tappe, K.A.; Mitchell, M.D.; Chapell, R.; Coates, V.; Turkelson, C.M. Prevention of pneumonia in elderly stroke patients by systematic diagnosis and treatment of dysphagia: an evidence-based comprehensive analysis of the literature. *Dysphagia* **2001**, *16*, 279-295.
13. Rugiu, M.G. Role of videofluoroscopy in evaluation of neurologic dysphagia. *Acta Otorhinolaryngol Ital* **2007**, *27*, 306-316.
14. Aviv, J.E.; Sataloff, R.T.; Cohen, M.; Spitzer, J.; Ma, G.; Bhayani, R.; Close, L.G. Cost-effectiveness of two types of dysphagia care in head and neck cancer: a preliminary report. *Ear, nose, & throat journal* **2001**, *80*, 553-556, 558.
15. Dziewas, R.; Glahn, J.; Helfer, C.; Ickenstein, G.; Keller, J.; Lapa, S.; Ledl, C.; Lindner-Pfleghar, B.; Nabavi, D.; Prosiegel, M.; et al. FEES für neurogene Dysphagien. *Der Nervenarzt* **2014**, *85*, 1006-1015, doi:10.1007/s00115-014-4114-7.
16. Lüttje, D.; Meisel, M.; Meyer, A.-K.; Wittrich, A. Änderungsvorschlag für den OPS 2010. **2010**.
17. Bohlender, J. Fiberendoskopische Evaluation des Schluckens – FEES. *Sprache Stimme Gehör* **2017**, *41*, 216-216, doi:10.1055/s-0043-120430.
18. Hey, C.; Pluschinski, P.; Pajunk, R.; Almahameed, A.; Girth, L.; Sader, R.; Stöver, T.; Zaretsky, Y. Penetration–Aspiration: Is Their Detection in FEES® Reliable Without Video Recording? *Dysphagia* **2015**, *30*, 418-422, doi:10.1007/s00455-015-9616-3.
19. Rosenbek, J.C.; Robbins, J.A.; Roecker, E.B.; Coyle, J.L.; Wood, J.L. A penetration-aspiration scale. *Dysphagia* **1996**, *11*, 93-98.
20. Colodny, N. Interjudge and Intrajudge Reliabilities in Fiberoptic Endoscopic Evaluation of Swallowing (Fees®) Using the Penetration–Aspiration Scale: A Replication Study. *Dysphagia* **2002**, *17*, 308-315, doi:10.1007/s00455-002-0073-4.
21. Curtis, J.A.; Borders, J.C.; Perry, S.E.; Dakin, A.E.; Seikaly, Z.N.; Troche, M.S. Visual Analysis of Swallowing Efficiency and Safety (VASES): A Standardized Approach to Rating Pharyngeal Residue, Penetration, and Aspiration During FEES. *Dysphagia* **2022**, *37*, 417-435, doi:10.1007/s00455-021-10293-5.
22. Butler, S.G.; Markley, L.; Sanders, B.; Stuart, A. Reliability of the Penetration Aspiration Scale With Flexible Endoscopic Evaluation of Swallowing. *Annals of Otology, Rhinology & Laryngology* **2015**, *124*, 480-483, doi:10.1177/0003489414566267.
23. Nienstedt, J.C.; Müller, F.; Nießen, A.; Fleischer, S.; Koseki, J.C.; Flügel, T.; Pflug, C. Narrow Band Imaging Enhances the Detection Rate of Penetration and Aspiration in FEES. *Dysphagia* **2017**, *32*, 443-448, doi:10.1007/s00455-017-9784-4.
24. Stanley, C.; Paddle, P.; Griffiths, S.; Safdar, A.; Phyland, D. Detecting Aspiration During FEES with Narrow Band Imaging in a Clinical Setting. *Dysphagia* **2022**, *37*, 591-600, doi:10.1007/s00455-021-10309-0.
25. Kim, J.K.; Choo, Y.J.; Choi, G.S.; Shin, H.; Chang, M.C.; Park, D. Deep Learning Analysis to Automatically Detect the Presence of Penetration or Aspiration in Videofluoroscopic Swallowing Study. *J Korean Med Sci* **2022**, *37*, e42, doi:10.3346/jkms.2022.37.e42.
26. Donohue, C.; Mao, S.; Sejdić, E.; Coyle, J.L. Tracking Hyoid Bone Displacement During Swallowing Without Videofluoroscopy Using Machine Learning of Vibratory Signals. *Dysphagia* **2021**, *36*, 259-269, doi:10.1007/s00455-020-10124-z.
27. Kuramoto, N.; Ichimura, K.; Jayatilake, D.; Shimokakimoto, T.; Hidaka, K.; Suzuki, K. Deep Learning-Based Swallowing Monitor for Realtime Detection of Swallow Duration. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 20-24 July 2020, 2020; pp. 4365-4368.

