Autonomous swallow segment extraction using deep learning in neck-sensor vibratory signals from patients with dysphagia

Yassin Khalifa, *Member, IEEE,* Cara Donohue, James L. Coyle, and Ervin Sejdić, *Senior, IEEE*

***Abstract*—Dysphagia occurs secondary to a variety of underlying etiologies and can contribute to increased risk of adverse events such as aspiration pneumonia and premature mortality. Dysphagia is primarily diagnosed and characterized by instrumental swallowing exams such as videofluoroscopic swallowing studies. videofluoroscopic swallowing studies involve the inspection of a series of radiographic images for signs of swallowing dysfunction. Though effective, videofluoroscopic swallowing studies are only available in certain clinical settings and are not always desirable or feasible for certain patients. Because of the limitations of current instrumental swallow exams, research studies have explored the use of acceleration signals collected from neck sensors and demonstrated their potential in providing comparable radiation-free diagnostic value as videofluoroscopic swallowing studies. In this study, we used a hybrid deep convolutional recurrent neural network that can perform multi-level feature extraction (localized and across time) to annotate swallow segments automatically via multi-channel swallowing acceleration signals. In total, we used signals and videofluoroscopic swallowing study images of 3144 swallows from 248 patients with suspected dysphagia. Compared to other deep network variants, our network was superior at detecting swallow segments with an average area under the receiver operating characteristic curve value of 0.82 (95% confidence interval: 0.807-0.841), and was in agreement with up to 90% of the gold standard-labeled segments.**

***Index Terms*—Swallowing, Accelerometry, Vibrations, Cervical Auscultation, Dysphagia, Segmentation, Signal Analysis, Deep Learning, Supervised Learning, Neural Networks.**

This study was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under award number R01HD092239, while the data was collected under award number R01HD074819. The computational resources utilized in this study were provided by Microsoft and its could service, Azure, through Microsoft's generous support to Pittsburgh CREATES. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding organizations.

Y. Khalifa is with Case Western Reserve University School of Medicine, Cleveland, OH, USA, Harrington Heart and Vascular Institute, University Hospitals, Cleveland, OH, USA and Biomedical Engineering, Cairo University, Giza, Egypt.

C. Donohue is with Aerodigestive Research Core Laboratory, University of Florida, Gainesville, FL, USA and Department of Speech, Language, and Hearing Sciences, University of Florida, Gainesville, FL, USA.

J. L. Coyle is with Department of Communication Science and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA, USA, Department of Otolaryngology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

E. Sejdić is with The Edward S. Rogers Department of Electrical and Computer Engineering, Faculty of Applied Science and Engineering, University of Toronto, Toronto, Ontario, Canada and North York General Hospital, Toronto, Ontario, Canada.E-mail: esejdic@ieee.org

I. Introduction

Aspiration pneumonia caused by swallowing disorders (dysphagia) is the most fatal category of pneumonia in patients 65 and older, eclipsing mortality rates of bacterial and pre-Covid-19 viral pneumonias [1]. Dysphagia is a swallowing disorder that can occur due to a variety of etiologies including stroke, neurodegenerative diseases, and/or head and neck cancer treatment [2], [3]. Dysphagia disrupts the patient's ability to generate a normal flow of solids and liquids through the upper aerodigestive tract. Dysphagia is characterized by difficulty in controlling and even initiating a swallow. A gold standard examination frequently used to diagnose swallowing disorders, is the videofluoroscopic swallow study (VFSS) [4]. During this exam, the patient is observed while swallowing materials impregnated with barium sulfate contrast while a trained clinician observes the swallow in real-time radiographic video frames. Though efficient in clinical assessment of swallowing, VFSSs expose patients to ionizing radiation and are not available in many care settings. This leads to leaving many patients undiagnosed and vulnerable to dysphagia-related complications [5], [6]. Therefore, there is a high demand for widely available and non-invasive artificial intelligence-powered dysphagia assessment tools that can deliver insights about the swallowing physiology to underserved patient populations [7].

High resolution cervical auscultation (HRCA) is an emerging sensor-based technology that utilizes a tri-axial accelerometer and a contact microphone attached to the anterior neck, to non-invasively assess several aspects of swallow function [8]. HRCA combines vibratory and sound signals collected from the neck-attached sensors with machine learning to characterize the patterns associated with swallowing physiology. For HRCA to work as a VFSS surrogate in swallow function assessment, it has to be able to characterize the main physiological events that contribute to safe swallowing. HRCA has demonstrated potential as a dysphagia screening method by classifying swallows into safe and unsafe based on the penetration-aspiration scale [8]–[13]. It has been proven effective also in demarcating multiple physiological events such as upper esophageal sphincter opening [14]–[16], laryngeal vestibule closure [17], [18], and hyoid bone motion [19], [20]. Moreover, HRCA was successfully employed for categorizing swallows between healthy and other patient populations [21]–[23], and clinically rating the swallow physiology in dysphagic patients based on the Modified Barium Swallow Impairment Profile (MBSImP) with a high degree of accuracy [15], [20],

[24]. The development of intelligent HRCA-based swallow function assessment methods offers a more objective way for early detection of swallowing impairments which may be extremely beneficial when expert personnel are not locally present, VFSSs are not immediately available or feasible, or in case of asymptomatic patients. In addition to being used for dysphagia screening/diagnosis, HRCA has potential as a biofeedback instrument for patients undergoing dysphagia rehabilitation.

To achieve the aspired outcome of HRCA as a subjective swallow assessment tool, it has to encompass a fully automated systematic analysis pipeline with the least human interference possible (Fig. 1). The process begins with extraction of swallow segments from the continuous HRCA signals as accurately as performed by experts using the gold standard, followed by the demarcation of kinematic events and anomalies such as aspiration. As can be seen in Fig. 1, accurate extraction of swallow segments is considered a critical step for any subsequent analysis to be performed on HRCA signals.

Traditional event detection methods that rely on statistical and non-sequence-aware classification models have been heavily investigated for the extraction of swallow segments in HRCA signals [25]–[31]. However, many of these methods either suffered from high computational complexity or lacked precision to detect the complete swallow segment which might have led to missing essential physiological events lying within. Deep learning is evolving to be a powerful approach for event detection in biomedical time series. Traditional methods relied on hand-crafted features and scanning time series for events and anomalies while lacking the ability to model long time dependencies [32]. Most recently, convolutional neural networks (CNNs) have been combined with recurrent neural networks (RNNs) for the detection and modeling of events of arbitrary lengths in time series [32], including arrhythmia detection in electrocardiography [33], [34] and epileptic seizure detection in electroencephalography [35], [36]. A CNN is a multi-stage trainable neural network that can automatically learn hierarchical representations and produce high levels of abstraction. RNN is another kind of neural networks that is specialized in processing sequential data one step at a time while controlling information transfer across time steps. In hybrid CNN/RNN models, CNN automatically extracts local features in short time contexts while RNN models the long temporal relationship between these contexts.

