Scientific Journal of King Faisal University:
Basic and Applied Sciences

Automated Oropharyngeal Dysphagia Assessment with Mask R-CNN and Kinematic Measures

Zirsha Riaz [1], Aniqa Dilawari [2], Sajid Iqbal [3] and Ahmed A. Alyahya [3]

[1] Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan
[2] Department of Computer Science & Information Technology, University of Home Economics, Lahore, Pakistan
[3] Department of Information Systems, College of Computer Science and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia

LINK	RECEIVED	ACCEPTED	PUBLISHED ONLINE	ASSIGNED TO AN ISSUE
https://doi.org/10.37575/b/med/250007	21/02/2025	21/08/2025	21/08/2025	01/12/2025

NO. OF WORDS	NO. OF PAGES	YEAR	VOLUME	ISSUE
7248	8	2025	26	2

ABSTRACT

Oropharyngeal dysphagia (OD) is characterised by difficulty swallowing liquids or food, significantly affecting an individual's quality of life and potentially leading to serious health issues such as poor nutrition, dehydration and pneumonia. Diagnosis typically involves the use of a video fluoroscopic swallowing study (VFSS), a method that, while effective, is expensive, time-consuming and requires expert interpretation. Recent advancements in artificial intelligence (AI) offer a promising alternative for enhancing dysphagia diagnosis by providing a more efficient and accurate solution. In this paper, we propose an AI-based system for diagnosing OD. The system processes multi-frame image data from VFSS videos using mask region-based convolutional neural network for object detection and segmentation. This method is based on a feature pyramid network and a ResNet101 backbone. It calculates five kinematic measures – ring measurement, hyoid displacement, bolus clearance ratio, pharyngeal constriction ratio and peak esophageal sphincter – to assess the presence or absence of the swallowing disorder. The system was evaluated in real time on 250 patients (150 males and 100 females), classifying them as either with or without dysphagia, and achieved an accuracy of 96.8%. This system is expected to significantly assist clinicians.

KEYWORDS

Artificial intelligence, fluoroscopic data, multi-frame image, oropharyngeal dysphagia, pyramid network, swallowing disorder

CITATION

Riaz, Z., Dilawari, A., Iqbal, S. and Alyahya, A.A. (2025). Automated oropharyngeal dysphagia assessment with mask r-cnn and kinematic measures. *Scientific Journal of King Faisal University: Basic and Applied Sciences*, **26**(2), 28–35. DOI: 10.37575/b/med/250007

1. Introduction

Oropharyngeal dysphagia (OD) refers to difficulty in swallowing a liquid or food from the mouth to the oesophagus (Kim *et al.*, 2024), which can lead to various health complications such as Parkinson's disease (López-Liria *et al.*, 2020), stroke (Labeit *et al.*,2024), or sclerosis (Sadeghi *et al.*, 2024). It adversely affects quality of life and well-being, and may result in social exclusion. Individuals suffering from dysphagia have an increased risk (Langmore *et al.*, 1998; Martin *et al.*, 1994; Jones *et al.*, 2020) of choking, malnutrition, dehydration and even pneumonia. Early identification of this condition is therefore crucial for enabling appropriate treatment planning and reducing adverse outcomes. Accurate diagnosis and timely detection ensure improved treatment results and enhance provider–patient interactions.

Conventional diagnostic techniques include clinical examination, video fluoroscopic swallowing studies (VFSSs) (Kim *et al.*, 2024; Min *et al.*, 2024) and fibreoptic endoscopic evaluation of swallowing (FEES) (Slovik *et al.*, 2025). While these are widely used tools for assessing dysphagia, each presents certain limitations. Clinical examination is highly dependent on the clinician's expertise and may be subjective. VFSS, also known as the modified barium swallow study, involves the use of fluoroscopy to visualise the swallowing process in real time while the patient consumes barium-coated food or liquid. VFSS presents several limitations (Inamoto *et al.*, 2024), including: 1) radiation exposure, 2) being resource-intensive and requiring specialised equipment, 3) the need for specialised personnel, 4) patient discomfort, 5) allergic reactions and 6) a limited assessment duration. FEES (Slovik *et al.*, 2025) involves the insertion of a flexible endoscope through the nasal passage to visualise the pharynx and larynx during swallowing. However, FEES also has drawbacks, such as discomfort, anxiety, limited visualisation, reliance on patient cooperation and the requirement for specialised training.

These limitations underscore the need for alternative methods to assess dysphagia. Recent advancements in artificial intelligence (AI) have opened new avenues for the intelligent detection and management of dysphagia. By leveraging advanced sensor technologies and machine learning algorithms, AI systems can provide more accurate, efficient, cost-effective and accessible diagnostic solutions, offering non-invasive alternatives that ultimately enhance patient care and outcomes (Verma *et al.*, 2025).

In Pakistan, dysphagia remains underdiagnosed due to several factors (Akhtar *et al.*, 2024), including a lack of awareness, limited access to specialised healthcare facilities and financial constraints. Research conducted in the Pakistani context has shown the use of upper gastric endoscopy to assess dysphagia (Kamran *et al.*, 2021; Rashid *et al.*, 2020). AI technology has the potential to address these challenges by enabling remote monitoring and supporting telemedicine, both of which are vital for reaching underserved populations. Machine learning and deep learning algorithms can analyse complex datasets. By learning from large volumes of data, subtle patterns and variations in swallowing activity can be detected, resulting in more accurate diagnosis.

This paper proposes a smart OD detection AI system based on video fluoroscopic data. The key contributions are as follows:

1. The proposed model utilises the AI framework mask region-based convolutional neural network (Mask R-CNN) for object detection. This framework is applied to X-ray images to enhance diagnostics. Mask R-CNN identifies and delineates regions of interest and enables accurate detection of dysphagia. Once the regions are identified, five kinematic measures – ring measurement, hyoid displacement (HD), bolus clearance ratio (BCR), pharyngeal constriction ratio (PCR) and peak oesophageal sphincter (PES_{max}) – are calculated to assess the presence or absence of the swallowing disorder. The system reduces the need for manual interpretation by clinicians and ensures reliable diagnosis.
2. Automating the detection process helps reduce the potential for human error. This is particularly valuable in clinical settings where accuracy is essential for effective and immediate treatment planning.
3. In Pakistan, developing and utilising this AI system for dysphagia detection promotes interdisciplinary collaboration among clinicians, computer scientists and researchers.

