Noninvasive Dynamic Characterization of Swallowing Kinematics and Impairments in High Resolution Cervical Auscultation via Deep Learning

by

Yassin Khalifa

Master of Science in Biomedical Engineering,

Cairo University, Egypt, 2013

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Yassin Khalifa

It was defended on

November 2, 2021

and approved by

Ervin Sejdić, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

James L. Coyle, Ph.D., Professor, Department of Communication Science and Disorders and Otolaryngology

Murat Akcakaya, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Liang Zhan, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Ahmed Dallal, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Dissertation Director: Ervin Sejdić, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Noninvasive Dynamic Characterization of Swallowing Kinematics and Impairments in High Resolution Cervical Auscultation via Deep Learning

Yassin Khalifa, PhD

University of Pittsburgh, 2021

Swallowing is a complex sensorimotor activity by which food and liquids are transferred from the oral cavity to the stomach. Swallowing requires the coordination between multiple subsystems which makes it subject to impairment secondary to a variety of medical or surgically related conditions. Dysphagia refers to any swallowing disorder and is common in patients with head and neck cancer and neurological conditions such as stroke. Dysphagia affects nearly 9 million adults and causes death for more than 60,000 yearly in the US. In this research, we utilize advanced signal processing techniques with sensor technology and deep learning methods to develop a noninvasive and widely available tool for the evaluation and diagnosis of swallowing problems. We investigate the use of modern spectral estimation methods in addition to convolutional recurrent neural networks to demarcate and localize the important swallowing physiological events that contribute to airway protection solely based on signals collected from non-invasive sensors attached to the anterior neck. These events include the full swallowing activity, upper esophageal sphincter opening duration and maximal opening diameter, and aspiration. We believe that combining sensor technology and state of the art deep learning architectures specialized in time series analysis, will help achieve great advances for dysphagia detection and management in terms of non-invasiveness, portability, and availability. Like never before, such advances will enable patients to get continuous feedback about their swallowing out of standard clinical care setting which will extremely facilitate their daily activities and enhance the quality of their lives.

Keywords: Swallowing, Dysphagia, Cervical Auscultation, Deep Learning, Signal Processing, Recurrent Neural Networks, Convolutional Neural Networks.

Table of Contents

List of Tables

List of Figures

Preface

This research has been driven by my passion for developing methods and tools that address society's greatest challenges, particularly in the healthcare field. This passion only grew as Artificial Intelligence massively advanced and created endless opportunities to automate and optimize healthcare delivery. Therefore, this dissertation highlights the use of Artificial Intelligence in developing low-cost and non-invasive technologies that can expedite the diagnosis of swallowing problems. At the first glance, it may sound like a not too serious disorder; however, swallowing is an essential process for human survival and any disruption in this process not only reduces the quality of life but also can lead to death if not early discovered.

The path towards this research has been circuitous and it was only completed with the divine grace of Almighty God and the great deal of support and assistance that I received from many people to whom I have to express my gratefulness and probably words will fall short.

First and foremost, I would like to thank my supervisor, **Professor Ervin Sejdić**, whose support and expertise were invaluable in formulating and conducting this research. Your insightful feedback and advice helped me sharpen my skills and push my thinking to a higher level.You are always very kind and patient and truly care about your students. Words are powerless to express my gratitude to you.

I would like to thank **Professor James Coyle** whose massive knowledge and clinical experience in swallowing helped me develop a firm grasp of the clinical concepts and experimentation. Working along side with you and your team was a priceless experience without which this research was never to see the light.

I extend my gratitude to the members of my dissertation committee, **Professor Murat Akcakaya**, **Professor Ahmed Dallal** and **Professor Liang Zhan**, not only for their time and patience, but for their guidance and support.

I would also like to thank my friends and colleagues, who were always there for me to provide the wise counsel and support. Graduate school can be a tough and draining experience, so thank you for the tons of fun we had together, and for making this journey memorable. It would have been intolerable without you.

Finally, but not least, I would like to express my deep and sincere gratitude to my parents and my sister for their unparalleled love, endless support and prayers. I am forever in your debt for selflessly giving me the experiences and opportunities that have made me who I am today. It was always your encouragement that raised me up when I got weary. Thank you for everything, for my life, I hope that I made you proud. This is dedicated to you.

1.0 Introduction

1.1 Motivation

Deglutition (swallowing) is a well-coordinated, yet complex process essential to ensure optimal nutrition and health, in which food and liquids are transported from the oral cavity to the stomach at a proper rate and speed [1, 2]. It involves the mechanical and neurological coordination between various anatomical structures such as swallowing, respiratory, and speech structures that are close in position and share several functions [2]. Being such a complex process makes it subject to a wide range of functional disorders secondary to many etiologies.

1.1.1 Dysphagia

Dysphagia is a symptom of deglutition dysfunction that provokes discomfort or difficulty during the progression of alimentary bolus from mouth to stomach [3, 4]. Dysphagia derives from the Greek terms *dys* (ill) and *phago* (eat). It can be characterized by a variety of sensations ranging from difficulty initiating the swallow to the perception of resistance or obstruction to bolus propagation through the esophagus [5]. Anatomically, dysphagia may rise from oropharyngeal or esophageal dysfunction; pathologically, it may result from various structural abnormalities that may alter the oropharyngeal reconfiguration from airway to alimentary canal or form a resistance towards bolus progression [3]. These abnormalities include osteophytes, esophageal or throat tumors, circopharyngeal bar, and Zenker's diverticulum. However, dysphagia frequently occurs secondary to systemic and neurological disorders or due to aging [3]. Aspiration, the entry of foreign materials into the airway below the true vocal folds, is the main clinically adverse outcome of oropharyngeal dysphagia, which causes nutritional and respiratory complications that may lead to morbidity and high mortality rates if misdiagnosed or untreated [6–8]. On the other hand, esophageal dysphagia is neither as severe symptomatically, prevalent, nor equally fatal as oropharyngeal dyspha-

gia; however, it is better discovered and managed clinically [3, 4, 9]. Due to its severity and wide prevalence, oropharyngeal dysphagia is to be the main focus of research provided in this dissertation.

1.1.2 Prevalence

Oropharyngeal dysphagia is one of the highest prevalent clinical conditions that impacts millions yearly. It is more common in three main populations, elderly people, patients with head/neck diseases, and patients with neurological or neurodegenerative diseases [4]. Table 1 shows the prevalence of oropharyngeal dysphagia among the described populations according to recent surveys and prevalence studies [4]. The prevalence presented in Table 1 is reported using either clinical exploration via water swallow test or volume-viscosity swallowing test, or instrumental exploration [4]. Although it is not nearly as common, dysphagia happens within younger populations, infants in particular. The incidence of dysphagia and swallowing dysfunction in such population is unknown, though it is obviously increasing given the rates of children with history of prematurity (less than 37 gestation), low birth weight, and complex medical conditions [10–15].

Table 1: Prevalence in main populations impacted by oropharyngeal dysphagia [4]

Population	Prevalence(%)	References
Elderly		
Independently-living older people	23	Serra-Prat et al. [16], 2011
Hospitalized in an acute geriatric unit	29.4–47.0	Lee et al. [17], 1999 Cabre et al. [18], 2014
Hospitalized with community acquired pneumonia	55.0–91.7	Cabre et al. [19], 2010 Almirall et al. [20], 2012
Institutionalized	38	Nogueira and Reis [21], 2013
Stroke: acute phase	51–55	Rosemary et al. [22], 2005
Stroke: chronic phase	25–45	Rosemary et al. [22], 2005
Neurodegenerative diseases		
Parkinson's disease	82	Kalf et al. [23], 2012
Alzheimer disease	57–84	Langmore et al. [24], 2007 Horner et al. [25], 1994
Dementia	57–84	Suh et al. [26], 2009 Langmore et al. [24], 2007 Horner et al. [25], 1994
Multiple sclerosis	34.3	Calcagno et al. [27], 2002
Amyotrophic lateral sclerosis	47–86	Chen and Garrett [28], 2005 Ruoppolo et al. [29], 2013
Head and neck diseases		
Head and neck cancer	50.6	Garcia-Peris et al. [30], 2007
Zenker diverticulum	86	Galli et al. [31], 2003
Osteophytes	17-28	Utsinger et al. [32], 1976 Resnick et al. [33], 1975

1.1.3 Complications

Regardless of the patient age or functional capacity, if present, oropharyngeal dysphagia is considered a risk factor for malnutrition, respiratory tract infections, and pneumonia and usually leads to a reduced overall quality of life [4, 34–36]. For instance, oropharyngeal dysphagia is the main reason that 41% of patients are anxious or panic during meals and 36% avoid eating with people. It was found that 66% of the elderly with oropharyngeal dysphagia suffer from severe muscular protein depletion and subclinical dehydration due to malnutrition which acts as co-risk factor with oropharyngeal dysphagia to raise the 1-year mortality rate in frail elderly patients up to 65.8% [37, 38]. Up to 25% of stroke patients were found to suffer from malnutrition caused by impaired swallowing function as well [39]. Aspiration pneumonia defined as the pneumonia occurring due to overt aspiration, is strongly believed to be the main cause of death in patients with Parkinson's disease, multiple types of dementia, and amyotrophic lateral sclerosis, and that it affects up to 20% of stroke patients and leads to mortality during one year following discharge [20, 36, 39, 40]. According to a 10-year survey, a 93.5% increase of aspiration pneumonia in hospitalized elderly patients, was noticed compared to other types of pneumonia [41]. Further, aspiration pneumonia occurs to 43-50% of nursing home residents during the first year and increases the mortality rate to 45% [42].

1.2 Physiology of Swallowing

Swallowing is essentially a motor activity that requires a precise coordination between several parts in the central nervous system, sensory and motor cranial nerves, and peripheral receptors [43]. This coordination is translated into the propulsion of the bolus from the oral cavity to the esophagus and hence to stomach. The anatomical integrity of the pharynx, which controls the shared functions between swallowing and respiration as shown in Fig. 1, and the sustainable neuromuscular function of cervical striated muscles are required for an unhindered motion of the bolus [4]. Normal swallowing is composed of three main phases:

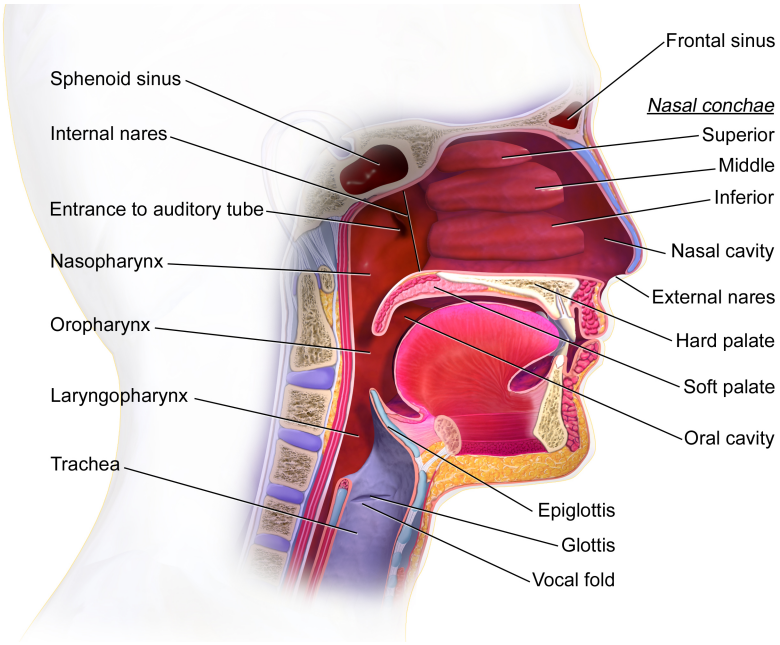


Figure 1: Anatomy of the pharynx and esophagus that shows the shared path taken by both air and food to trachea and esophagus respectively [45].

oral, pharyngeal, and esophageal phases. The three phases happens sequentially with crucial timing and only the oral phase is voluntary; however, the rest of the process occurs reflexly once triggered [44].

1.2.1 Swallowing Phases and Pathology

The oral phase is a preparatory phase that takes place voluntarily and changes according to the consistency of the bolus [4]. For fluid boluses, the soft palate is sealed down against the tongue to place the bolus in the anterior part of the mouth. In case of the failure of this seal, the bolus enters the pharynx ahead of triggering the pharyngeal phase which leads to aspiration [46]. For solid materials, the bolus is formed through mastication which involves cyclic jaw movement and material transport to the molars by tongue and cheeks to be then mixed with saliva and transported into the oropharynx where bolus resides for swallowing [47, 48]. The action the tongue squeezing against the palate, propels the bolus from the mouth into the pharynx. The mastication function can affected by aging due to tooth loss,

weak chewing, and reduced saliva production [49]. The propulsion efficiency is strongly affected in elderly patients with sarcopenia and patients with neuromuscular diseases which leads to the accumulation of oropharyngeal residue [39, 46, 50].

The pharyngeal phase of the swallow starts as the bolus crosses the faucial pillars in a complete involuntary act provided by a set of sequential neuromuscular events. This phase is mediated in the medulla and triggered when the bolus presence stimulates the glossopharyngeal and vagus nerves [44]. During this phase, the airway closes and respiration ceases for a fraction of a second. The hyoid bone and larynx are elevated in an anterosuperior direction due to the contraction of the suprahyoid muscles which paves the way for the bolus to move from the base of the tongue into the pharynx without entering the larynx [44]. This elevation creates negative pressure that helps the bolus to progress. Pressure asserted by tongue base in addition to the contraction of aryepiglottic muscles moves the epiglottis to direct the materials into the esophagus in a diversion act that helps prevent the laryngeal penetration [44]. The upper esophageal sphincter (UES) opening is one of the main biomechanical events during this phase, that allows the bolus passage into the esophagus [51]. A delayed laryngeal vestibule closure and/or UES opening is considered a key swallowing impairment that alters the early oropharyngeal reconfiguration from respiratory to alimentary pathway and leads to the formation of residue, penetration, and aspiration. Damage in the cortical or brainstem areas of swallowing, central or peripheral deafferentation, or sensory defects in the oropharyngeal area may cause delays occurring for different events of the pharyngeal phase in patients with neurological disorders and the elderly [39, 46, 52, 53].

The third phase is the esophageal phase and it represents the flow of the bolus through the esophagus and hence to the stomach. The esophagus is the inferior extension of the pharynx and it is basically a muscular tube that lies posterior to the trachea and anterior to the vertebral fascia [44]. The esophagus can be divided into three parts, the upper part and it is formed of the upper esophageal sphincter (UES), the middle part which contains straited muscle, and the lower part which forms the lower esophageal sphincter [44]. Upon triggering the pharyngeal phase the UES opens allowing the bolus to pass through the esophagus which represents the beginning of esophageal phase. Gravity and peristalsis help moving the bolus through of the esophagus. A successful esophageal phase requires a correct timing of UES

relaxation and subsequent contraction to prevent the laryngopharyngeal reflux [44]. Reduced UES opening may occur due to intrinsic, restrictive disorder that reduces the compliance of the sphincter in case of patients with Zenker diverticulum and/or circopharyngeal bar [54, 55]. Impaired suprahyoid traction or weak bolus propulsion in patients with neurological disorder or in elderly may lead to insufficient UES opening [56–58].

1.2.2 Swallowing Assessment

The evaluation of swallowing function is done over two stages. The first stage is called screening and it attempts to characterize the patient condition through evaluating the behavior during swallowing in a pass-fail manner [59–61]. However, sometimes this way of screening is deemed insufficient or inconclusive due to the absence of dysphagia overt signs such as coughing, and a more comprehensive diagnostic examination is required to assess the airway protection and identify the impairment mechanism[59–61]. Screening may as well lead to considerably variable findings due to the subjectivity between testers. On the other hand, diagnostic methods are more thorough in nature and involve direct observation of the swallowing process in action[44, 59–61].

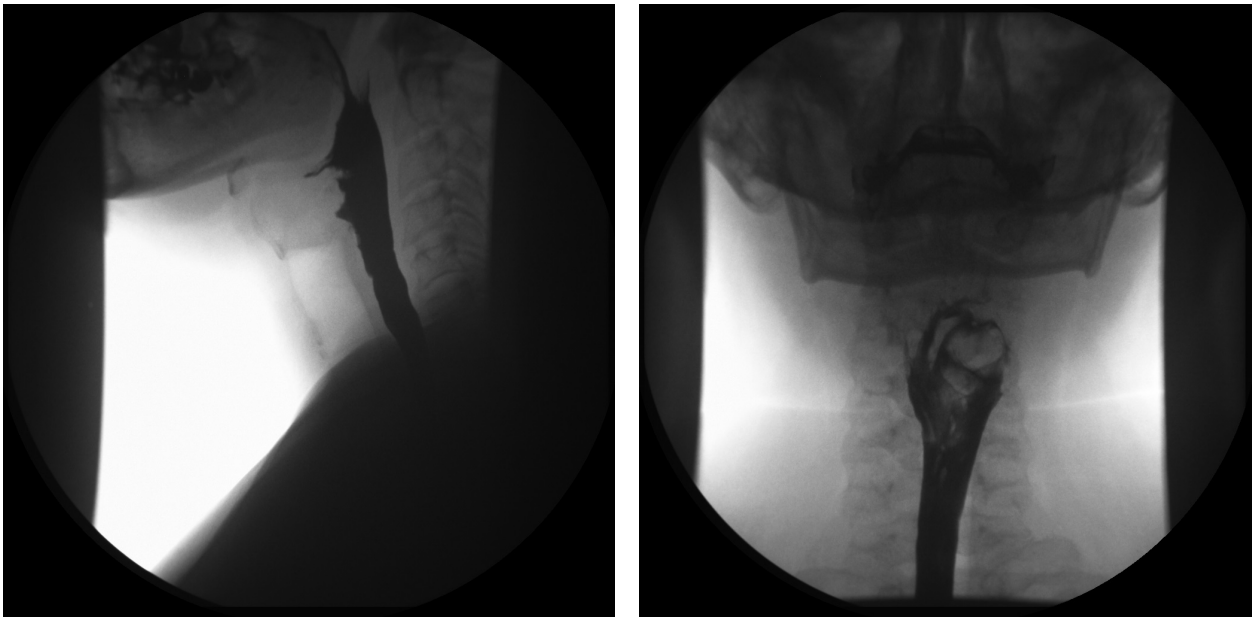
Swallowing screening is subdivided into two categories, non-instrumental and instrumental screening. Many methods have been introduced in the literature in the category of non-instrumental screening and while they may vary in detail, they all operate in the same manner. During such tests, the patient is asked to swallow a certain volume of some material while being observed by a trained examiner for any signs swallowing dysfunction such as dysphonia, choking, or coughing in addition to lips and tongue motility in some cases [62]. Examples for such screening methods include Toronto bedside test, the 3 ounce water challenge, and the modified MASA [63–66]. Such methods are widely used in the clinic despite of their limited accuracy and significantly high false positive rates. However, they are easy to administer and serve as a way to determine whether the patient needs further diagnostic examination [65].

