PLOS ONE

RESEARCH ARTICLE

Long-term performance assessment of fully automatic biomedical glottis segmentation at the point of care

René Groh[1], Stephan Dürr[2], Anne Schützenberger[2], Marion Semmler[2], Andreas M. Kist[1] *

1 Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Bavaria, Germany, 2 Division of Phoniatrics and Pediatric Audiology at the Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Erlangen, Friedrich–Alexander-Universität Erlangen–Nürnberg, Erlangen, Bavaria, Germany

* andreas.kist@fau.de

Abstract

Deep Learning has a large impact on medical image analysis and lately has been adopted for clinical use at the point of care. However, there is only a small number of reports of long-term studies that show the performance of deep neural networks (DNNs) in such an environment. In this study, we measured the long-term performance of a clinically optimized DNN for laryngeal glottis segmentation. We have collected the video footage for two years from an AI-powered laryngeal high-speed videoendoscopy imaging system and found that the footage image quality is stable across time. Next, we determined the DNN segmentation performance on lossy and lossless compressed data revealing that only 9% of recordings contain segmentation artifacts. We found that lossy and lossless compression is on par for glottis segmentation, however, lossless compression provides significantly superior image quality. Lastly, we employed continual learning strategies to continuously incorporate new data into the DNN to remove the aforementioned segmentation artifacts. With modest manual intervention, we were able to largely alleviate these segmentation artifacts by up to 81%. We believe that our suggested deep learning-enhanced laryngeal imaging platform consistently provides clinically sound results, and together with our proposed continual learning scheme will have a long-lasting impact on the future of laryngeal imaging.

1 Introduction

Laryngeal videoendoscopy is a major assessment tool to evaluate voice physiology qualitatively and quantitatively (Fig 1). Especially for quantifying voice physiology, high-speed videoendoscopy (HSV) [1, 2] is an emerging technique that is able to visualize each glottal cycle with high spatial and temporal resolution. As the vocal folds, the main source of our voice, are vibrating hundreds of times each second, high-speed recordings with at least 4,000 frames per second (fps) are needed to accurately record this motion [1, 3]. This rich laryngeal imaging data source has been assessed with various algorithms to understand voice physiology [4–6].

The glottal area, the opening between the two vocal folds, is a good proxy for assessing the oscillation behavior [1, 7]. Therefore, many works have focused on the segmentation of the glottal area (Fig 1), especially avoiding any manual intervention [7, 8]. This critical step is one of the main bottlenecks of the data analysis pipeline and has been the reason why HSV is barely applied in the clinic, because fully automatic data analysis solutions have not been available [1]. Lately, it has been shown that deep neural networks (DNNs) are highly suitable for solving the task of glottis segmentation [9–12], but were also employed in various other laryngeal imaging tasks [13–15].

DNNs are state-of-the-art in image segmentation [16–18]. These DNNs determine a given class on a pixel-by-pixel level in a given image. A common architectural design is the usage of contracting and expansion paths [18, 19]. The U-Net architecture is a major DNN in biomedical image segmentation [20] and has seen widespread adaptation. A plethora of U-Net-derived variants was proposed [21–23]. However, it has been shown that the base U-Net with appropriate hyperparameter selection is outperforming largely more specialized variants [24].

In terms of glottis segmentation, U-Net-derived DNNs were also optimized towards clinical applicability with barely sacrificing segmentation accuracy [9]. Together with the recent development of an open-source HSV system, namely OpenHSV, the latest hardware and software components were introduced [25] that features these lean optimized DNNs for potential clinical use. However, there is no record of how these DNNs perform actually on a broad spectrum of unseen data, as they have been validated on limited and selected data.

In this work, we report the performance of the AI-powered OpenHSV system together with the DNN in a two-year clinical environment. Our main contributions are summarized as follows: (1) we describe for the first time the overall image quality distribution during a two-year period of clinical use, and if lossy data compression is suitable for data storage and subsequent data analysis, (2) give an unbiased performance evaluation of previously proposed optimized DNNs for clinical use for different data origins and (3) employ a continual learning and fine-tuning strategy to allow continuous integration of novel data into the DNN. Taken together, we strongly believe that our study provides trust and shows the reliability of the OpenHSV system, positively impacting future clinical adaptation.

2 Methods

2.1 OpenHSV system

We are using the OpenHSV platform introduced in [25]. Subjects are typically examined using a rigid endoscope equipped with a high-speed camera (IDT CCmini-1540) running at 4,000
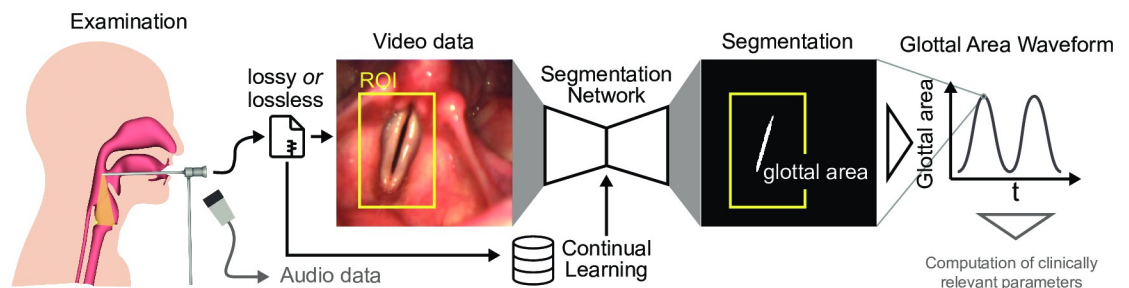
Fig 1. Data acquisition and analysis workflow. Each examination yields video and audio data, where the video data is stored either lossy or lossless compressed. In this study, we evaluate which data source and if cropping the video data to an ROI is sufficient for reliable glottis segmentation using previously proposed clinically optimized segmentation networks. The glottal area is computed for each video frame and plotted across time yielding the glottal area waveform (GAW), which is crucial for computing clinically relevant parameters. We investigate in this study how collected data can be used to allow constant fine-tuning of the segmentation network using continual learning schemes.