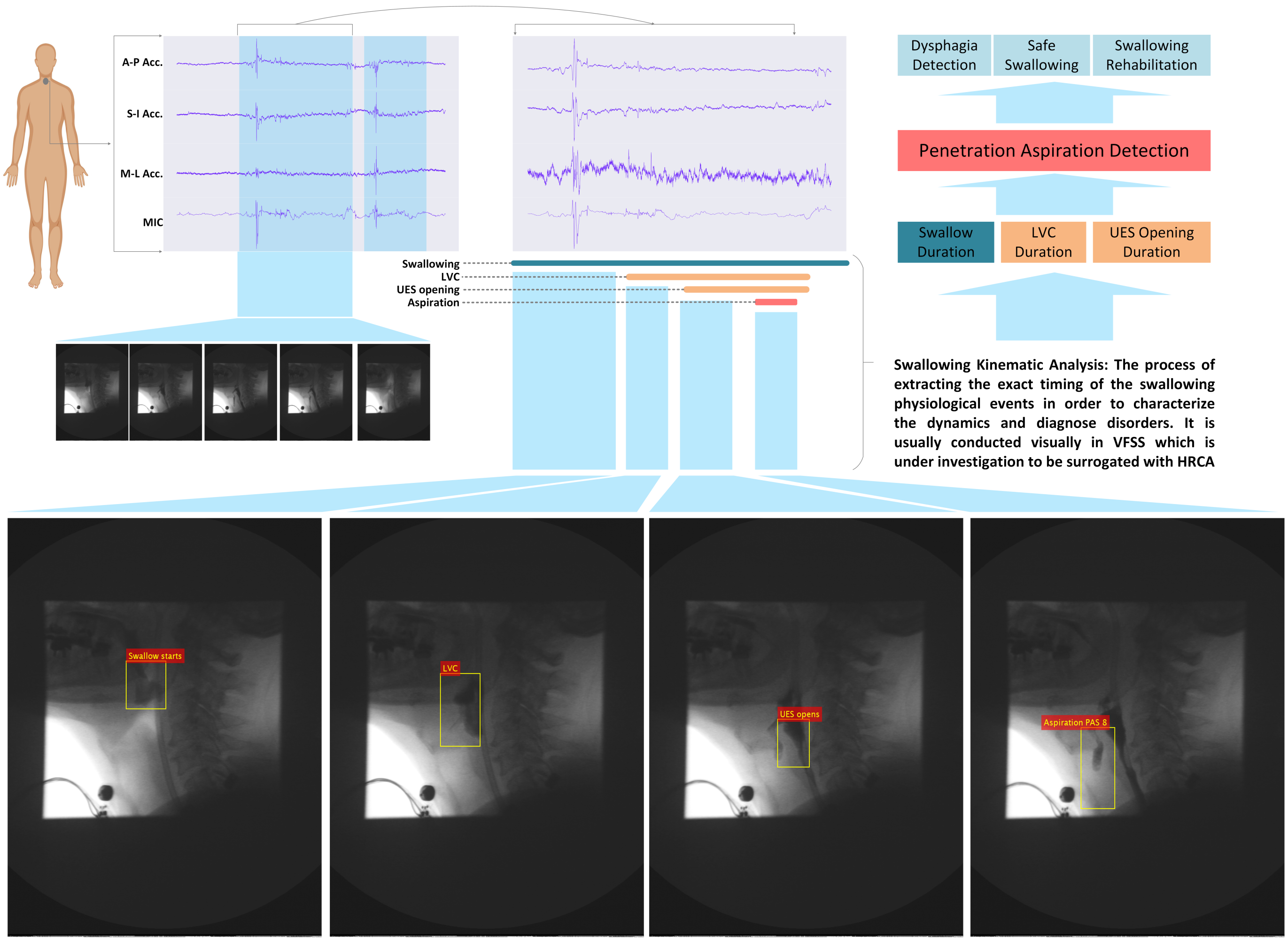
Here we introduce a hybrid CNN/RNN network, a deep learning framework that combines both CNNs and RNNs to automatically capture the swallowing activity in HRCA signals. The proposed framework overcomes many challenges in earlier adaptations of the swallowing segmentation in HRCA signals, including utilization of multi-channel input and automatic feature extraction. With a professional team of research clinicians and engineers, we established a diverse annotated dataset of concurrently collected HRCA signals and x-ray VFSS for more than 3000 swallows from 248 patients with suspected dysphagia. We focused on populations of patients who are most vulnerable to dysphagia such as patients post stroke, patients with neurodegenerative diseases and those suffering from iatrogenic dysphagia due to cardiothoracic surgeries. The dataset was used to validate the precision of swallowing segmentation using the proposed deep learning framework and compare its accuracy to other frameworks that have the potential of producing competing results in similar event detection problems. The alternative networks compared in this study, were chosen based on similar work in the literature to resemble the general types of models used for event detection in biomedical signals. The models included a sliding-window non-sequence-based feed-forward neural network and a hybrid sequence-based CNN/RNN that works directly on raw data. We tested also other variants of these models to explore the effect of network depth and residual learning on the performance.

II. Methods

A. Data collection protocol

This study was approved by the institutional review board of the University of Pittsburgh. All participating subjects provided informed written consents. All subjects were admitted to the University of Pittsburgh Medical Center Presbyterian Hospital where the experiment was conducted. The experiment included the collection of VFSS in addition to swallowing vibrations from an accelerometer attached to the anterior neck of the subject. Subjects were comfortably seated and imaged in the lateral plane. The detailed experimental setup has been described elsewhere [14]. Standard material consistencies were administered to the subjects over the course of a swallowing clinical evaluation that was altered to each subject based on their clinical manifestation of dysphagia. The administered materials included thin liquid (Varibar thin, Bracco Diagnostics, Inc., $<$ 5 cPs viscosity), mildly thick liquid (Varibar nectar, 300 cPs viscosity), puree (Varibar pudding, 5000 cPs viscosity), and Keebler Sandies Mini Simply Shortbread Cookies (Kellogg Sales Company).

VFSS was conducted using a Precision 500D system (GE Healthcare, LLC, Waukesha, WI) at a pulse rate of 30 pulses per second (PPS) [37]. The stream was digitized using an AccuStream Express HD video card (Foresight Imaging, Chelmsford, MA) at a resolution of 720×1080 and a sampling rate of 60 frame per second (FPS). Swallowing vibrations were collected through a tri-axial accelerometer (ADXL 327, Analog Devices, Norwood, Massachusetts) fixed on a small plastic case with a shape that rests well onto the neck curvature. The acclerometer case was attached to the skin overlying the cricoid cartilage with an adhesive tape; the reliability of this specific location in picking high quality swallowing vibrations was verified elsewhere [8], [38]. The accelerometer was placed such that it picks the swallowing vibrations in the anterior-posterior (A-P), superior-inferior (S-I), and medial-lateral (M-L) directions. The signals from the accelerometer were digitized at sampling rate of 20 kHz and temporally aligned with the VFSS stream through LabView (National Instruments, Austin, Texas). The accelerometer signals were properly down-sampled to 4 kHz to reduce the measurement errors and smooth the transient noise such as sudden head movements [14], [31], [39].



Swallowing Kinematic Analysis: The process of extracting the exact timing of the swallowing physiological events in order to characterize the dynamics and diagnose disorders. It is usually conducted visually in VFSS which is under investigation to be surrogated with HRCA

Fig. 1: HRCA signal analysis pipeline. The first step of swallow function evaluation using HRCA signals is the extraction of swallow segments, which is usually done by having expert judges examine VFSS images. Following this, the HRCA signal segments can be used for swallow kinematic analysis to identify the important physiological aspects that contribute to airway protection. VFSS is concurrently collected only in the development phase for the purpose of data labeling for HRCA-based supervised algorithms that perform the kinematic analysis.

B. Expert manual swallow segmentation (ground truth)

VFSS streams were inspected by two expert raters trained to perform swallow kinematic judgments, in order to identify the onset and offset of individual swallows (with random assignments). The onset of a swallow is defined as the frame at which the leading head of the bolus passes the shadow of the posterior border of the ramus of the mandible [31], [40], [41]. The offset is defined as the frame in which the hyoid bone returns to its resting location after completing the swallowing associated motion [31], [40]. The raters were blinded to participants' demographics and diagnoses. Inter- and intra-rater reliability were assessed with intra-class correlation coefficients (ICCs) [42]. Inter-rater reliability was performed on 10% of the swallows for both raters and the ICC values were computed. Ongoing intra-rater reliability was computed to assess the drift in each rater's measurements by randomly selecting one out of each 10 swallows to re-segment and calculate ICC values. Both raters maintained an inter-rater and intra-rater reliability with ICCs of 0.99 or higher during rating the swallows of the dataset. These ratings were used to label the concurrently collected swallowing vibratory signals.

C. Study data characteristics

This study relied on data from 248 adult patients with suspected dysphagia who underwent VFSSs as a part of their in-hospital clinical care. The mean age was 63.8 (standard deviation, s.d.= 13.7) years. The participants were admitted for evaluation with multiple conditions including but not limited to stroke, neurodegenerative diseases, lung transplant, lung lobectomy, heart disease, and head/neck surgeries (Table I). The data consisted of VFSSs simultaneously collected along with HRCA signals during a standard clinical swallowing evaluation procedure that was a part of patients' standard clinical care. The participants were examined under various bolus conditions (volume, consistency, mode of administration, etc.) and compensatory maneuvers (e.g. neutral head position and chin tuck) depending on the presentation of dysphagia during the examination. From the 248 patients, 3144 swallows were collected with a mean swallow segment duration of 862 msec (s.d.: 277). The duration of all swallows was around $6 - 10\%$ of the entire dataset duration. The characteristics of the collected swallows are detailed in Table III. Approximately 5% ($N = 165$) of swallows exhibited aspiration by patients (portions of the bolus entered the trachea) and only 3% ($N = 99$) of the aspiration events were asymptomatic/silent (no coughing).

D. Preparation of swallowing vibratory signals

Swallowing vibratory signals collected for this study, included three channels ($C = 3$). For models that utilized components of the power spectral estimate as input, the spectrogram is calculated for each of the channels of the vibratory signals using an M-point discrete Fourier transform ($M = 1024$) over a Hanning window of length N_1 and 50% overlap. The window length used in this study is $N_1 = 800 \equiv 0.2\ sec$ which was proved elsewhere to be effective in swallow extraction for the same dataset [31]. This window length configuration produced 24145 windows belonging to swallowing segments and 376427 windows belonging to non-swallowing/unidentified segments. Only the positive frequencies ($M/2$ bins) of the Fourier transform were used. Both phase and magnitude are extracted from the complex-value spectrogram and used as separate features with an overall dimension of $T \times M/2 \times 2C$ (C magnitude and C phase components, Fig. 2 A), where T is the sequence length (number of windows). For models that utilized the raw signals as input, the signals are split into windows of $N_2 = 66 \equiv 16\ msec \equiv 1\ VFSS\ frame$ in length with an overall dimension of $T \times N_2 \times C$.

E. Data partitioning

The dataset was partitioned for the training and testing of the proposed algorithms in two main schemes depending on the type of the model; however, both schemes are 10-fold cross validation-based schemes. In brief, we used the dataset to test two types of segmentation models, sliding window-based models and sequence-based models. The two types are similar except that the sequence-based models take sequence of windows as input to model recurrence instead of separate windows. Partitioning for the sliding window-based models relied on the total number of windows in the dataset while sequence-based models used partitioning performed on the total number of sequences. Sequence length (T) was chosen to be 2 sec (10 windows) with 50% sequence overlap (5 windows) for spectrogram-input and 1 sec (60 windows) with 50% sequence overlap for raw signal-input. For the first sequence configuration, a total of 21306 sequences were produced from the dataset.

