Dysphagia
https://doi.org/10.1007/s00455-025-10829-z

ORIGINAL ARTICLE

A Machine Learning Pipeline for Automated Bolus Segmentation and Area Measurement in Swallowing Videofluoroscopy Images of an Infant Pig Model

Max Sarmet[1,2] · Elska Kaczmarek[1] · Alexane Fauveau[1] · Kendall Steer[1] · Alex-Ann Velasco[1] · Ani Smith[1] · Maressa Kennedy[1] · Hannah Shideler[1] · Skyler Wallace[1] · Thomas Stroud[1] · Morgan Blilie[1] · Christopher J. Mayerl[1]

Received: 30 April 2024 / Accepted: 9 April 2025

Abstract

Feeding efficiency and safety are often driven by bolus volume, which is one of the most common clinical measures of assessing swallow performance. However, manual measurement of bolus area is time-consuming and suffers from high levels of inter-rater variability. This study proposes a machine learning (ML) pipeline using ilastik, an accessible bioimage analysis tool, to automate the measurement of bolus area during swallowing. The pipeline was tested on 336 swallows from videofluoroscopic recordings of 8 infant pigs during bottle feeding. Eight trained raters manually measured bolus area in ImageJ and also used ilastik's autocontext pixel-level labeling and object classification tools to train ML models for automated bolus segmentation and area calculation. The ML pipeline trained in 1h42min and processed the dataset in 2 min 48s, a 97% time saving compared to manual methods. The model exhibited strong performance, achieving a high Dice Similarity Coefficient (0.84), Intersection over Union (0.76), and inter-rater reliability (intraclass correlation coefficient = 0.79). The bolus areas from the two methods were highly correlated (R^2 = 0.74 overall, 0.78 without bubbles, 0.67 with bubbles), with no significant difference in measured bolus area between the methods. Our ML pipeline, requiring no ML expertise, offers a reliable and efficient method for automatically measuring bolus area. While human confirmation remains valuable, this pipeline accelerates analysis and improves reproducibility compared to manual methods. Future refinements can further enhance precision and broaden its application in dysphagia research.

Keywords Swallowing · Animal models · Machine learning · Artificial intelligence · Videofluoroscopy · Dysphagia

Introduction

Videofluoroscopic study of swallowing (VFSS) is an essential method for researching swallowing in animals and humans [1–7]. Quantitative VFSS measures, initially pioneered by Leonard and colleagues in the early 2000s [8, 9], have since led to various practical pixel-based applications in videofluoroscopy, like measurements related to bolus area, post-swallow residue, and airway safety [2, 3, 5, 7]. Pixel-based measurements are powerful tools for assessing swallowing physiology and impairment and can guide clinicians and researchers in swallowing diagnosis, management, and rehabilitation. For example, in pediatric animal models and in adult humans, bolus area from VFSS studies have been demonstrated to be the primary predictor of penetration and aspiration, with larger boluses resulting in decreased swallow safety [2–4, 10–14].

However, bolus measurement presents a complex task characterized by time-consuming procedures and susceptibility to errors, particularly when performed manually as in most swallowing research studies [9, 15]. Pixel-based measures are vulnerable to several sources of variability and are prone to poor inter-rater agreement [5]. Such variability can compromise the accuracy of bolus measurements, which can reduce their effectiveness as a management tool, and often requires the inclusion of multiple raters to improve

✉ Max Sarmet
Maxsarmet@gmail.com

[1] Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ 86011, USA

[2] Graduate Department of Health Science and Technology, University of Brasilia, Brasilia 70910-900, Brazil

Published online: 28 April 2025

reliability. This, in turn, can extend the project's duration and increase associated costs. Machine learning (ML) is a promising solution to these challenges.

ML models have been shown to be effective in assisting with dysphagia risk assessment and identification [16–26] and in monitoring aspects of swallowing physiology [27–32], including automated bolus segmentation and measurement [33–35]. However, there are challenges to incorporating ML in research and clinical practice. Developing ML models, particularly complex deep learning models, requires significant data, expertise, and ongoing validation [33]. Integrating ML into research projects can also demand technical expertise with robust tools [36], which may not be readily available in standard training programs [37]. Other challenges include high computing costs, infrastructure needs, staffing, energy consumption, and the necessity for frequent updates [38]. This results in low accessibility for researchers who are not experts in ML as well as protocols that cannot be adopted by the broader scientific community. Furthermore, the need for ML models to be validated using manual measures can make their implementation challenging, especially in clinical work with VFSS where radiation exposure is limited.

We used a ML program with a graphical user interface (GUI), ilastik [39], to evaluate whether non-experts could be trained with simple documentation to effectively use ML for bioimage analysis. To do this, we used infant pigs, a validated animal model for infant dysphagia [1–4, 10–13], to test whether ML procedures with ilastik could be used to automatically identify, segment, and measure the area of a bolus from VFSS images. Infant pigs are an excellent model to evaluate this procedure for a number of reasons. They allow for large numbers of swallows to be recorded at spatial and temporal precisions much higher than clinical populations with no concern for radiation exposure. Additionally, because they allow for investigation into otherwise healthy infants, we were able to manipulate the feeding conditions of the infants to increase variability in bolus conditions, which facilitates the ability to assess the sensitivity and precision of the ML model compared to manual bolus area measurement. This study aimed to assess the reliability of the ML pipeline compared to manual bolus measurement, with the hypothesis that the ML model will accurately measure bolus area in videofluoroscopy images.

Methods

Animal Model and Preparation

Animal care and procedures were approved by Northern Arizona University IACUC (22−010). We obtained 14 1-day-old full-term infant pigs, identified as CB01 to CB14 (Yorkshire/Landrace sows, Premier Biosource). While the cohort size is relatively small, it was determined to be sufficient to provide robust data for the ML pipeline, as individual swallows are the unit of analysis, and we recorded multiple (~20) swallows per individual. Throughout the course of the experiments, pigs received an infant pig milk replacer formula (Birthright Baby Pig Milk Replacer, Ralco Show, Marshall, MN USA) and were trained to feed from custom bottle-nipples. Additional housing and feeding followed previously outlined standards of care for this specific animal model [2, 3, 40]. To support a concurrent research project separate from the aims of this ML validation project, tantalum beads (0.8 mm diameter) were surgically placed within key oropharyngeal structures, encompassing the soft palate, palatopharyngeal arches, hard palate, and tongue [see refs [2, 3] for details]. Beads were implanted to enable X-ray Reconstruction of Moving Morphology (XROMM) analysis, a technique used to reconstruct 3D motion of anatomical structures [41]. Analgesia [buprenorphine (0.1 mg/kg) and meloxicam (0.4 mg/kg)] was provided before and for 48 h after surgery. Animals were monitored for any sign of discomfort after surgery every 3 to 5 h. We did not observe any impact of bead placement on feeding function, and pigs generated similar amounts of intraoral pressure, with similar feeding rates before and after bead placement.

Image Acquisition

Videofluoroscopic data was recorded in the lateral view using a GE 9400 C-Arm X-ray system (64 kV, 5.1 mA) and a high-speed Redwood camera (IO Industries, Ontario, Canada). Images were captured at 100 fps with a 7500 µs exposure and a 3048 × 3048 pixel resolution. The 100 fps frame rate, common in animal studies, ensures that rapid behaviors are captured in detail. The 7500 µs exposure minimized motion blur, particularly for fast-moving structures.