Instrumental screening is a more sophisticated method to observe the patient while swallowing some materials using an instrument that record enhance some physical signal like

sound or muscle waves. The most common methods used in instrumental screening are cervical auscultation and electromyography [67, 68]. In electromyography, surface electrodes are used to record the electrical activity of neck muscles involved in the swallowing process. Despite the fact that the muscle activity differs in cases where neuromuscular function is not well performed, it usually represents gross performance of muscles rather than the swallow itself [67, 68]. Cervical auscultation method on the other hand, is an analogy to the phonocardiogram method where heart sound is heard to detect murmurs happening due to the dysfunction of heart valves [69]. In a similar way, a stethoscope is placed over the thyroid cartilage to listen to the sounds generated by the hyolaryngeal excursion and bolus flow during the swallows [69]. Cervical auscultation is currently under development in order to establish their reliability and employ AI methods to detect the swalowing sub-physiological events [70].

Multiple diagnostic tools are available to assess the swallowing function such as fiberoptic endoscopic evaluation of the swallowing (FEES), pharyngeal manometry, fast pharyngeal CT/MRI, and the videofluoroscopic swallowing study (VFSS) [71–75]. VFSSs are the most widely used tool to clinically evaluate and diagnose swallowing disorders and is considered as the gold standard of swallowing function assessment. VFSS provides dynamic, radiographic evaluation of all three phases of swallowing which helps identify the disordered phases and conditions that pose challenge to swallowing; and determine possible strategies to rehabilitate or treat [72]. The examination involves asking the patients to swallow different bolus sizes and/or consistencies of materials mixed with contrast agent like barium sulfate. Optimally, the x-ray machine is aligned to produce images of the sagittal section for the pharynx and the upper esophagus; however, it is common also to acquire images of the coronal section for the same areas to serve different diagnostic purposes as shown in Fig. 2. During the exam, an expert observes the bolus path and speed during swallowing in addition any anatomical anomalies that might exist and cause a swallowing problem [72].

Given its efficiency and the multiple limitations associated with other diagnostic tools and relatively non-invasive action, VFSS is preferred over endoscopy and manometry [71–76]. However, VFSSs, which use ionizing radiation to produce radiographic images with full temporal resolution, are unavailable or undesirable to many patients, are relatively expensive,



(a) Sagittal plane (b) Coronal plane

Figure 2: A sample of VFSS frames taken at both (a) sagittal and (b) coronal planes for the same patient.

and require specialized instrumentation and trained clinicians to perform and interpret, leaving many patients undiagnosed or inaccurately diagnosed, and exposed to ongoing risk of dysphagia related complications.

1.2.3 Aspiration Evaluation

Biomechanical properties of the swallow are subjectively interpreted from VFSS by clinicians to determine the risk of penetration and aspiration. Severity of aspiration is rated using a standard scale called the Penetration Aspiration scale (PAS). PAS is an 8-point scale used to determine the depth of entry of the material into airway and attempts made by the patient to clear the material out of the airway [77]. Complete airway protection is given a PAS score of 1, while the impermanent entry of the material into the laryngeal vestibule (above vocal folds) is given a score of 2. Penetration happens if the material was not cleared

out of the laryngeal vestibule and is given a score of 3–5 according to the depth and amount. Scores 6–8 are given in case of aspiration, when the material crosses the vocal folds into the trachea [78]. A schematic representation of the PAS is shown in Fig. 3.

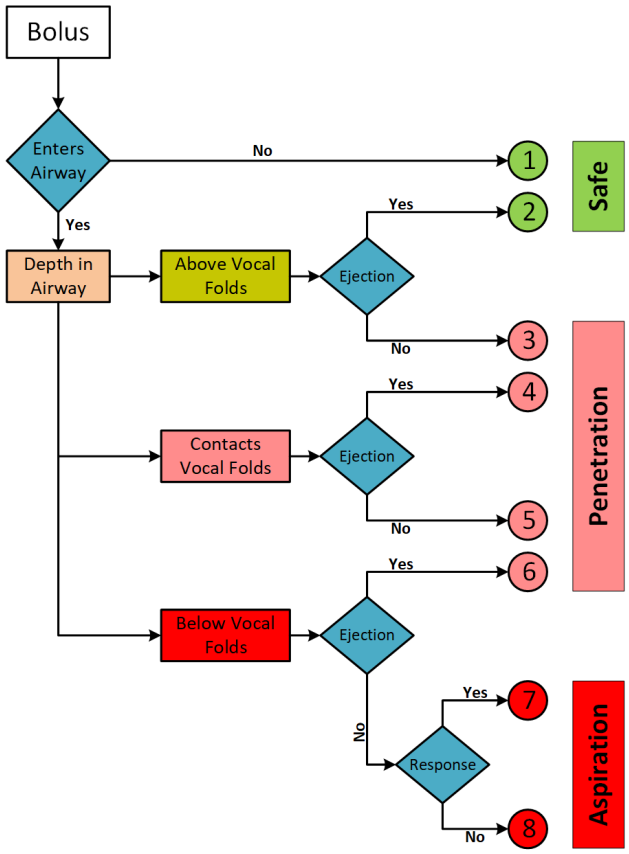


Figure 3: Schematic representation of the penetration aspiration rating procedure. The colors demonstrate the severity of the airway invasion with green as safe swallowing, light red as swallowing with penetration but the material was at the level of the vocal folds or below it but ejected and dark red as aspiration with no ejection efforts [77].

Despite the usefulness of the PAS in quantifying airway protection and aspiration severity, it remains subjective and needs skilled clinicians to be conducted. In addition, VFSS requires the exposure of patients to x-ray radiation which limits the period of the evaluation process and reduces the probability of capturing penetration aspiration events.

1.3 Dissertation Scope

The holy grail of dysphagia clinical evaluation methods has long been a noninvasive and clinically feasible method of accurately identifying the biomechanical events of swallowing that contribute to airway protection. Development of such methods would enable development of a screening tool that can differentiate between impaired and healthy swallowing with a high degree of sensitivity and specificity without the uncertainty of clinical examinations or the additional burden or lack of availability of imaging studies [79–83]. To address the obstacle of insufficient access to instrumental testing of swallowing function universally, high resolution cervical auscultation (HRCA) is currently being investigated as an affordable, feasible, non-invasive bedside assessment tool for dysphagia. HRCA combines the use of vibratory signals from an accelerometer with acoustic signals from a microphone attached to the anterior neck region during swallowing as shown in Fig. 4 & 5. Following collection of signals, advanced machine learning techniques are used to examine the association between HRCA signals and physiological events that occur during swallowing [84, 85].

Through advanced signal processing methods, HRCA has been shown to be highly correlated with hyolaryngeal excursion [86] which represent the core action of the swallowing process. HRCA has been successfully used also to non-invasively track the hyoid bone during swallowing [85] and showed promising results considering aspiration detection and identification of normal swallows in adults [87, 88]. Moreover, recent research studies have found that HRCA signals are associated with a variety of biomechanical events of swallowing including hyolaryngeal excursion and maximal hyoid bone displacement [89]. Another study explored the relationship between HRCA signals and laryngeal vestibule closure, UES opening, and tongue base contact with the posterior pharyngeal wall [90]. The relationship between HRCA signals and swallowing compensatory strategies such as a chin-down posture (chin tuck), and head rotation, has also been investigated and has revealed that HRCA signals contain unique information about the underlying alterations in pharyngeal physiology during deployment of the head positions during swallowing [91].

The need for a continuously available assessment tool to dynamically visualize the swallowing process without adding extra financial or relocation burdens to patients, alongside

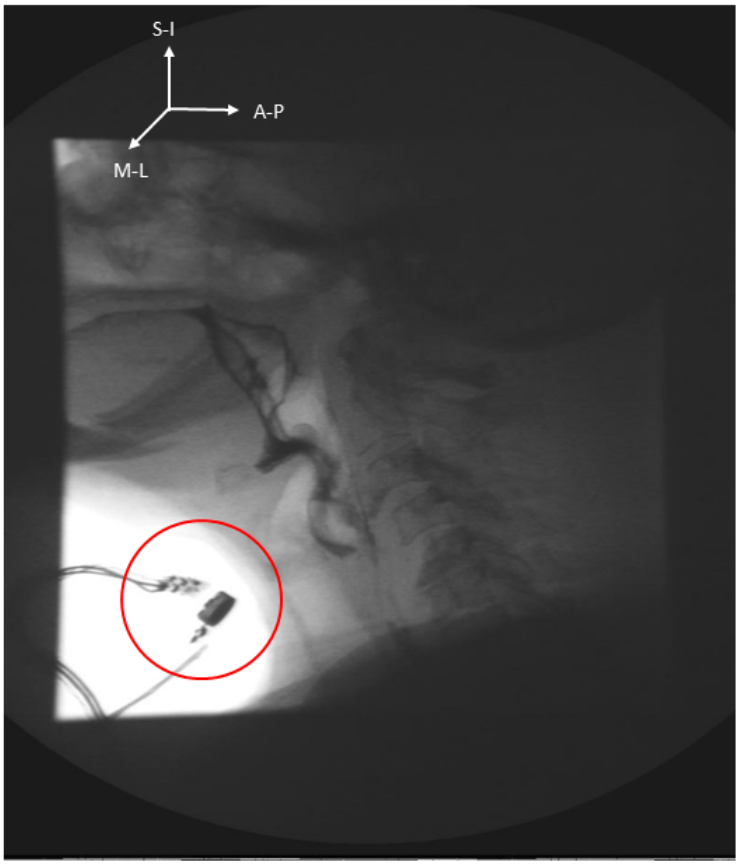
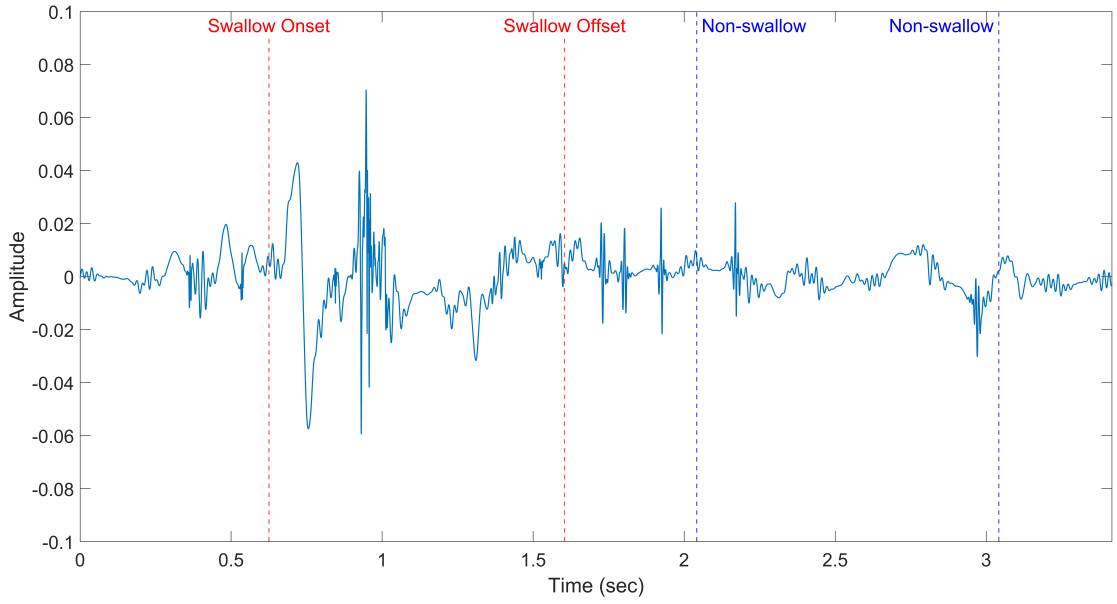
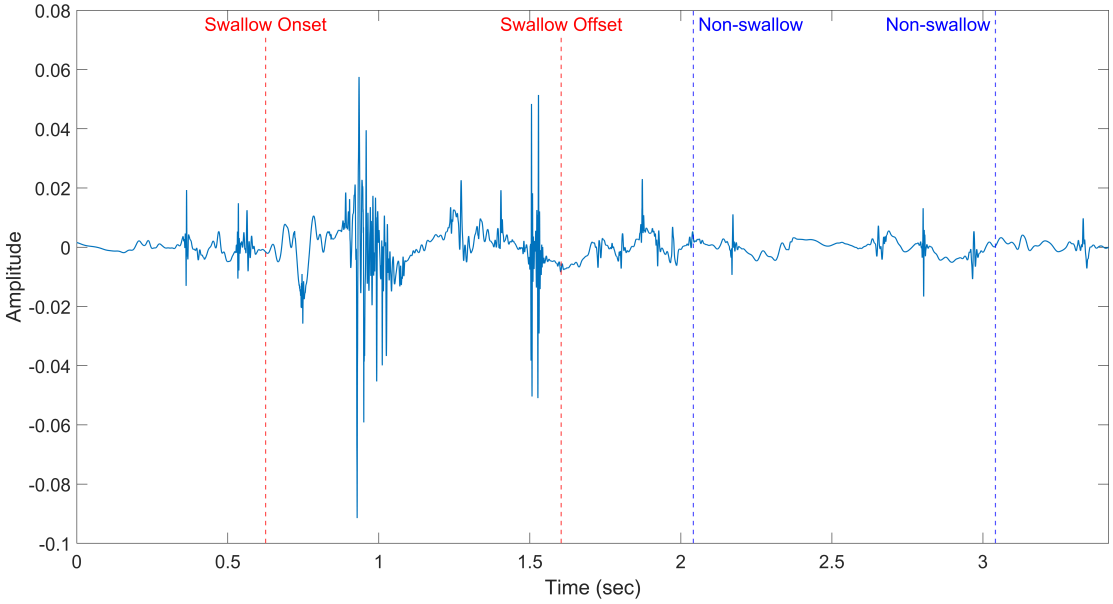


Figure 4: Placement of accelerometer and microphone on the anterior neck as seen in a typical VFSS exam.

with the recent advances in sensor-based technology specifically the positive indicators mentioned earlier about HRCA, has encouraged the development of more predictive profiles for HRCA that will help in the assessment and rehabilitation process of swallowing. One of the most critical issues for many patients, is the ability to get a consistent feedback about their swallowing, while they are swallowing. A feature that will not only help improve the clinic-based swallowing evaluation, but will also be a of a great benefit for the patients towards feeling the progress of the rehabilitation process and maintaining safe swallowing.



(a)



(b)

Figure 5: Sample raw sound and acceleration signals as recorded from the sensors attached on the anterior neck for each patient. The onset and offset of the swallow segment is marked in red dotted line and the rest are non-swallow segments with the segment marked with the blue dotted lines as an example. (a) Microphone signal (b) A-P acceleration signal.

1.4 Dissertation Contributions

The main objective of this work is to use unique approaches to achieve fundamental advancements in the field of signal processing and data analysis for swallowing difficulties while also translating this research into a clinically applicable tool. This will be achieved through the following contributions:

- **Major Contribution #1: Utilize deep learning architectures to develop algorithms that automatically demarcate the pharyngeal swallowing activity solely based on vibrations and sound signals collected from accelerators and microphones attached to the anterior neck.** We propose different methods that employ deep learning and spectral estimation to automatically demarcate the onset and offset of swallowing events in HRCA signals. The proposed methods investigate the following:
  - Single channel HRCA signal-based automatic segmentation.
  - Multi-channel HRCA signal-based automatic segmentation.

  The results from both methods will be compared to the manual gold standard segmentation by human experts from VFSS.
- **Major Contribution #2: Develop algorithms that can automatically extract the onset and offset of main swallowing physiological events that contribute to airway protection based on multi-dimensional swallowing vibrations.** We propose novel architectures that use convolutional recurrent neural networks in association with transfer and multi-task learning in order to extract the events of interest which include:
  - The opening and closure of upper esophageal sphincter.
  - The precise moment when the maximal upper esophageal sphincter opening diameter is achieved as well as the diameter itself.
- **Major Contribution #3: Noninvasive identification of silent aspiration based on swallowing vibrations and sounds.** We propose a deep learning system that uses HRCA signals in order to rate swallows based on the level of airway protection and the

extent to which aspiration happens as well as the preventive reflex to expel the aspirated materials if exists. The ratings will be based on PA scores given to swallows by expert clinicians through examining the concurrently collected VFSS.

1.5 Dissertation Organization

Chapter 2 explores the background of event detection in biomedical signals which represents the core for achieving the aims of this dissertation. Chapter 3 describes the main topics to be addressed for the final dissertation defense and the main approaches attempted. Chapter 4 presents the preliminary results for one of the proposed automatic segmentation algorithms for swallowing accelerometry and sounds and comparison to the gold standard manual segmentation in VFSS. Chapter 5 describes the results and implementation of using convolutional recurrent neural networks and multi-channel signal fusion to enhance the automatic segmentation of swallow segments in high resolution cervical auscultation. Chapter 6 presents the preliminary results of using a convolutional recurrent neural network for the upper esophageal sphincter opening detection in swallowing vibrations. Chapter 7 introduces a measurement protocol for the upper esophageal sphincter opening maximal A-P distension in VFSS and presents the results and the methods of using a hybrid convolutional recurrent neural networks supported by attention mechanisms to predict the upper esophageal sphincter opening maximal A-P distension using HRCA signals only. Chapter 8 summarizes the results of using a deep convolutional neural network supported by residual learning to detect the status of airway protection during swallowing using only HRCA signals by predicting the category of the swallow among three categories determined by the penetration-aspiration score of the swallow. Chapter 9 summarizes the conclusions of the research presented in this dissertation and the possible future directions to be investigated as the next steps for this research.