F. Sequence agnostic-based approach of segmentation

Deep neural networks have been used before for the extraction of swallows in swallowing vibrations. In this study, we utilize a fully connected deep network that was used in a previous study [31] to process the spectrogram of swallowing vibrations in a window-by-window manner. The spectrogram described previously is fed into a 3-layer ($size = 512$) fully connected network with a 4th sigmoid-activated layer for classification output. This model was implemented using Keras with a Tensorflow backend and evaluated using the window-based 10-fold cross validation. An Adam optimizer was used for the training process with a learning rate of 0.0001 and a binary cross entropy loss function [43]. Fig. 3 shows the architecture of the aforementioned model and its variants described later in text.

Another sequence-agnostic method that was considered for performance comparison in this study, included time-series feature extraction from the signal windows instead of spectrogram. The features of each window were then passed into traditional classifiers to determine whether the window belongs to a swallow or not. The analysis procedure in this method started with a multi-level denoising of the swallowing signals. The denoising procedure included modified covariance autoregressive modeling to generate finite impulse response (FIR) filters that remove the baseline noise or what's known as device noise [44], [45]. Fourth-order least square splines were then used to remove the low-frequency noise components and motion artifacts [46], [47] followed by tenth-order Meyer wavelet denoising to eliminate any additive noise (white Gaussian noise in particular) [48]. Following the denoising of the signals, several features were extracted from each signal window in different domains (time, frequency, time-frequency and information theoretic). The features are summarized in Table II. Multiple classifiers were used including support vector machines (SVM) and K-means to classify each window as a part of a swallow segment or not, based on the features calculated.

G. Sequence to sequence-based approach of segmentation

In this study, one of the approaches that we used to address the segmentation task of swallowing vibrations, included models that perform sequence to sequence mappings. Such models are capable of modeling the temporal dependencies across sequences due to the use of recurrent neural networks (RNN) [32]. The first part of the architecture in these models is a convolutional neural network (CNN) that extracts local features from input's time steps before passing them into the RNN to process the temporal dependencies. CNN is composed of repeated layers that feature successive convolutional filters with weights that are optimized during the training process. A typical CNN architecture uses sequential convolutional and pooling layers. CNNs can also perform 1D, 2D, or 3D convolution based on the specific problem addressed. The second part is a recurrent neural network (RNN) that takes the output of CNN for each time step and models the time dependencies a long the sequence. RNNs are known to be an effective architecture for learning time dependencies of arbitrary lengths which can be valuable for differentiating between swallowing events and other spontaneous or transient events such as coughing and head movement [32]. The last part of the model is a fully connected neural network that combines the temporal features generated by the RNN in order

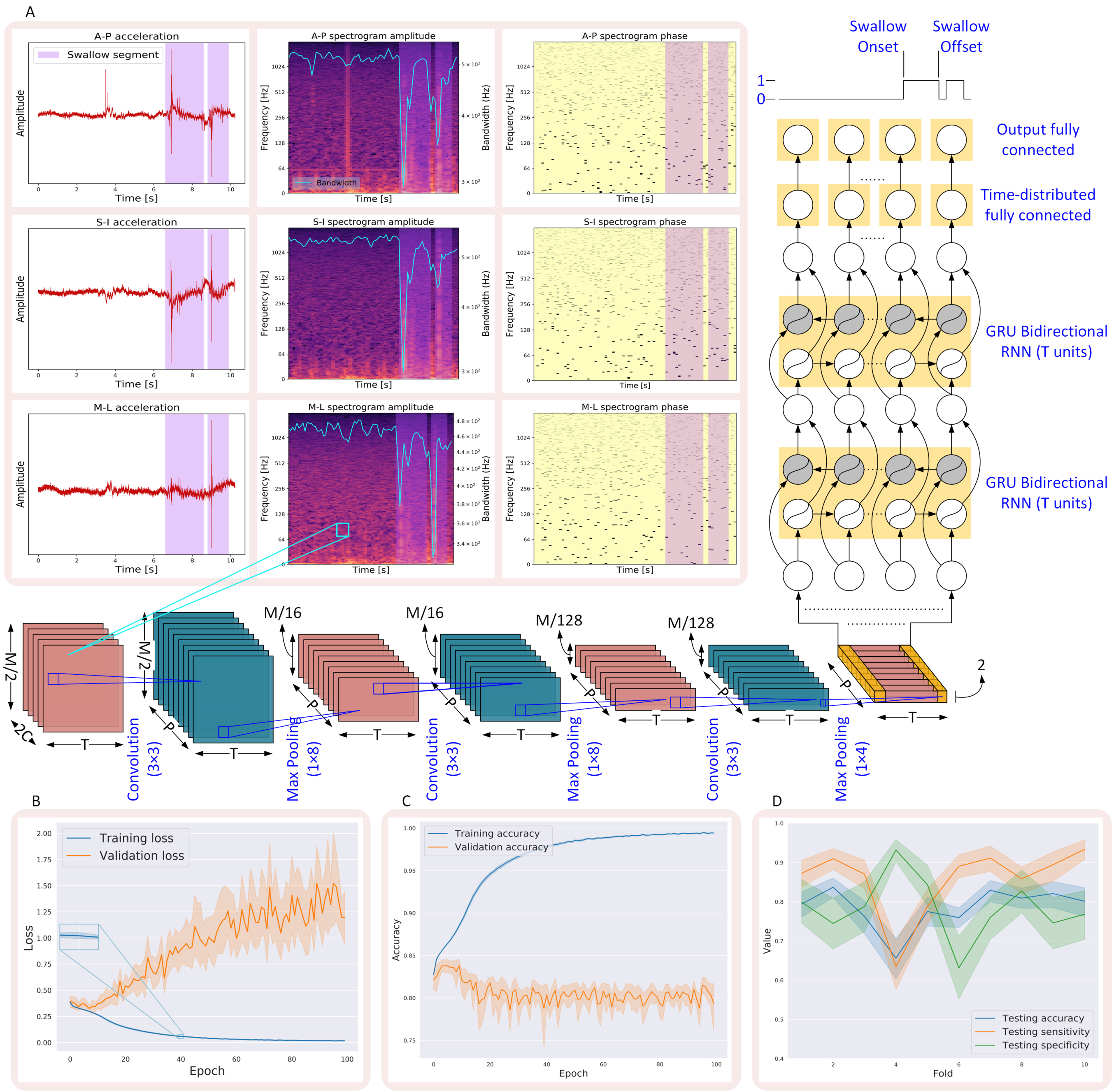


Fig. 2: The architecture of the main proposed deep network. **A.** shows a typical unfolded example of the network input of acceleration signals with two swallow segments as indicated by the purple shadows in the figures. The first column represents raw acceleration signals, and the second and third columns represent the spectrogram and phase for each of the acceleration axes. The drop in bandwidth can be clearly seen in the spectrogram during the swallow segments. **B.** represents the evolution of training and validation losses over 100 epochs of training and the variations across the 10-folds. **C.** represents the evolution of training and validation accuracy over 100 epochs of training and the variations across the 10-folds. **D.** shows accuracy, sensitivity and specificity and the variations across the 10-folds.

to generate a final segmentation sequence that represent the orientation of each window in the sequence. Fig. 2 shows the main proposed sequence-to-sequence architecture which takes spectrogram as input and is composed of a 2D CNN.

The 2D CRNN model shown in Fig. 2 features a 3-layer CNN. Each layer is composed of 64 filters with a kernel size of 3×3. Each layer is ReLU activated and followed by batch normalization and dropout rate of 20%. Max pooling is used as well after each CNN layer with the following sizes, $(8,\ 8,\ 4)$, and it is performed in this model only along the frequency axis of the spectrogram in order to preserve the all time steps. The final CNN output is fed into a 2-layer GRU-based bidirectional

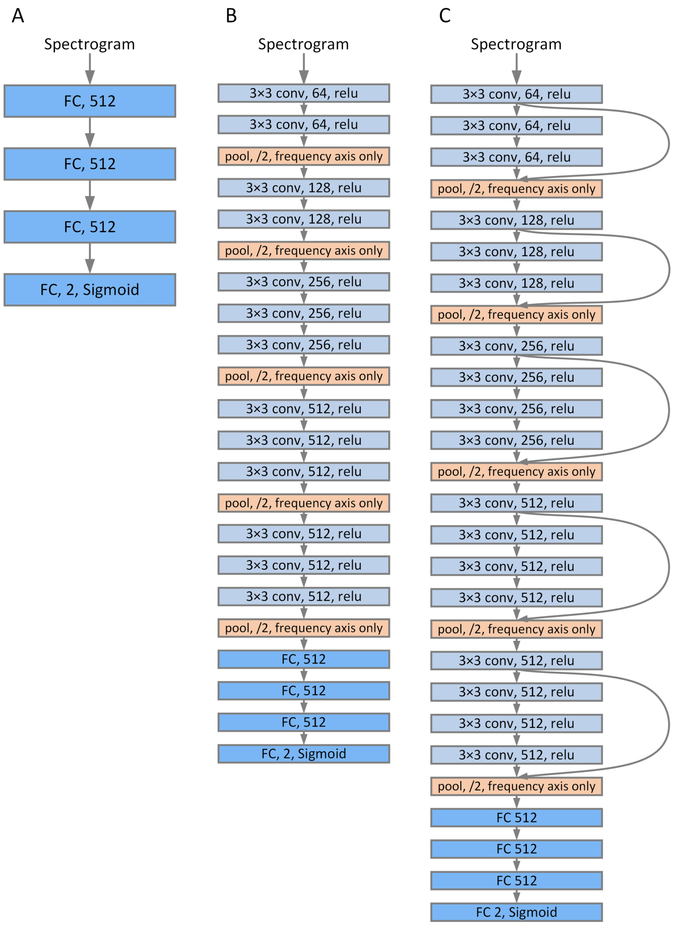


Fig. 3: Layer stacking in each of the network variants. **A.** shows the network that uses only fully connected layers to process the spectrogram. **B.** shows how the VGG16 CNN layers were stacked ahead of the fully connected layers. **C.** shows how the skip connection that perform the residual learning were introduced to the VGG16 design of the network.