A standard grid was used to remove image distortion [41]. We bottle-fed pigs a combination of infant milk replacer (Ralco Birthright, Marshall, MN, USA) and barium (75 g/1500 ml of milk) for X-ray contrast (E-Z-PAQUE, Bracco Diagnostics, Milan, Italy). Because these data were collected as part of another research project, we fed the pigs using two distinct bottle nipple types which were expected to result in variable bolus areas across individuals to fully capture potential for variation in bolus volume. Bottle nipples include a standard, hollow nipple, as well as a branching duct nipple similar to Mayerl et al. [11]. Excluding the initial 10 s of feeding, which differs markedly from the majority of the feeding bout in a variety of physiological parameters [42, 43], we recorded approximately 20 swallows per pig for each experimental condition (Fig. 1). This

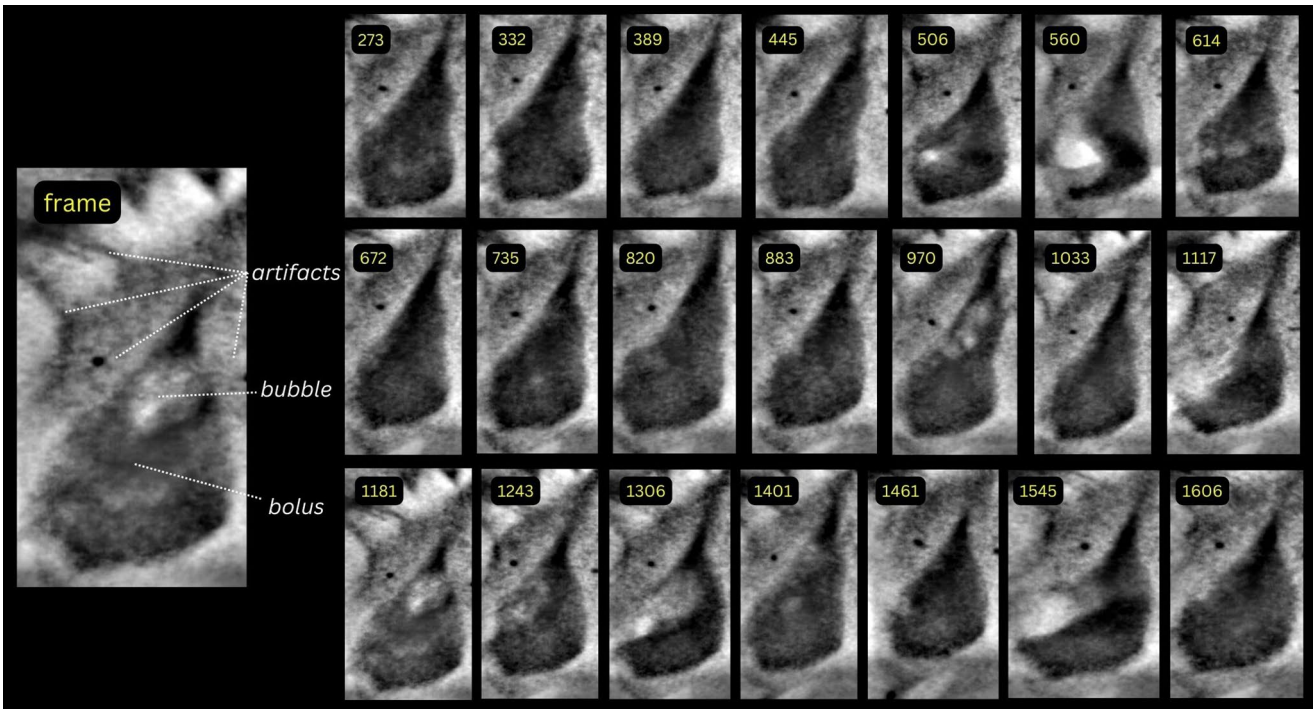

Fig. 1 Example of selected recording sequence. Pre-processed bolus images (frame number in yellow) from 21 consecutive swallows during a feeding session. These frames were recorded at 100fps, over 10.34 s. These frames display the variety in bolus features (shapes, bubbles, artifacts, brightness, and contrast)

approach was employed to ensure that similar sample sizes were collected for each nipple type, regardless of feeding duration or rate.

Dataset

For this project, we curated a set of 26 videos from 8 infant pigs, capturing a total of 336 swallows (Supplementary material 1). The initiation of each swallow was identified as the frame at which the bolus was accumulated in the back of the oropharynx prior to passing over the epiglottis following published procedures [2, 3, 40]. The raw dataset consisted of boluses that varied in their shape, brightness, and contrast, as well as the presence of bubbles inside the bolus and other artifacts in the images (Fig. 1, Supplementary material 1, page 3).

Manual Bolus Measurements

Before making measurements for this study, eight raters received training on measuring bolus area to ensure they attained both high intra- and inter-rater reliability. All raters were trained on a set of bolus images until each of their bolus measurements were within ±10% of the mean measurement of that bolus following standard published protocols (i.e. the average measurement of the eight raters who were being trained) [2, 3, 40].

These trained raters then utilized the free-hand selection tool in ImageJ (v. 1.53e National Institutes of Health, Bethesda, MD, USA) [44] to manually measure the area of bolus of the 336 raw swallow images following published protocols [3, 4, 10–13, 40, 45]. In short, raters outlined the bolus, including any bubbles at the frame at which the swallow was initialized following previously published protocols [3, 4, 10–13, 40, 45]. Raters calibrated the scale of the images (in pixels per millimeter) by measuring the diameter of a metal ring in the bottle lid (real diameter 6.45 cm) in the videofluoroscopic images. The bolus area in pixels was converted to mm² by dividing by the scale squared. While bolus area is a 2-D measure of a three-dimensional space, it is commonly used in clinical and basic science research, and is thought to correlate well with volume [46]. Each rater recorded the time it took to measure the area of 20 boluses. We used the average time (across raters) that it took to measure 20 boluses to calculate the approximate total time to measure the bolus area for all 336 images (Supplementary material 2).

ML Bolus Measurements Workflow

The workflow, from image acquisition to ML training and subsequent bolus predictions, is illustrated in Fig. 2, and described in detail below.