Corresponding Author: Sajid Iqbal 00966 594450107, siqbal@kfu.edu.sa

The remainder of the paper is structured as follows: Section 2 discusses related work on dysphagia detection using AI techniques. Section 3 details the proposed model architecture. Section 4 outlines the dataset used. Section 5 presents the results and findings. Sections 6 and 7 address patient privacy, ethical considerations and model limitations. Section 8 concludes the paper with insights drawn from the experiments and suggestions for future research.

2. Related work

OD is a medical condition that affects swallowing. If left untreated, it can lead to serious health problems such as malnutrition, dehydration, choking, coughing during meals and respiratory infections like pneumonia. The European Society for Swallowing Disorders and the European Union Geriatric Medicine Society classify OD as a geriatric syndrome (Sadeghi *et al.*, 2024).

The diagnosis of OD requires careful clinical observation in addition to relevant medical examination. It follows a multi-step approach that begins with screening and clinical history, and culminates in instrument-based evaluation (Sadeghi *et al.*, 2024). The doctor takes a detailed history and examines all reported symptoms and signs. This is followed by a clinical assessment of swallowing to evaluate whether choking or aspiration is present. The VFSS is considered the gold standard. It captures and stores sequential X-ray videos of the passage of the food bolus through the pharynx. Supplementary tests such as FEES, oesophagogastroduodenoscopy and manometry are performed when there is suspicion of a structural or functional abnormality. These tools collectively provide detailed information and visual insight into the swallowing process, aiding in the formulation of an effective intervention. Lately, advances have been made in imaging procedures, with greater reliance on technology. AI offers great value in identifying patterns in data that may be difficult for the human eye to detect. (Fattori *et al.*, 2016) compared VFSS with other modalities such as endoscopy and scintigraphy. Their study illustrates the diagnostic precision of these instruments for OD. Omari *et al.*, (2013) evaluated the impact of different food textures on the swallowing process using a technique called automated impedance manometry (AIM). Deep learning is now widely used in medical imaging. One popular technique is the convolutional neural network (CNN), which recognises patterns in images. Many CNN models exist – such as AlexNet (Krizhevsky *et al.*, 2017), VGGNet (Simonyan and Zisserman, 2014), GoogleNet (Szegedy, 2015), ResNet (He *et al.*, 2016) and DenseNet (Huang *et al.*, 2017) – and are used to classify medical images. Analysing videos, however, is more complex. Unlike single images, videos have a time-based sequence. To work with video data, individual frames are extracted and analysed. These frames are then processed using AI models like recurrent neural networks, 3D-CNNs or MoViNets. A major challenge is that most video analysis still depends on human interpretation. Future research is expected to focus on automating video-based OD detection. Machine learning and deep learning tools can assist doctors in making faster and more accurate decisions. Markus *et al.* (Gugatschka *et al.*, 2024) built a dysphagia risk prediction model using a machine learning method called Random Forest. It analysed the health records of more than 33,000 patients from 2011 to 2019 and included 800 different health-related features. Their system achieved an accuracy of 92.6%. In another study, Markus *et al.* (Martin-Martinez *et al.*, 2023) used feature selection and non-linear methods to create an expert system for predicting OD risk.

CNNs are especially useful for medical images. Jeong *et al.* (2024) created a web-based AI tool using the YOLOv7 model to analyse VFSS videos. Each video was broken down into about 300 frames. The system labelled them as oral, pharyngeal or oesophageal phases. It could also determine whether dysphagia involved penetration or aspiration. Their model achieved accuracy ranging from 0.79 to 0.96. Girardi *et al.*, (2023) reviewed how AI is used to study VFSS videos. They found that CNNs improve accuracy and help speech pathologists better understand swallowing problems. Reddy *et al.* (2023) used 2D-CNNs, LSTM networks and 3D-CNNs to detect aspiration in VFSS videos. They compared the performance of different models. Jeong *et al.* (2023) also developed a tool to automate the timing of swallowing phases in VFSS videos. Their ResNet3D model outperformed models such as VGG and I3D, offering faster and more reliable results. In addition to computer vision techniques applied to VFSS data, researchers have explored other approaches such as swallowing sound analysis (Dudik *et al.*, 2018; Miyagi *et al.*, 2020), monitoring neck vibrations and using wearable sensor devices (O'Brien *et al.*, 2021; Rafeedi *et al.*, 2023) to detect swallowing disorders. These techniques enable continuous real-time monitoring of swallowing function.

As far as is known, no existing systems use AI to detect OD specifically through VFSS video footage in the way proposed here. Although some paradigms exist for medical image analysis and swallowing assessment, they do not address the same multifaceted combination of objectives, input types and kinematic measurements. For this reason, fair comparisons with other models under the same conditions could not be established. Moreover, no statistical significance was tested against other methods, as no comparable baselines specifically designed for this solution currently exist. The proposed approach is therefore novel and sets the groundwork for future benchmarks and comparative analyses.

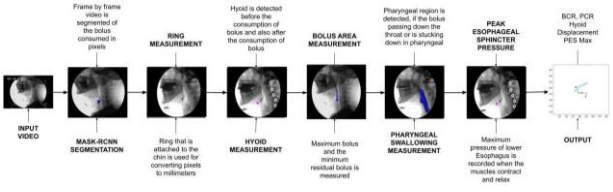
3. Methodology

The proposed AI-based smart tool is designed to assist in the evaluation of OD. It primarily uses video fluoroscopy to assess swallowing function. A comprehensive analysis of swallowing kinematics is performed using the following measures to capture both spatial and temporal aspects of swallowing:

- Ring measurement
- Hyoid displacement (Molfenter and Steele, 2013)
- Bolus clearance ratio (Leonard *et al.*, 2023)
- Pharyngeal constriction ratio (Stokely *et al.*, 2015)
- Peak oesophageal sphincter (Leonard *et al.*, 2000)

The workflow of the proposed system is illustrated in Figure 1.

Figure 1: Flow diagram of the proposed system. It begins with the input X-ray video of a person consuming syrup, proceeding to the final kinematic calculations.



The following steps outline the process of determining the presence or absence of dysphagia using video fluoroscopic data.

3.1. Video Acquisition and Frame Extraction:

The process begins with the acquisition of an X-ray video capturing a person consuming syrup, serving as the initial input for analysis. The most commonly used syrup for dysphagia assessment is nectar-thick syrup. This consistency is frequently selected because it is slightly thicker than water yet still easily drinkable, making it a practical and effective starting point for evaluating swallowing difficulties. Nectar-thick liquids are often preferred due to their balance between ease of swallowing and reduced risk of aspiration compared to thin liquids.