TABLE II: Summary of features extracted from swallowing signals [44], [49]–[52].

Time Domain Features	
Standard deviation	Describes the variance of a signal around its mean
Skewness	Describes the asymmetry of amplitude distribution about its mean
Kurtosis	Describes the tailedness/peakness of amplitude distribution relative to normal distribution
Frequency Domain Features	
Peak frequency (Hz)	Describes the frequency of maximum power
Spectral centroid (Hz)	Describes the center of mass of the frequency spectrum of a signal
Bandwidth (Hz)	Describes the frequency range of a signal
Time-Frequency Domain Features	
Wavelet entropy	Describes the disorderly behavior for non-stationary signal
Information-Theoretic Domain Features	
Lempel-Ziv Complexity	Describes the randomness of a signal
Normalized Entropy rate	Describes the degree of regularity of a signal distribution

RNN with 128 units per cell and a length that is equal to the input sequence length T. The output of the second RNN layer is fed into a 3-layer time-distributed fully connected network with the first two layers having the size of 128 and the third layer (output) having the size of T with Sigmoid activation to represent the network classification output per each time step in the input sequence.

Another 1D CRNN model was implemented which used raw signals as input instead of spectrograms. The model features a 3-layer ReLU-activated CNN with 64 filters per layer and a kernel size of 5. 20% dropout and batch normalization are adopted for this network following each CNN layer. The CNN is followed by a 2-layer GRU-based bidirectional RNN with 128 units per cell and a length that is equal to the input sequence length T. Similar to the previously described 2D CRNN model, a 3-layer time-distributed fully connected network is used to combine the recurrent output of the RNN and generate the final classification output per each time step. The size of the first two fully connected layers is 128 and the final layer is Sigmoid-activated with a size of T. Majority of the layers used in all models are ReLU-activated unless mentioned otherwise. Sequence-to-sequence models were all implemented using Keras with a Tensorflow backend and trained through an Adam optimizer with a learning rate of 0.0001 and a binary cross entropy loss function [43]. The sequence-based 10-fold cross validation scheme is used to evaluate all sequence-to-sequence-based models.

TABLE I: Characteristics of the participating patients with suspected dysphagia

Admitting diagnosis	Included conditions	Subject-level (N, %)	Age, year (mean ± s.d.)	Female (N, %)
Neurodegenerative disease	Amyotrophic lateral sclerosis (ALS) - Multiple sclerosis (MS) - Muscular dystrophy - Parkinson's disease - Myasthenia gravis - Motor neuron disease - Progressive muscle weakness - Progressive neurological deficits - Progressive supranuclear palsy - lingual atrophy - Myotonic dystrophy - Alzheimer's - Dementia	24, 9.7%	60.75 ± 13.5	9, 37.5%
Stroke	Right hemisphere - Left hemisphere - Brainstem - Bilateral frontal - Medulla	48, 19.4%	65.4 ± 11.4	10, 20.8%
Lung condition	COPD - Chronic bronchiectasis - Lung adenocarcinoma - Lung Cancer - Pulmonary fibrosis - Cystic fibrosis - Respiratory failure - Pulmonary embolism - Pneumonia - Lobectomy	51, 20.6%	64.9 ± 14.6	22, 43.1%
Cardiac condition	Cardiogenic shock - Heart failure - Cardiac arrest - Aortic valve replacement - Acute myocardial infection - Myocardial infarction - Heart transplant - Aortic abscess	16, 6.4%	58.2 ± 12.7	4, 25.0%
Organ Transplant	Multi-organ transplant - Liver transplant - Renal transplant - Lung/Double lung transplant	37, 14.9%	57.3 ± 11.9	12, 32.4%
Gastrointestinal condition	Paraesophageal hernia - Esophageal cancer - Esophagectomy - Esophagitis - Esophageal reflux	13, 5.2%	63.6 ± 13.1	7, 53.4%
Head & Neck condition	Spinal surgery - Anterior cervical fusion - Tonsil cancer radiation - Palatal hypoplasia	7, 2.8%	62.6 ± 9.4	5, 71.4%
Other conditions	Mental illness - Sleep Apnea - Cerebral palsy - Cerebellar ataxia - Sepsis - Cirrhosis - Diabetes - scleroderma	52, 21.0%	63.4 ± 17.3	37, 71.2.0%

TABLE III: Characteristics of the dataset

Bolus consistency	Utensil	Dataset-level (N, %)	Swallow type (consistency group-level)			Duration, msec (mean ± s.d.)
			Single (N, %)	Multiple (N, %)	Sequential (N, %)	
Thin	Spoon	448, 14.2%	164, 36.6%	281, 62.7%	3, 0.7%	878±303
	Cup	909, 28.9%	280, 30.8%	530, 58.3%	99, 10.9%	898±256
	Cup with straw	417, 13.3%	91, 21.8%	235, 56.4%	91, 21.8%	856±238
	NA	7, 0.2%	–	5, 71.4%	2, 28.6%	888±731
Thick	Spoon	401, 12.8%	98, 24.5%	300, 74.8%	3, 0.7%	874±320
	Cup	311, 9.9%	93, 29.9%	208, 66.9%	10, 3.2%	907±260
	Cup with straw	129, 4.1%	30, 23.3%	99, 76.7%	–	831±264
	NA	5, 0.2%	1, 20%	4, 80%	–	736±64
Pudding	Spoon	241, 7.7%	99, 41.1%	138, 57.3%	4, 1.6%	944±311
	Cup	3, 0.1%	1, 33.3%	2, 66.7%	–	794±164
	Cup with straw	1, 0.04%	–	–	1, 100%	683
Solids (Cookie or Peanuts butter sandwich	Spoon	108, 3.4%	48, 44.4%	60, 55.6%	–	898±271
	Cup	11, 0.35%	3, 27.3%	8, 72.7%	–	792±225
	NA	3, 0.1%	–	3, 100%	–	906±135
Saliva	NA	28, 0.9%	13, 46.4%	15, 53.6%	–	839±259
Tablet + Water	NA	6, 0.2%	–	6, 100%	–	739±255
Unreported consistency	NA	116, 3.7%	NA	NA	NA	731±162
Total		3144	921, 29.3%	1894, 60.2%	213, 6.8%	862±277

H. Deeper models and residual learning

Network depth has been proved, with substantial evidence, to be of crucial importance and led to some of the leading results in popular challenges especially with CNNs [53]–[55]. However, as the depth increases, the accuracy gets saturated and degrades rapidly [53]. Deep residual learning has been introduced to solve the degradation problem that evolves as the networks go deeper. In residual learning, instead of stacking layers directly to fit a certain mapping, these layers are stacked to fit a residual mapping through using skip (identity shortcut) connections which are easier to optimize than the unreferenced mapping [53]. In this study, we tried to employ both unreferenced layer stacking and residual mapping to create networks that have the potential to surpass the performance of the aforementioned models. Fig. 3 demonstrates how layers are stacked to modify the simple deep fully connected network model to be more deeper (Fig. 3 B) and to use residual learning represented by the introduced skip connections (Fig. 3 C) in order to learn a better network that achieves higher classification accuracy. The same stacking concept was used for building variants of sequence-to-sequence models presented earlier where the stacking happened only in the convolutional layers while the rest of the model's architecture (RNN and fully connected layers) remained the same.

For the unreferrenced layer stacking (can be called plain network), we used a VGG16 CNN architecture through stacking 16 weight convolutional layers as described for image recognition problems in [54]. For the residual network, we inserted skip connections into the VGG16 model which can be directly used when the dimensions of input and output are the same; however, in our case the identity shortcuts go across feature maps of different sizes which necessitates using projection or transformation to match dimensions. We used extra convolutional layers prior each identity shortcut to perform the matching (Fig. 3 C). For both deep plain and residual variants of the models, we adopted batch normalization after each network layer and before activation following the practice in [56]. All networks were trained from scratch with uniform initialization and a learning rate of 0.01. No dropout was used in the training of the deep plain and residual networks following [56].