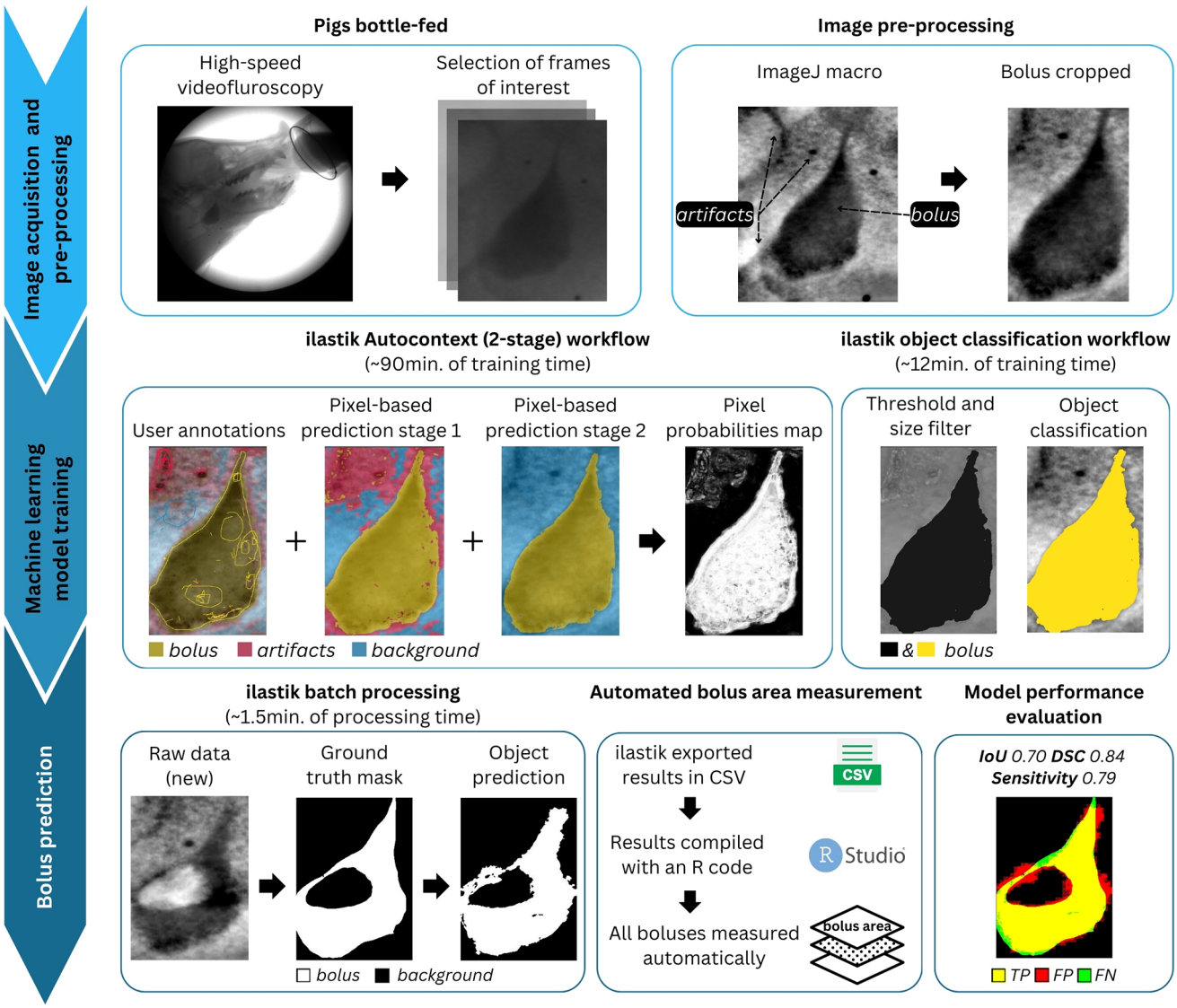

Fig. 2 Flowchart illustrating the automated bolus measurement process. The initial step encompasses image acquisition and image preparation, including image pre-processing and bolus cropping. The second step is training the ilastik machine learning model using two workflows: autocontext and object classification. In the autocontext workflow, yellow represents the bolus, blue the background, and red the artifacts. The last step is using batch processing to predict the boluses in new images. The resulting data are exported to CSV and compiled for analysis. To assess model performance, a comprehensive evaluation is conducted by comparing predicted bolus regions with ground truth masks. Specific metrics are calculated to quantify the accuracy of the predictions. Legend: DSC– Dice Similarity Coefficient; FN– false negative; IoU– Intersection over Union; TP– true positive; TN– true negative

Image Pre-processing

To minimize the impact of noise artifacts on algorithm training, each raw image was pre-processed using ImageJ by applying the "enhance contrast" function with a 0.35% adjustment and a fast Fourier transform band-pass filter (examples in Supplementary material 1, pages 1–2). The filter was set to reduce large structures down to 160 pixels and small structures up to 3 pixels while suppressing horizontal stripes with a 5% tolerance in direction. Additionally, after filtering, the "autoscaling" function was applied to improve image brightness and contrast. A macro was created in ImageJ to automatically pre-process batches of images (Supplementary Material 3). Finally, the pre-processed figures were manually cropped to the region of the bolus with as little background area as possible to reduce potential artifacts in the ML process.

Supervised ML-based Image Segmentation Using Ilastik Software

The software ilastik (version 1.4.0 for Windows 64-bit) was used to train the ML model (Fig. 2). ilastik is a Python-based interactive tool that employs ML-based bioimage analysis to classify pixels and objects based on user-provided annotations, requiring no machine learning expertise [39]. ilastik uses these user annotations, which can be as few as two labels (such as foreground or background) to predict the class of unannotated pixels or objects. This toolkit encompasses various 'workflows' designed to accomplish specific tasks [47], including the autocontext [48] and object classification workflows used in this study. When choosing ilastik for this project, our primary considerations were its existing validation for bioimages, user-friendly interface, and suitability for end users who may not have significant computational expertise [39, 49, 50].

Each rater who previously performed manual bolus measurements received brief ilastik software training and had access to a step-by-step tutorial (Supplementary Material 3). Each rater (1 to 7) independently selected 15 training images (less than 5% of the dataset) to capture variations in bolus shapes, bubbles, artifacts, and contrast levels. This ensured variability in the training data for the ML model. Rater 8 used a subset of 12 images to assess the model's performance with a smaller training dataset. They then used the autocontext and object classification ilastik workflows to identify and measure the bolus (described in detail in Supplementary Material 3). Using the image features computed by the object classification workflow, object predictions were batch-processed for the entire dataset. These predictions were exported in HDF5 (.h5) and converted to jpeg format using the data conversion workflow in ilastik. Additionally, we exported the object area data (in CSV format) from ilastik. Object areas for each bolus and for each rater were batch processed in R (v. 4.3.1), using the GUI RStudio (2023.06.2 Build 561) (Supplementary Material 3).

Hardware

In order to evaluate the ML model performance in different settings, two distinct hardware configurations were deployed. Both setups operated on the Windows 11 Pro operating system. The high-performance setup included an Intel® Core i9-13900 CPU running at a speed range of 1.5–5.6 GHz, 64 GB of RAM (operating at 4400 MHz DDR5), 1 TB SSD hard drive (M.2 2280, NVMe, C40), and a NVIDIA GeForce RTX3050 graphics processing unit with 8 GB of GDDR6 memory. Raters 1 to 7 used the same high-performance computer and lighting conditions. To evaluate the model's performance in more accessible settings, we replicated the analysis using a home-setup configuration, a common resource for many researchers. The home computer hardware setup, used by rater 8, included an Intel® Core i3-10100T CPU running at 3.00 GHz, 8 GB of RAM (operating at 2400 MHz DDR4), hard drive SSD M.2 Adata SX6000 256 GB, and integrated Intel® UHD Graphics.