https://doi.org/10.1371/journal.pone.0266989.g001

fps. Illumination is granted by a high-power LED light source (Storz 300 W LED). Each recording is at least 1,000 frames long and contains synchronously acquired video and audio data. We further record patient metadata consisting of the patient's age, gender, and condition. For each video, we save two files encoded with the h.264 codec: (1) lossy compression using common settings for video and (2) lossless compressed data to recover the original recorded raw data. For lossy compression, we use the libx264 codec, the yuv420p pixel format and set the quality to 5 resulting in a varying bitrate. For lossless compression, we used the libx264rgb codec, the rgb24 pixel format, set the '-crf' flag to 0 and used the ultrafast preset.

2.2 Subjects

We assessed a total of 583 recordings acquired between November 2019 and November 2021. All recordings were acquired in line with local regulations (Ethikkommission Medizinische Fakultät, Friedrich-Alexander-University Erlangen-Nürnberg, #290_15). All data was analyzed retrospectively. We first selected only those recordings that featured a complete set of data, such that 295 recordings remained. Next, we manually investigated the data quality. We ranked each video on an ordinary scale: 0 (insufficient), 1 (okay), 2 (excellent). Videos ranked 0 were showing insufficient data quality, such as non-visible glottis or foggy videos, and were discarded. Finally, 267 recordings from 202 unique patients remained and were subjected to further analyses (Fig 2A). We report the frequency of recordings across the last two years in Fig 2B. Additionally, COVID-19 cases for Germany were provided as a reference for how data generation was affected by lock-downs. The number of recordings were higher before the first lock-down, but their fluctuation remained constant when the general clinical activity was restored. The age distribution of the patients was largely spread (Fig 2C), where the mean age is 47.2 ± 20.2 (std) years. The reported gender for the analyzed subjects is 31.2% male, 66.8% female and 2.0% had no further specified gender. The average file size for lossless and lossy recordings was 2.06 ± 1.27 GB and 8.97 ± 5.16 MB, respectively (see Fig 2D). As the lossy files
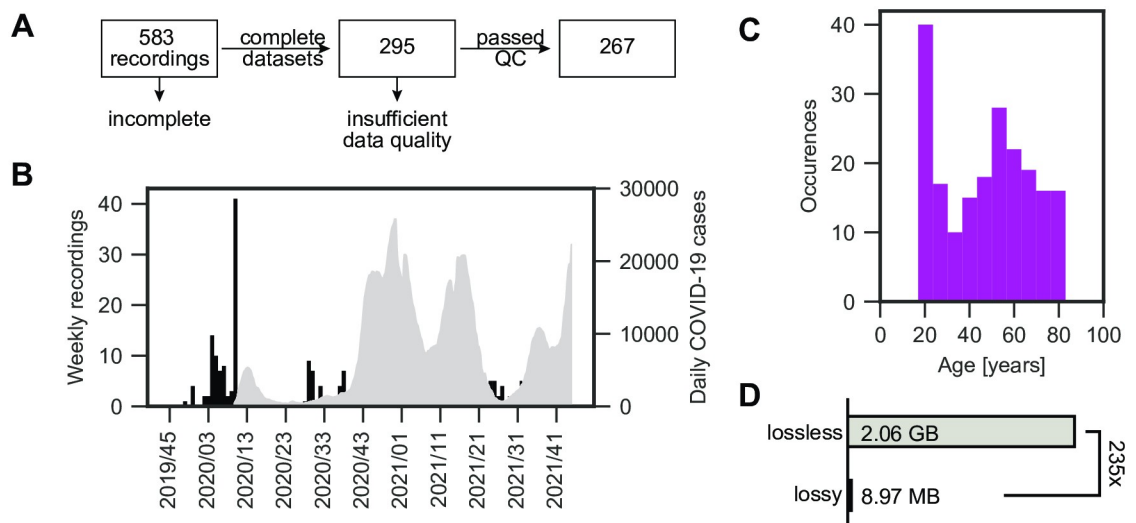
Fig 2. Acquired OpenHSV data overview. A: From initial 583 recordings, we excluded datasets that are missing any kind of data. From these, we excluded ones with insufficient data quality, such as occluded glottis. B: Frequency of recordings across the two year time frame is shown in black. Each bin corresponds to the recordings acquired in this week, denoted as year/calender week. The Germany COVID-19 cases are shown as a 7-day-average in gray. C: Age distribution across selected recordings from A. D: Comparison of file size between lossy and lossless compression.

https://doi.org/10.1371/journal.pone.0266989.g002

have an around 235-times lower file size than the lossless compressed counterparts, we investigated in this study if the lossy compression has an impact on the segmentation performance.