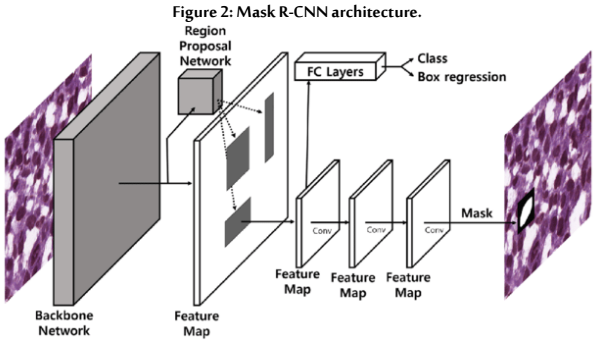
The video then undergoes frame extraction to dissect each stage of the swallowing process. Specific attention is directed towards the

oropharyngeal region, where swallowing occurs. Using segmentation techniques, individual frames are processed to isolate and delineate the bolus being swallowed, as well as any residual bolus present in the throat. To perform this segmentation, Mask R-CNN is employed, taking advantage of its capabilities in object identification and delineation within images. The model used for segmentation is Mask R-CNN.

3.2. Mask R-CNN Segmentation:

The Mask R-CNN (Bharati and Pramanik, 2020) architecture applied in this research was trained with three times (3×) the default number of iterations to improve convergence and performance. The implementation utilises a ResNet-101 backbone integrated with a feature pyramid network (FPN) for efficient multi-scale feature extraction. Mask R-CNN is designed for instance segmentation and object detection, automatically identifying objects in input images and generating corresponding masks. The base architecture is illustrated in Figure 2, adapted from (He *et al.*, 2017).

Figure 2: Mask R-CNN architecture.



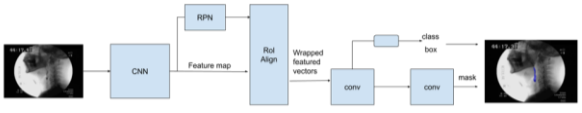
Mask R-CNN extends Faster R-CNN by incorporating a mask prediction branch, resulting in a robust deep learning model capable of object detection and pixel-level segmentation. It uses a Region Proposal Network (RPN) to generate candidate object regions, while separate branches (heads) handle classification, bounding box regression and mask prediction. The RoI Align operation ensures accurate spatial alignment between features and input data.

Training involves a combination of loss functions: binary cross-entropy for masks, smooth L1 loss for bounding boxes and cross-entropy for classification. Key hyperparameters include momentum (typically 0.9), batch size (2–16) and learning rate (e.g. 0.001). The dataset is generally divided into 70%–80% for training and 20%–30% for validation or testing.

The complexity of Mask R-CNN arises from its multi-branch architecture and deep backbone network, particularly when working with high-resolution images and multiple object classes. Training and inference require significant computational resources, including high-performance GPUs with large memory capacity. The number of region proposals, feature map sizes and backbone depth directly affect computational cost.

Moreover, the model's flexibility lies in its adaptability to varying input conditions. Through FPN and RoI Align, it effectively handles objects of different scales and aspect ratios. Its performance is also shaped by runtime factors such as data augmentation, pretraining and dataset diversity. The model further adjusts dynamically to changes in training conditions, including batch size and learning rate schedules.

Figure 3: Process starting with capturing an X-ray video of a person consuming syrup and ending with the final calculations.



3.2.1. Neural Network Architecture

The foundation of this model is a deep convolutional network known as ResNet-101 (He *et al.*, 2016). This network collects high-level features from the input image and creates a multi-scale feature map using an FPN. This makes it robust in detecting objects of varying sizes. FPN generates rich feature maps at various resolutions by integrating high-resolution features from earlier layers with low-resolution ones. This combination provides detailed spatial information and strong semantic context, which improves the network's accuracy in detecting objects of different sizes. In simple terms, the FPN combines the detailed information that the ResNet-101 backbone obtains from the input image at multiple scales, allowing the model to recognise and detect objects of any size.

ResNet-101 is a deep CNN with 101 layers. Its architecture includes an input layer, an initial convolutional layer, a max-pooling layer, residual blocks, fully connected layers and an output layer.

ResNet-101's input layer requires images sized at 224 × 224 pixels. The initial convolutional and pooling layers are the first stages of the network and are responsible for capturing basic features such as edges and textures. This convolutional layer applies 64 filters of size 7 × 7 with a stride of 2, reducing the spatial dimensions of the input. A 3 × 3 max-pooling layer, with a stride of 2, further reduces the spatial dimensions and helps make the model resistant to small shifts.

ResNet-101 includes a total of 33 bottleneck residual blocks. These blocks are organised into different stages that reduce the number of channels to lower dimensionality and facilitate the training of deep networks. After passing through the residual blocks, the feature map is average-pooled to reduce its dimensions down to 1 × 1. This is followed by a connected (dense) layer that links the pooled features to the number of output classes. A softmax activation function then generates class probabilities.

For dysphagia detection, this model helps by extracting detailed features from video frames to identify abnormalities in the swallowing process. It has the potential to be used in real-time systems to provide immediate feedback, assisting clinicians in making timely decisions.

3.3. Research Metrics:

Using sophisticated tools, several anatomical and functional parameters were evaluated to characterise and quantify the mechanics of swallowing. These measurements offered a thorough understanding of the swallowing process and the challenges associated with it, particularly in diagnosing dysphagia. Standardised reference objects, anatomical movement tracking and ratio calculations were used to assess oesophageal function, pharyngeal constriction and bolus clearance. Each technique was designed to extract key data from the X-ray videos, allowing for accurate and reliable evaluation of swallowing efficiency. The metrics are defined as follows:

3.3.1. Ring Measurement

In the X-ray video analysis, a reference object – usually a ring – is fixed to the subject's chin to precisely measure physical distances. The X-ray images are standardised using this reference, allowing pixel values to be converted into centimetres. The pixel distance is translated into real-world measurements using the known diameter of the ring. Equation 1 gives the pixel-to-centimetre formula for ring measurement:

Conversion Factor = Pixel Diameter of the Ring / Actual Diameter of the Ring (cm)

Real-World Distance (cm) = Pixel Distance × Conversion Factor (1)

3.3.2. Hyoid Displacement

To assess swallowing mechanics, the displacement of the hyoid region, a crucial anatomical marker, is examined. The initial position and displacement of the hyoid bone are monitored during and after the bolus is swallowed. The HD percentage is calculated based on the degree of movement observed during swallowing. The magnitude and direction of hyoid motion are determined by tracking its position across multiple frames. Equation 2 provides the formula for HD:

HD = initial hyoid position − displaced hyoid position (2)

3.3.3. Bolus Clearance Ratio Measurement

BCR is determined by comparing the area of bolus residue left in the throat with the total area of bolus successfully swallowed. This ratio measures bolus clearance during swallowing and is an important tool in dysphagia diagnosis. A lower BCR may indicate potential swallowing issues, increasing the risk of aspiration or other complications. A higher BCR reflects a more efficient swallowing process. The percentage value of residual bolus is defined in Equation 3.