Evaluation of Performance

To evaluate the ilastik model's performance, we employed standard machine learning metrics: Intersection over union (IoU), Dice similarity coefficient (DSC), and sensitivity [33–35, 51–55]. To calculate these metrics, we first determined the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) pixels. These pixel-level metrics formed the basis for calculating other performance indicators. Ground truth masks were manually created by the first author for all boluses. These masks were binarized, representing the bolus as a white foreground on a black background. The Mask Instant Comparator (MiC) plugin (*MIT*, *USA*) for Fiji [56], a specialized tool for segmentation mask comparison, was used to compute these metrics. IoU, a common metric for assessing segmentation overlap, measures the intersection between the predicted and ground truth masks divided by their union [54, 55]. DSC, another similarity measure, quantifies the overlap between two sets relative to their combined size [54, 57]. Sensitivity, or True Positive Rate, evaluates a model's ability to correctly identify positive instances [58]. In image validation literature, a DSC exceeding 0.70 generally indicates strong overlap between the predicted and ground truth segmentations [52, 53]. Similarly, an IoU greater than 0.7 is often considered indicative of high segmentation accuracy [54, 55]. These benchmarks provide a general context for interpreting the model's performance. The following formulas were used to calculate these metrics:

$$IoU = \frac{TP}{TP + FP + FN}$$

$$DSC = \frac{2TP}{2TP + FP + FN}$$

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

To adhere to common clinical research practices, we also evaluated the inter-rater reliability and model's performance using inter-rater intraclass correlation coefficient (ICC) and coefficient of variation (CV) [59, 60]. These statistics indicate how similar the bolus area measurements were across raters. We calculated the inter-rater ICC and CVs separately

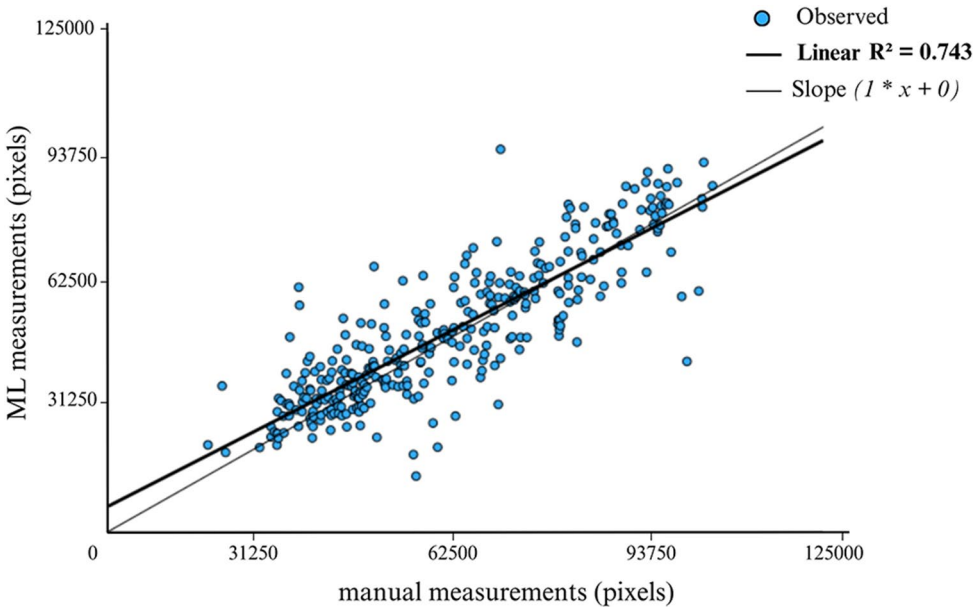

Fig. 3 Linear regression model. Linear regression model comparing the average bolus predictions generated by each machine learning model (y) to the average manual bolus measurements (x), measured in pixels

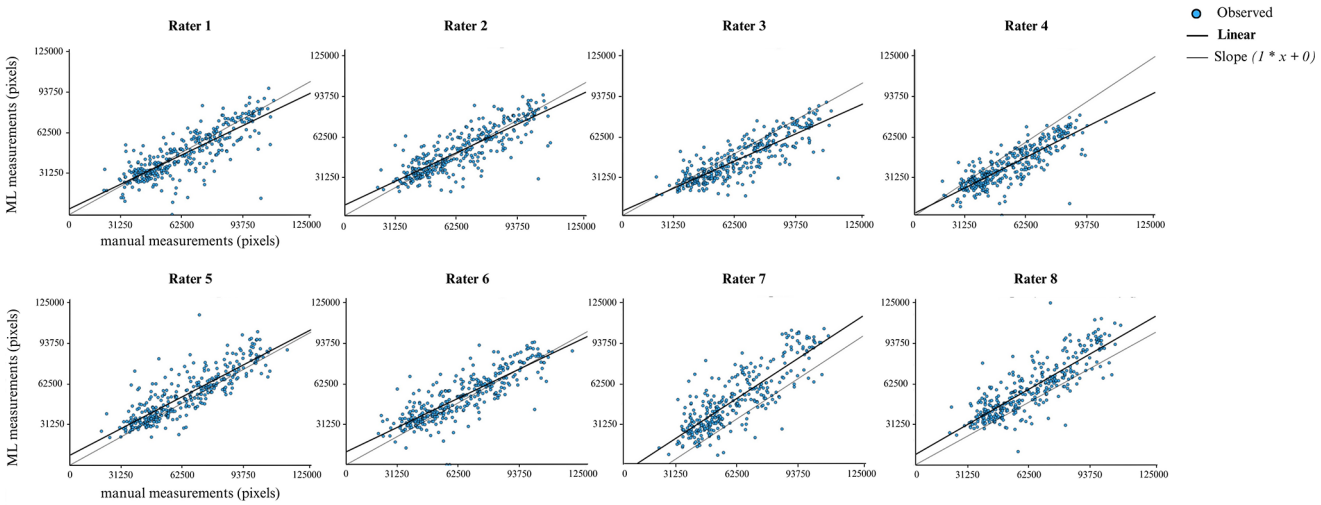

Fig. 4 Linear regression model. Linear regression model comparing each rater's bolus predictions generated by each machine learning model (y) to manual bolus measurements (x), measured in pixels. The independent variable is the average value of the manual bolus measurements

for the manual method and the ML method. Inter-rater ICCs and their 95% confidence intervals (CI) were calculated based on a single-rating, absolute agreement, two-way mixed-effects model. This model is based on the assumption that bolus measurements would be made using a single rater when either method (manual or ML) were applied for research. Reliability is considered excellent if the lower limit of the 95% CI of the ICC is greater than 0.90, good if between 0.75 and 0.90, moderate if between 0.5 and 0.75, and poor if below 0.5 [61]. The CV was calculated as standard deviation (in our case, the standard deviation of the bolus measurements from all raters) as a percentage of the mean value (in our case, the mean bolus measurement across all raters). The CV was calculated for each bolus and for each method. CV values $\leq$10% were considered excellent, 11–20% good, 21–30% acceptable, and $\geq$30% poor.

We also evaluated intra-rater reliability through linear regression of a rater's manual measurements against their ML measurements. The R^2 values indicate how correlated the measurements of the two methods were. Although we do not know the true sizes of the boluses in this study, manual measurements are standard practice and are therefore a useful baseline for comparison. The mean bolus areas were compared using the same approach (Figs. 3 and 4). Given the inherent challenges in measuring boluses that contain bubbles, we separated boluses with bubbles from boluses without bubbles. We also employed a linear mixed-effects model analysis to explore potential distinctions between manual and ML measurements. This test aimed to understand

the relationship between mean bolus area and measurement method (fixed effect, manual or ML), while accommodating the variability introduced by individual pig subjects as a random effect with an intercept. Additionally, Levene's test assessed the equality of variances between the two methods. All statistical analyses were performed using SPSS statistics (Version 29, IBM Corp., Armonk, NY, USA).

Qualitative Analyses

We also investigated potential avenues through which manual and ML methods might differ, as it is possible that differences between the measurements made by these methods may reflect the inaccuracy of the manual method, the ML method, or both methods combined. In other words, differences between the methods (i.e. low intra-rater reliability) do not necessarily indicate that the ML method is performing poorly–the manual method may be performing poorly instead. Through in-depth analysis of these cases, we identified recurrent patterns in the data and model behavior, paving the way for targeted improvements.