2.3 Image quality assessment

As we saved the data in two different compression modes, namely lossy and lossless, we evaluated if there are any compression artifacts. First, we investigated the dynamic range of the image, where 0 is no dynamic range and 1 is full dynamic range by computing the normalized absolute difference of the minimum and maximum pixel value. We determined the no-reference, completely blind image quality score "Natural Image Quality Evaluator" (NIQE) [26] to provide an unbiased score for objective image quality assessment. We further assessed full reference metrics, such as peak signal to noise ratio (PSNR) and structural similarity index metric (SSIM) [27] to compare lossy compression against the lossless stored data. PSNR can be described as follows:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\max(I)^2}{\text{MSE}}\right) \tag{1}$$

where $\max(I)$ is the maximum value corresponding to the highest pixel value with respect to images, and MSE is the mean squared error between two images I and K, defined as follows:

$$\text{MSE} = \frac{1}{mn}\sum_{i=1}^{m-1}\sum_{j=0}^{n-1}\left[I(i,j) - K(i,j)\right]^2 \tag{2}$$

SSIM is defined as follows:

$$\text{SSIM}(I,K) = \left[l(I,K)\right]^{\alpha} \cdot \left[c(I,K)\right]^{\beta} \cdot \left[s(I,K)\right]^{\gamma} \tag{3}$$

where $l(I,K)$ is a luminance comparison function, $c(I,K)$ is a contrast comparison function and $s(I,K)$ is a structure comparison function between image I and image K. $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are weighting factors of each of the metrics. The respective details are described elsewhere [27].

We use respective implementations of the metrics in scikit-video (NIQE) and scikit-image (PSNR, SSIM).

2.4 Deep neural network

The OpenHSV system is shipped with a deep neural network (DNN) based on the U-Net architecture [20], that was optimized for clinical use as described previously [9]. The used DNN is openly available on the OpenHSV Github account at https://github.com/anki-xyz/openhsv. The training process is described in [25]. Briefly, the DNN is setup in TensorFlow 2.2 and the high-level Keras package. The DNN was pretrained in a supervised fashion on the full training dataset (55,750 images) of the open benchmark for automatic glottis segmentation (BAGLS, [10]). The pretrained network has never been exposed to OpenHSV data during training.

2.4.1 Region of interest. A rectangular region of interest (ROI) was drawn manually for each recording. We saved the ROI coordinates for further use in JSON format. Each ROI was adjusted as such that the width and height is divisible by 32 to ensure proper DNN propagation.

2.4.2 Segmentation. Lossy or lossless compressed endoscopic frames were first converted to grayscale by extracting the luminance channel using standard procedures, as it has been shown that color information is not essential for glottis segmentation [10]. In some

experiments, only an ROI around the vocal folds was used for inference. The input image intensity was normalized between -1 and 1. The segmentation mask gained from the DNN provides values in the range of 0 (background) and 1 (glottis) by a sigmoid function in the output layer. For further use and due to memory limitations, we multiplied the predicted segmentation masks by 255 in order to save the data in uint8 data format. The glottal area waveform (GAW) is computed by summing the segmented pixels in each frame for every timepoint of a given recording (see Fig 1).

2.4.3 Continual training. We performed retrospective continual training on the original OpenHSV segmentation DNN. We preprocessed new images as described above and used two continual learning strategies: We either selected a fixed time period for data collection (7, 14, 30 days) or a fixed video quantity (every 10, 20 or 40 videos, see also Fig 5A) starting from November 2019. At each continual learning point, we trained the model for ten epochs using the previously predicted segmentation masks as ground truth for the training process. We chose a low learning rate of 10^{-6} combined with the Adam optimizer to fine-tune the model. After each continual training step, we evaluated the occurrence of artifacts by visual assessment (number of artifacts, shown in Fig 5B and S1 File) and computed the achieved Intersection over Union (IoU, [28]) score (Fig 5C and 5D) among others. The IoU score, also referred to mask IoU score in terms of image segmentation, is a measure how well the prediction segmentation mask overlaps with the ground-truth segmentation mask and ranges between 0 (no overlap) and 1 (perfect overlap). It can be described as follows:

$$IoU = \frac{|G \cap P|}{|G \cup P|} \tag{4}$$

where G is the ground-truth segmentation mask and P is the predicted segmentation mask. For the purpose of evaluation, we calculated several other metrics, namely boundary IoU [29], pixel accuracy, and Dice coefficient.

3 Results

3.1 Recording quality is consistent across time

To evaluate the performance of the segmentation DNN, we first assessed the overall image quality for both, lossless and lossy recordings, as this is a major confounding source for segmentation success. First, we determined the maximum dynamic range of each given image. We found that the maximum dynamic range is constantly high for both, lossless and lossy recordings (Fig 3A, left panel), however, the dynamic range is significantly lower for lossy recordings as for lossless recordings (paired Student's t-test, p <0.01, Fig 3A, right panel). Using a complete blind, non-reference metric, the NIQE score [26], we could show that lossless recordings have lower, therefore better NIQE scores until beginning of 2021 (Fig 3B), afterwards the NIQE scores were highly overlapping. In [25], a mean NIQE for the OpenHSV system of 13.19 was reported, showing that image quality has been consistent since the introduction of the system.