BCR = (area of residue bolus left in throat / total area of bolus) * 100 (3)

3.3.4. Pharyngeal Constriction Ratio Measurement

The PCR is calculated after the BCR. Pharyngeal contraction occurs when the bolus moves towards the oesophagus, which is measured using the PCR. This is calculated by examining how the pharyngeal region changes both prior to and during swallowing.

A higher PCR value may imply less constriction, possibly indicating dysphagia. A lower PCR value implies better pharyngeal constriction and more efficient bolus propulsion. Equation 4 shows the formula for PCR:

PCR = ((total area of bolus − area of residual bolus) / total area of bolus) * 100 (4)

3.3.5. Peak Oesophageal Sphincter Measurement

PES pressure was determined using sensors positioned in the lower oesophageal region. These sensors recorded pressure both before and during the swallowing of the bolus. Peak pressure was identified when the sensors detected the maximum pressure exerted in the lower oesophagus. The pressure readings, initially measured in millimetres of mercury (mmHg), were converted to centimetres.

The greatest opening of the upper oesophageal sphincter (UES) during swallowing is measured by the PES_{max} opening. This parameter is essential for the bolus to move from the pharynx into the oesophagus. The effectiveness of the swallowing mechanism was evaluated by examining the frames to determine the maximum UES opening.

Minimum PES_{max} values may indicate compromised UES function, which increases the risk of aspiration and contributes to dysphagia. The PES_{max} formula is shown in Equation 5:

PESmax = pressure in mmHg/10 (5)

- DATASETS

In this research, 250 VFSS cases were randomly selected from 1,000 cases. The patients were between 24 and 85 years of age, including 150 males and 100 females. Dysphagia was diagnosed in 130 patients through VFSS video readings by medical doctors. The data were collected from specialised medical centres, with patient confidentiality carefully maintained. Two physicians with more than five years of experience participated in conducting the VFSS examinations and validating the results produced by the proposed system.

The dataset used consists of videos showing individuals swallowing a bolus. Frames from these videos were extracted to distinguish cases with dysphagia from those without any signs. This dataset played a key role in training the model, which focuses on analysing bolus clearance during swallowing.

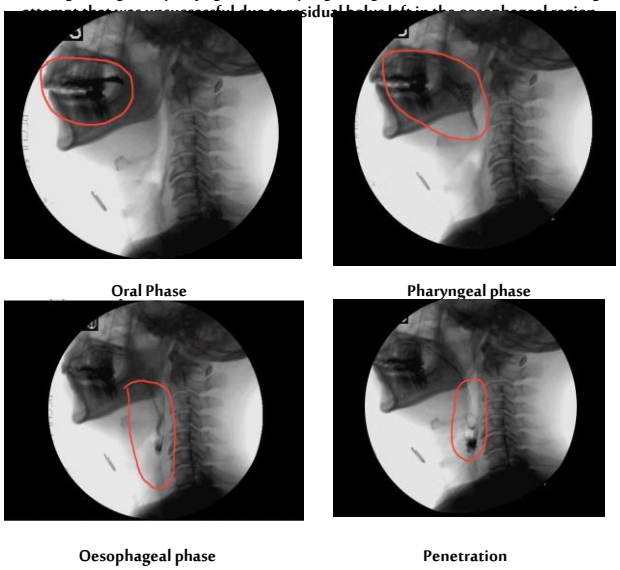
4. Experimental Results

This section discusses two cases in detail. Case 1 illustrates the processing steps used to demonstrate the presence of dysphagia in VFSS data, while Case 2 depicts the absence of the swallowing disorder.

4.1. Case 1 – Presence of Dysphagia:

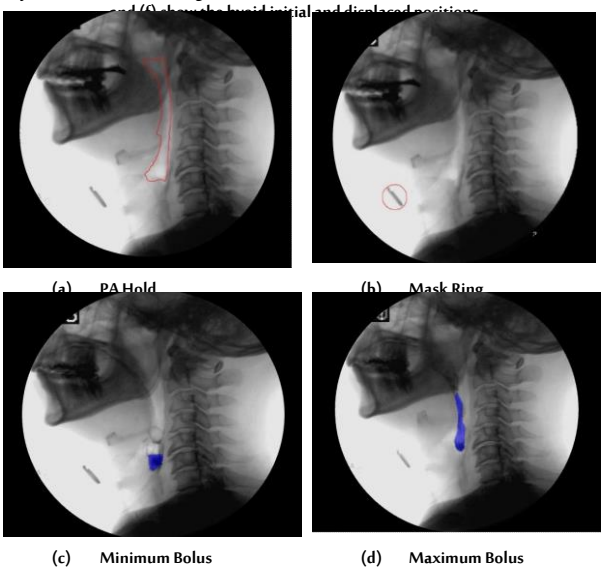
This is a case where a person shows signs of dysphagia, as demonstrated through images of the bolus passing through the oesophagus and the residue left behind. Figure 4 shows the labelling of the bolus as it moves from the mouth to the pharyngeal phase, frame by frame, which was later approved by the medical officer.

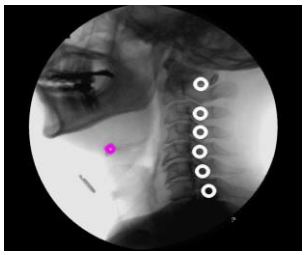
Figure 4: Images (a)–(d) show the patient holding a bolus in the mouth and then consuming it. The bolus passes down the throat, but a residual amount is left in the pharyngeal region, indicating the presence of dysphagia. (a) shows the patient holding bolus/syrup in the mouth; (b) shows the swallowing of bolus; (c) shows the bolus moving through the pharyngeal to oesophageal region; and (d) shows a swallowing



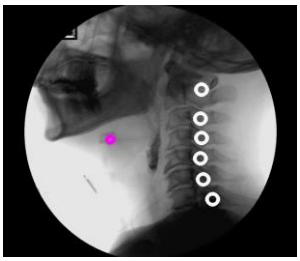
The step-by-step outcome of the proposed AI system showing how the bolus moves from the mouth to the oesophagus is illustrated in Figure 5.