The qualitative analysis involved a 4-step workflow:

a) A thorough examination of artifacts, achieved by individually scrutinizing the original bolus images, the pre-processed images, and their corresponding ML predictions (Supplementary material 1, page 3).
b) We evaluated the influence of individual pigs on ML results. To do this, we compared the first swallow image of each recording (trial) to its ML prediction (Supplementary material 1, pages 1–2). For these initial swallow images, we then calculated the mean bolus area (in both pixels and mm^2) across boluses for each recording and for each method. Additionally, we computed the percent difference in bolus areas (in mm^2) between the manual and ML methods for each recording. Lastly, we calculated the mean coefficient of variation of bolus area for each recording and for each method.
c) Utilizing the linear regression of manual and ML measurements in pixels, we computed the difference between the expected ML value and the observed ML measurement. This enabled the identification of the top-performing predictions (10%, $n=34$) (Supplementary material 1, pages 4–16) as well as the most discrepant predictions (10%, $n=34$) (Supplementary material 1, pages 17–29).
d) We assessed the impact of artifacts on the ML predictions by comparing ML predictions made before and after the manual removal of artifacts. In other words, we introduced a human-supervised component into the ML workflow. To begin, we selected the results of one rater (rater 8). Using ImageJ's "color picker" and "brush" tools, we then excluded the obvious artifacts before measuring the area of the revised bolus prediction with the "set threshold", "make binary," and "analyze particles" features. Following this process, we assessed the impact of the artifacts by comparing the coefficient of determination (R^2) of the linear regression between the average manual measurements and the ML predictions before and after artifact correction.

Results

The ML pipeline took an average of 1 h and 42 min for training and 2 min and 48 s for batch processing the dataset. Raters spent an average of 23 s to manually measure each bolus, while the ML pipeline took only 0.5 s (Table 1, Supplementary material 2). This translates to a 97% time savings compared to manual measurements. Seven raters used high-performance computers (raters 1–7), while one used a home computer (rater 8). The ilastik batch processing time was 1.5 min on average for high-performance setups and 12 min for the home setup. Table 1 summarizes the results of bolus area measurement for manual and ML workflows, including analysis time. Table 2 presents the ICCs and CVs for each method, comparing performance on boluses with and without bubbles. Manual measurements exhibited low variability (CV: 5.3%), with 94.6% rated excellent, 4.8% good, and 0.3% poor. ML measurements exhibited higher inter-rater variability (CV: 17.8%). While 76.5% of ML measurements were rated as good or excellent, a significant proportion (23.6%) were deemed acceptable or poor.

Model Performance Evaluation

The model exhibited strong performance, achieving a mean DSC of 0.84 and an IoU of 0.76 (Table 3). Within each rater, there was a strong relationship between manual and ML bolus areas (R^2 ranged from 0.66 to 0.72, $p<0.001$) (Table 3; Figs. 3 and 4). This was especially true for boluses without bubbles ($R^2 = 0.78$, $p<0.001$). The relationship was strong,

Table 1 Bolus area and analysis time by method and nipple type

	Bolus area			Analysis time per bolus
Method	All nipples, SD, range (mm^2) $n=336$	Standard nipple, SD, range (mm^2) $n=170$	Duct nipple, SD, range (mm^2) $n=166$	All nipples, SD, range (seconds) manual $n=20$, ML $n=336$
Manual	185 ± 62 (54–325)	199 ± 58 (90–322)	171 ± 63 (54–322)	23.98 ± 10.86 (9.65–46.10)
ML	192 ± 66 (52–359)	215 ± 65 (60–407)	190 ± 68 (81–350)	0.50 ± 0.66 (0.26–2.14)

Legend: ML: machine learning; SD: standard deviation

Table 2 Inter-rater reliability and variability for manual and ML methods

All boluses					Boluses without bubbles				Boluses with bubbles			
Method	ICC	95% C.I.	p-value	CV ± SD	ICC	95% C.I.	p-value	CV ± SD	ICC	95% C.I.	p-value	CV ± SD
Manual	0.97	0.97, 0.98	<0.001	5.3 ± 3.5	0.97	0.97, 0.98	<0.001	4.5 ± 2.5	0.95	0.93, 0.96	<0.001	6.6 ± 4.5
ML	0.79	0.58, 0.88	<0.001	17.8 ± 8.0	0.81	0.61, 0.90	<0.001	16.5 ± 8.4	0.70	0.45, 0.83	<0.001	20.0 ± 6.7

CI: confidence interval; CV: coefficient of variation; ICC: Intraclass Correlation Coefficient; SD: standard deviation

Table 3 Model performance evaluation: manual vs. machine learning measurements

Rater	DSC	IoU	Sensitivity	R^2
1	0.81	0.72	0.80	0.67*
2	0.84	0.73	0.86	0.69*
3	0.87	0.76	0.86	0.72*
4	0.83	0.75	0.85	0.70*
5	0.88	0.77	0.87	0.72*
6	0.85	0.75	0.83	0.70*
7	0.80	0.71	0.81	0.66*
8	0.86	0.76	0.87	0.71*
Mean ± SD	0.84 ± 0.02	0.74 ± 0.02	0.84 ± 0.02	0.69 ± 0.02

Legend: Values indicate similarity between manual and machine learning measurements. The R^2 values represent the coefficient of determination for linear regressions performed between manual and machine learning measurements, calculated individually for each rater. DSC: Dice similarity coefficient; IoU: Intersection over Union; SD: standard deviation; *$p < 0.001$

but less so, when boluses with bubbles were compared ($R^2 = 0.67$, $p < 0.001$). The linear mixed effects model revealed no significant difference in bolus area between manual and ML methods across pigs ($p = 0.09$), and Levene's test found no significant difference in variances between the measurement methods ($p = 0.48$).

Qualitative Analysis

Across the 26 recordings, the average percent difference in bolus area between manual measurements and ML predictions for each trial ranged from 4.91 to 20.50%. The qualitative analysis revealed the existence of artifacts primarily attributed to bubbles and the surgical placement of beads, integral components of the original research project (Supplementary material 1, page 3). These artifacts were consistently observed across all recordings, and removing them using the integrated human-supervised workflow improved the ML predictions. The R^2 value of the linear regression between ML predictions (after artifacts were removed) and the average manual measurements was 0.85 ($p < 0.001$), a remarkable enhancement compared to the R^2 of 0.74 for the average ML predictions without artifact correction. For visual reference, examples illustrating the appearance of boluses after undergoing this workflow can be found in Supplementary material 1 (pages 1 and 2).

The examination of top-performing predictions ($n = 34$, Supplementary material 1, pages 4–16), and those displaying significant deviations (most discrepant predictions, $n = 34$, Supplementary material 1, pages 17–29), revealed occurrences of both underestimations and overestimations across both methods. The comparison between the predictions and the probable cause investigation can be seen in Supplementary material 1, page 30.

Discussion

The ML-based pipeline demonstrated high accuracy in measuring bolus area from videofluoroscopy images, as evidenced by the high DSC scores. The model demonstrated good predictive performance, even when trained by non-expert ML users on a small fraction of the dataset (less than 5%) and processing images with diverse artifacts.

ML has been used extensively in swallowing research for different applications [16–22, 27–30, 33–35]. One of the major advantages of employing these methods is their accuracy, inter-rater reliability, and time-saving potential. Previous studies have demonstrated that ML models created using ilastik for different biomedical tasks can significantly reduce analysis time, with reported reductions of up to 80 to 90% compared to manual methods [62–64] Our ML model further supports this, reducing analysis time by 97% compared to manual workflow, making it ideal for large-scale studies.