To compare images retrieved from lossy recordings to their lossless counterparts, we relied on two reference metrics, peak-signal-to-noise-ratio (PSNR) and structural similarity index measure (SSIM). PSNR has a constant high value above 90 dB, suggesting a high-quality compression (Fig 3C). In contrast to PSNR, SSIM also takes the perceptual change in structural information into account. In Fig 3D we show that the lossy compressed videos nevertheless are in almost perfect agreement with the lossless reference.
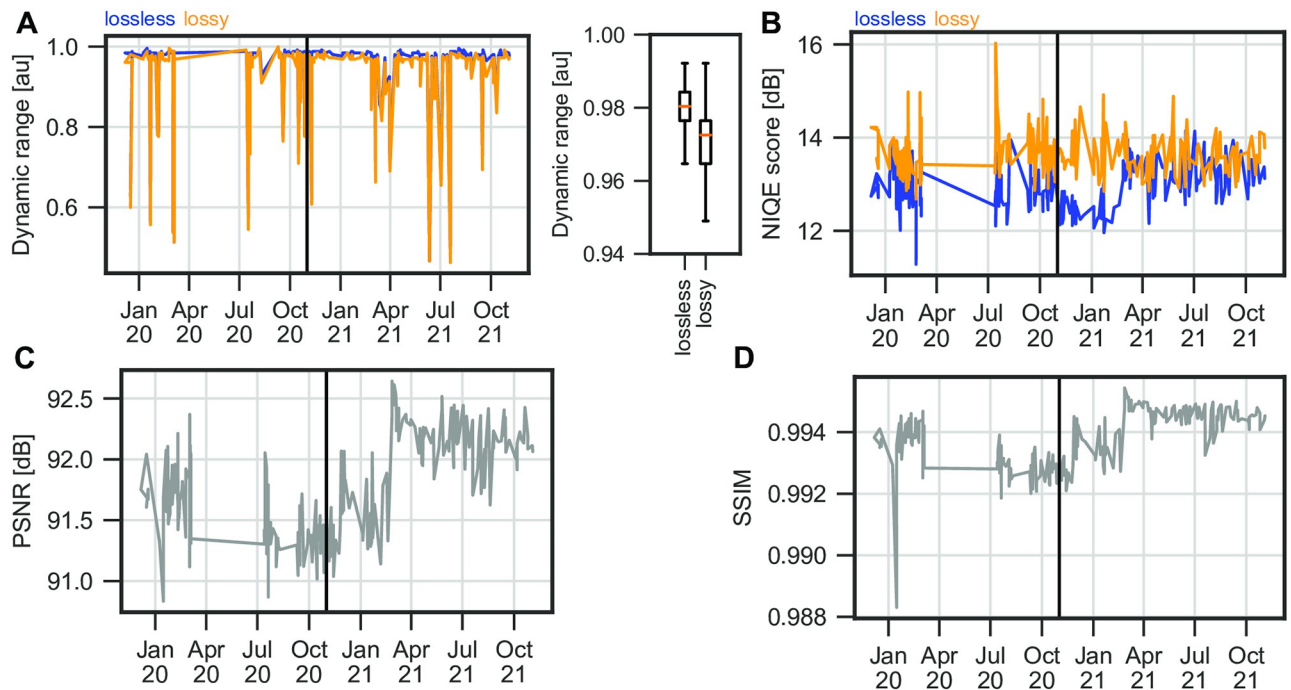
Fig 3. Image quality is relatively stable across time. A: Dynamic range per recording. Lossless compressed data in blue, lossy in orange. Boxplots show the 95-percentile of the data. B: NIQE score for lossy (orange) and lossless (blue) compressed videos across time. C: Peak-Signal-To-Noise-Ratio (PSNR) across time. D: Structural Similarity Index Measure (SSIM) across time.

In summary, we found that the image quality has single outliers, but overall is highly consistent across time. Further, high PSNR and SSIM values indicate an overall accurate conversion from lossless to lossy image content and high quality in lossy compressed videos.