Figure 5: Images (a)–(f) show the area outlining measures to capture swallowing. (a) Posterior–anterior (PA) hold shows that the patient is positioned so that the fluoroscopy images can be taken from the back (posterior) to the front (anterior) of the body; (b) shows the mask ring; (c) and (d) show the minimum and maximum bolus; (e)





(e) Hyoid Initial



(f) Hyoid Displaced

The discovery of a residual bolus in the pharyngeal area despite partial passage down the throat in this example emphasises the presence of dysphagia in the subject. The amount of this residual bolus relative to the total bolus swallowed is calculated, shedding light on the severity of the swallowing impediment. Furthermore, the evaluation measures the quantity of bolus still present, particularly in the pharyngeal area, which suggests compromised clearance mechanisms at this critical phase of swallowing.

These metrics provide an extensive assessment of swallowing ability, which is essential for identifying and treating dysphagia. More can be learned about biomechanical anomalies and swallowing efficiency by measuring HD. On the other hand, measuring the highest pressure in the lower oesophageal area provides insight into physiological factors, such as the health of the oesophageal sphincter and possible causes of dysphagia symptoms. When combined, these measures offer valuable information that can help pinpoint deficiencies and guide focused interventions for managing dysphagia.

The measures calculated to capture swallowing are as follows:

BCR = 50.98%

PCR = 11.57%

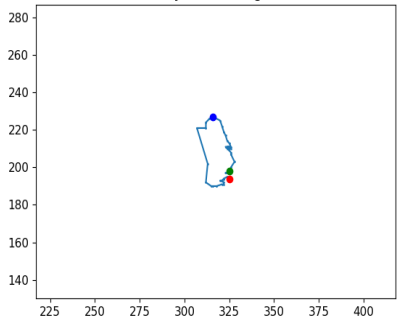
HD = 1.55 cm

PESmax = 0.61 cm

These measures indicate the percentage of the bolus that has been successfully cleared from the throat, represented by the BCR. Similarly, the PCR expresses the extent of pharyngeal constriction during swallowing. The pharyngeal region is contrasted before and during the swallow. HD reflects the actual physical movement of the hyoid bone. PES_{max} is measured in centimetres to represent the maximum physical opening of the UES during swallowing.

The final output of the proposed AI system is shown in the graph in Figure 6.

Figure 6. Graph showing the movement of the hyoid during the swallowing process.

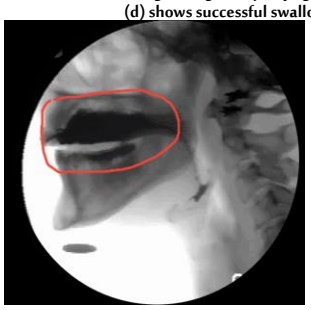


The graph shows the movement of the hyoid during the swallowing process. The green point indicates the starting position, red indicates the lowest point, and blue indicates the highest point. It was generated using the x- and y-coordinates of the hyoid movement during the process. The purpose was to observe how much the hyoid was displaced throughout the entire swallowing event.

4.2. Case 2 – Absence of Dysphagia:

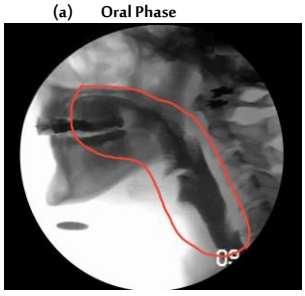
This is the case where a person shows no signs of dysphagia, as shown through the images – the bolus passes through the oesophagus, and no residue is left behind. Figure 7 shows the labelling of the bolus as it moves from the mouth to the pharyngeal phase, frame by frame, which was later approved by the medical officer.

Figure 7: Images (a)–(d) show the patient holding the bolus in the mouth and then consuming it. The bolus passes down the throat without getting stuck or leaving residue in the pharyngeal region, showing the absence of dysphagia. (a) shows the patient holding bolus/syrup in the mouth, (b) shows the swallowing of the bolus, (c) shows the bolus moving through the pharyngeal region to the oesophageal region, and (d) shows successful swallowing with no residual.

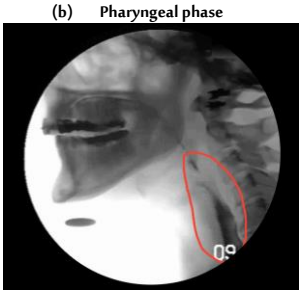


(a) Oral Phase

(b) Pharyngeal phase



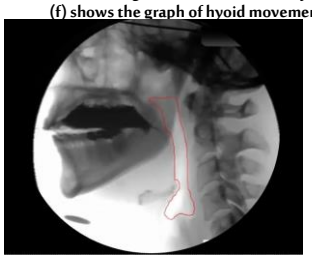
(c) Oesophageal phase



(d) Penetration

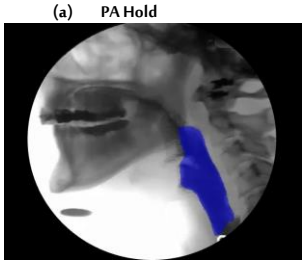
The area outlining the step-by-step outcome of our proposed AI system, which shows the bolus moving from the mouth to the oesophagus, is presented in Figure 8.

Figure 8: Images (a)–(f) show the steps used to capture swallowing. (a) Posterior-Anterior (PA) hold shows the patient positioned for fluoroscopy images taken from the back (posterior) to the front (anterior) of the body; (b) shows maximum bolus; (c) shows the mask ring; (d) and (e) show the hyoid in initial and displaced positions; and (f) shows the graph of hyoid movement during the swallowing process.

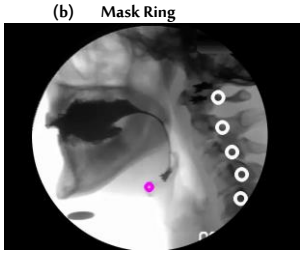


(a) PA Hold

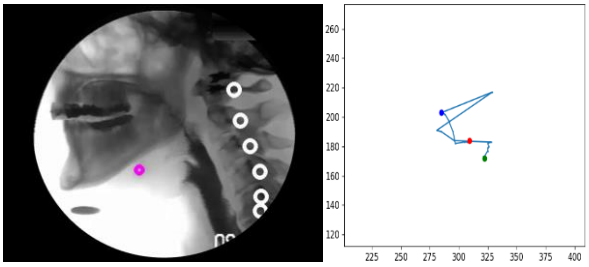
(b) Mask Ring



(c) Maximum Bolus



(d) Hyoid Initial



(e) Hyoid Displaced (f) Graph

The measures calculated to capture swallowing are as follows:

BCR = 0.1%

PCR = 0.0%

HD = 1.64 cm

PES_{max} = 1.37 cm

The findings show that the patient does not have dysphagia. There is no material remaining in the pharynx, and the bolus travels down the throat easily and smoothly. This indicates that the swallowing function is effective and free from any obstruction or impairment, giving confidence in the individual's swallowing ability.