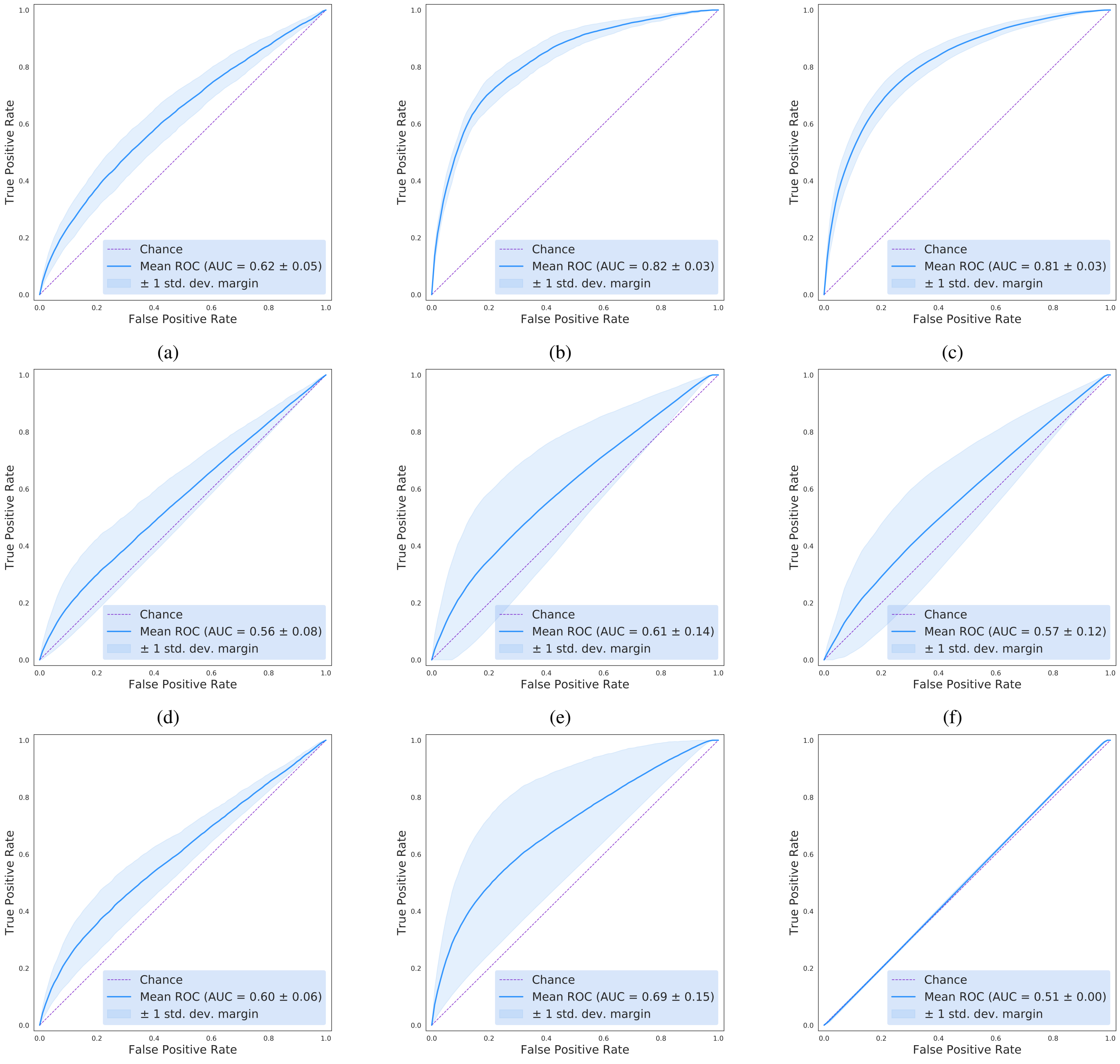
I. Performance metrics

The main segmentation problem in this study is a binary classification task, for which the area under the curve (AUC) of receiver operating characteristic curves (ROC) was calculated as the primary performance metric for all the developed models. In addition, we used the average accuracy, sensitivity, and specificity values as secondary performance metrics. For models that worked directly over windows, the metrics were calculated on the window level which means that we aggregate all the windows in each fold and calculate a single value for accuracy, sensitivity and specificity in addition to a single ROC curve. The average and standard deviation for these metrics were also calculated across the folds of cross validation. For sequence-based models, the performance metrics were calculated per sequence and averaged across sequences of the fold. Although AUCs and other binary classification metrics visualize the overall performance of the algorithms in terms of true and false positive rates, they don't show the temporal prediction quality of the detected swallow segments which are composed of multiple consecutive binary-classified windows. For that, we calculated the overlapping ratio between the predicted swallow segments (after discontinuity post-processing) and their ground truth counterparts [31].

III. Results

A. Identification of swallow segments solely using HRCA signals

We tested multiple deep networks to detect the swallow segments solely from the 3D acceleration component of HRCA signals. The signals were prepared according to the model used for the experiment. We adopted a single structure of a deep network as the main contribution of this work and compared its performance with other base models that were all inspired by the literature of event detection in time series. In total we tested three base models to extract the swallow segments from the HRCA signals. Two more variants were created for each of the base models to make the total number of tested



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 4: Receiver operating characteristic curves of the window-wise predictions of swallow segments. The nine models are (1) a 4-layer fully connected neural network with the spectral estimate as input (2) a 2D shallow CRNN with the spectral estimate as input (3) a 1D shallow CRNN with the raw signals as input (4) a VGG16 adjustment of model 1 (5) a VGG16 adjustment of model 2 (6) a VGG16 adjustment of model 3 (7) residual learning-based variant of the VGG16 adjustment of model 1 (8) residual learning-based variant of the VGG16 adjustment of model 2 (9) residual learning-based variant of the VGG16 adjustment of model 3. Panels a-i correspond to ROC curves and AUC for the models 1-9 respectively.

models, nine. The first base model was inspired by the work developed on the same dataset, which used the power spectral estimate as an input of a deep fully connected network that demarcates the parts of the signal that belong to a swallow segment in a window-by-window fashion [31]. The second base model, which represents the main contribution of this work, employs the power of RNNs in modeling sequences and long-range dependencies to convert the problem into sequence-to-sequence decoding. The model is comprised of a shallow 2D CNN that extracts the local features from input and then feeds the features from multiple successive time steps into a bi-directional GRU-based RNN that models the dependencies between features in time. The outputs are then combined to form predictions through fully connected layers. Such model takes a sequence of windows (power spectral estimate) as an input and produces a sequence of predictions that correspond to the sequence of windows. The third base model is similar to the second base model in concept; however, it uses raw signals as input and 1D convolution instead of 2D convolution [14]. This model uses sequence of raw signal windows as input and produces the corresponding sequence of predictions.

For each of the three base models, two modifications were deployed in order to enhance the detection performance of the models. The first variant was a deeper model created by stacking 16 weight convolutional layers (called VGG16 network [54]) before the base model layers (Fig. 3 B). 2D convolutional layers were stacked in the case of power spectral estimate inputs while 1D convolutional layers were used for the models using raw signals as input. The second variant of the base models was based on the aforementioned VGG16-based models; however, residual learning was emphasized through adding skip connections (Fig. 3 B) which was described elsewhere [53] in order to reduce the training error in the case of very deep models.

The power spectral estimate of HRCA signals from the dataset, was calculated based on the window size that was proven effective for the same dataset in previous studies [31]. For the models utilizing raw data, the window size used to split signals was also calculated based on a similar study developed on the same dataset [14]. The nine proposed deep learning models were all evaluated through a 10-fold cross validation procedure by partitioning the data into 10 equal splits (folds) based on the number of windows/sequences extracted from the dataset. The performance of the proposed CNN-based architectures surpassed the ordinary feed-forward-based network's performance with an average AUC of 0.82 over the 10-folds compared to an average AUC of 0.62 for the feed-forward network (Fig. 4 and Table IV). Adding more layers to the CNN parts of the network did not improve the performance as can be seen in Fig. 4d, 4e, and 4f. On the other hand, residual learning achieved a performance that was between the base models and the VGG16 variant models (Fig. 4g-4h) except for the model that used raw signals as input (Fig. 4i).

The traditional machine learning classifiers tested with features extracted from the signals showed poor performance in comparison with all the deep learning models tested in this study. This part of the study was implemented just to compare the performance of the proposed deep learning models and traditional classifiers that work on handcrafted features. The maximum classification accuracy achieved, did not exceed 73% in addition to extremely low sensitivity values.

TABLE IV: Performance for window-level prediction for each of the nine tested models.

Model	Accuracy	Sensitivity	Specificity
4-layer fully connected network + spectrogram input	$0.793 \pm 0.0.056$	0.128 ± 0.100	0.937 ± 0.089
2D shallow CRNN + spectrogram input	0.832 ± 0.117	0.633 ± 0.242	0.901 ± 0.125
1D shallow CRNN + raw signals input	0.849 ± 0.097	0.336 ± 0.277	0.954 ± 0.072
2D VGG16 CNN + spectrogram input	0.808 ± 0.053	0.137 ± 0.178	0.945 ± 0.093
2D VGG16 CRNN + spectrogram input	0.801 ± 0.132	0.220 ± 0.360	0.943 ± 0.133
1D VGG16 CRNN + raw signals input	0.832 ± 0.114	0.045 ± 0.159	0.991 ± 0.039
2D Residual CNN + spectrogram input	0.799 ± 0.030	0.192 ± 0.145	0.928 ± 0.061
2D Residual CRNN + spectrogram input	0.817 ± 0.121	0.307 ± 0.342	0.943 ± 0.101
1D Residual CRNN + raw signals input	0.837 ± 0.105	0.0	1.0

B. Interpretation of detection accuracy: Which model performs better temporally?

Achieving high performance on the window level doesn't necessarily mean that the model fully detects swallow segments as defined by the gold standard as it may have detected only a part of the swallow segments. The portion of the swallow segment detected by the proposed models compared to the full swallow segment defined by the gold standard must be as close as possible to 100% in order to guarantee that the detected portion includes the major pharyngeal swallow events such as the upper esophageal sphincter opening and the laryngeal vestibule closure. Generally, the proposed models label each window of the signals as being a part of a swallow segment or not. Then a post processing algorithm that combines these labels to get the start and end of each swallow segment is applied.