3.2 Segmentation performance is not affected by lossy data compression

To further evaluate the performance of the segmentation DNN, we manually annotated the glottal area in the first one hundred frames of ten randomly selected videos, which were manually rated as quality 2 (excellent). Using this ground truth and the predicted segmentation masks by the DNN, we computed the Intersection over Union (IoU) for each frame across all videos. We found that lossy and lossless saved videos achieve a similar performance with a median IoU of 0.756/0.742 and 0.770/0.768 for segmentation masks computed with and without ROI, respectively (Fig 4B). These IoU values are comparable to other works [9], where IoU scores between 0.741 and 0.769 were achieved, and are sufficient for clinical reliability. The use of an ROI enhances the segmentation speed as smaller images are used, however, this leads to worse results. We hypothesize that this is caused due to the loss of global spatial information. We further mined the computed IoU scores to determine why very low IoU scores are obtained. Fig 4C shows that the low IoU scores emerge with a small segmented area, i.e. when the glottis is closed. This is in line with previous reports [10], and has a negligible effect on the data analysis. We next investigated if any configuration has an impact on the clinically relevant glottal area waveform (GAW) signal. We were able to confirm that all combinations despite their deviation in the IoU score have little to no effect on the GAW, as they do not deviate (S1A Fig in S1 File) and almost perfectly correlate with each other (S1B Fig in S1 File), which is important for downstream computation of quantitative parameters.
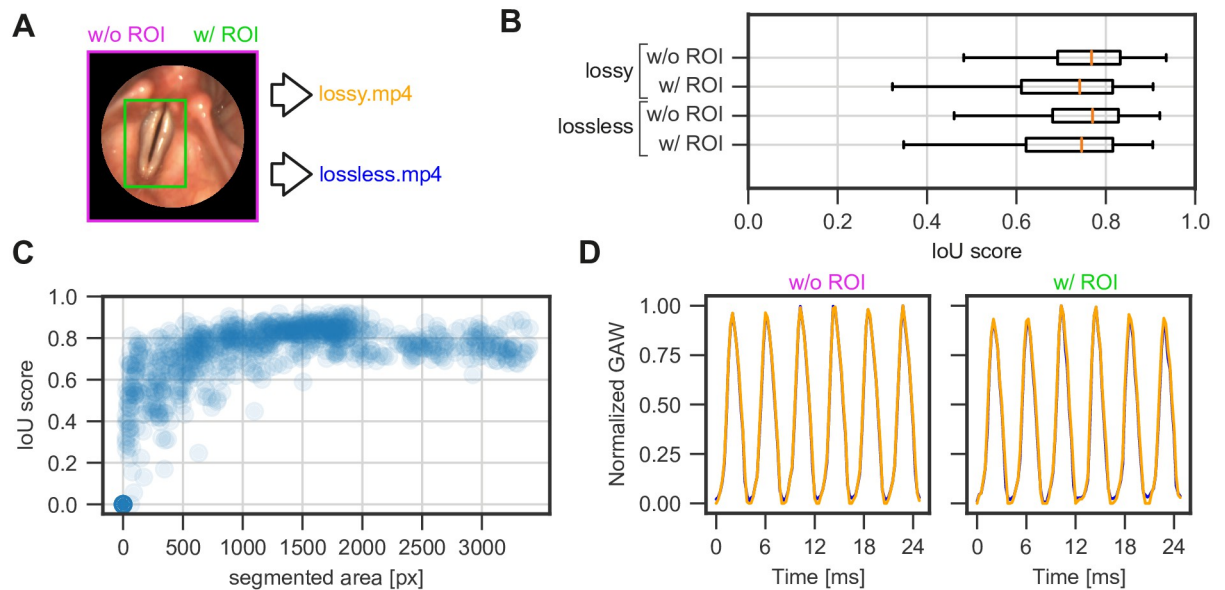
Fig 4. Lossy compression does not affect segmentation performance. A: Input data for DNN inference. Either full-frame (magenta) or ROI-constrained (green) image data were used for inference. Image data were stored either lossless or lossy compressed as .mp4-files (see Methods). B: Intersection over Union (IoU) score per configuration shown in A. C: Dependency of IoU scores on the segmented glottal area. D: Exemplary glottal area waveform for full frame (w/o ROI) and ROI-based (w/ ROI) segmentations. Lossy and lossless compressed data are plotted in orange and blue, respectively.

https://doi.org/10.1371/journal.pone.0266989.g004

To further emphasize the practical relevance of our model, we measured the inference time on a CPU (Intel Xeon Silver 4110) and a GPU (NVIDIA GeForce RTX 2080 Ti) with and without ROI. For this, we calculated the average ROI size of all annotated ROIs, resulting in an input image size of 224x160 px. Predicting 100 images with an image size of 1024x1024 px on CPU and GPU takes 38.55 ± 0.35 seconds and 1.73 ± 0.08 seconds, respectively. Using the average ROI frame size on CPU and GPU results in a reduction of the inference time to 1.79 ± 0.23 seconds and 0.12 ± 0.009 seconds, respectively. Even on a CPU the inference is quite fast when using ROIs instead of full frames. This underpins the possibility of using our model in a clinical environment where computing resources are limited.