Understanding how bolus volume affects swallowing is critical in dysphagia research [65]. Studies reveal patients with oropharyngeal dysphagia or aspiration exhibit altered responses to bolus volume and viscosity changes, from premature infants to post-stroke adults [10, 45, 65–67]. This reflects a reduced ability to adapt swallow response to different volumes, increasing the risk of penetration or aspiration. Bolus volume and fluid viscosity influence oral sensory receptors for touch, kinesthesia, and proprioception, affecting oral and pharyngeal kinematics, upper esophageal sphincter opening, and hyolaryngeal excursion, and correspond to differences in cortical activation in the brain [68, 69]. Precise bolus measurement is crucial for swallowing studies, and refined ML models can streamline this process, enabling more efficient research.

Our ML model demonstrated strong performance, achieving a DSC of 0.84 and an IoU of 0.74, even when tested on a simplified home setup with a reduced dataset. Notably, this configuration achieved comparable or superior results to the ilastik model using more robust hardware (Table 3). These findings highlight the feasibility of our approach for researchers with limited computational resources.

While Ariji et al. achieved a higher DSC of 0.94 using deep learning for bolus segmentation in VFSS, their model required 15 h and 43 min of training using more robust hardware [33]. In contrast, our model was trained by non-ML experts and completed in just 1 h and 42 min, significantly faster than other approaches. Similarly, Li et al. proposed a model with lower performance metrics, achieving a DSC of 0.81 and an IoU of 0.68, but with a faster inference speed of 49.34 ms [35]. In comparison, our model's inference time was slightly longer—260 ms on a high-performance setup and 2.14 s on a home-computer setup (Supplementary material 2) —while delivering superior segmentation accuracy. Shaheen et al. reported a less accurate model, with a mean DSC of 0.67 for bolus segmentation, but did not disclose training or inference times [34]. While deep learning approaches such as those above can achieve high accuracy, they often demand significant computational resources and technical expertise [36]. These results imply the use of ilastik for bolus identification is accurate and precise in comparison to more computationally expensive models, with minimal ML expertise required for use, suggesting that its use may prove clinically relevant for a variety of researchers in the field of dysphagia.

The qualitative analysis revealed the presence of various artifacts, including bubbles, which can arise from factors such as incomplete seals or high flow rates during bottle feeding, and are likely a component of most infant feeding studies. While previous studies have not explicitly addressed the impact of bubbles on bolus measurement, our findings suggest that boluses with bubbles can lead to increased inter-rater variability for both manual and ML methods. To address this, we propose a multi-pronged approach: enhanced rater training, model refinement, and integrated human-supervised workflows. Future research should focus on refining these methods to improve accuracy and broaden their application in swallowing research.

Limitations

The ilastik software may produce imprecise predictions, potentially impacting research results. To address this, we recommend a multi-pronged approach combining ilastik with human expertise. Additionally, the model's performance can be influenced by factors such as user experience and training data. The use of an animal model also allowed us to record at higher frame rates than in common clinical settings (100 fps vs. 15–60 fps in clinical settings). Lower frame rates in clinical settings might induce higher variability in the bolus, due to swallow initiation occurring between rather than within frames. Researchers should use the ML pipeline cautiously and consider a pilot study validating the model to manual measures to assess its performance in their specific context.

Conclusions

The ML pipeline using ilastik accurately measured boluses, demonstrating strong performance with a mean DSC of 0.84, an IoU of 0.76, and a 97% reduction in analysis time. The ML workflow offers several advantages to address large datasets, including reduced analysis time, smooth learning curve, and free access, complemented by an array

of tutorial resources. In summary, this method offers an efficient, reproducible, and low-cost approach for measuring bolus area from videofluoroscopy images of an animal model, potentially making it versatile and applicable across species. ML pipeline predictions ought to undergo human confirmation, and by refining both the raters and the ML model, we can achieve even greater precision. This work has the potential to facilitate the ability to evaluate bolus area, a critical component of swallow performance, with more swallows in less time than manual measures, and represents an important step forward in our ability to diagnose and evaluate dysphagia.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00455-025-10829-z.

Author Contributions Conceptualization: MS, CM; Methodology: MS, EK, MS; Validation: MS; Software: MS; Investigation: MS, AF, KS, AV, AS, MK, HS, SW, TS, MB, EK, CM; Formal analysis: MS, EK, CM; Data Curation and Visualization: MS; Writing - original draft preparation: MS, EK, CM; Writing - review and editing: MS, AF, KS, AV, AS, MK, HS, SW, TS, MB, EK, CM; Funding acquisition: CM; Resources: MS, CM; Project administration: MS, Supervision: CM.

Funding Funding was provided to Christopher J Mayerl by the National Institutes of Health (NIH) R00 HD205922 and R21 HD105294 to Christopher J Mayerl and Rebecca German.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics Approval Animal care and procedures were approved by the Northern Arizona University IACUC (22−010).

Consent for Publication All the authors of the study have consented for publication.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. German RZ, Crompton AW, Gould FDH, Thexton AJ. Animal models for dysphagia studies: what have we learnt so Far. Dysphagia. 2017;32:73–7.
2. Mayerl CJ, Steer KE, Chava AM, Bond LE, Edmonds CE, Gould FDH et al. Anatomical and physiological variation of the hyoid musculature during swallowing in infant pigs. J Exp Biol. 2021;224.
3. Mayerl CJ, Myrla AM, Gould FDH, Bond LE, Stricklen BM, German RZ. Swallow safety is determined by bolus volume during infant feeding in an animal model. Dysphagia. 2021;36:120–9.
4. Edmonds CE, German RZ, Bond LE, Mayerl CJ. Oropharyngeal capsaicin exposure improves infant feeding performance in an animal model of superior laryngeal nerve damage. J Neurophysiol. 2022;128:339–49.
5. Steele CM, Peladeau-Pigeon M, Nagy A, Waito AA. Measurement of pharyngeal residue from lateral view videofluoroscopic images. J Speech Lang Hear Res. 2020;63:1404–15.
6. Waito AA, Tabor-Gray LC, Steele CM, Plowman EK. Reduced pharyngeal constriction is associated with impaired swallowing efficiency in amyotrophic lateral sclerosis (ALS). Neurogastroenterol Motil. 2018;30.
7. Plowman EK, Tabor-Gray L, Rosado KM, Vasilopoulos T, Robison R, Chapin JL, et al. Impact of expiratory strength training in amyotrophic lateral sclerosis: results of a randomized, sham-controlled trial. Muscle Nerve. 2019;59:40–6.
8. Leonard R, Kendall KA, McKenzie S. Structural displacements affecting pharyngeal constriction in nondysphagic elderly and nonelderly adults. Dysphagia. 2004;19:133–41.
9. Dharmarathna I, Miles A, Allen J. Predicting penetration–aspiration through quantitative swallow measures of children: a videofluoroscopic study. Eur Arch Otorhinolaryngol. 2021;278:1907–16.
10. Mayerl CJ, Myrla AM, Bond LE, Stricklen BM, German RZ, Gould FDH. Premature birth impacts bolus size and shape through nursing in infant pigs. Pediatr Res. 2020;87:656–61.
11. Mayerl CJ, Kaczmarek EB, Smith AE, Shideler HE, Blilie ME, Edmonds CE et al. A Ducted, Biomimetic Nipple Improves Aspects of Infant Feeding Physiology and Performance in an Animal Model. Dysphagia [Internet]. 2024 [cited 2024 Dec 11];1–10. Available from: https://link.springer.com/article/10.1007/s00455-024-10780-5
12. Steer KE, Johnson ML, Edmonds CE, Adjerid K, Bond LE, German RZ et al. The Impact of Varying Nipple Properties on Infant Feeding Physiology and Performance Throughout Ontogeny in a Validated Animal Model. Dysphagia [Internet]. 2024 [cited 2024 Dec 11];39:460–7. Available from: https://link.springer.com/article/10.1007/s00455-023-10630-w
13. Mayerl CJ, Edmonds CE, Gould FDH, German RZ. Increased viscosity of milk during infant feeding improves swallow safety through modifying sucking in an animal model. J Texture Stud [Internet]. 2021 [cited 2024 Dec 11];52:603–11. Available from: https://onlinelibrary.wiley.com/doi/full/https://doi.org/10.1111/jtxs.12599
14. Kawamura LR, de SM, Sarmet M, de Campos PS, Takehara S, Kumei Y, Zeredo JLL. Apnea behavior in early- and late-stage mouse models of Parkinson's disease: Cineradiographic analysis of spontaneous breathing, acute stress, and swallowing. Respir Physiol Neurobiol [Internet]. 2024 [cited 2024 Dec 13];323. Available from: https://pubmed.ncbi.nlm.nih.gov/38395210/
15. Dharmarathna I, Miles A, Allen J. Quantifying bolus residue and its risks in children: A videofluoroscopic study. Am J Speech Lang Pathol. 2021;30:687–96.
16. Jauk S, Kramer D, Veeranki SPK, Siml-Fraissler A, Lenz-Waldbauer A, Tax E, et al. Evaluation of a machine Learning-Based dysphagia prediction tool in clinical routine: A prospective observational cohort study. Dysphagia. 2023;38:1238–46.
17. Kim JK, Choo YJ, Choi GS, Shin H, Chang MC, Park D. Deep learning analysis to automatically detect the presence of penetration or aspiration in videofluoroscopic swallowing study. J Korean Med Sci. 2022;37.
18. Martin-Martinez A, Miró J, Amadó C, Ruz F, Ruiz A, Ortega O, et al. A systematic and universal artificial intelligence screening method for oropharyngeal dysphagia: improving diagnosis through risk management. Dysphagia. 2023;38:1224–37.
19. Sakai K, Gilmour S, Hoshino E, Nakayama E, Momosaki R, Sakata N et al. A machine learning-based screening test for sarcopenic dysphagia using image recognition. Nutrients. 2021;13.