We compared the detected swallow segments by each of the proposed models to the corresponding defined swallow segments by the gold standard in order to measure the average overlap ratio and determine which model performs better temporally when considering the length of swallow segments. The 2D shallow CRNN model that used spectrogram of the signals as input was the best model considering the detected portion of the swallow segments (Fig. 5). The indicated model consistently detected around 79% (s.d.: 11% and 95% CI: 77.8-79.6%) of the swallow segment across all folds. The number of false positive swallow segments produced by the 2D shallow CRNN model was 299 segments across all validation folds (less than 10% of the total number of swallows in the dataset). On the other hand, the rest of the models performed poorly and/or with strong variations in the quality of detection in the same fold and across folds as indicated in Fig. 5. The closest performance was achieved by the 1D shallow CRNN that uses raw signals as input. It detected approximately 49% (s.d.: 32% and 95% CI: 46.5-50.6%) of the swallow segment when considering all folds. A sample of swallow segments as detected by the best model, the 2D shallow CRNN, is presented in Fig. 6 with an overlap with the gold standard labels of 91.6% and 76.9% (left to right).

IV. Discussion

Here we outlined the development of a swallow segment extraction framework for HRCA signals as an initial step in the pipeline of HRCA-based dysphagia characterization. The proposed framework overcomes the limitations of older segmentation models including high false positive rates and the low temporal detection accuracy. In contrast to ordinary

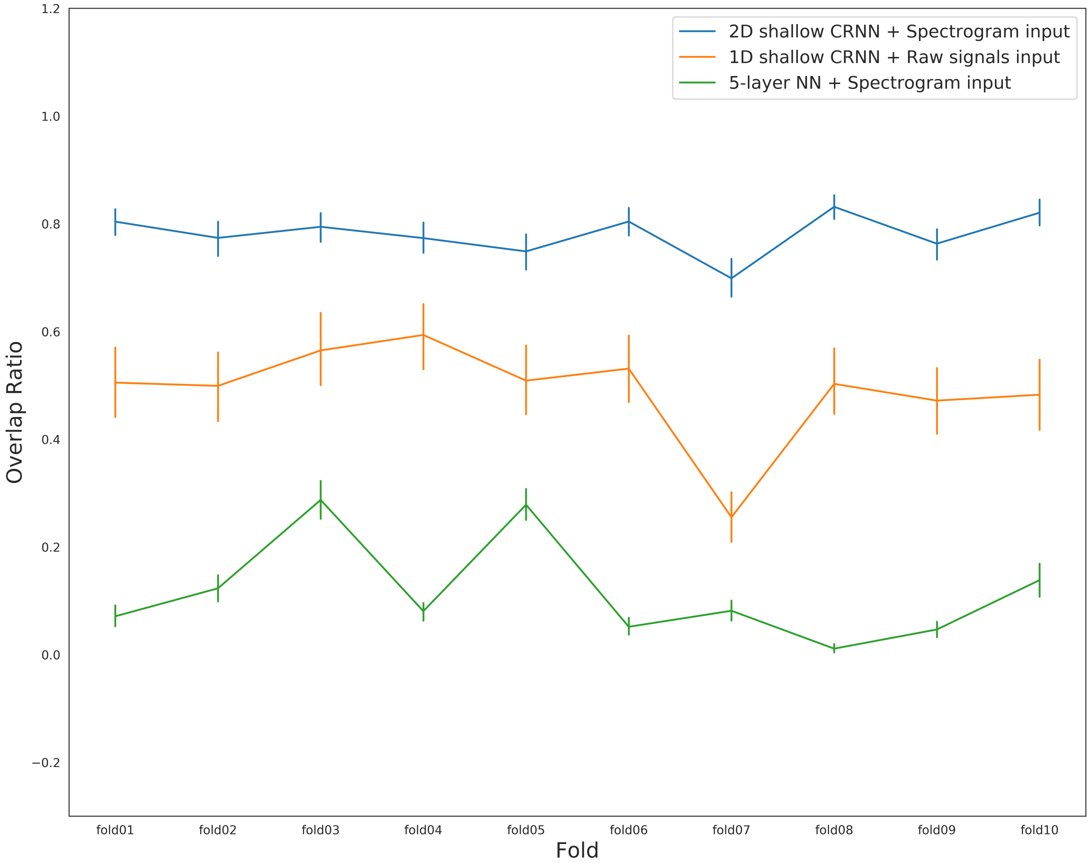


Fig. 5: Average overlap ratio between detected swallow segments by the three best performing models and the reference swallow segments labeled by the gold standard across the 10 folds of the cross-validation process.

machine learning signal segmentation models, the proposed deep learning framework relies on CNNs for local feature extraction and RNNs for modeling time dependencies which significantly contribute to the separation of swallow segments and swallow-like noise such as coughing. The work proposed here, is also different from previous work because it considered only HRCA signals with labels from VFSS for the evaluation process in contrast to other studies that used signals with blind segments [31]. Blind segments are segments of the signals that are recorded while the VFSS is turned off and sometimes include unlabeled swallow segments as blank or non-swallow segments due to the lack of visual evidence of the swallow from VFSS images. Since our study included only labeled signals, this guarantees the credibility and superiority of the presented results. Although the proposed framework was specifically introduced for swallow segment extraction, the same architecture is being broadly applied for event detection problems in multiple types of signals and will help further improve detection quality over traditional methods including probabilistic and non-sequence-based models. On the basis of our results, the proposed segmentation framework is easily applicable for swallowing evaluation devices to be used out of standard clinical care settings and provides accurate swallow segment extraction that is comparable to clinicians' ratings for VFSS.

Among the experimented frameworks in this study, the main proposed framework achieved high detection accuracy-sensitivity combination (see Table IV) with an overall average accuracy of 83.2% (s.d.: 11.7%) and average sensitivity of 63.3% (s.d.: 24.2%). It also achieved the best AUC under the ROC with an average AUC of 0.82 (s.d.: 0.03 and 95% CI: 0.807-0.841) across the 10-folds of the entire dataset (see Fig. 4). In addition to the AUC values and direct window level accuracy for the 10-fold cross validation, we were able to calculate the average overlap between the swallow segments detected by the model and the human labeled swallow segments. This overlap refers to the percentage of the swallow segment that was detected by the model. On average, the proposed framework was able to detect 79% (s.d.: 11% and 95% CI: 77.8-79.6%) of each swallow segment in the dataset. The closest performing framework was the 1D shallow CRNN that used raw signals as input with an average overlap percentage of 49% (s.d.: 32% and 95% CI: 46.5-50.6%). Fig.6 shows that the agreement between the swallow segments detected by the proposed framework and the ground truth labels from the gold standard is highly achieved through including most of the major components of swallow vibrations and sounds within the detected segments.

The proposed segmentation model among the rest of the tested model showed unbalanced sensitivity/specificity combinations with relatively lower sensitivity values. This can be explained by the unbalanced nature of the signal recordings

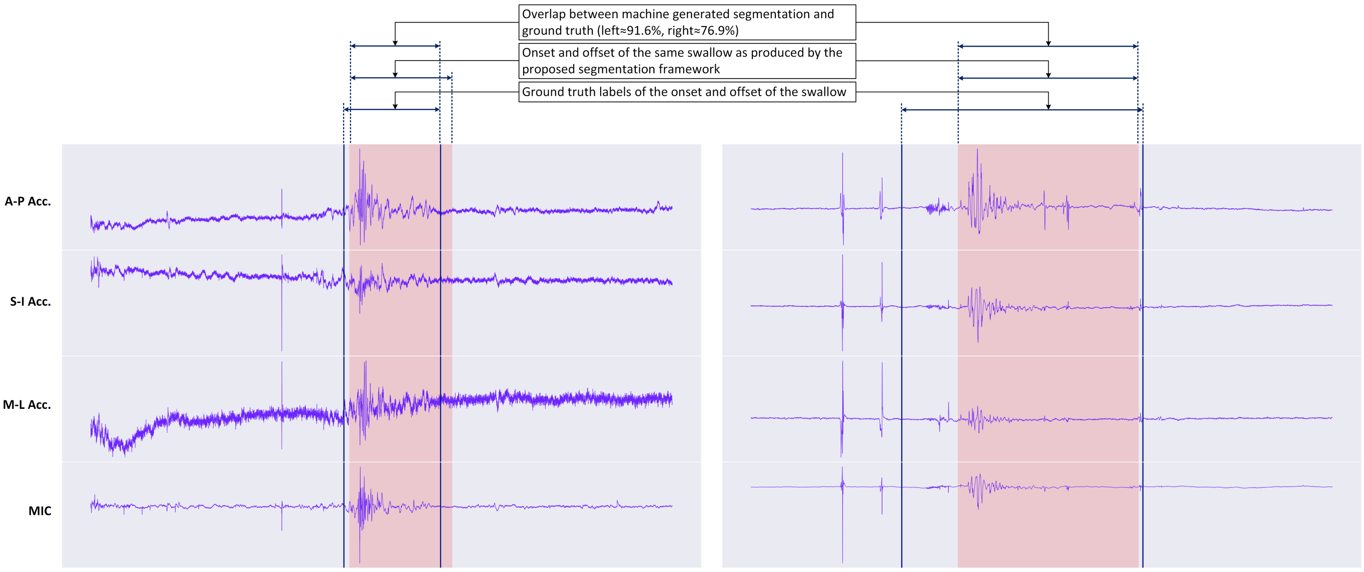


Fig. 6: This figure shows two swallows from two different subjects, a male (age: 44) who developed dysphagia secondary to stroke (left panel) and a female (age:69) who developed dysphagia secondary to subdural hematoma (right panel). The onset and offset of the swallow segments are marked with dark blue vertical lines as labeled by the gold standard while the swallow segments detected by the proposed framework is highlighted in light red. The agreement (overlap) between the gold standard and the machine-based segments is 91.6% for the segment in the left panel and 76.9% for the segment in the right panel.