3.3 Continual training for DNN fine-tuning improves performance

Despite the fact that we gained mostly successful and accurate segmentations, we manually reviewed all segmentation masks and found two common issues in a minority of videos (9%): (1) artifacts in the segmentation masks and (2) empty segmentation masks (S2A Fig in S1 File). We hypothesized that continuous integration of new, system-specific data using continual training [30, 31] is increasing the DNN performance. Here, we evaluated two strategies for integrating new data (Fig 5A). We either used a fixed time period or fixed quantities of videos. We used artifact-free, full-frame, lossy compressed videos for continual training, as full frames resulted in higher IoU scores (Fig 4A). We found that all strategies were able to reduce the number of artifacts after only the first iteration of continual training and removed artifacts on average by 38–48% (Fig 5B). Additionally, the more data is used for each training step, the higher the impact on artifact removal, and that a fixed quantity is preferable to a fixed time period (Fig 5B). In particular, when determining the median, i.e. what is the reduction in 50% of all cases, we found that the fixed amount of 40 videos has the best overall performance of reducing the artifacts of by 48%. In general, we can confirm that a fixed data amount seems to
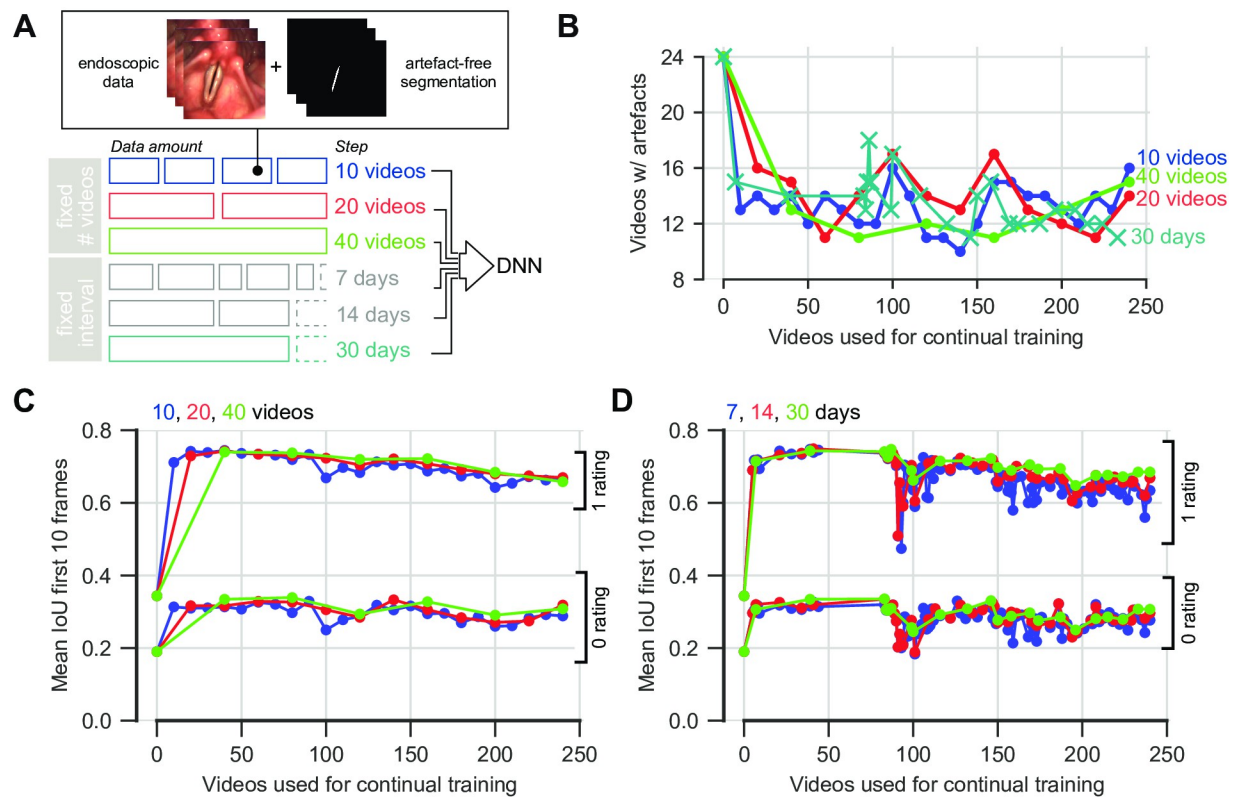
Fig 5. Continual training alleviates segmentation artifacts. A: Continual training strategies. B: Reduction of artifacts depending on continual learning strategy. C: IoU score of the first ten frames of the evaluated data after continual learning with a fixed amount of videos. D: Same as in C, but with a fixed time interval.

https://doi.org/10.1371/journal.pone.0266989.g005

be preferable to a fixed time frame, as the latter also shows an unstable performance behavior across continual learning epochs (S1 File). To further quantify this effect more objectively, we annotated the first 30 frames of each artifact video to compute IoU scores after each continual learning step (Fig 5C and 5D). Each strategy was able to increase the IoU score already after the first continual learning iteration, whereas a fixed data amount has a more stable and constant performance (Fig 5C) compared to a fixed time interval (Fig 5D), similar to our observations with manually scored artifacts (S1 File). In addition, we determined the performance of several other image segmentation metrics after each continual learning iteration. We found consistent results across assessed metrics compared to the IoU score reported in Fig 5C, strengthening our continual learning scheme (S1 File). In contrast, the fixed-time-period strategy shows no apparent during the first COVID-19 wave, which again suggests that the use of fixed quantities is preferable to fixed time periods. Taken together, our data shows that the segmentation DNN is not only able to quickly adapt to new data, but also that continual learning is an important feature in using DNNs in a clinical context.

4 Discussion

Laryngeal high-speed videoendoscopy is a major tool in quantifying laryngeal physiology. In this study, we show that clinically optimized DNNs have an overall high performance due to a constantly high data quality and a well-pre-trained, generalized DNN. Although the DNN has never be trained on this system's data, we still gain in most (91%) cases satisfactory results. The remaining videos (i.e. 9%) show segmentation artifacts or have completely empty

segmentations while there is an open glottis in the endoscopic images. We further show that lossy compressed videos are on par with lossless compressed videos in terms of segmentation performance (Fig 4). Using continuous integration of novel data, we can also show that the DNN is able to adapt such that previously identified artifacts are reduced and ensure a constantly improving segmentation environment (Fig 5).

4.1 Acquisition circumstances as potential confounders

We could show that the analyzed data has a relatively high image quality throughout the analyzed time frame (Fig 3). However, other systems have been shown to suffer from low lighting, which can be rescued with Deep Learning methods [32]. As the data acquisition is a manual procedure, the motion of the patient or the examiner can be a confounder. Motion correction techniques were proposed [33, 34] that can be used as a pre-processing step that have not been employed in this analysis.