5. Performance Measurement Method

To calculate the accuracy of dysphagia classification, the true labels and predicted labels for each VFSS instance were determined. True labels indicate whether dysphagia is present (positive) or absent (negative) based on expert diagnosis, while predicted labels are the output from our proposed AI system. A confusion matrix was constructed, including the following:

- True positives (TP): Cases where the system correctly predicts the presence of dysphagia.
- True negatives (TN): Cases where the system correctly predicts the absence of dysphagia.
- False positives (FP): Cases where the system incorrectly predicts the presence of dysphagia.
- False negatives (FN): Cases where the system incorrectly predicts the absence of dysphagia.

Accuracy, recall, and F1-score were calculated using Equations (5), (6), and (7), respectively. These metrics provide insight into how effectively the system classifies the presence or absence of dysphagia.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1-score = \frac{2*precision*recall}{precision+recall} \quad (7)$$

Based on Equation (5), accuracy was calculated as:

$$\text{Accuracy} = \frac{242}{250} * 100 = 96.8\%$$

Substituting values into Equation (6), recall was calculated as:

$$\text{Recall} = \frac{122}{122+4} = 0.968$$

Substituting into Equation (7), the F1-score was calculated as:

$$\text{F1-score} = \frac{2*0.968*0.968}{0.968+0.968} = 0.968$$

In our study, the average processing time for detecting dysphagia using the proposed system for an entire VFSS video ranged from 1 to 1.5 minutes. One factor contributing to this variation was the hardware specifications of the PC used. Table 1 presents the accuracy scores reported by state-of-the-art methodologies used for dysphagia detection.

Table 1: State of The Art Approaches in Dysphagia Detection

Methods	Sample	Findings
ResNet3D variant (Jeong *et al.*, 2023)	547 VFSS video clips	0.901
2D-CNNs with short temporal windows (Bandini and Steele, 2021)	78 participants	0.93
Three CNNs (Iida *et al.*, 2023) Simple layer Multiple layer Modified LeNet	130 participants	 0.973 0.890 0.95
Deep CNN using U-Net (Lee *et al.*, 2020)	106 participants with dysphagia	0.932 (video files)
CNN (MobileNet with fine-tuning) (Kim *et al.*, 2022)	190 participants	0.936
Mask RCNN, feature pyramid network and ResNet101 (present study)	250 participants	0.968

This study has certain limitations: the system was developed based on patients from a single medical centre, necessitating additional validation tests from other hospitals. Furthermore, VFSS has limitations related to poor inter- and intra-rater reliability. Inexperienced medical examiners might misinterpret results due to the complex anatomy of the human neck or poor video quality, which could result from uncooperative patients. Future research should aim to advance the study by considering the characteristics of disease-specific swallowing disorders.

6. Patient Privacy and Ethical Approvals in Medical Research

In medical research, especially studies involving AI and sensitive data such as VFSSs, ensuring patient privacy and obtaining ethical approval are critical. To maintain the privacy of the patient, all personally identifiable information requires anonymisation or pseudonymisation prior to use. The researchers comply with relevant laws such as HIPAA or GDPR, which stipulate the deletion of names, contact information and other identifying information. Confidentiality is also maintained through secure data storage, encryption and restricted access to authorised personnel. Patient data must be used only after an institutional review board or ethics committee grants the necessary approval. These entities assess the study's data handling, the design of the study and informed consent processes, and then develop an opinion on the ethical validity of the study. Participants must be effectively informed about how their data are utilised while agreeing to provide consent without duress. Ethical and patient privacy compliance enhances the trustworthiness of the study and the protection of human rights, notably in innovations involving AI and healthcare technologies.

7. Misclassification Cases and Model Limitations

Even though our AI-based system for detecting OD showed accurate results, some issues stood out, such as the model's limitations. Misclassification or incomplete classification — in our case, false negative scenarios — are among the biggest concerns. False negatives refer to cases where the system did not detect dysphagia in patients who were clinically diagnosed and confirmed to have the issue. These cases pose a problem, as missing an emerging health disorder, such as swallowing difficulties, can result in serious complications like aspiration pneumonia or malnutrition. On closer analysis, numerous problematic areas were located among the video files, some of which were suspected to contain one or two binders supplying only video streams. Matching videos to descriptions posed a challenge, even for seasoned professionals. Incomplete collaboration, such as not fully swallowing, also contributed to these hurdles. In other instances, the

overlap of certain bone structures and anatomical features may have interfered with the tracking of key kinematic elements, including HD and pharyngeal constriction. From the results, we realise that although the model performed to our satisfaction, there is a need for improved sensitivity in borderline cases. Additional steps that could augment the model's potential, such as incorporating patient history or audio data, may also include increasing the frame rate of animations to ensure that relevant features are more recognisable in dynamic images. Focus should also be placed on the patient's clinical information to improve preprocessing protocols.

8. Conclusion and Future Work

This study presents an AI-based detection system that interprets VFSS videos to identify patterns and anomalies that may not be visible to the human eye. The processing and interpretation of VFSS are adept at recognising visual patterns in medical images, allowing the detection of dysphagia-related abnormalities in swallowing mechanics. To achieve this, an AI-based system was developed for diagnosing OD. The system processes multi-frame image data from VFSS videos using Mask R-CNN for object detection and segmentation, which is based on FPN and ResNet101 architecture. It then calculates five kinematic measures — ring measurement, HD, BCR, PCR and PES_{max} — to assess the presence or absence of the swallowing disorder. A total of 250 patients were screened automatically in real time and compared with clinician assessments. An accuracy of 96.8% was achieved through this proposed system. This algorithm proved to be an excellent tool for classifying patients with or without dysphagia. The software has potential utility in regions where medical resources are limited, such as Pakistan. Future developments include the preparation of additional datasets and testing the model using advanced AI systems capable of generating clinical reports to support physicians.

Data Availability Statement

The data that support the findings of this study are available within the article and/or its supplementary materials. The author can provide further details with reasonable requests.

Acknowledgement

The author extends sincere appreciation to King Faisal University, Saudi Arabia, for its support and encouragement during this research.

Funding

This work was supported by the Deanship of ScientificResearch, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU252544].

Conflict of Interest

The author states that there is no conflict of interest. The author declares no financial or non-financial competing interests.