28. Lee, J.; Blain, S.; Casas, M.; Kenny, D.; Berall, G.; Chau, T. A radial basis classifier for the automatic detection of aspiration in children with dysphagia. *Journal of neuroengineering and rehabilitation* **2006**, *3*, 14, doi:10.1186/1743-0003-3-14.
29. Mao, S.; Zhang, Z.; Khalifa, Y.; Donohue, C.; Coyle, J.L.; Sejdic, E. Neck sensor-supported hyoid bone movement tracking during swallowing. *Royal Society Open Science* **2019**, *6*, 181912, doi:10.1098/rsos.181982.
30. Feng, S.; Shea, Q.-T.-K.; Ng, K.-Y.; Tang, C.-N.; Kwong, E.; Zheng, Y. Automatic Hyoid Bone Tracking in Real-Time Ultrasound Swallowing Videos Using Deep Learning Based and Correlation Filter Based Trackers. *Sensors* **2021**, *21*, 3712.
31. Lee, J.C.; Seo, H.G.; Lee, W.H.; Kim, H.C.; Han, T.R.; Oh, B.M. Computer-assisted detection of swallowing difficulty. *Comput Methods Programs Biomed* **2016**, *134*, 79-88, doi:10.1016/j.cmpb.2016.07.010.
32. Zhang, Z.; Coyle, J.L.; Sejdić, E. Automatic hyoid bone detection in fluoroscopic images using deep learning. *Scientific Reports* **2018**, *8*, 12310, doi:10.1038/s41598-018-30182-6.
33. Frakking, T.T.; Chang, A.B.; Carty, C.; Newing, J.; Weir, K.A.; Schwerin, B.; So, S. Using an Automated Speech Recognition Approach to Differentiate Between Normal and Aspirating Swallowing Sounds Recorded from Digital Cervical Auscultation in Children. *Dysphagia* **2022**, doi:10.1007/s00455-022-10410-y.
34. Khalifa, Y.; Coyle, J.L.; Sejdić, E. Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings. *Scientific reports* **2020**, *10*, 8704-8704, doi:10.1038/s41598-020-65492-1.
35. Steele, C.M.; Mukherjee, R.; Kortelainen, J.M.; Polonen, H.; Jedwab, M.; Brady, S.L.; Theimer, K.B.; Langmore, S.; Riquelme, L.F.; Swigert, N.B.; et al. Development of a Non-invasive Device for Swallow Screening in Patients at Risk of Oropharyngeal Dysphagia: Results from a Prospective Exploratory Study. *Dysphagia* **2019**, doi:10.1007/s00455-018-09974-5.
36. Hadley, A.J.; Krival, K.R.; Ridgel, A.L.; Hahn, E.C.; Tyler, D.J. Neural Network Pattern Recognition of Lingual–Palatal Pressure for Automated Detection of Swallow. *Dysphagia* **2015**, *30*, 176-187, doi:10.1007/s00455-014-9593-y.
37. Jayatilake, D.; Ueno, T.; Teramoto, Y.; Nakai, K.; Hidaka, K.; Ayuzawa, S.; Eguchi, K.; Matsumura, A.; Suzuki, K. Smartphone-Based Real-time Assessment of Swallowing Ability From the Swallowing Sound. *IEEE J Transl Eng Health Med* **2015**, *3*, 1-10, doi:10.1109/JTEHM.2015.2500562.
38. Jones, C.A.; Hoffman, M.R.; Lin, L.; Abdelhalim, S.; Jiang, J.J.; McCulloch, T.M. Identification of swallowing disorders in early and mid-stage Parkinson's disease using pattern recognition of pharyngeal high-resolution manometry data. *Neurogastroenterology & Motility* **2018**, *30*, e13236, doi:10.1111/nmo.13236.
39. Kritas, S.; Dejaeger, E.; Tack, J.; Omari, T.; Rommel, N. Objective prediction of pharyngeal swallow dysfunction in dysphagia through artificial neural network modeling. *Neurogastroenterology and motility : the official journal of the European Gastrointestinal Motility Society* **2016**, *28*, 336–344, doi:10.1111/nmo.12730.
40. Lee, J.; Steele, C.M.; Chau, T. Swallow segmentation with artificial neural networks and multi-sensor fusion. *Medical Engineering & Physics* **2009**, *31*, 1049-1055, doi:10.1016/j.medengphy.2009.07.001.
41. Lee, J.T.; Park, E.; Hwang, J.-M.; Jung, T.-D.; Park, D. Machine learning analysis to automatically measure response time of pharyngeal swallowing reflex in videofluoroscopic swallowing study. *Scientific Reports* **2020**, *10*, 14735, doi:10.1038/s41598-020-71713-4.
42. Sakai, K.; Gilmour, S.; Hoshino, E.; Nakayama, E.; Momosaki, R.; Sakata, N.; Yoneoka, D. A Machine Learning-Based Screening Test for Sarcopenic Dysphagia Using Image Recognition. *Nutrients* **2021**, *13*, doi:10.3390/nu13114009.
43. Roldan-Vasco, S.; Orozco-Duque, A.; Suarez-Escudero, J.C.; Orozco-Arroyave, J.R. Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia. *Computer Methods and Programs in Biomedicine* **2021**, *208*, 106248, doi:10.1016/j.cmpb.2021.106248.
44. Regulation 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/e. **2016**.