4.2 Comparison to other glottis segmentation platforms

In this study, we investigated the performance of a single DNN. Offline image analysis platforms, such as the Glottis Analysis Tools [34] (GAT), can further serve as a reference for segmentation performance and may allow segmentations of higher quality. In comparison, the analyzed OpenHSV DNN has similar performance as the smallest GAT neural network, however, larger and more elaborate networks have superior performance (S1 File). Notably, this effect is largely compensated by our continual training scheme (Fig 5). Nevertheless, the average IoUs obtained in this study are in an acceptable range (0.742–0.770) that do not impact the clinical soundness of downstream quantitative parameter computation as shown previously [9, 25].

4.3 Continual learning strategies

Our proposed continuous integration of more data is straightforward and effortless. We have seen that using a fixed data mount is beneficial to a fixed time interval, which is maybe due to the irregular patient stream (Fig 2). We further observe a decline in IoU scores across data and time (Fig 5C and 5D) and a tendency to more artifacts after integrating a large amount of data (Fig 5B, S1 File). To counteract catastrophic forgetting [30], a common problem in continual learning schemes, additional precautions can be introduced, such as integrating the BAGLS dataset [10] in the continual learning training dataset or testing on previous artifact-free images if the current model performs better than before. Combining both aforementioned examples would incorporate external and internal regularization factors that are maybe beneficial in a more elaborate continual learning paradigm. In addition, we have not investigated how more sophisticated human-in-the-loop strategies [35], such as manual segmentation and retraining, perform. With wider adoption of OpenHSV, we also believe that federated learning techniques [36] will further boost the DNN segmentation performance.

5 Conclusion

In this work, we evaluated a collection of videos recorded over a two-year period by an AI-powered laryngeal high-speed videoendoscopy imaging system. The videos were stored in lossless and lossy formats. First, we examined recording quality over time, which we show is consistent throughout the two-year period by calculating the reference metrics PSNR and SSIM. We found that both, lossless and lossy compressed videos achieve high scores illustrating the high image quality even for lossy compressed videos. To further investigate the effect of

compression in storing laryngeal high-speed videoendoscopy data, we performed the glottis segmentation task with a previously trained and openly available DNN based on the U-Net architecture. Our results show that lossy compression does not affect the DNNs segmentation performance. Determining the ROI (the glottal area) prior to the segmentation task enhances the segmentation speed, however, affects IoU scores. We found two common segmentation artifacts that we could improve using a continual learning scheme using a fixed data mount across evaluation metrics. Future research should investigate human-in-the-loop strategies combined with continual learning strategies to provide a constantly high segmentation quality.

Supporting information

S1 File. This file contains further S1-S4 Figs and S1 Table.
(PDF)

Author Contributions

Conceptualization: René Groh, Andreas M. Kist.

Data curation: René Groh, Stephan Dürr, Anne Schützenberger.

Formal analysis: René Groh.

Funding acquisition: Anne Schützenberger.

Investigation: René Groh, Stephan Dürr, Anne Schützenberger.

Methodology: René Groh, Stephan Dürr, Anne Schützenberger, Andreas M. Kist.

Project administration: Marion Semmler, Andreas M. Kist.

Resources: René Groh, Andreas M. Kist.

Software: René Groh.

Supervision: Anne Schützenberger, Marion Semmler, Andreas M. Kist.

Validation: Stephan Dürr, Anne Schützenberger, Andreas M. Kist.

Visualization: René Groh, Andreas M. Kist.

Writing – original draft: Marion Semmler, Andreas M. Kist.

Writing – review & editing: René Groh, Marion Semmler, Andreas M. Kist.

References

1. Deliyski DD, Petrushev PP, Bonilha HS, Gerlach TT, Martin-Harris B, Hillman RE. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. Folia Phoniatrica et Logopaedica. 2008; 60(1):33–44. https://doi.org/10.1159/000111802 PMID: 18057909

2. Zacharias SR, Deliyski DD, Gerlach TT. Utility of laryngeal high-speed videoendoscopy in clinical voice assessment. Journal of Voice. 2018; 32(2):216–220. https://doi.org/10.1016/j.jvoice.2017.05.002 PMID: 28596101

3. Kunduk M, Döllinger M, McWhorter AJ, Lohscheller J. Assessment of the variability of vocal fold dynamics within and between recordings with high-speed imaging and by phonovibrogram. The Laryngoscope. 2010; 120(5):981–987. https://doi.org/10.1002/lary.20832 PMID: 20422695

4. Yousef AM, Deliyski DD, Zacharias SR, de Alarcon A, Orlikoff RF, Naghibolhosseini M. Spatial segmentation for laryngeal high-speed videoendoscopy in connected speech. Journal of Voice. 2020. https://doi.org/10.1016/j.jvoice.2020.10.017 PMID: 33257208

5. Naghibolhosseini M, Deliyski DD, Zacharias SR, de Alarcon A, Orlikoff RF. Temporal segmentation for laryngeal high-speed videoendoscopy in connected speech. Journal of Voice. 2018; 32(2):256–e1. https://doi.org/10.1016/j.jvoice.2017.05.014 PMID: 28647431

6. Döllinger M, Hoppe U, Hettlich F, Lohscheller J, Schuberth S, Eysholdt U. Vibration parameter extraction from endoscopic image series of the vocal folds. IEEE Transactions on Biomedical Engineering. 2002; 49(8):773–781. https://doi.org/10.1109/TBME.2002.800755 PMID: 12148815

7. Andrade-Miranda G, Stylianou Y, Deliyski DD, Godino-Llorente JI, Henrich Bernardoni N. Laryngeal image processing of vocal folds motion. Applied Sciences. 2020; 10(5):1556. https://doi.org/10.3390/app10051556