Biographies

Zirsha Riaz

Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan, 0303-4843811, zirshariaz40@gmail.com

Zirsha is a computer engineer driven by a passion for leveraging technology across diverse domains. Her research interests span medical imaging, chatbot development, optical character recognition (OCR), and virtual try-on systems. She holds a bachelor's degree in electrical engineering with a concentration in Computer Engineering. Currently, she is focused on exploring the convergence of these technologies to develop innovative and impactful solutions.

ORCID: 0009-0000-6207-6412

Aniqa Dilawari

Department of Computer Science & Information Technology, University of Home Economics, Lahore, Pakistan, 0322-4958179, aniqa.dilawari@gmail.com

Aniqa is a renowned computer science and AI expert, currently leading the Department of Computer Science & IT at the University of Home Economics, Lahore. Her research spans image processing, NLP, medical imaging, pattern recognition, and deep learning. With significant contributions to AI research projects, she is recognized as a prominent figure in the field, committed to advancing innovative technologies through interdisciplinary research and academic leadership.

ORCID: 0000-0003-4821-7199

Sajid Iqbal

Department of Information Systems, College of Computer Science and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia, +966 594450107, siqbal@kfu.edu.sa

Sajid is an Assistant Professor of Information Systems at King Faisal University, Al-Ahsa, Saudi Arabia. He holds a Ph.D. in Computer Science with a focus on Artificial Intelligence. His research interests include machine learning, data science, and digital image processing, with over 2,000 citations on Google Scholar. Prior to this, he served as an Assistant Professor at Bahauddin Zakariya University (BZU), Multan, Pakistan.

ORCID: 0000-0002-8464-2275

Ahmed A. Alyahya

Department of Information Systems, College of Computer Science and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia, +966 500504234, aaalyahya@kfu.edu.sa

Ahmed is an assistant professor at King Faisal University in Saudi Arabia, specializing in Information Security. His research interests include building an Information Security Culture, cybersecurity awareness, social engineering, network and web application security, risk management, digital forensics, and penetration testing.

ORCID: 0009-0003-7877-986X

References

Akhtar, R.N., Behn, N. and Morgan, S. (2024). Understanding dysphagia care in Pakistan: A survey of current speech language therapy practice. *Dysphagia*, **39**(3), 484–94. DOI: 10.1007/s00455-023-10633-7.

Bandini, A. and Steele, C.M. (2021). The effect of time on the automated detection of the pharyngeal phase in videofluoroscopic swallowing studies. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. DOI: 10.1109/EMBC46164.2021.9629562.

Bharati, P. and Pramanik, A. (2020). Deep learning techniques R-CNN to mask R-CNN: A survey. In: A., Das, J., Nayak, B., Naik, S., Pati, D., Pelusi (eds) *Computational Intelligence in Pattern Recognition. Advances in Intelligent Systems and Computing*. Springer, Singapore. DOI: 10.1007/978-981-13-9042-5_56.

Dudik, J.M., Kurosu, A., Coyle, J.L. and Sejdić, E. (2018). Dysphagia and its effects on swallowing sounds and vibrations in adults. *Biomedical Engineering Online*, **17**(1), 69. DOI: 10.1186/s12938-018-0501-9.

Fattori, B., Giusti, P., Mancini, V., Grosso, M., Barillari, M.R., Bastiani, L. and Nacci, A. (2016). Comparison between videofluoroscopy, fiberoptic endoscopy and scintigraphy for diagnosis of oro-pharyngeal dysphagia. *Acta Otorhinolaryngologica Italica*, **36**(5), 395. DOI: 10.14639/0392-100X-829.

Girardi, A.M., Cardell, E.A. and Bird, S.P. (2023). Artificial intelligence in the interpretation of videofluoroscopic swallow studies: implications and advances for speech–language pathologists. *Big Data and Cognitive Computing*, **7**(4), 178. DOI: 10.3390/bdcc7040178.

Gugatschka, M., Egger, N.M., Haspl, K., Hortobagyi, D., Jauk, S., Feiner, M. and Kramer, D. (2024). Clinical evaluation of a machine learning-based dysphagia risk prediction tool. *European Archives of Oto-Rhino-Laryngology*, **281**(8), 4379–84. DOI: 10.1007/s00405-024-08678-x.

He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017). Mask r-cnn. In: *Proceedings of the IEEE international Conference on Computer Vision*. DOI: 10.1109/iccv.2017.322.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* . DOI: 10.1109/cvpr.2016.90.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/CVPR.2017.243.

Iida, Y., Näppi, J., Kitano, T., Hironaka, T., Katsumata, A. and Yoshida, H. (2023). Detection of aspiration from images of a videofluoroscopic swallowing study adopting deep learning. *Oral Radiology*, **39**(3), 553–562. DOI: 10.1007/s11282-023-00669-8.

Inamoto, Y., Ueha, R. and Gonzalez-Fernandez, M. (2024). Use of CT for dysphagia evaluation: Advantages and disadvantages in the study of swallowing. *Current Physical Medicine and Rehabilitation Reports*, **12**(3), 250–255. DOI: 10.1007/s40141-024-00451-9.

Jeong, C.W., Lee, C.S., Lim, D.W., Noh, S.H., Moon, H.K., Park, C. and Kim, M.S. (2024). The development of an artificial intelligence video analysis-based web application to diagnose oropharyngeal Dysphagia: A pilot study. *Brain Sciences*, **14**(6), 546. DOI: 10.2196/preprints.53738.

Jeong, S.Y., Kim, J.M., Park, J.E., Baek, S.J. and Yang, S.N. (2023). Application of deep learning technology for temporal analysis of videofluoroscopic swallowing studies. *Scientific Reports*, **13**(1), 17522. DOI: 10.21203/rs.3.rs-2311543/v1.

Jones, C.A., Colletti, C.M. and Ding, M.C. (2020). Post-stroke dysphagia: recent insights and unanswered questions. *Current neurology and Neuroscience Reports*, **20**(12), 61. DOI: 10.1007/bf00262751.

Kamran, M., Fawwad, A., Haider, S.I., Hussain, T. and Ahmed, J. (2021). Upper gastrointestinal endoscopy; A study from a rural population of Sindh, Pakistan. *Pakistan Journal of Medical Sciences*, **37**(1), 9. 10.12669/pjms.37.1.3297.

Kim, H.T., Min, H.J. and Kim, H.J. (2025). Reliability and validity analyses of the practical assessment of dysphagia test in stroke. *Dysphagia*, **40**(1), 110–7. DOI: 10.1007/s00455-024-10708-z.

Kim, J.K., Choo, Y.J., Choi, G.S., Shin, H., Chang, M.C. and Park, D. (2022). Deep learning analysis to automatically detect the presence of penetration or aspiration in videofluoroscopic swallowing study. *Journal of Korean Medical Science*, **37**(6), n/a. DOI: 10.3346/jkms.2022.37.e42.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, **60**(6), 84-90. DOI: 10.1145/3065386.