2.0 Background

The majority of this chapter has been previously published in and reprinted with permission from [92]. © 2021 Elsevier. 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.
DOI: 10.1016/j.inffus.2020.11.008

Biomedical signals carry signature rhythms of complex physiological processes that control our daily bodily activity. The properties of these rhythms indicate the nature of interaction dynamics among physiological processes that maintain a homeostasis. Abnormalities associated with diseases or disorders usually appear as disruptions in the structure of the rhythms which makes isolating these rhythms and the ability to differentiate between them, indispensable. Computer aided diagnosis systems are ubiquitous nowadays in almost every medical facility and more closely in wearable technology, and rhythm or event detection is the first of many intelligent steps that they perform. How these rhythms are isolated? How to develop a model that can describe the transition between processes in time? Many methods exist in the literature that address these questions and perform the decoding of biomedical signals into separate rhythms. In here, we demystify the most effective methods that are used for detection and isolation of rhythms or events in time series and highlight the way in which they were applied to different biomedical signals and how they contribute to information fusion. The key strengths and limitations of these methods are also discussed as well as the challenges encountered with application in biomedical signals.

2.1 Introduction

Physiological processes are complex tasks performed by the different systems of the human body in a rarely periodic but rather irregular manner to deliver an action that could

be biochemical, electrical, or mechanical [93, 94]. Some of these actions are obvious like heart beating, breathing, and other physical activities and some are not as obvious like hormonal stimulation that regulates multiple body functions. The action produced can be usually manifested as some sort of a signal that holds information about the parent physiological process [94]. Disruptions in these physiological processes associated with diseases, lead to the development of pathological processes that alter the performance of the human body. Both normal and pathological processes in addition to other artifacts from the environment and surrounding processes, are all held in the manifested signals and the associated changes in their waveform. These signals are called biomedical signals and can be of many forms including the electrical form (potential or current changes) or physical (force or temperature) [94].

Artificial intelligence is currently taking over to empower a variety of assistive technologies that help solve the problems of the healthcare sector given the continuously increasing cost and shortage of professional caregivers. These technologies are advancing to perform not only diagnosis but also intervention and curing due to the superior sensitivity, adaptability, and fast response. Of these assistive technologies, computer aided diagnosis and wearable systems are powered by the virtual side of artificial intelligence (machine learning techniques) and play a vital role in anomaly detection, monitoring, and even emergency response [95]. The rise of such systems has led to the evolution of biomedical signal analysis which has been the focus of researchers for the last couple of decades. This evolution not only included the macro-analysis of gross processes but also the detection and analysis of micro-events within each gross process [95]. As mentioned before, biomedical signals carry the signatures of many processes and artifacts, which makes the extraction/identification of the specific part of interest (called event or epoch), the first step of any systematic signal analysis or monitoring [96]. Further, the need for robust event extraction algorithms for biomedical signals is driven by the exponential growth of the amount and complexity of data generated by biomedical systems [97]. Moreover, reducing the human-dependent steps in the analysis, mitigates the reliability and subjectivity issues associated with human tolerance.

Epoch extraction is not only essential for systematic signal analysis, but also substantial to information fusion for multi-channel systems and/or sensor networks which represent a

large portion of biomedical-signal-based decision-making systems nowadays. Multiple fusion models can employ epoch extraction and event detection to overcome different obstacles including but not limited to signal synchronization and feature fusion [98, 99]. In complementary data-level fusion, events can be used to align signals as preparation for feature extraction such as using heart beats to align the signals from multiple electrocardiography (ECG) leads. In feature-level fusion models, event detection can be used to combine features from different signals during only the events of interest that contribute to morphology analysis and the decision-making process [99, 100].

Epoch extraction algorithms have been used repeatedly in segmentation of many biomedical signals, including, but not limited to, heart sound and ECG [101, 102], electroencephalography (EEG) [103–105], and swallowing vibrations [106–109]. Such algorithms immensely depend on modeling time-series, the paradigm that is not explicitly provided by regular machine learning and sequence-agnostic models such as support vector machines, regression, and feed forward neural networks [110]. These models depend on a major assumption that the training and test examples are independent and not related in time or space which in result initiates a reset to the entire state of the model [110]. Particularly speaking, splitting time series into data chunks and using consecutive chunks independently in building models is unacceptable because even in the case of modeling a time series with iid processes, the underlying processes might be longer than a single chunk which induces dependency between consecutive chunks.

Sliding window approach has been introduced to tackle the problem of dependence between consecutive chunks through using an overlap which guarantees that a part of each chunk will be carried over to the next chunk. Although this might be useful in modeling many processes, it fails to model long range dependencies and requires the optimization of both data chunk and overlap lengths to best represent the target processes. Additionally, using windowing in time domain provokes a sort of distortion to the frequency representation due to the leakage effect and can only be used for modeling fixed-length input/output scenarios [110]. All of this raised the need for models capable of selectively transferring states across time, processing sequences of not necessarily independent elements, and yielding a computational paradigm that can handle variable-length inputs and outputs [111]. It was

not that long before the researchers started to bring stochastic-based models [112] and design deep recurrent networks [111] to perfectly fit the event extraction problems and overcome the limitations of regular machine learning methodologies.

Multiple models have been offered for time dependency representation including Hidden Markov models (HMMs) and Recurrent Neural Networks (RNNs). HMMs were introduced as an extension to Markov chains to probabilistically model a sequence of observations based on an unobserved sequence of states [112]. On the other hand, RNNs generalize the feed-forward neural networks with the ability to process sequential data one step at a time while selectively transferring information across sequence elements [110]. Hence, RNNs are successful in modeling sequences with unknown length, components that are not independent, and multi-scale sequential dependencies [111, 113, 114]. Further, RNNs overcame a major HMM limitation in modeling the long-range dependencies within the sequences [110, 115].

In this manuscript, we review the fundamental methods developed for event extraction in biomedical signals and unravel the key differences between these methods based on the state-of-the-art practices and results. We show the theoretical and practical aspects for most of the methods and the way in which they were used to handle the time modeling in event detection problems. Further, we discuss the recent major machine learning applications in biomedical signal processing and the anticipated advances for future implementations.

2.2 Hidden Markov Models

A time series can be characterized using either deterministic or statistical models. Deterministic models usually describe the series using some specific properties such as being the sum of sinusoides or exponentials and aim to estimate the values of the parameters contributing to these properties (e.g. amplitude, frequency, and phase of the sinusoides) [112]. On the other hand, statistical models assume that the series can be described through a parametric random process whose parameters can be estimated in a well defined way [112, 116]. HMMs belong to the statistical models category and usually are referred to as probabilistic functions of Markov chains in the literature [112, 117].

2.2.1 Markov Chains

Markov chain is a stochastic process modeled by a finite state machine that can be described at any instance of time to be one of N distinct states. These states can be tags or symbols representing the problem of interest. The machine may stay at the same state or switch to another state at regularly spaced discrete times according to a set of transition probabilities associated with each state [112, 116] and the transition probabilities are assumed to be time independent. The initial state is deemed to be known and the transition probabilities are described using the transition matrix: $A = \{a_{ij}\}$; where a_{ij} is the transition probability from state S_i to state S_j and both i, and j can take values from 1 to N. The actual state at time t is denoted as q_t and for a full description of the probabilistic model, the current state as well as at least the state previous to it (for a first order Markov chain), need to be specified. The first order Markov chain assumes that the current state depends only on the previous state: $P(q_t = j|q_{t-1} = i, q_{t-2} = k, \dots) = P(q_t = j|q_{t-1} = i)$. This results in the following properties for the transition probabilities:

$$\begin{aligned} a_{ij} &= P(q_t = j|q_{t-1} = i); \qquad i \geq 1,\ j \leq N \\ a_{ij} &\geq 0 \\ \sum_{j=1}^{N} a_{ij} &= 1 \end{aligned}$$

The probability of being at state S_i at $t = 1$ is denoted as π_i, and the initial probability distribution as:

$$\begin{aligned} \pi_i &= P[q_1 = S_i]; \qquad 1 \leq i \leq N \\ \Pi &= [\pi_1, \pi_2, \dots, \pi_N]^T \end{aligned}$$

An example of a 4-states Markov chain is shown in Fig. 6. This stochastic process is called the observable Markov model since each state corresponds to a visible (observable) event.

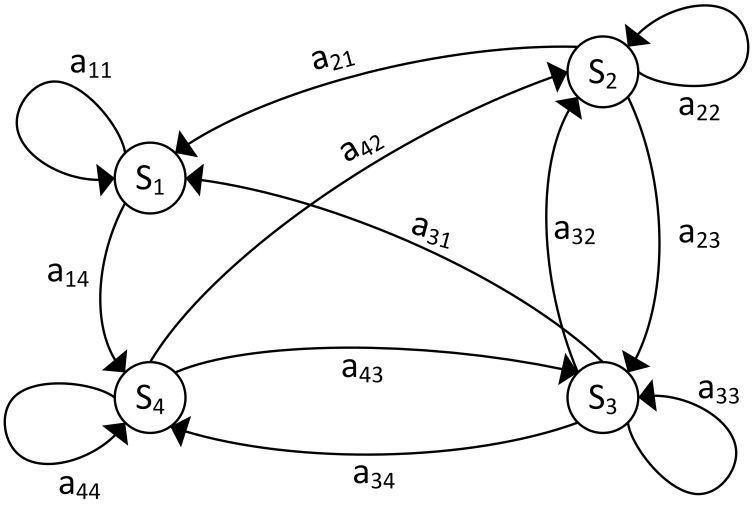


Figure 6: An example of a Markov chain with 4 states, S_1 to S_4, and selected state transitions. A set of probabilities is associated with each state to indicate how the system undergoes state change from one state to itself or another at regular discrete times.

2.2.2 Hidden Markov Models

So far, we introduced Markov chains in which each state corresponds to an observable event, however this is insufficient for most of the applications where the states cannot always be observable. Therefore, Markov chain models are extended to HMMs which can be widely used in many applications [112]. HMM is considered a doubly stochastic process with one of them hidden or not observable; states, in this case, are hidden from the observer [112]. An HMM is characterized through the following properties [112, 116]:

1. The number of states, N, included in the model. As mentioned before, the states are usually hidden in HMMs but sometimes they have a physical significance.

$$q_t \in \{S_1, S_2, \ldots, S_N\}$$

2. The number of distinct observations, a state can take, M.
3. The state transition matrix or distribution $A = \{a_{ij}\}$.

4. The observation probability distribution for each state $B = \{b_j(k)\} = P[v_k \ at \ t | q_t = S_j]$; where v_k represents an element of the distinct observations that a state can take and $1 \leq j \leq N, \ 1 \leq k \leq M$.
5. The initial state distribution $\Pi = \{\pi_i\}$.

When known, the previously mentioned parameters can be used to fully describe the HMM ($\lambda(A, \ B, \ \Pi)$) and generate an observation sequence $O = \{O_1, O_2, \ldots, O_T\}$ as in the algorithm shown in Algorithm 1.

Algorithm 1: HMM as observations generator

```
Set t = 1;
Choose an initial state q_1 = S_i according to Π;
while t ≤ T do
    Choose O_t = v_k according to the observation distribution in the current state (b_i(k));
    Move from the current state S_i to the new state q_{t+1} = S_j according to a_ij;
    set t = t + 1;
end
```

Result: $O = \{O_1, O_2, \ldots, O_T\}$

For the model to be useful for trending applications, it must address three fundamental problems [118]:

- **Likelihood:** Computing the probability of an observation sequence $O = \{O_1, O_2, \ldots, O_T\}$, given the model ($P(O|\lambda)$).
- **Decoding:** Choosing the optimal hidden state sequence $Q = \{q_1, q_2, \ldots, q_T\}$ that best represents a given observation sequence ($O = \{O_1, O_2, \ldots, O_T\}$).
- **Estimation:** Adjusting the model parameters $\lambda(A, \ B, \ \Pi)$ to maximize the likelihood of a given sequence of observations O.

2.2.3 Likelihood Problem Solution

In the case of Markov chains, where the states are not hidden, the computation of the likelihood is much easier as it narrows the computational burden to just multiplying the transition probabilities within the underlying state sequence. In HMMs, states are hidden which necessitates including all possible state sequences in computing the joint probability (N^T possible hidden state sequences). A dynamic programming solution called the forward-backward algorithm was created for the likelihood problem with a simple time complexity

[112]. The forward-backward algorithm sums the probabilities of all possible state sequences that could be included in generating the target observation sequence. The algorithm considers an efficient way to calculate the probability through defining and inductively computing the forward variable $\alpha(t, i)$ which represents the probability of the partial observation sequence $P(O_1\ O_2\ \ldots\ O_t, q_t = S_i|\lambda)$ [112, 119, 120]. The forward algorithm for the likelihood problem is fully described as follows:

Algorithm 2: The forward algorithm

$O = \{O_1, O_2, \ldots, O_T\}$;
$S \in \{S_1, S_2, \ldots, S_N\}$;
Create the forward probability table $\alpha[T, N]$;
foreach *state* $S \in \{S_1, S_2, \ldots, S_N\}$ **do**
  $\alpha[1, S] \leftarrow \pi_S \times b_S(O_1)$; `// Initialization`
end
foreach *time step* $t \in 2, 3, \ldots, T$ **do**
  foreach *state* $S \in \{S_1, S_2, \ldots, S_N\}$ **do**
    $\alpha[t, S] \leftarrow \sum_{\hat{S}=S_1}^{S_N} \alpha[t-1, \hat{S}] \times a_{\hat{S},S} \times b_S(O_t)$; `// Induction`
  end
end
$P(O|\lambda(A,\ B,\ \Pi)) \leftarrow \sum_{S=S_1}^{S_N} \alpha[T, S]$; `// Termination`
Result: $P(O|\lambda(A,\ B,\ \Pi))$

As a part of the forward-backward algorithm, another variable is considered that will be of help in the solution of the estimation problem. The variable is called the backward probability table, $\beta(t, i) = P(O_{t+1},\ O_{t+2},\ \ldots,\ O_T|q_t = S_i, \lambda(A,\ B,\ \Pi))$, which represents the probability of the partial observation sequence that starts one time step after the current observation, given the current state S_i and the model. The backward probability can be calculated in a similar way as the forward probability (Algorithm 3).

2.2.4 Decoding Problem Solution: The Viterbi Algorithm

Finding the optimal hidden states sequence that best represents a sequence of observations is more challenging compared to the likelihood problem. Unlike the likelihood problem, the decoding problem does not have an exact solution unless the model is degenerate, which makes it hard to choose the optimality criterion that judges the state sequence [112]. For example, one may choose states based on the individual likelihood of occurrence which

Algorithm 3: Computing the backward probability

Create the backward probability table $\beta[T, N]$;
foreach *state* $S \in \{S_1, S_2, \ldots, S_N\}$ **do**
  $\beta[T, S] \leftarrow 1$; // Initialization
end
foreach *time step* $t \in T-1, T-2, \ldots, 1$ **do**
  foreach *state* $S \in \{S_1, S_2, \ldots, S_N\}$ **do**
    $\beta[t, S] \leftarrow \sum_{\hat{S}=S_1}^{S_N} \beta[t+1, \hat{S}] \times a_{S,\hat{S}} \times b_{\hat{S}}(O_{t+1})$; // Induction
  end
end
Result: $\beta[T, N]$

achieves the maximum number of correct states individually but not for the overall computed sequence [112]. Another way to solve the decoding problem can be achieved through running the forward-backward algorithm for all possible hidden state sequences and choose the sequence with the maximum likelihood probability, however this is computationally unfeasible [118].

In the same way as the forward-backward algorithm, the Viterbi algorithm solves the decoding problem using dynamic programming. The algorithm recursively computes the probability of being in a state S_j at time t taking in consideration the most probable state sequence (path) $q_1,\ q_2,\ \ldots,\ q_{t-1}$ that leads to this state. The Viterbi algorithm is shown in Algorithm 4.

2.2.5 Model Estimation Problem Solution

The third problem can be formulated as finding HMM's model parameters (A, B, Π) to maximize the conditional probability of observation sequence, given that model [112]. Such a problem doesn't have an analytical solution, however, iterative methods can be used to find a local maxima for $P(O|\lambda)$. Here, we focus on the Baum-Welch algorithm that is based on the expectation-maximization method [121, 122]. The algorithm is based on maximizing Baum's auxiliary function over the updated model parameters λ,

$$Q(\bar{\lambda},\ \lambda) = \sum_{\forall q} P(O_{1:T}, q_{1:T}|\bar{\lambda}) \log P(O_{1:T}, q_{1:T}|\lambda),$$

Algorithm 4: The Viterbi algorithm

$O = \{O_1, O_2, \ldots, O_T\}$;
$S \in \{S_1, S_2, \ldots, S_N\}$;
Create the best path probability table $\delta[T, N]$;
Create the state index table (the index of state that by adding to the path, maximizes δ) $\psi[T, N]$;
foreach *state* $S \in \{S_1, S_2, \ldots, S_N\}$ **do**
  $\delta[1, S] \leftarrow \pi_S \times b_S(O_1)$; `// Initialization`
  $\psi[1, S] \leftarrow 0$;
end
foreach *time step* $t \in 2, 3, \ldots, T$ **do**
  foreach *state* $S \in \{S_1, S_2, \ldots, S_N\}$ **do**
    $\delta[t, S] \leftarrow \max_{\hat{S}=S_1}^{S_N} \delta[t-1, \hat{S}] \times a_{\hat{S},S} \times b_S O_t$; `// Induction`
    $\psi[t, S] \leftarrow \arg\max_{\hat{S}=S_1}^{S_N} \delta[t-1, \hat{S}] \times a_{\hat{S},S} \times b_S O_t$;
  end
end
$P^* \leftarrow \max_{S=S_1}^{S_N} \delta[T, S]$; `// Termination`
$q_T^* \leftarrow \arg\max_{S=S_1}^{S_N} \delta[T, S]$;
for $t \in \{T,\ T-1,\ T-2,\ \ldots,\ 2\}$ **do**
  $q_{t-1}^* \leftarrow \psi[t, q_t]$; `// Backtracking`
end
Result: The optimal state sequence: $q_1^*,\ q_2^*,\ \ldots,\ q_T^*$

where $P(O_{1:T}, q_{1:T}|\lambda) = \pi \prod_{t=1}^{T-1} a_{q_t,q_{t+1}} b_{q_{t+1}}(O_{t+1})$, and $\bar{\lambda}$ is the initial model. The iterations are performed based on the calculations by the forward-backward probabilities described previously in the solution of the first two problems, and they go as described in Algorithm 5.