20. Santoso LF, Baqai F, Gwozdz M, Lange J, Rosenberger MG, Sulzer J, et al. Applying machine learning algorithms for automatic detection of swallowing from sound. Annu Int Conf IEEE Eng Med Biol Soc. 2019;2019:2584–8.
21. Frakking TT, Chang AB, Carty C, Newing J, Weir KA, Schwerin B, et al. Using an automated speech recognition approach to differentiate between normal and aspirating swallowing sounds recorded from digital cervical auscultation in children. Dysphagia. 2022;37:1482–92.
22. O'Brien MK, Botonis OK, Larkin E, Carpenter J, Martin-Harris B, Maronati R, et al. Advanced machine learning tools to monitor biomarkers of dysphagia: A wearable sensor Proof-of-Concept study. Digit Biomark. 2021;5:167–75.
23. Ye F, Cheng L-L, Li W-M, Guo Y, Fan X-FA, Machine-Learning. Model Based on Clinical Features for the Prediction of Severe Dysphagia After Ischemic Stroke. Int J Gen Med [Internet]. 2024 [cited 2024 Dec 13];17:5623–31. Available from: https://pubmed.ncbi.nlm.nih.gov/39624613/
24. Shu K, Mao S, Zhang Z, Coyle JL, Sejdić E. Recent advancements and future directions in automatic swallowing analysis via videofluoroscopy: A review. Comput Methods Programs Biomed [Internet]. 2025 [cited 2024 Dec 13];259. Available from: https://pubmed.ncbi.nlm.nih.gov/39579458/
25. Kim JM, Kim MS, Choi SY, Lee K, Ryu JS. A deep learning approach to dysphagia-aspiration detecting algorithm through pre- and post-swallowing voice changes. Front Bioeng Biotechnol [Internet]. 2024 [cited 2024 Dec 13];12. Available from: https://pubmed.ncbi.nlm.nih.gov/39157445/
26. Shin B, Lee SH, Kwon K, Lee YJ, Crispe N, Ahn SY et al. Automatic Clinical Assessment of Swallowing Behavior and Diagnosis of Silent Aspiration Using Wireless Multimodal Wearable Electronics. Adv Sci (Weinh) [Internet]. 2024 [cited 2024 Dec 13];11. Available from: https://pubmed.ncbi.nlm.nih.gov/38981027/
27. Hsiao MY, Weng CH, Wang YC, Cheng SH, Wei KC, Tung PY, et al. Deep learning for automatic hyoid tracking in videofluoroscopic swallow studies. Dysphagia. 2023;38:171–80.
28. Donohue C, Mao S, Sejdić E, Coyle JL. Tracking hyoid bone displacement during swallowing without videofluoroscopy using machine learning of vibratory signals. Dysphagia. 2021;36:259–69.
29. Donohue C, Khalifa Y, Perera S, Sejdić E, Coyle JL. How closely do machine ratings of duration of UES opening during videofluoroscopy approximate clinician ratings using Temporal kinematic analyses and the MBSImP? Dysphagia. 2021;36:707–18.
30. Lee JT, Park E, Hwang JM, Jung T, Du, Park D. Machine learning analysis to automatically measure response time of pharyngeal swallowing reflex in videofluoroscopic swallowing study. Sci Rep. 2020;10.
31. Riebold B, Seidl RO, Schauer T. Electromyography- and Bioimpedance-Based Detection of Swallow Onset for the Control of Dysphagia Treatment. Sensors (Basel) [Internet]. 2024 [cited 2024 Dec 13];24. Available from: https://pubmed.ncbi.nlm.nih.gov/39460005/
32. Heo S, Uhm KE, Yuk D, Kwon BM, Yoo B, Kim J et al. Deep learning approach for dysphagia detection by syllable-based speech analysis with daily conversations. Sci Rep [Internet]. 2024 [cited 2024 Dec 13];14. Available from: https://pubmed.ncbi.nlm.nih.gov/39217249/
33. Ariji Y, Gotoh M, Fukuda M, Watanabe S, Nagao T, Katsumata A et al. A preliminary deep learning study on automatic segmentation of contrast-enhanced bolus in videofluorography of swallowing. Sci Rep. 2022;12.
34. Shaheen N, Burdick R, Peña-Chávez R, Ulmschneider C, Yee J, Kurosu A et al. Use of deep learning to segment bolus during videofluoroscopic swallow studies. Biomed Phys Eng Express. 2024;10.
35. Li W, Mao S, Mahoney AS, Petkovic S, Coyle JL, Sejdić E. Deep learning models for bolus segmentation in videofluoroscopic swallow studies. J Real Time Image Process [Internet]. 2024 [cited 2024 Dec 11];21:1–10. Available from: https://link.springer.com/article/https://doi.org/10.1007/s11554-023-01398-1
36. Oliveira AM, Coelho L, Carvalho E, Ferreira-Pinto MJ, Vaz R, Aguiar P. Machine learning for adaptive deep brain stimulation in Parkinson's disease: closing the loop. J Neurol [Internet]. 2023 [cited 2024 Dec 13];270:5313–26. Available from: https://pubmed.ncbi.nlm.nih.gov/37530789/
37. Turgeon S, Lanovaz MJ, Tutorial. Applying machine learning in behavioral research. Perspect Behav Sci. 2020;43:697–723.
38. Pareek A, Martin RK. Editorial commentary: machine learning in medicine requires clinician input, faces barriers, and High-Quality evidence is required to demonstrate improved patient outcomes. Arthrosc - J Arthroscopic Relat Surg. 2022;38:2106–8.
39. Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, et al. Ilastik: interactive machine learning for (bio)image analysis. Nat Methods. 2019;16:1226–32.
40. Mayerl CJ, Edmonds CE, Catchpole EA, Myrla AM, Gould FDH, Bond LE, et al. Sucking versus swallowing coordination, integration, and performance in preterm and term infants. J Appl Physiol. 2020;129:1383–92.
41. Brainerd EL, Baier DB, Gatesy SM, Hedrick TL, Metzger KA, Gilbert SL et al. X-ray reconstruction of moving morphology (XROMM): precision, accuracy and applications in comparative biomechanics research. J Exp Zool A Ecol Genet Physiol [Internet]. 2010 [cited 2024 Dec 11];313:262–79. Available from: https://pubmed.ncbi.nlm.nih.gov/20095029/
42. Gierbolini-Norat EM, Holman SD, Ding P, Bakshi S, German RZ. Variation in the timing and frequency of sucking and swallowing over an entire feeding session in the infant pig Sus scrofa. Dysphagia. 2014;29:475–82.
43. McGrattan KE, McGhee HC, McKelvey KL, Clemmens CS, Hill EG, DeToma A, et al. Capturing infant swallow impairment on videofluoroscopy: timing matters. Pediatr Radiol. 2020;50:199–206.
44. Schneider CA, Rasband WS, Eliceiri KW. NIH image to imageJ: 25 years of image analysis. Nat Methods. 2012;9:671–5.
45. Gould FDH, Mayerl CJ, Adjerid K, Edmonds C, Charles N, Johnson M, et al. Impact of volume and rate of milk delivery on coordination of respiration and swallowing in infant pigs. J Exp Zool Ecol Integr Physiol. 2023;339:1052–8.
46. Bayona HHG, Inamoto Y, Saitoh E, Aihara K, Kobayashi M, Otaka Y. Prediction of Pharyngeal 3D Volume Using 2D Lateral Area Measurements During Swallowing. Dysphagia. 2024.
47. Haubold C, Schiegg M, Kreshuk A, Berg S, Koethe U, Hamprecht FA. Segmenting and tracking multiple dividing targets using Ilastik. Adv Anat Embryol Cell Biol. 2016;219:199–229.
48. Kreshuk A, Zhang C. Machine Learning: Advanced Image Segmentation Using ilastik. Methods in Molecular Biology [Internet]. 2019 [cited 2024 Dec 11];2040:449–63. Available from: https://link.springer.com/protocol/10.1007/978-1-4939-9686-5_21
49. Bayer D, Antonucci S, Müller HP, Saad R, Dupuis L, Rasche V et al. Disruption of orbitofrontal-hypothalamic projections in a murine ALS model and in human patients. Transl Neurodegener. 2021;10.
50. MacO B, Cantoni M, Holtmaat A, Kreshuk A, Hamprecht FA, Knott GW. Semiautomated correlative 3D electron microscopy of in vivo-imaged axons and dendrites. Nat Protoc. 2014;9:1354–66.
51. Peng Q, Chen X, Zhang C, Li W, Liu J, Shi T, et al. Deep learning-based computed tomography image segmentation and volume measurement of intracerebral hemorrhage. Front Neurosci. 2022;16:965680.

52. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging. 1994;13:716–24.
53. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ et al. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index: Scientific Reports. Acad Radiol [Internet]. 2004 [cited 2024 Dec 12];11:178. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC1415224/
54. Tao R, Gavves E, Smeulders AWM. Siamese instance search for tracking. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016;2016-December:1420–9.
55. Takahashi T, Nozaki K, Gonda T, Mameno T, Ikebe K. Deep learning-based detection of dental prostheses and restorations. Sci Rep [Internet]. 2021 [cited 2024 Dec 12];11. Available from: https://pubmed.ncbi.nlm.nih.gov/33479303/
56. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T et al. Fiji: an open-source platform for biological-image analysis. Nature Methods 2012 9:7 [Internet]. 2012 [cited 2024 Dec 11];9:676–82. Available from: https://www.nature.com/articles/nmeth.2019
57. Zhao ZQ, Zheng P, Xu ST, Wu X. Object detection with deep learning: A review. IEEE Trans Neural Netw Learn Syst. 2019;30:3212–32.
58. Liu Y, Liang P, Liang K, Chang Q. Automatic and efficient pneumothorax segmentation from CT images using EFA-Net with feature alignment function. Sci Rep [Internet]. 2023 [cited 2024 Dec 13];13. Available from: https://pubmed.ncbi.nlm.nih.gov/37714871/
59. Adams V, Mathisen B, Baines S, Lazarus C, Callister R. Reliability of measurements of tongue and hand strength and endurance using the Iowa Oral Performance Instrument with healthy adults. Dysphagia [Internet]. 2014 [cited 2024 Dec 13];29:83–95. Available from: https://pubmed.ncbi.nlm.nih.gov/24045852/
60. Pisegna JM, Borders JC, Kaneoka A, Coster WJ, Leonard R, Langmore SE. Reliability of Untrained and Experienced Raters on FEES: Rating Overall Residue is a Simple Task. Dysphagia [Internet]. 2018 [cited 2024 Dec 13];33:645–54. Available from: https://pubmed.ncbi.nlm.nih.gov/29516172/
61. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155–63.
62. Davies BK, Hibbert AP, Roberts SJ, Roberts HC, Tickner JC, Holdsworth G, et al. A machine Learning-Based image segmentation method to quantify in vitro osteoclast culture endpoints. Calcif Tissue Int. 2023;113:437–48.
63. Garcia A, Talavera-Mateo L, Santamaria ME. An automatic method to quantify trichomes in Arabidopsis thaliana. Plant Sci. 2022;323.
64. Li C, Ma X, Deng J, Li J, Liu Y, Zhu X, et al. Machine learning-based automated fungal cell counting under a complicated background with Ilastik and ImageJ. Eng Life Sci. 2021;21:769–77.
65. Nollet JL, Cajander P, Ferris LF, Ramjith J, Omari TI, Savilampi J. Pharyngo-Esophageal modulatory swallow responses to bolus volume and viscosity across time. Laryngoscope. 2022;132:1817–24.
66. Lazarus CL, Logemann JA, Rademaker AW, Kahrilas PJ, Pajak T, Lazar R, et al. Effects of bolus volume, viscosity, and repeated swallows in nonstroke subjects and stroke patients. Arch Phys Med Rehabil. 1993;74:1066–70.
67. Ekberg O, Olsson R, Sundgren-Borgström P. Relation of bolus size and pharyngeal swallow. Dysphagia. 1988;3:69–72.
68. Jestrović I, Coyle JL, Perera S, Sejdić E. Influence of attention and bolus volume on brain organization during swallowing. Brain Struct Funct. 2018;223:955–64.
69. Perlman AL, Schultz JG, VanDaele DJ. Effects of age, gender, bolus volume, and bolus viscosity on oropharyngeal pressure during swallowing. J Appl Physiol. 1993;75:33–7.