processed in this study. The utilized dataset, in general, includes less than 10% of its duration as swallow segments which makes an unbalanced input nearly unavoidable especially when dealing with models that process sequences for real-time event detection. In testing, input will always include hundreds of successive sequences that don't include events. It's also worth mentioning that swallow segments are variable in duration, so the number of windows that represent swallow segments can go to as low as 4 windows per sequence for the extremely short swallow segments. Sensitivity and specificity were calculated per sequence and averaged over all sequences in each fold. The fluctuations in sensitivity values were anticipated especially with sequences including short swallows which pushed the overall average down to a lower value. While global sensitivity/specificity across sequences can be considered as an overall indicator in terms of false positive and false negative rates, they don't show how well the detection is aligned with the ground truth of the entire swallow segment. Given the temporal accuracy of the models shown in Fig. 5, we can clearly see how well the proposed model can detect the swallow segment despite of the biased sensitivity values that resulted from the unbalanced input sequences.

The clinical importance of the proposed network is threefold. It promotes the use and development of HRCA-based devices as a surrogate for VFSS in swallowing evaluation. This, not only contributes to reducing the costs and unnecessary radiation exposure of VFSS in many cases, but also increases the accessibility of swallowing evaluation methods in care settings and/or areas where VFSS is unavailable or undesirable. In addition to being important as a first step for any subsequent algorithms that analyze swallow function [9], [13], [14], [17], [19], the proposed automated segmentation framework mitigates the unavoidable human error in manual segmentation on which most of dysphagia characterization algorithms are reliant [40]. We also find it promising that the proposed algorithm works directly on the spectral estimate derived from raw signals without any preprocessing or denoising despite of the presence of multi-source noise in the data which makes it perfect to a non-standard clinical operation where patients may be constantly moving or speaking.

Swallow function analysis aims to detect everything about a swallow starting with its onset and offset to a full kinematic analysis for each of the physiological aspects contributing to a safe swallow. Among these aspects, hyoid bone displacement, upper esophageal sphincter opening and laryngeal vestibule closure were recently measured in HRCA signals using similar deep learning architectures to the proposed framework that employ CNNs and RNNs for the detection of these events [14], [17], [19]. Now that the segmentation process can be performed in the same way with reasonable precision, the entire process can be combined in a single multi-task deep learning framework which wasn't possible when segmentation needed a separate statistical or classification module to perform. Therefore, this work integrates well with the state-of-the-art developments in swallowing signal analysis and uses an architecture that is widely employed in event detection.

Although the work presented in this study represents a necessary step for the automation of swallow function analysis, it can't work as a standalone system because swallow segment extraction doesn't provide any diagnostic value on its own. The next logical step for this research is to combine it with the existing research that depicts swallow safety and can be used to give feedback to patients about their swallowing while they are actually swallowing. Such integrated systems that rely only on non-invasive sensors can provide a complete picture about swallow function in terms of airway protection status, presence of pharyngeal residue, and whether the swallow is within normal limits or impaired. Furthermore, there is a growing evidence in the literature now that points towards the ability to figure out the patient condition from just HRCA signals [22], [23], [57]. This means that not only can these systems provide a diagnostic profile of the swallow but also tell the origin of the abnormality if exists.

In summary, This work showed that deep learning-based architectures could be used to automatically extract the onset and offset of swallows in HRCA signals. The combined use of CNNs and RNNs can achieve good detection accuracy when it comes to modeling sequences for event extraction which is considered one of the setbacks in the traditional machine learning techniques. Deep learning continues to show its ability to play a vital role in clinical decision making and rehabilitation support of dysphagia and swallowing function through creating widely accessible and cheap tools that provide the same diagnostic value as the currently utilized clinical exams. Such tools could help identify dysphagia in early stages before the development of severe complications like pneumonia and recommend referral for a specialist who can conduct more diagnostic exams thus leaving no patient undiagnosed or incorrectly diagnosed.

Data availability

The entire dataset analyzed in this manuscript is available on Zenodo: (https://doi.org/10.5281/zenodo.4539695). The dataset includes the raw swallowing acceleration signals as well as the onset and offset labels for each swallow.

Code availability

Both the implementation of all deep models described in this manuscript and direct instructions to replicate the findings can be found in the GitHub repository at (https://github.com/yassinkhalifa/pHRCA_AutoSeg).

References

[1] W. B. Baine, W. Yu, and J. P. Summe, “Epidemiologic trends in the hospitalization of elderly Medicare patients for pneumonia, 1991-1998,” *American Journal of Public Health*, vol. 91, no. 7, pp. 1121–1123, Jul. 2001.

[2] M. R. Spieker, “Evaluating dysphagia,” *American Family Physician*, vol. 61, no. 12, pp. 3639–3648, Jun. 2000.

[3] J. A. Logemann, “The evaluation and treatment of swallowing disorders,” *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 6, no. 6, p. 395, Dec. 1998.

[4] J. L. Coyle and J. Robbins, “Assessment and behavioral management of oropharyngeal dysphagia,” *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 5, no. 3, p. 147, Jun. 1997.

[5] I. Zammit-Maempel, C.-L. Chapple, and P. Leslie, “Radiation Dose in Videofluoroscopic Swallow Studies,” *Dysphagia*, vol. 22, no. 1, pp. 13–15, Jan. 2007.

[6] H. S. Bonilha, K. Humphries, J. Blair, E. G. Hill, K. McGrattan, B. Carnes, W. Huda, and B. Martin-Harris, “Radiation Exposure Time during MBSS: Influence of Swallowing Impairment Severity, Medical Diagnosis, Clinician Experience, and Standardized Protocol Use,” *Dysphagia*, vol. 28, no. 1, pp. 77–85, Mar. 2013.

[7] H. Zhao, Y. Jiang, S. Wang, F. He, F. Ren, Z. Zhang, X. Yang, C. Zhu, J. Yue, Y. Li, and Y. Liu, “Dysphagia diagnosis system with integrated speech analysis from throat vibration,” *Expert Systems with Applications*, vol. 204, p. 117496, 2022.

[8] J. M. Dudik, J. L. Coyle, and E. Sejdić, “Dysphagia screening: Contributions of cervical auscultation signals and modern signal-processing techniques,” *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 465–477, Aug. 2015.

[9] E. Sejdić, C. M. Steele, and T. Chau, “Classification of penetration-aspiration versus healthy swallows using dual-axis swallowing accelerometry signals in dysphagic subjects,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1859–1866, Jul. 2013.

[10] J. M. Dudik, I. Jestrovic, B. Luan, J. L. Coyle, and E. Sejdić, “Characteristics of dry chin-tuck swallowing vibrations and sounds,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 2456–2464, Oct. 2015.

[11] J. M. Dudik, I. Jestrovic, B. Luan, J. L. Coyle, and E. Sejdić, “A comparative analysis of swallowing accelerometry and sounds during saliva swallows,” *Biomedical Engineering Online*, vol. 14, no. 1, p. 3, Jan. 2015.

[12] J. M. Dudik, J. L. Coyle, A. El-Jaroudi, Z. H. Mao, M. Sun, and E. Sejdić, “Deep learning for classification of normal swallows in adults,” *Neurocomputing*, vol. 285, pp. 1–9, Apr. 2018.

[13] C. Yu, Y. Khalifa, and E. Sejdić, “Silent aspiration detection in high resolution cervical auscultations,” in *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics*, May 2019, pp. 1–4.

[14] Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdić, “Upper esophageal sphincter opening segmentation with convolutional recurrent neural networks in high resolution cervical auscultation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 493–503, Feb. 2021.

[15] C. Donohue, Y. Khalifa, S. Perera, E. Sejdić, and J. L. Coyle, “How Closely do Machine Ratings of Duration of UES Opening During Videofluoroscopy Approximate Clinician Ratings Using Temporal Kinematic Analyses and the MBSImP?” *Dysphagia*, vol. 36, no. 4, pp. 707–718, Aug. 2021.

[16] Y. Khalifa, C. Donohue, J. Coyle, and E. Sejdić, “On the robustness of high-resolution cervical auscultation-based detection of upper esophageal sphincter opening duration in diverse populations,” in *Proceedings of SPIE 11730, Big Data III: Learning, Analytics, and Applications*, vol. 11730. SPIE, Apr. 2021.