8. Gloger O, Lehnert B, Schrade A, Völzke H. Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions. IEEE Transactions on Biomedical Engineering. 2014; 62(3):795–806. https://doi.org/10.1109/TBME.2014.2364862 PMID: 25350912

9. Kist AM, Döllinger M. Efficient biomedical image segmentation on EdgeTPUs at point of care. IEEE Access. 2020; 8:139356–139366. https://doi.org/10.1109/ACCESS.2020.3012722

10. Gómez P, Kist AM, Schlegel P, Berry DA, Chhetri DK, Dürr S, et al. BAGLS, a multihospital benchmark for automatic glottis segmentation. Scientific data. 2020; 7(1):186. https://doi.org/10.1038/s41597-020-0526-3 PMID: 32561845

11. Fehling MK, Grosch F, Schuster ME, 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(2):e0227791. https://doi.org/10.1371/journal.pone.0227791 PMID: 32040514

12. Laves MH, Bicker J, Kahrs LA, 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(3):483–492. https://doi.org/10.1007/s11548-018-01910-0 PMID: 30649670

13. Azam MA, Sampieri C, Ioppi A, Africano S, Vallin A, Mocellin D, et al. Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection. The Laryngoscope. 2021. PMID: 34821396

14. Yousef AM, Deliyski DD, Zacharias SR, de Alarcon A, Orlikoff RF, Naghibolhosseini M. A Deep Learning Approach for Quantifying Vocal Fold Dynamics During Connected Speech Using Laryngeal High-Speed Videoendoscopy. Journal of Speech, Language, and Hearing Research. 2022; p. 1–16. PMID: 35605603

15. Ren J, Jing X, Wang J, Ren X, Xu Y, Yang Q, et al. Automatic recognition of laryngoscopic images using a deep-learning technique. The Laryngoscope. 2020; 130(11):E686–E693. https://doi.org/10.1002/lary.28539 PMID: 32068890

16. Jiang F, Grigorev A, Rho S, Tian Z, Fu Y, Jifara W, et al. Medical image semantic segmentation based on deep learning. Neural Computing and Applications. 2018; 29(5):1257–1265. https://doi.org/10.1007/s00521-017-3158-6

17. Guo Y, Liu Y, Georgiou T, Lew MS. A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval. 2018; 7(2):87–93. https://doi.org/10.1007/s13735-017-0141-z

18. Lateef F, Ruichek Y. Survey on semantic segmentation using deep learning techniques. Neurocomputing. 2019; 338:321–348. https://doi.org/10.1016/j.neucom.2019.02.003

19. Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review. 2021; 54(1):137–178. https://doi.org/10.1007/s10462-020-09854-1

20. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.

21. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; 2018. p. 3–11.

22. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:180403999. 2018.

23. Alom MZ, Yakopcic C, Taha TM, Asari VK. Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net). In: NAECON 2018-IEEE National Aerospace and Electronics Conference. IEEE; 2018. p. 228–233.

24. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021; 18(2):203–211. https://doi.org/10.1038/s41592-020-01008-z PMID: 33288961

25. Kist AM, Dürr S, Schützenberger A, Döllinger M. OpenHSV: an open platform for laryngeal high-speed videoendoscopy. Scientific Reports. 2021; 11(1):1–12. https://doi.org/10.1038/s41598-021-93149-0 PMID: 34215788

26. Mittal A, Soundararajan R, Bovik AC. Making a "completely blind" image quality analyzer. IEEE Signal processing letters. 2012; 20(3):209–212. https://doi.org/10.1109/LSP.2012.2227726

27. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004; 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861 PMID: 15376593

28. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat. 1901; 37:547–579.

29. Cheng B, Girshick R, Dollár P, Berg AC, Kirillov A. Boundary IoU: Improving Object-Centric Image Segmentation Evaluation; 2021. Available from: https://arxiv.org/abs/2103.16562.

30. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: A review. Neural Networks. 2019; 113:54–71. https://doi.org/10.1016/j.neunet.2019.01.012 PMID: 30780045

31. Lee CS, Lee AY. Clinical applications of continual learning machine learning. The Lancet Digital Health. 2020; 2(6):e279–e281. https://doi.org/10.1016/S2589-7500(20)30102-3 PMID: 33328120

32. Gómez P, Semmler M, Schützenberger A, Bohr C, Döllinger M. Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network. Medical & biological engineering & computing. 2019; 57(7):1451–1463. https://doi.org/10.1007/s11517-019-01965-4 PMID: 30900057

33. Deliyski DD. Endoscope motion compensation for laryngeal high-speed videoendoscopy. Journal of Voice. 2005; 19(3):485–496. https://doi.org/10.1016/j.jvoice.2004.07.006 PMID: 16102674

34. Kist AM, Gómez P, Dubrovskiy D, Schlegel P, Kunduk M, Echternach M, et al. A deep learning enhanced novel software tool for laryngeal dynamics analysis. Journal of Speech, Language, and Hearing Research. 2021; 64(6):1889–1903. https://doi.org/10.1044/2021_JSLHR-20-00498 PMID: 34000199

35. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis. 2021; 71:102062. https://doi.org/10.1016/j.media.2021.102062 PMID: 33901992

36. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research. 2021; 5(1):1–19. https://doi.org/10.1007/s41666-020-00082-4 PMID: 33204939