Labeit, B., Michou, E., Trapl-Grundschober, M., Suntrup-Krueger, S., Muhle, P., Bath, P.M. and Dziewas, R. (2024). Dysphagia after stroke: research advances in treatment interventions. *The Lancet Neurology*, **23**(4), 418–28. DOI: 10.1016/s1474-4422(24)00053-x.

Langmore, S.E., Terpenning, M.S., Schork, A., Chen, Y., Murray, J.T., Lopatin, D. and Loesche, W.J. (1998). Predictors of aspiration pneumonia: how important is dysphagia?. *Dysphagia*, **13**(2), 69–81. DOI: 10.1007/pl00009559.

Lee, S.J., Ko, J.Y., Kim, H.I. and Choi, S.I. (2020). Automatic detection of airway invasion from videofluoroscopy via deep learning technology. *Applied Sciences*, **10**(18), 6179. DOI: 10.3390/app10186179.

Leonard, R., Miles, A. and Allen, J. (2023). Bolus clearance ratio elevated in patients with neurogenic dysphagia compared with healthy adults: A measure of pharyngeal efficiency. *American Journal of Speech-Language Pathology*, **32**(1), 107–14. DOI: 10.1044/2022_ajslp-22-00199.

Leonard, R.J., Kendall, K.A., McKenzie, S., Gonçalves, M.I. and Walker, A. (2000). Structural displacements in normal swallowing: A videofluoroscopic study. *Dysphagia*, **15**(3), 146–52. DOI: 10.1007/s004550010017.

López-Liria, R., Parra-Egeda, J., Vega-Ramírez, F.A., Aguilar-Parra, J.M., Trigueros-Ramos, R., Morales-Gázquez, M.J. and Rocamora-Pérez, P. (2020). Treatment of dysphagia in Parkinson's disease: a systematic review. *International Journal of Environmental Research and Public Health*, **17**(11), 4104. DOI: 10.3390/ijerph17114104.

Martin, B.J., Corlew, M.M., Wood, H., Olson, D., Golopol, L.A., Wingo, M. and Kirmani, N. (1994). The association of swallowing dysfunction and aspiration pneumonia. *Dysphagia*, **9**(1), 1–6. DOI: 10.1007/BF00262751.

Martin-Martinez, A., Miró, J., Amadó, C., Ruz, F., Ruiz, A., Ortega, O. and Clave, P. (2023). A systematic and universal artificial intelligence screening method for oropharyngeal dysphagia: improving diagnosis through risk management. *Dysphagia*, **38**(4), 1224–37. DOI: 10.1007/s00455-022-10547-w.

Min, I., Woo, H., Kim, J.Y., Kim, T.L., Lee, Y., Chang, W.K. and Seo, H.G. (2024). Inter-rater and intra-rater reliability of the videofluoroscopic dysphagia scale with the standardized protocol. *Dysphagia*, **39**(1), 43–51. DOI: 10.1007/s00455-023-10590-1.

Miyagi, S., Sugiyama, S., Kozawa, K., Moritani, S., Sakamoto, S.I. and Sakai, O. (2020, April). Classifying dysphagic swallowing sounds with support vector machines. In: *Healthcare*. MDPI. DOI: 10.3390/healthcare8020103.

Molfenter, S.M. and Steele, C.M. (2013). Variation in temporal measures of swallowing: sex and volume effects. *Dysphagia*, **28**(2), 226–33. DOI: 10.1007/s00455-012-9437-6.

O'Brien, M.K., Botonis, O.K., Larkin, E., Carpenter, J., Martin-Harris, B., Maronati, R. and Jayaraman, A. (2021). Advanced machine learning tools to monitor biomarkers of dysphagia: a wearable sensor proof-of-concept study. *Digital Biomarkers*, **5**(2), 167–75. DOI: 10.1159/000517144.

Omari, T.I., Dejaeger, E., Tack, J., Van Beckevoort, D. and Rommel, N. (2013). Effect of bolus volume and viscosity on pharyngeal automated impedance manometry variables derived for broad dysphagia patients. *Dysphagia*, **28**(2), 146–52. DOI: 10.1007/s00455-012-9423-z.

Rafeedi, T., Abdal, A., Polat, B., Hutcheson, K.A., Shinn, E.H. and Lipomi, D.J. (2023). Wearable, epidermal devices for assessment of swallowing function. *Npj Flexible Electronics*, **7**(1), 52. DOI: 10.1038/s41528-023-00286-9.

Rashid, H., Bakht, K., Arslan, A. and Ahmad, A. (2020). Endoscopic Findings and Their Association With Gender, Age and Duration of Symptoms in Patients With Dysphagia. *Cureus*, **12**(10),n/a. DOI: 10.7759/cureus.11264.

Reddy, C.S., Park, E. and Lee, J.T. (2023). Comparative analysis of deep learning architectures for penetration and aspiration detection in videofluoroscopic swallowing studies. *IEEE Access*, **11**(n/a), 102843–102851. DOI: 10.1109/access.2023.3315342.

Sadeghi, Z., Afshar, M., Memarian, A. and Flowers, H.L. (2024). Risk factors and long-term outcomes of oropharyngeal dysphagia in persons with multiple sclerosis: A systematic review protocol. *Systematic Reviews*, **13**(1), 121. DOI: 10.1186/s13643-024-02530-3.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. DOI: 10.48550/arXiv.1409.1556.

Slovik, Y., Kaminer, B.M., Revital, G., Ron, A., Harris, M., Ziv, O. and Cohen, O. (2025). A Modified Fiberoptic Endoscopic Evaluation of Swallowing Evaluating Esophageal Dysphagia by a Capsule: A Pilot Study. *Dysphagia*, **40**(1), 263–70. DOI: 10.1007/s00455-024-10724-z.

Stokely, S.L., Peladeau-Pigeon, M., Leigh, C., Molfenter, S.M. and Steele, C.M. (2015). The relationship between pharyngeal constriction and post-swallow residue. *Dysphagia*, **30**(3), 349–56. DOI: 10.1007/s00455-015-9606-5.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D. and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. DOI: 10.1109/cvpr.2015.7298594.

Verma, S., Devarajan, G.G. and Sharma, P.K. (2025). Modified efficient net of chest x-ray images for lung disease classification using transfer learning approach. *Scientific Journal of King Faisal University: Basic and Applied Sciences*, **26**(1), 35–42. DOI: 10.37575/b/eng/240032

Copyright