[17] S. Mao, A. Sabry, Y. Khalifa, J. L. Coyle, and E. Sejdić, “Estimation of laryngeal closure duration during swallowing without invasive X-rays,” *Future Generation Computer Systems*, vol. 115, pp. 610–618, Feb. 2021.

[18] A. Sabry, A. S. Mahoney, S. Mao, Y. Khalifa, E. Sejdić, and J. L. Coyle, “Automatic Estimation of Laryngeal Vestibule Closure Duration Using High-Resolution Cervical Auscultation Signals,” *Perspectives of the ASHA Special Interest Groups*, vol. 5, no. 6, pp. 1647–1656, Dec. 2020.

[19] S. Mao, Z. Zhang, Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdić, “Neck sensor-supported hyoid bone movement tracking during swallowing,” *Royal Society Open Science*, vol. 6, no. 7, p. 181982, Jul. 2019.

[20] C. Donohue, S. Mao, E. Sejdić, and J. L. Coyle, “Tracking Hyoid Bone Displacement During Swallowing Without Videofluoroscopy Using Machine Learning of Vibratory Signals,” *Dysphagia*, vol. 36, no. 2, pp. 259–269, Apr. 2021.

[21] A. Kurosu, J. L. Coyle, J. M. Dudik, and E. Sejdić, “Detection of swallow kinematic events from acoustic high-resolution cervical auscultation signals in patients with stroke,” *Archives of Physical Medicine and Rehabilitation*, vol. 100, no. 3, pp. 501–508, Mar. 2019.

[22] C. Donohue, Y. Khalifa, S. Perera, E. Sejdić, and J. L. Coyle, “A Preliminary Investigation of Whether HRCA Signals Can Differentiate Between Swallows from Healthy People and Swallows from People with Neurodegenerative Diseases,” *Dysphagia*, vol. 36, no. 4, pp. 635–643, Aug. 2021.

[23] C. Donohue, Y. Khalifa, S. Mao, S. Perera, E. Sejdić, and J. L. Coyle, “Characterizing Swallows From People With Neurodegenerative Diseases Using High-Resolution Cervical Auscultation Signals and Temporal and Spatial Swallow Kinematic Measurements,” *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 9, pp. 3416–3431, Sep. 2021.

[24] B. Martin-Harris, M. B. Brodsky, Y. Michel, D. O. Castell, M. Schleicher, J. Sandidge, R. Maxwell, and J. Blair, “MBS measurement tool for swallow impairment-MBSImp: Establishing a standard,” *Dysphagia*, vol. 23, no. 4, pp. 392–405, Dec. 2008.

[25] E. Sejdić, C. M. Steele, and T. Chau, “Segmentation of dual-axis swallowing accelerometry signals in healthy subjects with analysis of anthropometric effects on duration of swallowing activities,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1090–1097, Apr. 2009.

[26] S. Damouras, E. Sejdić, C. M. Steele, and T. Chau, “An online swallow detection algorithm based on the quadratic variation of dual-axis accelerometry,” *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3352–3359, Jun. 2010.

[27] J. M. Dudik, A. Kurosu, J. L. Coyle, and E. Sejdić, “A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals,” *Computers in Biology and Medicine*, vol. 59, pp. 10–18, Apr. 2015.

[28] J. Lee, C. M. Steele, and T. Chau, “Swallow segmentation with artificial neural networks and multi-sensor fusion,” *Medical Engineering & Physics*, vol. 31, no. 9, pp. 1049–1055, Nov. 2009.

[29] J. Lee, S. Blain, M. Casas, D. Kenny, G. Berall, and T. Chau, “A radial basis classifier for the automatic detection of aspiration in children with dysphagia,” *Journal of Neuroengineering and Rehabilitation*, vol. 3, no. 1, p. 14, Jul. 2006.

[30] T. Chau, D. Chau, M. Casas, G. Berall, and D. J. Kenny, “Investigating the stationarity of paediatric aspiration signals,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 1, pp. 99–105, Mar. 2005.

[31] Y. Khalifa, J. L. Coyle, and E. Sejdić, “Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings,” *Scientific Reports*, vol. 10, no. 1, p. 8704, May 2020.

[32] Y. Khalifa, D. Mandic, and E. Sejdić, “A review of Hidden Markov models and Recurrent Neural Networks for event detection and localization in biomedical signals,” *Information Fusion*, vol. 69, pp. 52–72, May 2021.

[33] M. Limam and F. Precioso, “Atrial fibrillation detection and ECG classification based on convolutional recurrent neural network,” in *Computing in Cardiology*, Sep. 2017, pp. 1–4.

[34] S. P. Shashikumar, A. J. Shah, G. D. Clifford, and S. Nemati, “Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’18. London, United Kingdom: ACM, Jul. 2018, pp. 715–723.

[35] A. M. Abdelhameed, H. G. Daoud, and M. Bayoumi, “Deep Convolutional Bidirectional LSTM Recurrent Neural Network for Epileptic Seizure Detection,” in *2018 16th IEEE International New Circuits and Systems Conference (NEWCAS)*, Jun. 2018, pp. 139–143.

[36] H. Daoud and M. Bayoumi, “Deep Learning based Reliable Early Epileptic Seizure Predictor,” in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct. 2018, pp. 1–4.

[37] H. S. Bonilha, J. Blair, B. Carnes, W. Huda, K. Humphries, K. McGrattan, Y. Michel, and B. Martin-Harris, “Preliminary investigation of the effect of pulse rate on judgments of swallowing impairment and treatment recommendations,” *Dysphagia*, vol. 28, no. 4, pp. 528–538, Dec. 2013.

[38] J. A. Cichero and B. E. Murdoch, “The physiologic cause of swallowing sounds: Answers from heart sounds and vocal tract acoustics,” *Dysphagia*, vol. 13, no. 1, pp. 39–52, 1998.

[39] R. Schwartz, Y. Khalifa, E. Lucatorto, S. Perera, J. Coyle, and E. Sejdić, “A Preliminary Investigation of Similarities of High Resolution Cervical Auscultation Signals Between Thin Liquid Barium and Water Swallows,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–9, 2022.

[40] G. L. Lof and J. Robbins, “Test-retest variability in normal swallowing,” *Dysphagia*, vol. 4, no. 4, pp. 236–242, Dec. 1990.

[41] A. S. Mahoney, Y. Khalifa, E. Lucatorto, E. Sejdić, and J. L. Coyle, “Cervical Vertebral Height Approximates Hyoid Displacement in Videofluoroscopic Images of Healthy Adults,” *Dysphagia*, Mar. 2022.

[42] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: Uses in assessing rater reliability,” *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, Mar. 1979.

[43] Y. Bengio, “Practical recommendations for gradient-based training of deep architectures,” in *Neural Networks: Tricks of the Trade*, 2nd ed., ser. Lecture Notes in Computer Science, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Springer, 2012, vol. 7700, pp. 437–478.

[44] E. Sejdić, V. Komisar, C. M. Steele, and T. Chau, “Baseline characteristics of dual-axis cervical accelerometry signals,” *Annals of Biomedical Engineering*, vol. 38, no. 3, pp. 1048–1059, Mar. 2010.

[45] L. Marple, “A new autoregressive spectrum analysis algorithm,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 441–454, Aug. 1980.

[46] E. Sejdić, C. M. Steele, and T. Chau, “A method for removal of low frequency components associated with head movements from dual-axis swallowing accelerometry signals,” *PloS One*, vol. 7, no. 3, p. e33464, Mar. 2012.

[47] E. Sejdić, C. M. Steele, and T. Chau, “The effects of head movement on dual-axis cervical accelerometry signals,” *BMC Research Notes*, vol. 3, p. 269, Oct. 2010.

[48] E. Sejdić, C. M. Steele, and T. Chau, “A procedure for denoising dual-axis swallowing accelerometry signals,” *Physiological Measurement*, vol. 31, no. 1, pp. N1–N9, Jan. 2010.

[49] C. Rebrion, Z. Zhang, Y. Khalifa, M. Ramadan, A. Kurosu, J. L. Coyle, S. Perera, and E. Sejdić, “High-resolution cervical auscultation signal features reflect vertical and horizontal displacements of the hyoid bone during swallowing,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, p. 1800109, Feb. 2019.

[50] J. M. Dudik, A. Kurosu, J. L. Coyle, and E. Sejdić, “A statistical analysis of cervical auscultation signals from adults with unsafe airway protection,” *Journal of Neuroengineering and Rehabilitation*, vol. 13, no. 1, p. 7, Jan. 2016.

[51] A. Porta, S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti, “Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series,” *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1282–1291, Nov. 2001.

[52] M. Aboy, R. Hornero, D. Abasolo, and D. Alvarez, “Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 11, pp. 2282–2288, Nov. 2006.

[53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun. 2016, pp. 770–778.

[54] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, Sep. 2014.

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *arXiv preprint arXiv:1409.4842*, Sep. 2014.

[56] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, Feb. 2015.

[57] C. Donohue, Y. Khalifa, S. Mao, S. Perera, E. Sejdić, and J. L. Coyle, “Establishing Reference Values for Temporal Kinematic Swallow Events Across the Lifespan in Healthy Community Dwelling Adults Using High-Resolution Cervical Auscultation,” *Dysphagia*, May 2021.