2.2.6 Continuous Density HMM

The previously described adaptations for HMM problems are based on the requirement that the observations are discrete which is considered restrictive because in most cases they are continuous. Therefore, a necessary first step will be the transformation of continuous observation sequence into a discrete vector. This can be done through dividing the observations' space into sub-spaces and using codebooks to give discrete symbol/value for each sub-space [116]; however, this introduces quantization errors into the problem. One way to overcome this, is using continuous observation densities in HMM's. The finite mixture

Algorithm 5: The estimation algorithm

$O = \{O_1, O_2, \ldots, O_T\}$;
$S \in \{S_1, S_2, \ldots, S_N\}$;
Initialize $\bar{\lambda} = \lambda(A,\ B,\ \Pi)$;
repeat
  Using the forward-backward algorithm and $\bar{\lambda}$ calculate $\alpha[T, N]$ and $\beta[T, N]$;
  Create the probability tables $\xi[T, N, N]$ (the probability of being in a state S_i at time t and a state S_j at time $t+1$) and $\gamma[T, N]$ (the probability of being in a state S_i at time t);
  foreach *time step* $t \in 2, 3, \ldots, T$ **do**
    foreach *state* $S \in \{S_1, S_2, \ldots, S_N\}$ **do**
      foreach *state* $S^* \in \{S_1, S_2, \ldots, S_N\}$ **do**
        $\xi[t, S, S^*] \leftarrow \frac{\alpha[t,S] \times a_{S,S^*} \times b_{S^*}(O_{t+1}) \times \beta[t+1,S^*]}{\sum_{S=S_1}^{S_N} \sum_{S^*=S_1}^{S_N} \alpha[t,S] \times a_{S,S^*} \times b_{S^*}(O_{t+1}) \times \beta[t+1,S^*]}$;
      end
      $\gamma[t, S] \leftarrow \sum_{\bar{S}=S_1}^{S_N} \xi[t, S, \bar{S}]$;
    end
  end
  $\bar{\pi}_S \leftarrow \gamma[1, S]$;
  $\bar{a}_{S,S^*} \leftarrow \frac{\sum_{t=1}^{T-1} \xi[t,S,\bar{S}]}{\sum_{t=1}^{T-1} \gamma[t,S]}$;
  $\bar{b}_S(k) \leftarrow \frac{\sum_{t=1}^{T} \gamma[t,S]}{\sum_{t=1}^{T} \gamma[t,S]}$;
  $\bar{\lambda} = \lambda(\bar{A}, \bar{B}, \bar{\Pi})$;
until *Convergence*;
Result: $\lambda(A,\ B,\ \Pi)$

representation of the observation density function, is one of the representations that has a formulated re-estimation procedure: $b_j(O) = \sum_{m=1}^{M} c_{jm} \mathfrak{N}[O, \mu_{jm}, U_{jm}]$, where $1 \leq j \leq N$, O is the observation vector, c_{jm} is the mixture coefficient for the m^{th} mixture in state j, and $\mathfrak{N}$ is an elliptically or long-concave symmetric density with a mean vector of μ_{jm} and a covariance matrix of U_{jm} [123–125]. A Gaussian density function is usually used for $\mathfrak{N}$; however, other non-Gaussian models have been considered as well in many applications [126, 127]. The pdf is guaranteed to be normalized, given that the mixture coefficients satisfy the following stochastic conditions: $\sum_{m=1}^{M} c_{jm} = 1$ and $c_{jm} \geq 1$, where $1 \leq j \leq N,\ 1 \leq m \leq M$. The parameters of the observation density function $(c_{jm}, \mu_{jm}, U_{jm})$ can be estimated through the modified Baum-Welch algorithm [112]. Using continuous density in HMM makes it more accurate; however, it requires a larger dataset and a more complex algorithm to train.

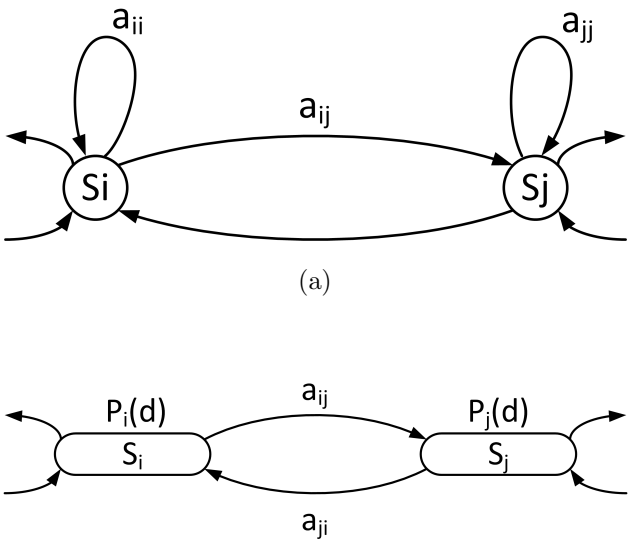


Figure 7: An illustration of interstate connections in HMMs. (a) represents a normal HMM with self transitions from each state back to itself. (b) represents a variable duration HMM with no self state transition and specified state duration densities.

2.2.7 State Duration in HMM

One of the convenient ways to include state duration in HMMs, especially with physical signals, is through explicitly modeling the duration density and setting the self-transition coefficients into zeros [112]. The transition from a state to another only occurs after a certain number of observations, specified by duration density, is made in the current state as shown in Fig. 7. In normal HMMs, the states have exponential duration densities that depend on the self transition coeeficients a_{ii} and a_{jj} as in Fig. 7(a). In HMMs where state duration is modeled by explicit duration densities, there is no self transition and the transition happens only after a specific number of observations determined by the duration density as in Fig. 7(b). The re-estimation formulae needed for model estimation can be defined through including the state duration in the calculation of forward and backward variables. The re-estimation formulae can be found in detail in the tutorial of Rabiner [112].

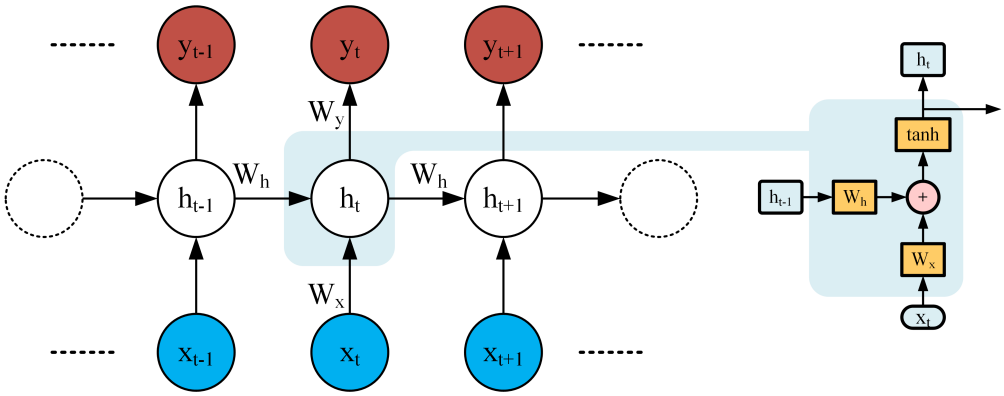


Figure 8: A simple RNN with a single hidden layer. At each time step t, output is produced through passing activations as in a feedforward network. Activations are passed to next node at time $t+1$ as well to achieve recurrence.

2.3 Recurrent Neural Networks

Neural networks are biologically-inspired computational models that are composed of a set of artificial neurons (nodes) joined with directed weighted edges which recently became popular as pattern classifiers [110, 128]. The network is usually activated by feeding an input that then spreads throughout the network along the edges. Many types of neural networks have evolved since its first appearance; however, they will fall under two main categories, the networks whose connections form cycles and the ones that are acyclic [128]. RNNs are the type of neural networks that introduces the notion of time by using cyclic edges between adjacent steps. RNNs have been proposed in many forms including Elman networks, Jordan networks, and echo state networks [129–132].

As shown in Fig. 8, the hidden units at time t receive input from the current input x_t and the previous hidden unit value h_{t-1}. The output y_t is calculated using the current hidden unit value h_t. Time dependency is created between time steps by means of recurrent connections between hidden units. In a forward pass, all the computations are specified using the following two equations: $h_t = \sigma_h\left(W_x x_t + W_h h_{t-1} + b_h\right)$, $y_t = \sigma_y\left(W_y h_t + b_y\right)$; where W_x and W_y represent the matrices of weights between the hidden units and both input and output respectively and W_h is the matrix of weights between adjacent time steps. b_h and b_y

are bias vectors which allow offset learning at each node. Nonlinearity is introduced through the activation functions σ_h and σ_y which can be hyperbolic tangent function (tanh), sigmoid, or rectified linear unit (ReLU). In a simple RNN unit, tanh is usually used.

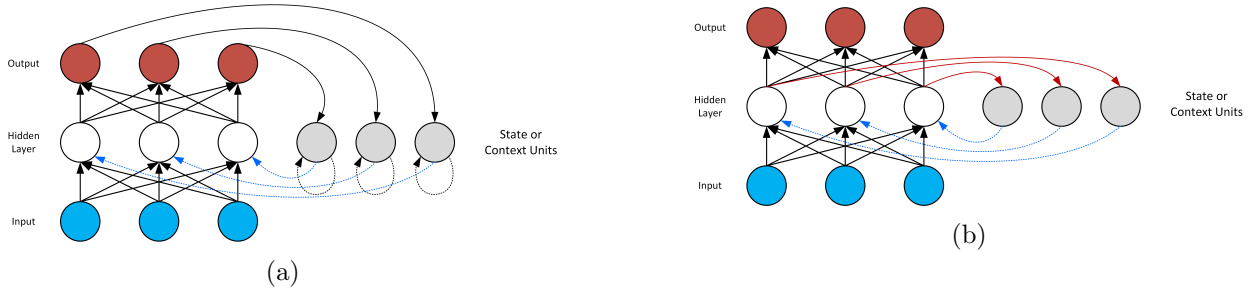


Figure 9: Early designs of RNNs. The dotted arrows represent the edges feeding at the next time step. (a) Jordan network. Output units are connected to context units that provide feedback at next time step to hidden units and themselves. (b) Elman network. Hidden units are connected to the context units that provide feedback to the hidden units only at the next time step.

2.3.1 Early RNN Architectures

Jordan [133] introduced an early form of recurrence in networks by adding extra "special" units called context or state units that feed values to the hidden units in the following time step. The network was as simple as a multi-layer feed-forward network with the context units taking input from the network output at the current time step and feed them back to themselves and the hidden units at the next time step as shown in Fig. 9 (a). The context units allow the network to remember its outputs at previous time steps and being self connected enables sending information across time steps without intermediate output perturbation [110]. Elman [129] also introduced a simple architecture in which the context units are associated with each each hidden layer unit at the current time step and give feedback to the same hidden unit at the next time step as shown in Fig. 9 (b). This notation of self-connected hidden units became the basis for the work and design of long-short term memory (LSTM) units [111]. This type of recurrence has been demonstrated to learn time dependencies by Elman [129].

2.3.2 Training of RNNs

The expression of a generic RNN can be represented as $h_t = \mathcal{F}(h_{t-1}, x_t, \theta) = W_h \sigma_h(h_{t-1}) + W_x x_t + b_h$[1], where θ refers to the network parameters W_h: recurrent weight matrix, W_x: input weight matrix, and b_h: the bias. Initial state h_0, is usually set to zero, provided by user, or learned. Network performance on a certain task is measured through a cost function $\varepsilon = \sum_{1 \leq t \leq T} \varepsilon_t$, where $\varepsilon_t = \mathcal{L}(h_t)$, T is the sequence length (total number of time steps), and $\mathcal{L}$ is the cost operator that measures the performance of the network (e.g. squared error and entropy). Necessary gradients for optimization can be computed using backpropagation through time (BPTT), where the network is unrolled in time so that the application of backpropagation is feasible as shown in Fig. 10.

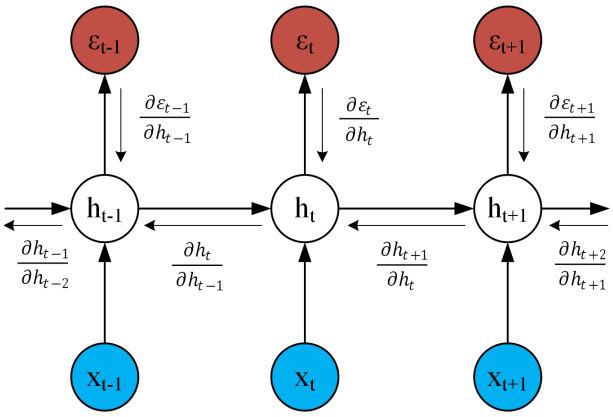


Figure 10: Unfolded recurrent neural network in time [134]. ε_t denotes the error calculated from the output, h_t represents the hidden state, and x_t represents the input at time t.

A gradient component $\frac{\partial \varepsilon}{\partial \theta}$ is calculated through the summation of temporal components as follows:

$$\frac{\partial \varepsilon}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\partial \varepsilon_t}{\partial \theta}$$

$$\frac{\partial \varepsilon_t}{\partial \theta} = \sum_{1 \leq k \leq t} \left(\frac{\partial \varepsilon_t}{\partial h_t} \times \frac{\partial h_t}{\partial h_k} \times \frac{\partial h_k}{\partial \theta} \right)$$

$$\frac{\partial h_t}{\partial h_k} = \prod_{t \geq i > k} \frac{\partial h_i}{\partial h_{i-1}} = \prod_{t \geq i > k} W_h^T \, diag\left(\sigma_h'\left(h_{i-1}\right)\right) \tag{2.1}$$

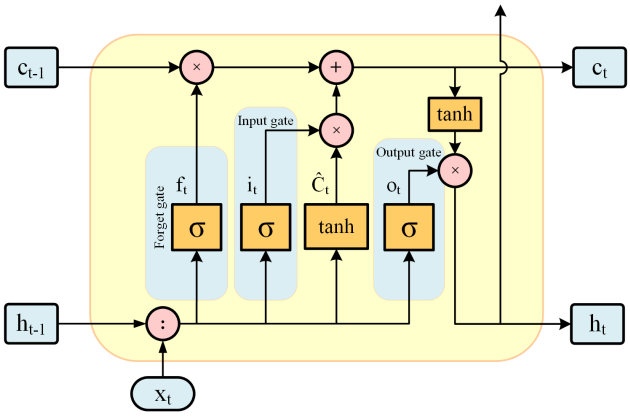
[1]This formulation doesn't contradict with the previously mentioned formulation ($h_t = \sigma_h\left(W_x x_t + W_h h_{t-1} + b_h\right)$) and both have the same behavior [134].

The effect that the network parameters (θ) at step k have over the cost at subsequent steps ($t > k$), can be measured through the temporal gradient component $\frac{\partial \varepsilon_t}{\partial h_t} \times \frac{\partial h_t}{\partial h_k} \times \frac{\partial h_k}{\partial \theta}$. In Eq. 2.1, the matrix factors are in the form of a product of $t-k$ Jacobian matrices which will either explode or shrink to zero depending on whether the recurrent weights are greater or smaller than one [134]. The vanishing gradient is common when using sigmoid activations, while the exploding gradient is more common when using rectified linear unit activations [110, 134]. Enforcing the weights through regularization to values that help avoid gradient vanishing and exploding, is one of the solutions to such a problem. Truncated backpropagation through time (TBPTT) is also used as another solution for exploding gradient through setting a maximum number of time steps through which the error is propagated [110].

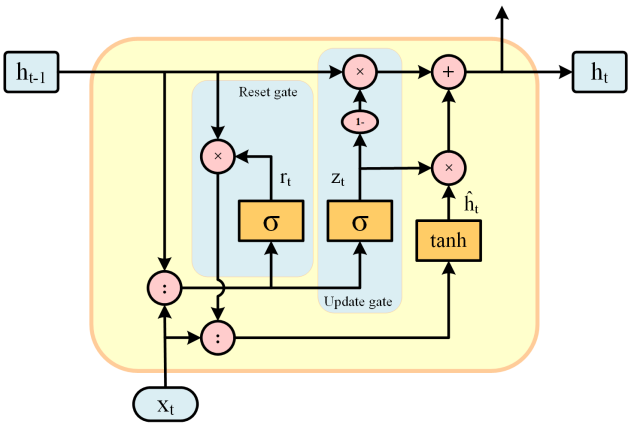
2.3.3 Current RNN Designs

Although early designs of RNNs helped to map input into output sequences through using contextual information, this contextual mapping had limited range and the influence of input on hidden layers and thus output, either vanishes or blows up due to cycling through the network recurrent connections as described previously [135]. Gradient vanishing/exploding problem has led to the emergence of new network designs that improve convergence [136, 137]. Of these designs, LSTM, gated recurrent units (GRU), and bidirectional RNNs (BRNN) have proved superiority in long-range contextual mappings and employing both future and past contexts to determine the output of the network [110]. Both LSTM and GRU resemble a standard RNN but with each hidden node replaced by a complete cell as shown in Fig. 11. They also employ a unity-weighted recurrent edge to ensure the transfer of gradient across time steps without decaying or exploding. LSTM forms the long-term memory through the weights which change slowly during training. On the other hand, short term memory is formed by transient activations that pass between successive node [110]. GRU is an LSTM alternative that has a simpler structure and is faster to train; however, it still provides comparable performance to LSTM [138].

In an LSTM unit, a forget gate is an adaptive gate whose output is squashed through a sigmoid activation in order to reset the memory blocks once they are out of date and prevent



(a)



(b)

Figure 11: Current designs of RNNs. The symbols used in both diagrams are as follows, : represents concatenation, + represents element-wise summation, × represents element-wise multiplication, σ represents a sigmoid activation, and *tanh* represents a hyperbolic tangent (tanh) activation. (a) Schematic of an LSTM unit which is typically composed of three main parts, input, output, and forget gates. (b) Schematic of a GRU unit which is a simplified version of LSTM with only reset and update gates.

information storage for arbitrary time lags [139]. The input gate is a sigmoid activated gate whose function is to regulate the new information to be written to the cell state. The output gate is also a sigmoid activated gate that regulates the internal state after being dynamically customized through a *tanh* activation to be forwarded as the unit output. In the same way,

the GRU unit has a similar design; however, it doesn't have an output gate. It has a reset gate that works as a forget gate and an update gate to regulate the write operation into the unit output from both the state of the past time step and the input from the current time step.

On the other hand, BRNNs resemble a standard RNN architecture as well but with two hidden layers instead of one and each hidden layer is connected to both input and output . One hidden layer passes activations in the forward directions (from the past time steps) and the other layer passes the activations in the backward direction (from future time steps). BRNN is in fact a wiring method for RNN hidden layers regardless of the type of the nodes, which makes it compatible with most RNN architectures including LSTM and GRU [110, 136].