45. Holzinger, A.; Biemann, C.; Constantinos; Douglas. What do we need to build explainable AI systems for the medical domain? *arXiv pre-print server* **2017**, doi:arxiv:1712.09923.
46. Fehling, M.K.; Grosch, F.; Schuster, M.E.; Schick, B.; Lohscheller, J. Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. *PLoS One* **2020**, *15*, e0227791, doi:10.1371/journal.pone.0227791.
47. Laves, M.-H.; Bicker, J.; Kahrs, L.A.; Ortmaier, T. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International Journal of Computer Assisted Radiology and Surgery* **2019**, *14*, 483-492, doi:10.1007/s11548-018-01910-0.
48. Matava, C.; Pankiv, E.; Raisbeck, S.; Caldeira, M.; Alam, F. A Convolutional Neural Network for Real Time Classification, Identification, and Labelling of Vocal Cord and Tracheal Using Laryngoscopy and Bronchoscopy Video. *J Med Syst* **2020**, *44*, 44, doi:10.1007/s10916-019-1481-4.
49. Meine, H.; Moltz, J.H. SATORI. AI collaboration toolkit. Available online: https://www.mevis.fraunhofer.de/en/research-and-technologies/ai-collaboration-toolkit.html (accessed on 27.09.2022).
50. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging* **2018**, *9*, 611-629, doi:10.1007/s13244-018-0639-9.
51. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for BiomedicalImage Segmentation. **2015**, doi:10.48550/arXiv.1505.04597.
52. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **2021**, *18*, 203-211, doi:10.1038/s41592-020-01008-z.
53. Srivastava, N.; Hinton, G.E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **2014**, *15*, 1929-1958.
54. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning* **2015**, *37*, 448-456.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), 7-13 Dec. 2015, 2015; pp. 1026-1034.
56. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. 2015.
57. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), 25-28 Oct. 2016, 2016; pp. 565-571.
58. Powers, D.M.W. Evaluation: From Precision, Recall And F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* **2011**, *2*, 37-63, doi:10.48550/arXiv.2010.16061.
59. Sasaki, Y. The truth of the F-measure. **2007**.
60. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825-2830, doi:10.48550/arXiv.1201.0490.
61. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **2020**, *17*, 261-272, doi:10.1038/s41592-019-0686-2.
62. Inoue, K.; Yoshioka, M.; Yagi, N.; Nagami, S.; Oku, Y. Using Machine Learning and a Combination of Respiratory Flow, Laryngeal Motion, and Swallowing Sounds to Classify Safe and Unsafe Swallowing. *IEEE Transactions on Biomedical Engineering* **2018**, *65*, 2529-2541, doi:10.1109/TBME.2018.2807487.

63. O'Brien, M.K.; Botonis, O.K.; Larkin, E.; Carpenter, J.; Martin-Harris, B.; Maronati, R.; Lee, K.; Cherney, L.R.; Hutchison, B.; Xu, S.; et al. Advanced Machine Learning Tools to Monitor Biomarkers of Dysphagia: A Wearable Sensor Proof-of-Concept Study. *Digital Biomarkers* **2021**, *5*, 167-175, doi:10.1159/000517144.

64. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv:1712.09923* **2017**.

65. Jayatilake, D.; Ueno, T.; Teramoto, Y.; Nakai, K.; Hidaka, K.; Ayuzawa, S.; Eguchi, K.; Matsumura, A.; Suzuki, K. Smartphone-Based Real-time Assessment of Swallowing Ability From the Swallowing Sound. *IEEE J Transl Eng Health Med* **2015**, *3*, 2900310-2900310, doi:10.1109/JTEHM.2015.2500562.