2.4 Critical Differences between HMMs and RNNs

As demonstrated in the previous sections, construction of hidden Markov models relies on a representing state space from which states are drawn. Scaling such system has long been considered to be difficult or infeasible even with the presence of dynamic programming solutions such as the Viterbi algorithm due to the quadratic complexity nature of the inference problem and transition probability matrix which causes the model parameter estimation and inference to scale in time as the size of the state space grows [140]. Modeling long range dependencies also is impractical in HMMs as transitions occur from a state to the following with no memory of the previous state unless a new space is created with all possible cross-transitions at each time window which leads to exponential growth of the state space size [110, 115]. On the other hand, the number of states that can be represented by a hidden layer in RNNs increases exponentially with the number of nodes in the layer leading to nodes that can carry information from contexts of arbitrary lengths. Moreover, despite of the exponential growth of the expressive power of the network, training and inference complexities only grow quadratically at most [110]. From a theoretical point of view, RNNs can be efficient in the perception of long contexts; however, this comes at the cost of error

propagation. Highly sampled inputs as in the case of raw waveforms, can lead to elongation of the range through which the error signal propagates, thus making the network hard to optimize and reducing the efficiency of computational acceleration tools such as GPUs [141, 142].

2.5 Event Detection in Electrocardiography

ECG is the graphical interpretation of skin-recorded electrical activity of the electric field originating in the heart [143]. ECG provides information that is not readily available through other methods about heart activity and is considered the most commonly used procedure in the diagnosis of cardiac diseases due to the fact that it is non-invasive, simple, and cost-effective. This makes ECG subject to intense research related to the automatic analysis to reduce the subjectivity and the time spent on interpreting hours of recordings [144–146]. ECG is a time periodic signal, which allows to mark out an elementary beat that constitutes the basis for ECG signal analysis [145]. For instance, heart rate can be estimated through the detection of QRS-complex from an ECG signal and the time interval between successive QRS-complexes (also known as R-R interval) can be used to detect premature ectopic beats [143]. In that sense, ECG beat detection is considered fundamental for most of the automated analysis algorithms. A detailed description of the recent publications that cover event detection in ECG using different methods, is included in Table 2.

Table 2: Summary of event detection work done in ECG event detection.

Publication	Event under Investigation	Implementation details	Dataset
Gersch et al. [147], 1975	Premature Ventricular Contraction (PVC) through R-R intervals	A three states Markov chain was used to model R-R interval (quantized as short, regular, or long) sequences and then the model is used to characterize rhythms through the probability that the observed R-R symbol sequence is generated by any of a set of models generated from multiple cardiac arrhythmias. Theb manuscript used a maximum likelihood approach to determine the arrhythmia type.	Clinical test data from patients with atrial fibrillation (AF)
Coast et al. [148], 1990	Beat detection for arrhythmia analysis	A parallel combination of HMMs (one for each arrhythmia type), is used to classify arrhythmia. The classification process is inferred through determining the most likely path through the parallel models. All ECG waveform parts were included in the states of each model. The results reported in this study relied on single ECG channel and didn't include multi-channel ECG fusion.	The American Heart Association (AHA) ventricular arrhythmia database [149]

Continued on next page

Table 2 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Andreao et al. [145], 2006	ECG beat detection and segmentation	An HMM was constructed for ECG beat with each waveform part represented in the model including the isoelectric parts (ISO, P, PQ, QRS, ST, T). Model parameters were estimated using Baum-Welch method and the number of states in each model were specified empirically to achieve a good complexity-performance compromise. The proposed segmentation in this study was based on a single channel but the authors provided insights about the possibility of adaptation with multi-channel fusion.	QT database [150]
Sandberg et al. [151], 2008	Atrial fibrillation frequency tracking	An HMM is used for frequency tracking to overcome the corruption of residual ECG by muscular activity or insufficient beat cancellation. States of the HMM were used to represent the underlying frequencies in short-time Fourier transform while observations corresponded to the estimated frequency of specific time intervals from the signal. Experiments were performed on single channel simulated signals with inclusion of mutli-channel fusion.	Simulated atrial fibrillation signals with four different frequency trends: constant frequency, varying frequency, gradually decreasing frequency, and stepwise decreasing frequency.
Oliveira et al. [152], 2017	Automatic segmentation (beat) of ECG and Phonocardiogram (PCG)	An ECG channel along with a phonocardiogram were fused in a single coupled HMM for beat detection. The coupled HMM was constructed to consider the high dynamics and non-stationarity of the signals where the channels were assumed to be co-dependent through past states and observations. Each of ECG and phonocardiogram was modeled using 4 states. This study introduced a decision-level fusion through combining two channels in a single HMM. The study also experimented two different coupled HMMs, a fully connected where transition can happen between any two states from both channels and a partially connected model where certain limitations were added over transitions through considering the prior knowledge of the relationship between heart sounds and ECG components.	A self-recorded dataset from healthy male adults.
Übeyli [153], 2009	Arrhythmia detection/classification	An Elman-based RNN is used for beat classification with the Levenberg-Marquardat algorithm for training (a least-squares estimation algorithm based on the maximum neighborhood idea). This model used power spectral density (calculated with three different methods; Pisarenko, MUSIC, and Minimum-Norm) of ECG signals as input. All the models trained in this study, used feature-level fusion.	Four types of ECG beats obtained from Physiobank Database [154].
Zhang et al. [155], 2017	Supraventriular and verntricular ectopic beat detection (SVEB and VEB)	An LSTM-based RNN preceded by a density-based clustering for training data selection from a large data pool. In this implementation, the authors fed the RNN with the current ECG beat and the T wave part from the former beat to automatically learn the underlying features. The RNN layers were followed by two fully connected layers in order to combine the temporal features and generate the desired output. This study only used a single channel ECG (limb lead II) with no multi-channel fusion.	MIT-BIH Arrhythmia database (MITDB) [156].
Xiong et al. [157], 2017	Atrial fibrillation automatic detection	A 3 layer RNN was implemented to extract the temporal features from the raw ECG signals. No multi-channel fusion was performed in this study and only a single ECG channel was employed.	The 2017 PhysioNet/CinC Challenge dataset [154].
Schwab et al. [141], 2017	Different cardiac arrhythmia classification/detection	In this work a combination of GRU and bidirectional LSTM (BLSTM) based RNNs and nonparameteric Hidden Semi-Markov Models (HSMM), was used for building the beat classification model and then a blender [158] was used to combine the predictions from the models. No multi-channel fusion was performed in this study and only a single ECG lead was employed.	The 2017 PhysioNet/CinC Challenge dataset [154].
Zihlmann et al. [159], 2017	Atrial fibrillation detection	A single layer LSTM-based convolutional RNN (CRNN) was constructed for atrial fibrillation detection in arbitrary length ECG recordings. This work employed the log spectrogram as an input to the CRNN to increase the accuracy. No multi-channel fusion was performed in this study and only a single ECG lead was used.	The 2017 PhysioNet/CinC Challenge dataset [154].
Limam and Precioso [160], 2017	Atrial fibrillation detection	A two layer LSTM-based CRNN was used for atrial fibrillation detection from single-lead ECG and heart rate. Feature-level fusion was performed after the convolutional neural network (CNN) layers to combine features from both inputs. The output from the RNN was used to either feed a dense layer to perform classification directly or train an SVM for classification and the results from both models were compared.	The 2017 PhysioNet/CinC Challenge dataset [154].

Continued on next page

Table 2 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Chang et al. [161], 2018	Atrial fibrillation detection	A single layer LSTM-based RNN was constructed for atrial fibrillation detection in multi-lead ECG. This model also used spectrograms of the input ECG signals to feed the network. Feature-level fusion was performed to combine spectrograms of multi-lead ECG before feeding into the LSTM units.	Multiple datasets for atrial fibrillation and normal sinus rhythms [154, 156, 162–165].
Lui and Chow [166], 2018	Myocardial infarction classification	A deep single-layer LSTM based CRNN was used for classifying ECG beats from single-lead ECG. Multiple models were performed including a direct 4-class beat classifier from the LSTM CRNN via dense layers and 4-class beat classifier via the fusion of multiple one-versus-one binary classification networks using stacking.	The Physikalisch-Technische Bundesanstalt (PTB) diagnostic ECG database [165] and the 2017 AF-Challenge [167].
Singh et al. [168], 2018	Arrhythmia detection	3 models were built for arrhythmia detection, each of them is based on a different type of RNN. Regular RNNs, GRU, and LSTM were used for each of the three models. Each model included 3 layers of different unit sizes with a dense layer to generate a classification output (normal/abnormal). No multi-channel fusion was performed in this study and only a single ECG lead (ML2) was employed.	MIT-BIH Arrhythmia database (MITDB) [156].

2.6 Event Detection in Electroencephalography

EEG is mostly a non-invasive technique to measure the electrical activity of the brain through a set of electrodes placed on the subject's scalp. EEG exhibits highly non-stationary behavior and significant non-linear dynamics [169]. The excitatory and inhibitory postsynaptic potentials of the cortical nerve cells are considered the main source of EEG signals [170]. EEG can be invasive if acquired using subdural electrode grids or using depth electrodes and is called intracranial EEG (iEEG); however, typical EEG signals are recorded from scalp locations specified by the 10-20 electrode placement criterion designed by the International Federation of Societies for Electroencephalography and have an amplitude of 10-100 μV and a frequency range of 1-100 Hz [169, 170]. EEG signals are used in the diagnosis of multiple neurological disorders including epilepsy, lesions, tumors, and depression and their characteristics depend strongly on the age and state of the subject. There are multiple events that influence EEG and require the tedious job of analyzing hours of recordings to be extracted. These events range from the diagnosis/detection of certain seizures and syndromes to the tasks of brain computer interface (BCI). These events include the different sleep stages and sleep disorders, epileptic seizures, the effect of music or other artifacts, and the motor imagery tasks.

2.6.1 Sleep Staging in EEG

Sleep is an essential part of the human life cycle and plays a vital role in maintaining most of the body functionality [171]. Sleep disorders include problems with initiating sleep, insomnia, and sleep apnea syndrome (SAS) [172]. Diagnosis of sleep disorders can be done through identifying sleep stages in an overnight polysomnogram (PSG) which utilizes EEG as one of its sensing modalities [173]. Visual scoring of the PSG components is the basic way to categorize sleep epochs and as any manual rating, it suffers from subjectivity and inter-rater tolerance. Many attempts have been proposed in the literature to remedy the problems of expert-based visual scoring of the different components of PSG. The attempts employed multiple algorithms to achieve automatic sleep staging including Markov models and neural networks. Here, we list the recent publications (Table 3) for sleep staging and the detailed description of the methods used within the scope of our review.

Table 3: Summary of EEG-based sleep staging.

Publication	Event under Investigation	Implementation details	Dataset
Flexerand et al. [174], 2002	Sleep staging in combined EEG and EMG	A three state (wakefulness, deep sleep, and rapid eye movement sleep) Gaussian observation HMM (GOHMM) was used and sleep stages were represented as mixtures of the basic three states. The probability of being in any of the three states was computed for 1 sec windows so that a continuous probability monitoring can be achieved. Expectation-maximization algorithm was used for parameter estimation and the Viterbi algorithm was used to calculate the posteriori estimate for being in each state. Feature-level fusion was performed on features from EEG channels (C3 and C4) and EMG.	Nine whole-night sleep recordings from a group of nine healthy adults.
Flexer et al. [175], 2005	Sleep staging in single channel EEG (C3)	A three state (wakefulness, deep sleep, and rapid eye movement sleep) Gaussian observation HMM (GOHMM) was used and sleep stages were represented as mixtures of the basic three states. The probability of being in any of the three states was computed for 1 sec windows so that a continuous probability monitoring can be achieved. Expectation-maximization algorithm was used for parameter estimation and the Viterbi algorithm was used to calculate the posteriori estimate for being in each state. No multi-channel fusion was performed in this study and only a single EEG channel was used.	Two datasets were used, the first consists of 40 whole night sleep recordings from healthy adults and the second consists of 28 whole night sleep recordings of healthy adults.
Doroshenkov et al. [176], 2007	Sleep staging using two channel EEG (Fpz-Cz and Pz-Oz)	A six state HMM was constructed for the purpose of sleep staging. Baum-welch algorithm was used for model's parameter estimation and the Viterby algorithm for state sequence decoding. Feature-level fusion was performed for features calculated from the two EEG channels.	Sleep-EDF database [177].
Bianchi et al. [178], 2012	Sleep cycle (quantifying probabilistic transitions between stages and multi-exponential dynamics) and fragmentation in case of apnea in PSG	An eight state HMM was constructed for sleep-wake activity. The connectivity between states was inferred through exponential fitting of subsets of the pooled bouts and adjacent-stage analysis.	Sleep Heart Health Study database [179].

Continued on next page

Table 3 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Pan et al. [180], 2012	Sleep staging using central EEG (C3-A2), chin electromyography (EMG), and electrooculogram (EOG)	A six state transition-constrained discrete HMM was constructed for sleep staging. Thirteen features were utilized including temporal and spectrum analyses of the EEG, EOG and EMG signals with feature-level fusion employed.	PSG including six channel EEG, EOG, EMG, and ECG signals, was obtained from 20 healthy subjects.
Yaghouby and Sunderam [181], 2015	Sleep staging and scoring (quasi-supervised) in PSG	A five state Gaussian HMM was constructed for sleep staging with Baum-Welch algorithm for parameter estimation. In this implementation, feature-level fusion was achieved through feeding augmented vector of PSG features and human rated scores into the estimation algorithm in order to obtain the parameters to maximize the likelihood that a model with larger number of states explains the data.	Sleep-EDF database [177].
Onton et al. [182], 2016	Sleep staging in 2-channel home EEG (FP1-A2 and FP2-A2) and electrodermal activity (EDA)	A five state Gaussian HMM was constructed for sleep staging with expectation-maximization algorithm for parameter estimation and the Viterbi algorithm to find the maximum a posteriori estimate of state sequence. In this implementation, the relative power across the entire night was averaged in five frequency bands and fed into the model (feature-level fusion).	A self recorded data from 51 participants who were medication-free and self-reported asymptomatic sleepers and wit no history of neurologic or psychiatric disorders.
Davidson et al. [183], 2005	Behavioral microsleep detection in EEG (P3-01 and P4-02)	This study utilized an LSTM-based RNN to detect the lapses in visuomotor performance associated with behavioral microsleep events. The network used the power spectral density of 1 sec windows of the used two channels (calculated using the covariance method) with feature-level fusion in place to combine data. The network included 6 LSTM blocks of 3 memory cells each.	A self-recorded dataset from 15 subjects performing visuomotor tracking task.
Hsu et al. [184], 2013	Automatic deep sleep staging in single channel EEG (Fpz-Cz)	This study utilized an Elman recurrent neural network that works on the energy features extracted from a single channel EEG to perform 5-level sleep staging. No multi-channel fusion was employed in this study.	Sleep-EDF database [177].
Supratak et al. [185], 2017	Automatic sleep staging in single channel EEG (Fpz-Cz or Pz-Oz)	A convolutional RNN (CRNN) was constructed to work directly of the raw signal data. Two branches of CNN, each of 4 layers, were used for representation learning and their outputs were combined and fed into a two layer LSTM-based BRNN with skip branch to generate the sleep stage. No multi-channel fusion was employed in this study.	Montreal Archive of Sleep Studies (MASS) [186] and Sleep-EDF database [177].
Biswal et al. [187], 2017	Automatic sleep staging	Raw EEG signals were split into 30-seconds windows, then the spectrogram and expert defined features were extracted and fused at the feature-level. The best accuracy reported among different RNN architectures, was reported for a 5-layer LSTM-based RNN. This study presented also an LSTM-based CRNN architecture to extract spatial features automatically and then pass them to the RNN part for temporal context extraction.	10,000 PSG studies with multi-channel EEG data (F3, F4, C3, C4, O1 and O2 referenced to the contralateral mastoid, M1 or M2).
Phan et al. [188], 2018	Automatic deep sleep staging in single channel EEG (Fpz-Cz)	A two-layer GRU-based BRNN was constructed to learn temporal features from the single channel EEG. This implementation included an attention mechanism that was applied on the BRNN output features. The weighted output was then used to feed a linear SVM classifier. No multi-channel fusion has been employed in this study.	Sleep-EDF database [177].
Bresch et al. [189], 2018	Sleep staging in single-channel EEG	An LSTM-based CRNN with 3 CNN layers and 3 LSTM layers, was built to process 30-seconds windows of raw EEG data (FPz, left EOG, and right EOG referenced to M2). No multi-channel fusion has been employed in this study.	The SIESTA database [190] and a self-recorded dataset with 147 recordings from 29 healthy subjects.
Phan et al. [191], 2019	Automatic sleep staging	This study featured multi-modality fusion on the feature level between EEG, EOG, and EMG. All were split into windows and converted into time-frequency representation using filter banks. The fused data were fed into a BRNN that is used to encode the features, then the output is passed through an attention layer followed by another BRNN that performs the cclassification of the sleep stage.	Montreal Archive of Sleep Studies (MASS) Dataset [186].
Michielli et al. [192], 2019	Automatic sleep staging in single channel EEG	A dual branch LSTM-based RNN was constructed for the classification of 5 different sleep stages. the network starts with a preprocessing and feature extraction stages and then the data is distributed over two branches. The first branch uses mRMR for feature selection followed by a one layer LSTM and fully connected layer to classify between 4 classes only (W, N1-REM, N2 and N3). The second branch uses PCA for feature selection followed by a 2 layer LSTM and a fully connected layer for binary classification. The LSTM in the second branch takes the classification output from the first branch to consider only the combined stage N1-REM for separation. No multi-channel fusion has been employed in this study.	Sleep-EDF database [154].

Continued on next page

Table 3 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Sun et al. [193], 2019	Sleep staging in single channel EEG	A two stage network was built to perform the classification. The first stage is time distributed stage that included two parallel branches, the first included a window deep belief network for feature extraction followed by a dense layer and a second branch with hand-crafted features extraction then a dense layer. The two branches were then fused through another dense layer and fed as an input to an LSTM-based BRNN (the second stage) to generate the classes.	Sleep-EDF database [154].

2.6.2 Epilepsy Detection in EEG

Epilepsy is one of the episodic disorders of the brain that is characterized by recurrent seizures, unjustified by any known immediate cause [194, 195]. Epileptic seizure is the clinical manifestation that results from the abnormal excessive discharge of some set of neurons in the brain [194]. The seizure consists of transient abnormal alterations of sensory, motor, consciousness, or psychic behavior [194, 195]. Around 80% of the epileptic seizures can be effectively treated if early discovered [196]. Although seizure activity can be easily distinguished in EEG as transient spikes and relatively quiescent periods, it is a time-consuming process and needs clinicians to devote a tremendous amount of time going through hours and days of EEG activity [197]. An efficient and reliable seizure prediction/detection method can be of a great help for the diagnosis, treatment, and even early warning for patients to stop activities that might be of a significant danger during an episode like driving. Several methods have been proposed for seizure prediction, at which EEG signal features are temporally analyzed and compared to heuristic thresholds to trigger a warning for seizures; however, these methods lack generalization when investigated on extensive datasets [198–203]. This can be referred to using feature sets that are not highly affected by the transition from seizure-free to peri-ictal or seizure states or simply the effect cannot be tracked using low-order statistics [198]. Therefore, stochastic-based models, multivariate analysis, and long-range analysis methods were investigated to provide better performance and generalization for EEG-based epileptic seizure prediction. In Table 4, we review the recent publications that use HMMs and RNNs for seizure prediction.

Table 4: Summary of EEG-based seizure prediction.

Publication	Event under Investigation	Implementation details	Dataset
Wong et al. [204], 2007	Evaluation framework for seizure prediction in iEEG	A three state HMM (baseline, detected, and seizure) was constructed to evaluate the prediction algorithms of epileptic seizures. The prediction algorithm is used to generate a binary sequence which is combined with the ground truth (binary detector outputs plus gold-standard human seizure markings) and converted into a trinary observation sequence. The trinary vector is used to train the HMM using Baum-Welch which is then used to Viterbi decode the observation sequences into the hidden states sequence. A hypothesis test that a statistical association exists between the detected and seizure states, is performed through counting the transitions from detected state into seizure states in the HMM output.	iEEG data collected from patients diagnosed with mesial temporal lobe epilepsy using 20-36 surgically implanted electrodes on the brain or brain substance [205].
Santaniello et al. [198], 2011	Early detection of seizures in iEEG from a rat model	Multichannel iEEG were used and Welch's cross power spectral density was calculated over windows of 3 sec for each pair of channels which were used as input for the detection model. A two state HMM was constructed to map the iEEG signals into either normal or peri-ictal states. Baum-Wlech algorithm was used for parameter estimation and a Bayesian evolution model was used determine the time of state transition.	Data collected from male Sprague-Dawley rats with four implanted skull screw EEG electrodes placed bifrontally and posteriorly behind bregma and a fifth depth electrode placed in hippocampus, were collected and used for this study.
Direito et al. [206], 2012	Identification of the different states of epileptic brain	The relative power in EEG sub-bands (delta, theta, alpha, beta, and gamma) was calculated and used for computing the topographic maps of each sub-band. The maps were then segmented and used overtime to train a 4 state (preictal, ictal, postictal and interictal) HMM. The Baum–Welch algorithm was used to train the model and the Viterbi algorithm to decode the state-sequence.	EPILEPSIAE database [207].
Abdullah et al. [196], 2012	Seizure detection in iEEG	A three state discrete HMM was built to classify iEEG segments into one of three states (ictal, preictal, and interictal). Seven level decomposition stationary wavelet transform (SWT) was applied on the signals (as input features for the model) and a code book was created to perform vector quantization. Baum-Welch algorithm was used for model parameter estimation and the Viterbi algorithm for recognition. This study employed a feature-level fusion model to feed the data into the prediction model.	Freiburg Seizure Prediction EEG (FSPEEG) database [200].
Smart and Chen [197], 2015	Seizure detection in scalp EEG	This study used a 5 sec sliding window with 1 sec increments to process the EEG signals. A set of 45 measurements was calculated for each sliding window then principal component analysis (PCA) was used to reduce dimensionality. One of the used models was HMM, particularly a two state (seizure and non-seizure) HMM was constructed to perform the detection. Baum-Welch was used here as well to estimate the model parameters.This study used a feature-level fusion model for multi-channel EEG data to feed the data into the prediction model.	CHB-MIT Scalp EEG Database [208].
Petrosian et al. [209], 2000	Onset detection of epileptic seizures in both scalp and intracranial EEG	Both raw EEG data and their wavelet transform "daub4" were used in training an Elman RNN. This study used a feature-level fusion model for multi-channel EEG data to provide an input for the RNN.	Scalp and iEEG data were collected from two patients who were undergoing long-term electrophysiological monitoring for epilepsy.
Güler et al. [210], 2005	Identification of subject condition in terms of epilepsy (healthy, epilepsy patient during seizure-free interval, and epilepsy patient during seizure episode) using surface and intracranial EEG	Lyapunov exponents of the EEG signals were used to train an Elman RNN for the identification task. This study used a feature-level fusion model for multi-channel EEG data to train the RNN.	Publicly available epilepsy dataset by University of Bonn [211].
Kumar et al. [212], 2008	Automatic detection of epileptic seizure in surface and intracranial EEG	Wavelet and spectral entropy were extracted from the EEG signals and used to train an Elman RNN. This study used a feature-level fusion model for multi-channel EEG data to train the RNN.	Publicly available epilepsy dataset by University of Bonn [211].

Continued on next page

Table 4 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Minasyan et al. [213], 2010	Automatic detection of epileptic seizures prior to or immediately after clinical onset in scalp EEG	A set of time domain, spectral domain, wavelet domain, and information theoretic features were used to train an ELman RNN per each channel of the EEG and the output is combined in time and space through a decision making module that performs a decision-level fusion in order to declare a seizure event if N out of M channels declared it.	EEG dataset from 25 patients hospitalized for long-term EEG monitoring in five centers including Thomas Jefferson University, Dartmouth University, University of Virginia, UCLA and University of Michigan medical centers.
Naderi and Mahdavi-Nasab [214], 2010	Automatic detection of epileptic seizure in surface and intracranial EEG	Power spectral density was calculated for EEG signals using Welch method then a dimensionality reduction algorithm was applied and the output was used to train an ELman RNN. This study used a feature-level fusion model for multi-channel EEG data to train the RNN.	Publicly available epilepsy dataset by University of Bonn [211].
Vidyaratne et al. [215], 2016	Automated patient specific seizure detection using scalp EEG	The preprocessed (denoised) EEG signals were segmented into 1 sec non overlapping epochs and used to train a BRNN. Data from all channels were used simultaneously (feature-level fusion model).	CHB-MIT Scalp EEG Database [208].
Talathi [216], 2017	Epileptic seizures detection	Single-channel EEG data (no multi-channel fusion) were used to train a GRU-based RNN that classifies each EEG segment into one of three states: healthy, inter-ictal, or ictal. Two layers of GRU were used, the first was followed by a fully connected layer and the second was followed by a logistic regression classification layer.	Publicly available epilepsy dataset by University of Bonn [211].
Golmohammadi et al. [217], 2017	Epileptic seizure detection	Linear frequency cepstral coefficient feature extraction was performed for the EEG data and used to feed a CRNN that is based on a bidirectional LSTM. Features from multi-channel EEG were fused prior to feeding into the CRNN. The network used in this study employed both 2D and 1D CNN at different stages. Another network where LSTM was replaced with GRU was devloped as well for comparison.	A subset of the TUH EEG Corpus (TUEEG) [218] that has been manually annotated for seizure events [219].
Raghu et al. [220], 2017	Epileptic seizures classification	This study developed two techniques that are based on Elman RNN that works on features extracted from EEG signals. The first technique used wavelet decomposition with the estimation of log energy and norm entropy to feed the RNN classifier (normal vs preictal). The second way extracted the log energy entropy to feed the RNN classifier.	Publicly available epilepsy dataset by University of Bonn [211].
Abdelhameed et al. [221], 2018	Epileptic seizure detection	This study used raw EEG signals to feed a 1D CRNN that is based on bidirectional LSTM to classify EEG segments into one of two states (normal-ictal and normal-ictal-interictal).	Publicly available epilepsy dataset by University of Bonn [211].
Daoud and Bayoumi [222], 2018	Epileptic seizure prediction	This study used raw EEG signals to feed a 2D CRNN that is based on a bidirectional LSTM to classify EEG segments into one of two classes (preictal and interictal).	A dataset recorded at Children's Hospital Boston which is publicly available [154, 208].
Hussein et al. [223], 2019	Epileptic seizures detection	This study developed an LSTM-RNN that takes raw EEG signals as input in order to create predictions. The network was composed of a one layer LSTM followed by a fully connected layer and an average pooling layer to combine the temporal features and then an output softmax layer.	Publicly available epilepsy dataset by University of Bonn [211].

2.6.3 BCI Tasks in EEG

Motor imagery alters the the neural activity of the brain's sensorimotor cortex in a way that is as observable as if the movement was really executed [224]. Identification of the transient patterns in EEG signals during the different motor imagery tasks like imagining the movement of one of the limbs, is recognized among the most promising and widely used techniques of BCI [225–228]. This is referred to the relatively low cost of the systems used and the high temporal resolution [227]. This type of BCI is called asynchronous BCI

because the subject is free to invoke specific thought [224]. On the other hand, synchronous BCI includes the generation of specific mental states in response to external stimuli [224]. EEG analysis for BCI applications includes the processing of EEG oscillatory activity and the different shifts in its sub-bands in addition to the event-related potentials like VEP and P300 [224, 229]. Many modeling schemes have been introduced to solve the of multi-class BCI problem; however, most of them process EEG signals in short windows where stationarity is assumed, which limits the modeling process and excludes the dynamic EEG patterns such as desynchronization [224]. To overcome such a limitation, probabilistic models like HMMs and models capable of representing long range dependencies have been proposed into the implementation of BCI systems. As follows in Table 5, we list the recent work the relies on HMMs and RNNs in BCI systems and uses EEG as the source signal.

Table 5: Summary of EEG-based BCI systems.

Publication	Event under Investigation	Implementation details	Dataset
Obermaier et al. [230], 2001	5 tasks BCI system (imagining left-hand, right-hand, foot, tongue movements, or simple calculation).	A 5 state HMM with 8 (max) Gaussian mixtures per state, was used to model the spatiotemporal patterns in each signal segment. Features were extracted from all electrodes and fused into a combined feature vector and it had its dimensionality reduced before use in building the model. The expectation-maximization algorithm was used for the estimation of the transition matrix and the mixtures.	Data from 3 male subjects were collected for motor imagery tasks with the particpants free of any medical or central nervous system conditions.
Obermaier et al. [231], 2001	Two class motor imagery (left and right hands) BCI	Two 5 state HMMs (one for each class) with 8 (max) Gaussian mixtures per state, was used to model the spatiotemporal patterns in each signal segment. The Hjorth parameters of two channels (C3 and C4) were fused and fed into the HMM models to calculate the single best path probabilities for both models. The expectation-maximization algorithm was used for the estimation of the transition matrix and the mixtures.	Data from 4 male subjects were collected for motor imagery tasks with the particpants free of any medical or central nervous system conditions.
Pfurtscheller et al. [232], 2003	Two class motor imagery BCI for virtual keyboard control	Two HMMs, one for each class, were trained and the maximal probability achieved by the respective HMM-model represents the chosen class.	Signals from two bipolar channels were acquired from three able-bodied subjects.
Solhjoo et al. [233], 2005	EEG-based mental task classification (left or right hand movement)	Discrete HMM and multi-Gaussian HMM -based classifiers have been used for raw EEG signals.	Dataset III of BCI Competition II (2003) provided by the BCI research group at Graz University [234].
Suk and Lee [235], 2010	Multi-class motor imagery classification	In this study, dynamic patterns in EEG signals were modeled using two layers HMM. First time-domain patterns were extracted from the signals and have dimension reduced using PCA. Second, the likelihood for each channel is computed in the first layer of HMM and assembled in vector whose dimension is reduced with PCA as well. finally, the class label is calculated through the largest likelihood in the upper layer of HMM. Baum-Welch algorithm was used to estimate the parameters of the initial state distribution, the state transition probability distribution, and the observation probability distribution and Viterbi algorithm was used for decoding the state sequence.	Dataset IIa of BCI Competition IV (2008) provided by the BCI research group at Graz University [236].
Speier et al. [237], 2014	P300 speller	An HMM was used to model typing as a sequential process where each character selection is influenced by previous selections. The Viterbi algorithm was used to decode the optimal sequence of target characters.	Data were collected from 15 healthy graduate students and faculty with normal or corrected to normal vision between the ages of 20 and 35.

Continued on next page

Table 5 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Erfanian and Mahmoudi [238], 2005	Real-time adaptive noise canceler for ocular artifact suppression in EEG	A recurrent multi-layer perceptron with a single hidden layer was trained for the noise canceling with the inputs as the contaminated EEG signal and the reference EOG.	A simulated EEG dataset was used for this study, generated through Gaussian white noise-based autoregressive process.
Forney and Anderson [239], 2011	EEG signal forecasting and mental tasks classification	An Elman RNN was trained for forecasting EEG a single time step ahead then an Elman RNN-based classifier was trained to classify the mental task associated with the EEG signals.	4 class dataset was collected from 3 subjects including combinations of the following mental tasks: clenching of right hand, shaking of left leg, visualization of a tumbling cube, counting backward from 100 by 3's, and singing a favorite song.
Balderas et al. [240], 2015	EEG classification for 2 class motor imagery (left hand and right hand)	An LSTM based classifier was trained and evaluated for EEG oscillatory components classification and compared with the regular neural network implementations.	BCI competition IV (2007) dataset 2b [241]
Maddula et al. [242], 2017	P300 BCI classification	A 3D CNN in conjunction with a 2D CNN were combined with an LSTM-based RNN to capture spatio-temporal patterns in EEG.	Data from P300 segment speller were collected, where the subjects mentally noted whenever the flashed letter is part of their target [243].
Thomas et al. [244], 2017	Steady-state visual evoked potential (SSVEP)-based BCI classification	A single layer BRNN was used to perform classification and compared to different architecture and traditional classifying techniques.	5-class SSVEP dataset [245].
Spampinato et al. [246], 2017	Visual object classifier using EEG signals evoked by visual stimuli	An LSTM based encoder to learn high order and temporal feature representations from EEG signals and then a classifier is used for identifying the visual object tat generated the stimuli. The authors here tested different architectures for the encoder including a common LSTM for all channels, channel LSTMs + common LSTM, and Common LSTM + fully connected layer. The authors also trained a CNN-based regressor for generating the EEG features to replace the whole EEG module and work only using source images of visual stimuli.	A subset of ImageNet dataset (40 classes) [247] was used to generate visual stimuli for six subjects while EEG data is recorded.
Hosman et al. [248], 2019	Intercortical BCI for cursor control	An single layer LSTM-based decoder was built with three outputs to generate the cursor speed in x and y directions in addition to the distanc to target.	Intercortical neural signals recorded from three participants, each with 2 96-channel micro-electrode arrays [249].
Zhang et al. [250], 2020	EEG-Based Human Intention Recognition	In this study, multi-channel raw EEG sequences into mesh-like representations that can capture spatiotemporal characteristics of EEG and its acquisition. These meshes are then fed into deep neural networks that perform the recognition process. Multiple network architectures were investigated including a CRNN that starts with a 2D CNN that processes the meshes followed by a two-layer LSTM-based RNN to extract the temporal features, then a fully connected layer and an output layer. The second network investigated was composed of two parallel branches the first was a two layer LSTM-based RNN to extract the temporal features and the second was a multi-layer 2D/3D CNN to extract the spatial features and the output from the two branches is fused and used for recognition. This study used fusion on both data-level and feature-level.	EEG Motor Movement/Imagery Dataset [154, 251].
Tortora et al. [252], 2020	BCI for gait decoding from EEG	EEG data were preprocessed to remove motion artifacts through high pass filtration and independent component analysis. Different frequency bands were then extracted and a separate classifier is trained based on each frequency band. The classifiers were based on a two-layer LSTM-based RNN followed by a fully connected layer, a softmax layer, and an output layer that manifests the prediction output.	EEG data were recorded from 11 subjects walking on a treadmill using a 64-channel amplifier and 10/20 montage.

2.7 Event Detection in EMG

Electromyography (EMG) is the method of sensing the electric potential evoked by the activity of muscle fibers as driven by the spikes from spinal motor neurons. EMGs are recorded either using surface electrodes or via needle electrodes; however, surface EMG (sEMG) is rarely used clinically in the evaluation of neuromuscular function and its use is limited to the measurement of voluntary muscle activity [253]. Routine evaluation of the neuromuscular function is typically performed using needle (invasive) EMG that, despite of its effectiveness and the availability of several electrode types that suite many clinical questions, is often painful and traumatic and may lead to the destruction of several muscle fibers [253, 254]. sEMG has been widely used as control signals for multiple applications especially in rehabilitation including but not limited to body-powered prostheses, grasping control, and gesture based interfaces [255]. A myoelectric signal usually has its manifested events as two states, the first is the transient state which emanates as the muscle goes from the resting state to voluntary contraction. The second is the steady state which represents maintaining the contraction level in the muscle [255]. It has been shown that the steady state segments are more robust as control signals compared to the transient state due to longer duration and better classification rates [256]. As follows in Table 6, we give a review about the recent advances in the detection of myoelectric events in EMG signals.

Table 6: Summary of event detection in EMG signals.

Publication	Event under Investigation	Implementation details	Dataset
Chan and Englehart [257], 2005	Continuous identification of six classes hand movement in sEMG	An HMM with uniformly distributed initial states and Gaussian observation probability density function whose parameters can be completely estimated from the training data, was constructed for the detection process. The expectation-maximization algorithm wasn't used here due to the assumption of uniform initial state probabilities and directly estimating the Gaussian parameters from the training data. Overlapping 256 ms observation windows were used and in each observation window the root mean square value and the first 6 autoregressive coefficients were computed as features.	4-channel sEMG collected from the forearm of 11 subjects for six distinct motions (wrist flexion, wrist extension, supination, pronation, hand open, and hand close) [258].

Continued on next page

Table 6 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Zhang et al. [259], 2011	Hand gesture recognition in acceleration and sEMG	In this work, the authors actually identified the active segments via processing and thresholding of the average signal of the multichannel sEMG. The onset is when the energy is higher than a certain threshold and the offset when the energy is lower than another threshold. Features from time, frequency, and time-frequency domains were extracted from both acceleration signals and sEMG, and fed to five-state HMMs for classification. Baum-Welch algorithm was used for training with Gaussian multivariate distribution for observations. Decision making here is done in a tree-structure (decision-level fusion) through four layers of classifiers with the last layer as the HMM.	sEMG and 3d acceleration were collected from two right-handed subjects who performed 72 Chinese sign language words in a sequence with 12 repetitions per motion, and a predefined 40 sentences with 2 repetitions per sentence.
Wheeler et al. [260], 2006	Hand gesture recognition in sEMG	Moving average was used on the sEMG signals to provide the input for continuous left-to-right HMMs with tied Gaussian mixtures. The training was performed using the Baum-Welch algorithm and the real-time recall was performed with The Viterbi algorithm. The models were also initialized using K-means clustering so that the states were partitioned to equalize the amount of variance within each state. This study employed feature-level fusion to combine multi-channel data.	Data from one participant repeating 4 gestures on a joystick (left, right, up, and down) for 50 times per gesture, were collected using four pairs of dry electrodes. Another portion of data was collected using 8 pairs of wet electrodes on gestures of typing on a number pad keyboard (0-9) for 40 strokes on each key.
Monsifrot et al. [261], 2014	Extraction of the activity of individual motor neurons in single channel intramuscular EMG (iEMG)	The iEMG signal was modeled as a sum of independent filtered spike trains embedded in noise. A Markov model of sparse signals was introduced where the sparsity of the trains was exploited through modeling the time between spikes as discrete weibull distribution. An online estimation method for the weibull distribution parameters was introduced as well as an implementation of the impulse responses of the model.	The method introduced was tested over both simulated and experimental iEMG signals. the simulated signals were generated via Markov model under 10 kHz sampling frequency and with filter shapes obtained from experimental iEMG for more realistic simulation. The experimental iEMG signals were acquired from the extensor digitorum of a healthy subject with teflon coated stainless steel wire electrodes.
Lee [262], 2008	sEMG-based speech recognition	A continuous HMM was constructed with Gaussian mixtures model adopted for sEMG-based word recognition based on log mel-filter bank spectrogram of the windowed EMG signals. The segmental K-means algorithm was used for optimal HMM parameters estimation where HMM parameters for the i^{th} state and k^{th} word are estimated from the observations of the corresponding state of the same word. Viterbi algorithm was used for the decoding process.	EMG signals were collected from articulatory facial muscles from 8 Korean male subjects. The subjects were asked to pronounce each word from a 60-word vocabulary in a consistent manner in addition to generating a random set of words based on this vocabulary.
Chan et al. [263], 2002	sEMG-based automatic speech recognition	A six state left-right HMM with single mixture observation densities, was constructed for identifying the words based on three features extracted from sEMG that included the first two autoregressive coefficients and the integrated absolute value. HMM was trained in this work using the expectation-maximization algorithm.	sEMG from five articulatory facial muscles were collected. The dataset used here was a subset of the dataset described in [264] with ten-English word vocabulary.
Li et al. [265], 2014	Identification/ prediction of functional electrical stimulation (FES)-induced muscular dynamics with evoked EMG (eEMG)	A nonlinear ARX-type RNN was used to predict the stimulated muscular torque and track muscle fatigue. The model takes the eEMG as an input and produces the predicted torque.	The experiments were conducted on 5 subjects with spinal cord injuries.
Xia et al. [266], 2018	Hand motion estimation from sEMG	A CRNN with 3 CNN layers and 2 LSTM layers was used for the prediction and the model used the power spectral density as input.	sEMG signals were collected from 8 healthy subjects using 5 pairs of bipolar electrodes placed on shoulder to record EMG from biceps brachii, triceps brachii, anterior deltoid, posterior deltoid, and middle deltoid. The hand position in 3D space was tracked as the objective for this system.

Continued on next page

Table 6 – *Continued from previous page*

Publication	Event under Investigation	Implementation details	Dataset
Quivira et al. [267], 2018	Simple hand finger movement identification in sEMG	An LSTM-based RNN was used to implement a recurrent mixture density network (RMDN) [268] that probabilistically model the output of the Network in order to capture the complex features present the hand movement.	8 channel EMG signals were collected from the proximal forearm region, targeting most muscles used in hand manipulation. The hand pose tracking was performed with a Leap Motion sensor and the subjects were asked to perform 7 hand gestures with repetitions per gesture.
Hu et al. [269], 2018	Hand gesture recognition in sEMG	sEMG signals from all channels were segmented into windows of fixed size and transformed into an image representation that was then fed into a CNN with two convolutional layers, two locally connected layers, and three fully connected layers followed by an LSTM-based RNN and an attention layer to enhance the output of the network.	Experiments were performed over the first and second subdatabases of NinaPro (Non Invasive Adaptive Prosthetics) database [270].
Samadani [271], 2018	EMG-Based Hand Gesture Classification	Different RNN architectures were tested in this study to chose the best performing architecture. The evaluated models included uni and bidirectional LSTM- and GRU-based RNNs with attention mechanisms. The models worked on the preprocessed (denoised) raw EMG signals.	Publicly-available NinaPro hand gesture dataset (NinaPro2) was used [272].
Simão et al. [273], 2019	EMG-based online gestures classification	Features were extracted from multi-channel EMG (standard deviation along each time frame) and fed into a dynamic RNN model that is composed of a dense layer followed by an LSTM-based RNN layer and another dense layer followed by the output layer. This model was compared to a similar GRU-based model and another static feed forward neural network model. This study used combined feature vector as an input for the models.	the synthetic sequences of the UC2018 DualMyo dataset [274] and a similar subset of the NinaPro DB5 dataset [275]

2.8 Event detection in other biomedical signals

Physiological monitoring is an essential part of all care units nowadays and it is not limited to the aforementioned biomedical signals only. Tens of variables are collected in the form of time series containing hundreds of events that are of importance to the diagnosis and treatment/rehabilitation. Event detection methods have had a strong presence in the analysis of such series. For instance, cardiovascular disorders are not only assessed through ECG but also phonocardiogram is used as an easier way for general practitioner to identify the changes in heart sounds. Extracting the cardiac cycle has been one of the major problems in phonocardiogram as well and was addresses using HMMs in multiple pieces of work [276–279]. On the other hand, most of RNN based methods in phonocardiogram, have been used for pure classification purposes and anomaly recognition [244].

3D acceleration is an emerging technology as well, that has been extensively used in the assessment and detection of many medical conditions in swallowing [84] and human gait analysis [280]. In swallowing, acceleration signals have been used for the detection of

pharyngeal swallowing activity via maximum likelihood methods with minimum description length in [108] and using short time Fourier transform and neural networks in [106]. RNNs were also employed for event detection in swallowing acceleration signals including the upper esophageal sphincter opening in [107, 281], laryngeal vestibule closure [282], and hyoid bone motion during swallowing [85]. In gait analysis, HMMs were used for recognition and extraction in multiple occasions [283–286] as well as RNNs [287–289].

2.9 Challenges and Future Directions

Event detection in biomedical signals is a critical step for diagnosis and intervention procedures that are extensively used on a daily basis in nearly every standard clinical setting. It also represents the core of various eHealth technologies that employ wearable devices and regular monitoring of physiological signs. Being such a fundamental operation that controls the clinical decision making process, it necessitates precise detection in a fairly complex environment that contains multiple events occurring concurrently. Particularly, false positive rate in clinical testing is an important indicator for how well the detection model generalizes and differentiates between the event of interest and the background noise. Building such highly accurate models depends on many factors that include the diversity in the used dataset and labels in addition to model capacity.

2.9.1 Classical Models Scaling: Challenges

As mentioned before, biomedical signals are the manifestation of well-coordinated, yet complex physiological processes which involve various anatomical structures that are close in position and share several functions. Hence, the collected signals pick not only the target physiological process but also other unavoidable neighbor processes. An example of that is the detection of the combined activation for multiple muscles in sEMG, eye blinking along with neural activity in EEG, and head movement along with swallowing vibrations in swallowing accelerometry. Extraction of the event of interest in this case requires the exhausting

labeling of the underlying set of processes in order to be able to build the predefined state space for classical stochastic methods such as HMM, from which the state sequence is drawn. Manual labeling or interpretation of the biomedical signals is not only an exhausting task, but also requires extensive domain knowledge and expertise to perform.

One way that can be used to enhance the expressive power of stochastic models such as HMM, is the inclusion of non-Gaussian mixtures which can boost the performance in many cases because Gaussianity is not always a reasonable assumption in many applications. One of the mixtures that was proposed as an extension for non-Gaussian mixtures, is independent component analyzers mixture model (ICAMM) and it has been applied in multiple biomedical signal applications such as sleep disorders detection and classification of neuropsychological tasks in EEG [126, 127].

An additional way to increase the model capacity and its ability to model the underlying sequence of events, is through using strongly representing domain features. One of the most popular domains representations, is wavelet decomposition which has proven its superiority to provide high level representation of events in a wide variety of biomedical signals such as phonocardiograms [101, 279], EEG [196, 209, 212, 213], and EMG [256]. Handcrafting features, however, is not an easy task and requires an extensive domain knowledge and significant efforts to come up with cues that trigger the identification of specific signal components. Furthermore, mapping the feature space into a more comprehensive space of less dimensionality is often a paramount operation prior to building the model. Given the previous factors, models that are able to learn high level representations simultaneously from raw signals and have the massive expressive power to model tasks involving long time lags, can be of a great benefit [290].

2.9.2 High Capacity Models Embedding Feature Extraction

The evolution of deep learning has revolutionized the way in which problems are addressed and instead of classification and detection systems that solely relied on handcrafted features, end-to-end systems are being trained to take care of all steps from the raw input till the final output. End-to-end systems are complex, although rich, processing pipelines

that make the most of the available information through using a unified scheme that trains the system as a whole from the input till the output is produced [291]. It has been shown that deep architectures can replace handcrafted feature extraction stages and work directly on raw data to produce high levels of abstraction. RNNs have been introduced in 1996 for the identification of arm kinematics during hand drawing from raw EMG signals [292] and then the same architecture was adopted for lower limb kinematics in [293]. In both studies, the authors verified that an RNN was able to map the relationship between raw EMG signals and limbs' kinematics during drawing for the arm and human locomotion for the lower limb. Chauhan and Vig [294] and Sujadevi et al. [295] have also used more sophisticated multi-layer LSTM-based RNN architectures on raw ECG signals for arrhythmia detection. Spampinato et al. [246] have employed RNNs as well to extract discriminative brain manifold for visual categories from EEG signals. Further, Vidyaratne et al. [215] used RNNs for seizure detection in EEG; however, they used a denoised and segmented version of the signals. As mentioned earlier, although RNNs are efficient in modeling long contexts, they tend to have the error signals propagate through a tremendous number of steps when being fed highly sampled inputs such as raw signals which affects the network optimizability and training speed [141, 142].

In this regard, convolutional neural networks (CNNs) have been utilized to perceive small local contexts which then are propagated to an RNN for the perception of temporal contexts or a feed-forward network for a classification or prediction target. CNNs were introduced as a solution to enable recognition systems to learn hierarchical internal representations that form the scenes in vision applications (pixels form edglets, edglets form motifs, motifs form parts, parts form objects and objects form scenes) [296, 297]. Thus, CNNs are basically multi-stage trainable architectures that are stacked on top of each other to learn each level of the feature hierarchy [290, 296]. Each stage is usually composed of three layers, a filter bank layer, a non-linear activation layer, and a pooling layer. A filter bank layer extracts particular features at all locations on the input. The non-linear activation works as a regulator that determines whether a neuron should fire or not through checking the its value and deciding if the following connections should consider this neron activated [290]. A pooling layer represents a dimensionality reduction procedure that processes the feature maps in order

to produce lower resolution maps that are robust to the small variations in the location of features [296]. The coefficients of the filters are the trainable parameters in the CNNs and they are updated simultaneously by the training algorithm to minimize the discrepancy between the actual output and the desired output [296].

The design concept of CNNs first evolved for vision applications; but since then, the same concept is being adopted for pattern analysis and recognition in biomedical signals [266, 298–302]. For instance, Shashikumar et al. [300] used a 5-layer 2D CNN followed by a BRNN in association with soft attention mechanism to process the wavelet transform of ECG signals for the detection of atrial fibrillation. Tan et al. [301] also used a 1D 2-layer CNN with a 3-layer LSTM-based RNN for the detection of coronary artery disease in ECG. Further, Xiong et al. [302] used a residual convolutional recurrent neural network for the detection of cardiac arrhythmia in ECG. All these experiments using RNNs on top of CNNs for biomedical signal analysis were successful to produce extremely high levels of abstraction and rich temporal representation that can perceive long range contexts without human intervention in addition to being easier to optimize computationally. CNNs have been also utilized in association with fully connected networks to increase the capacity of HMMs in connectionist hybrid DNN-HMM models due to the ability of CNNs to process high-dimensional multi-step inputs [303]. Such hybrid systems provided state of the art performance especially in the field of handwriting recognition [304, 305].

2.9.3 Transfer Learning

Despite the fact that most of the previously mentioned methods are achieving great results on certain datasets, it is popular that they can easily overfit the data, resulting in poor generalization. Thus, it requires not only very large but also diverse datasets to train and validate models that well generalize. In biomedical signal processing field, the collection of such datasets may pose a challenge towards developing reliable models. Strictly speaking, it may not be feasible to find a large population of subjects when studying a rare disease and yet if it is feasible, it is extremely difficult to acquire the expert reference annotations for the underlying dataset [306]. Many factors contribute to this, as mentioned

before, the noisy nature of biomedical signals increases the difficulty of manual interpretation and necessitates the presence of reference modalities to acquire accurate information about the processes such as collection of x-ray videofluoroscopy simultaneously with swallowing accelerometry [87]. Another factor is that the experts annotating the data need to maintain high record of reliability across time and to be compared to peer experts which might be difficult to achieve or require continuous training and checking of the experts' reliability.

One way to overcome limited- size and/or diversity datasets, is to utilize the the pre-trained models from relatively different domains and apply them to solve the particular targeted problem or so-called transfer learning [307]. In transfer learning, the pretrained model's weights are used as initialization and then fine-tuned accordingly to fit the new dataset. In most cases, retraining happens in a much lower (10 times smaller) learning rate than the original. Transfer learning has been used for event detection and classification tasks in multiple biomedical signals including ECG for cardiac arrhythmia detection [308], EEG for drowsiness detection [309] and driving fatigue detection [310], and EMG for hand gesture classification [311]. However, one thing worth mentioning is that transfer learning sometimes may not help perform better than the originally trained model if there exist huge differences between the datasets or deterioration in inter-subject variability [307].

2.10 Conclusion

In this paper, we provided a comprehensive review of event extraction methods in biomedical signals, in particular hidden Markov models and recurrent neural networks. HMM is a probabilistic model that represents a sequence of observations in terms of a hidden sequence of states and sets the concepts and methods on how to find the optimal state sequence that best describes the observations. RNN is a type of neural networks that was introduced to model the time dependency and perform contextual mapping in sequences. This review showed that the presence of dynamic programming algorithms like the EM and Viterbi, led to the wide spread of HMMs which were used to dynamically transcribe the context of many biomedical signals. It wasn't too long until HMMs became insufficient for time series model-

ing needs, specifically modeling long range dependencies and larger state spaces, and RNNs started to gradually replace HMMs in time-dependent contextual mappings. So far, RNNs have proven superiority in time series modeling especially in biomedical signals and continue to expand their domination in building automatic detection and diagnosis systems through the emerging designs and practices experimented in nearly every field.

3.0 Areas of Investigation

3.1 Automatic Segmentation of Swallowing Vibrations

3.1.1 Motivation and Scope

Automated identification of vibratory and acoustic signals demarcating individual swallows is a critical first step that for many applications that rely on swallowing sounds and vibrations which have been suggested as alternative bedside tools for dysphagia screening [108, 312–316], to discriminate between patients with healthy and dysphagic swallows [312, 313]. Many swallowing event detection methods have been introduced in the literature especially for swallowing accelerometry. Sejdić et al. [108] developed a segmentation algorithm that yielded over 90% accuracy for identifying individual segments for both simulated and real data. Their algorithm used sequential fuzzy partitioning of the acceleration signal based on its variance [108]. The output of partitioning from two orthogonal axes of acceleration (anterior posterior and superior inferior) was logically combined to achieve better detection of individual swallows and the algorithm was designed to deal with non-stationary long signals [108]. Damouras et al. [109] proposed a volatility-based online swallow detection algorithm that works on raw acceleration signals. This algorithm achieved precision and recall values that are comparable to the results in [108] and outperformed k-means and density-based spatial clustering of applications with noise (DBSCAN) algorithms [317]. Moreover, Lee et al. [314], introduced a pseudo-automatic detection algorithm that depends on simple empirical thresholding of dual-axis accelerometry. They achieved high sensitivity, however the temporal accuracy of the detected segments was unacceptable compared to the expert manual segmentation. Other methods used manual segmentation either through inspection of acceleration by human experts [318] or synchronizing with reference events in simultaneous videofluoroscopic studies [312, 319]. Multi-sensor fusion was also used in swallowing segmentation by identifying the most useful signal combinations among three types of signals (dual-axis accelerometry, submental MMG, and nasal flow) achieving accuracies up

to 89.6% [320]. However, most of these studies considered limited or controlled datasets that didn't include the common dysphagia screening conditions such as different swallowing maneuvers, materials, and consistencies. A robust swallowing segmentation algorithm should be able to achieve high detection quality over a wide dataset that simulates the standard clinical procedure for swallowing function evaluation.

3.1.2 Plan of Action

In order to develop and test an automated swallowing signals segmentation algorithm, more than 3000 swallows were collected from 248 patients with different conditions and 20 healthy subjects. The swallows were collected in a standard clinical setup for swallowing function assessment using multiple materials with different consistencies and performed in multiple head positions (maneuvers). 3D swallowing acceleration and sounds were collected simultaneously with videofluoroscopy which was used as a reference by the speech language pathologists to manually annotate the onset and offset of each swallow. Swallowing signals were then segmented and labeled accordingly and the short-time Fourier transform was calculated based on a predefined window size in order to feed a deep feed-forward neural network that annotates each window as being a part of a swallow or not. Different window sizes were selected and tested based on the statistics of the swallow duration from the collected dataset to determine the optimal size. The proposed system was evaluated using a ten fold cross validation procedure and the segmentation results were then validated manually against the expert manual segmentation in order to assess the detection quality both numerically and temporally.

3.2 Automatic Swallowing Segmentation using Convolutional Recurrent Neural Networks and Sensor Fusion

3.2.1 Motivation and Scope

In a previous part, we introduced a segmentation algorithm for swallowing signals, that uses the power spectral estimate as an input and is based on deep feed-forward neural network as the main core of implementation. Although the algorithm achieved high detection accuracy quantitatively and qualitatively compared to the other methods in the literature, it has a slight uncertainty regarding the false positives rate due to the long uncovered/unlabeled periods of the signals while the x-ray machine is paused for no-dose mode. As a result of this, we tried to validate the results against the logs and videos in order to come up with a solid idea about the performance of the algorithm. Out of the validation, the algorithm was found to detect some of the events such as head movement and coughing as swallowing; the thing which reduces the algorithm's overall performance. Thus, we thought to further investigate this problem through using a different architecture that takes the temporal characteristics of the swallow into consideration and process longer time chunks simultaneously. In addition to that, we started to further label the pause periods in the x-ray videos to induce extra certainty into both the training and evaluation processes. Here, we investigate the use of recurrent neural networks and convolutional neural networks to provide a concrete solution for the swallowing segmentation problem in swallowing signals.

3.2.2 Plan of Action

For this study, the same data from the first segmentation work will be used and the short-time Fourier transform will be calculated using the same window size used before. On the other hand, the unlabeled parts of signals corresponding to pause periods of videofluoroscopy will not be included in the training or evaluation of this new system. For this to be achieved, a trained rater extracted the segments where the x-ray machine was stopped during the recording process. A 3-layer convolutional neural network is constructed to transform the input spectral estimate from a number of consecutive windows into higher level representation

that is more abstract. The time dependent series of outputs are fed into a recurrent neural network to model the long range dependencies between different time frames and then a 3 layer fully connected network is used to generate the segmentation mask that indicates which frames are part of a swallow. The same segmentation evaluation procedure used in the previous implementation will be used again here in order to assess the performance of the new architecture.

3.3 Non-Invasive Detection of Upper Esophageal Sphincter Opening

3.3.1 Motivation and Scope

Among the most important physiologic correlates of healthy swallowing function is the duration of upper esophageal sphincter (UES) opening (DUESO). UES opening enables food and liquid to enter the esophagus [79, 321–323]. Reduced UES opening diameter, delayed onset of opening, or premature closure attenuate DUESO and can result in pharyngeal residue that in turn can enter the upper (laryngeal penetration) or lower (aspiration) airway, which are known risk factors for pneumonia and airway obstruction [324]. UES opening is the product of hyolaryngeal excursion, bolus propulsion, and neural inhibitory relaxation of the UES itself [79, 324]. UES dysfunction may occur due to neurological diseases that alter the timing of UES relaxation and the delivery of muscular traction forces that act to distend the relaxed UES during swallowing, or due to impaired propulsive forces applied by the oropharyngeal pump [322, 324].

Table 7 summarizes the different diagnostic modalities that can generate images and signals for the assessment of UES function [322, 326, 327]. The modalities include videofluoroscopic swallow studies (VFSSs), fast pharyngeal CT/MRI, fiberoptic endoscopic evaluation of swallowing (FEES), and non-imaging instrumental tests such as pharyngeal manometry and Electromyography (EMG). Most of these modalities require expertise to perform and highly trained clinicians to interpret. VFSSs are most frequently and actually the best modality to clinically assess swallow kinematic events such as UES opening, because of the

Table 7: Summary of tools used for diagnostic assessment of UES.

Modality	Strengths	Weaknesses
VFSS[322]	- Dynamically visualize UES during all phases of swallowing - Provides the exact moments when UES opens and closes	- Subjective interpretation - Radiation exposure
FEES[325]	- Direct visualization of swallowing pharyngeal stage	- Limited in describing UES activity (either probe is covered with bolus or already through UES)
CT/MRI[322]	- Panoramic and full-thickness visualization of oropharyngeal structures	- Hard to conduct - Radiation exposure (CT) - Require synchronization with patient behavior (MRI) - Limited availability
Manometry[322]	- Monitor UES pressure during swallowing - Detect UES impaired relaxation/distension	- Invasive - Subjective interpretation - Limited availability
EMG[322]	- Monitor muscle activations during swallowing - Detect UES impaired relaxation/distension	- Can't tell the exact moments when UES opens/closes - Subjective interpretation

ability to dynamically visualize the UES during all phases of the swallow and give exact estimates of the moments when UES opens and closes [321, 322]. However, VFSSs, which use x-ray to produce radiographic images with full temporal resolution, are unavailable or undesirable to many patients, are relatively expensive, and require specialized instrumentation and trained clinicians to perform and interpret, leaving many patients undiagnosed or inaccurately diagnosed, and exposed to the risk of dysphagia-related complications [321].

A non-invasive automatic detection method for UES duration can be of a great benefit for both swallowing evaluation and rehabilitation. Neck-based swallowing accleremetry has shown many indicators that proved association with multiple swallowing physiological events including hyolaryngeal excursion and maximal hyoid bone displacement [89], laryngeal vestibule closure, UES opening, and tongue base contact with the posterior pharyngeal wall [90]; however, no studies tried to actually detect any of these events based on the swallowing accelerometry. Hence, in this study we investigate the ability of UES opening detection using 3D swallowing acceleration.

3.3.2 Plan of Action

Three axis of swallowing acceleration were collected from 116 patients with different conditions who performed swallows in different maneuvers and using multiple bolus sizes and consistencies. Another set of swallows were collected from 15 healthy subjects in order to test the proposed system in an independent identical clinical setup. Swallowing vibrations were collected simultaneously with videofluoroscopy here as well for end-to-end synchronization between signals and videos. Speech language pathologists trained on performing swallowing kinematic analysis, marked the onset and offset of the separate swallows and the opening and closure moments of UES in videofluoroscopy data which were then mapped to mark the same moments on signals. The signals were then segmented according to the onset and offset of each pharyngeal swallow and labeled temporally according to the opening of UES. The acceleration signals were then divided into windows of predefined length and preprocessed to remove the multi-source noise components existing in the signals such as head movement and sensor noise. The raw signals were then fed into a convolutional recurrent neural network composed of two 1D convolutional layers followed by 3 layers of GRU-based recurrent network and the UES opening segmentation mask is generated through three layers of a fully connected neural network. The proposed system was evaluated using a ten fold cross validation procedure and the UES opening was compared for each swallow, to the expert annotations to compute the prediction error relative to the inter-human raters error.

3.4 Upper Esophageal Sphincter Opening Maximal Distension Detection and Localization

3.4.1 Motivation and Scope

As mentioned earlier in sec. 3.3, The upper esophageal sphincter opening is considered one of the most important physiological aspects of swallowing. Acting as the gateway of esophagus, it allows the passage of ingested materials from the hypopharyngeal cavity into the esophagus and any dysfunction may lead to inefficient clearance and contribute

to penetration and/or aspiration. Three parameters characterize the opening of the upper esophageal sphincter during swallowing, from which we addressed the onset/offset times and duration, and the third parameter is the diameter of opening and its timing during the swallow. The gold standard for the upper esophageal sphincter opening diameter evaluation is videofluoroscopy as well; however, judgments are scale-based subjective and limited to either present, absent, or incomplete [328, 329]. Although being clinically expedient, a more objective quantification of the opening diameter and the time of achieving the maximal diameter is necessary for the evaluation of impaired sphincter recovery and clinical intervention procedures. In a pilot study, we investigated the correlation between swallowing signals (acceleration and sounds) and the diameter of maximal upper esophageal sphincter opening and the study showed strong association between the diameter and many swallowing signal features which motivated our endeavor here to build an algorithm to actually measure the diameter of the maximal opening and locate it through the swallowing duration.

3.4.2 Plan of Action

In order to establish the ground truth of this study, a UES maximal diameter measurement tool and protocol were developed and tested through collaboration with clinicians experienced in swallowing kinematic analysis. The tool included a drawer graphical user interface that allows the rater to draw segments representing the components of the measurement process on the videofluoroscopic frames such as C3 vertebra, C2-C4 inter-vertebrae segment and the UES opening diameter itself, in addition to depicting the actual distance of the segments in terms of pixels. The measurement protocol includes defining the rules of using the drawing tool to actually measure the diameter such as which frame in the swallow to use and the exact point of UES at which the diameter will be measured. The measurement sequence and protocol were developed based on the common clinical practices and prior studies about UES diameter ratings. Judges were trained and had their reliability tested in order to perform the UES diameter measurement based on the developed tool and protocol. The same set of swallows used for UES opening study was used for this study as well as both are investigating the dynamics of a common structure (UES) and the opening

duration prediction is a necessary prerequisite for UES diameter measurement. A multi-task convolutional recurrent neural network framework is then constructed to use the spectrogram of the raw 3D swallowing acceleration signals in order to predict the normalized UES diameter with respect to the C3 vertebral length and the 3-frame time window within which the maximal diameter is achieved. A multi-dimensional attention mechanism may also be added to focus on the UES opening duration and the anterior-posterior axis of acceleration. The reason for using attention is that a pilot study discovered stronger association between the signals during the UES opening rather than the whole swallowing segment and along anterior-posterior axis of acceleration and spikes were observed in the spectrogram during UES opening period.

3.5 Non-Invasive Detection of Unsafe Airway Protection

3.5.1 Motivation and Scope

Despite the usefulness of penetration aspiration scores [77] in quantifying airway protection and aspiration severity, it remains subjective and needs skilled clinicians to be conducted. In addition, videofluoroscopy requires the exposure of patients to x-ray radiation which limits the period of the evaluation process and reduces the probability of capturing penetration aspiration events. Therefore, the development of an objective method that can quantify penetration-aspiration, will be of great benefit to clinicians. HRCA signals have demonstrated high potential in aspiration detection using classical classification techniques and showed strong association with the penetration-aspiration score [87]. Although the previous studies addressed the same problem, in this study we extend the classification problem to the full range of the penetration-aspiration score (8 categories) instead of tearing down into sub-categories such as 1-2, 3-6, and 7-8 [87]. In turn, this should give a complete description of the airway protection during swallowing including the depth of penetration if any and the presence of patient reaction if any. A deep learning framework hasn't been used to address this problem yet, which gives another reason to further investigate this problem.

3.5.2 Plan of Action

For this study, swallows were collected from 265 patients suspected with dysphagia with different conditions including stroke and some neurodegenerative diseases. The study yielded more 3000 swallows of different types and maneuvers. The swallows were analyzed by experienced speech language pathologists i order to annotate the onset and offset of each swallow and rate each swallow in terms of airway protection. The airway protection is rated based on the penetration-aspiration score that ranges from 1-8 for 1 to represent safe airway protection and 8 to represent silent aspiration. swallowing acceleration and sounds are to be segmented and denoised based on the onset and offset of each swallow demarcated by experts and then fed into a convolutional recurrent neural network to classify the penetration-aspiration score. Here, we will try different network architectures including a 2D convolutional network that works on the spectral estimate of the signals then a recurrent neural network processes the temporal characteristics of this representation followed by fully connected layers that generate the classification. The other architecture to be tried is a 1D convolutional network that works directly on the raw signals and the rest of the architecture is similar to the previous one with recurrent and fully connected neural networks.

4.0 Automatic Segmentation of Swallowing Vibrations

The majority of this chapter has been previously published in and reprinted with permission from [106]. © 2021 Springer Nature Limited. 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.
DOI: 10.1038/s41598-020-65492-1

4.1 Objective

The purpose of this study was to evaluate the accuracy of spectral estimation and deep neural networks (DNNs) in automatic swallowing activity detection in both swallowing accelerometry signals and swallowing sounds. Three axes of acceleration and a single channel of swallowing sounds were investigated individually as standalone event detectors after which the best system was chosen according to detection quality when compared to the expert manual segmentation. Moreover, the used dataset overcomes the limitations of controlled data acquisition in the past segmentation studies, including number of subjects, swallowing maneuvers, swallowed materials and bolus size which represent most of the conditions common in dysphagia screening. This makes the dataset investigated in this study, optimal for the validation of such segmentation algorithm. We hypothesize that the proposed method will be able to correctly identify around 95% of the swallowing segment in more than 90% of attempts, irrespective of the texture or volume of the swallowed material, swallowing maneuver, or patient diagnosis.

4.2 Methods

4.2.1 Materials and Methods

This study was approved by the Institutional Review Board of the University of Pittsburgh. All participating patients gave informed consent to join the study. A total of 248 patients (148 males, 100 females, age: 63.8 ± 13.7) served as the sample for this experiment. They were recruited from the population of patients referred to the Speech Language Pathology service for an oropharyngeal swallowing function assessment with videofluoroscopy at the University of Pittsburgh Medical Center (Pittsburgh, PA), due to clinical suspicion of dysphagia. Of the sample, 44 patients (32 males, 12 females, age: 66.6 ± 13.7) were diagnosed with stroke while the remaining 204 patients (116 males, 88 females, age: 63.0 ± 14.3) had medical conditions unrelated to stroke. Patients were asked to swallow multiple materials of different viscosities and volumes including chilled ($5°C$) Varibar thin liquid (Bracco Diagnostics Inc., Monroe Township, NJ), chilled ($5°C$) Varibar nectar, honey thick liquid, barium tablets (EZ Disk, Bracco Diagnostics Inc., Monroe Township, NJ), Varibar pudding, or a cookie coated with Varibar pudding. The swallows were performed with and without verbal command and in multiple maneuvers including neutral, chin down, left and right head rotation, combined chin down and head rotation, Supraglottic swallow (SGS), and modified SGS. The vibrations of each swallow were recorded as a separate task by the LabView Signal Express and exported in a plain text format to be used for subsequent analysis. A total of 3144 swallows (603 from stroke diagnosed patients and 2541 from other patients) were recorded with an average duration of 862.6 ± 277 msec. The collected swallows included 1038 single swallows, 1893 multiple swallows (several swallows to swallow a single bolus) and 213 sequential swallows (swallows of more than one bolus one at a time in a rapid sequence). The swallowing event start (onset) and end (offset) times taken as gold standard for the experiment were obtained through manual segmentation of videofluoroscopy sequences by experienced speech language pathologists (SLPs) in our Swallowing Research Lab along with the penetration aspiration (PA) scores [77] of the swallows as described in [330]. PA scale scores indicate the depth of entry of swallowed material into the patient's airway when swal-

lowing, and the quality of the patient's airway protective response to airway penetration (material remaining above the true vocal folds) or aspiration (material coursing through the larynx and entering the trachea). The number and type of swallows in each PA score are summarized in Fig. 12.

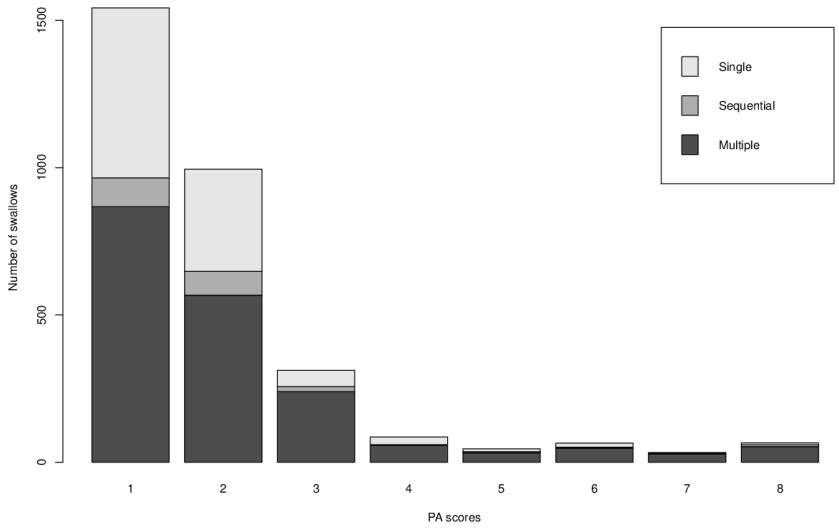


Figure 12: Number of swallows for each PA score

4.2.2 Data Acquisition

Data acquisition was performed per previous work published by Dudik et al. [331]. The swallowing vibrations were recorded during a routine videofluoroscopy with two types of sensors, a tri-axial accelerometer (ADXL 327, Analog Devices, Norwood, Massachusetts) and a lapel microphone (model C 411L, AKG, Vienna, Austria) attached to the subject's anterior neck. The accelerometer complex (sensor in a plastic case) was attached to the skin overlying the cricoid cartilage for the best signal quality [332]. The first two axes of accelerometer were aligned to the anterior-posterior (A-P) and superior-inferior (S-I) directions which can be described as perpendicular to the coronal plane and parallel to the cervical spine respectively. The third axis of accelerometer (medial lateral axis or M-L) was parallel to the axial/transverse plane of the patient's head and neck. The sensor was powered using

a 3V power supply (model 1504, BK Precision, Yorba Linda, California) and had its output signals hardware band-limited to 0.1-3000 Hz and amplified with a gain of 10 (model P55, Grass Technologies, Warwick, Rhode Island).

The microphone was mounted towards the right lateral side of the larynx with no contact with the accelerometer to avoid any friction noise and to avoid obstructing the upper airway radiographic view, and powered via a microphone specific power supply (model B29L, AKG, Vienna, Austria) with the maximum possible volume level (9 for this device). The conditioned signals from the microphone and accelerometer were fed into a National Instruments 6210 DAQ, sampled at a 20 kHz rate, and acquired by LabView's Signal Express (National Instruments, Austin, Texas). The previous setup for both accelerometer and microphone has proven to be effective in collecting swallowing vibrations [332–335]. A video capture card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA) was used to feed the output of the videofluoroscopy instrument (Ultimax system, Toshiba, Tustin, CA) into LabView for recording. All signals fed to the DAQ were acquired and recorded simultaneously for a complete start-to-end synchronization.

4.2.3 System Design

All the acquired signals (swallowing sounds and acceleration) from the microphone, and the A-P, S-I, and M-L axes of the accelerometer were sampled at 20 kHz. Since numerous physiologic and kinematic events occur simultaneously during swallowing recordings (e.g.: breathing, coughing), collected signals contain vibratory and acoustic information from multiple sources [109]. To overcome these and other measurement errors, we downsampled the entire dataset to 20% of the recorded sampling rate (i.e. 4 kHz instead of 20 kHz) [336]. All four signal streams (microphone, and accelerometer A-P, S-I, and M-L) were independently considered for swallowing segmentation.

To simulate the online processing scheme, and since we sought to determine whether automated segmentation could replicate gold-standard manual segmentation, a sliding window of size N samples was used to partition the signals into time samples with no overlap between successive windows. The window size N is considered as the predefined segmenta-

tion resolution of the system; therefore, we tested different values of N to see the effect of window size on the overall performance of the segmentation process. We used sizes of 500 to 1500 (125 *to* 375 *msec*) with a step of 100 samples and the selection of this range of values came from the fact that the acquired swallowing segments can be represented with the used window sizes. Moreover, a typical swallow segment can range in duration from 1 second (4k samples in this case) to more than 3 seconds which makes the selected window sizes robust to statistical error and efficient to detect the shortest swallows [109, 337]. So, four different segmentation models were trained and tested based on the four signal lines from microphone and accelerometer, each dependent on the spectrogram of underlying signal in order to determine the best window size and the best performing line as in Fig. 13.

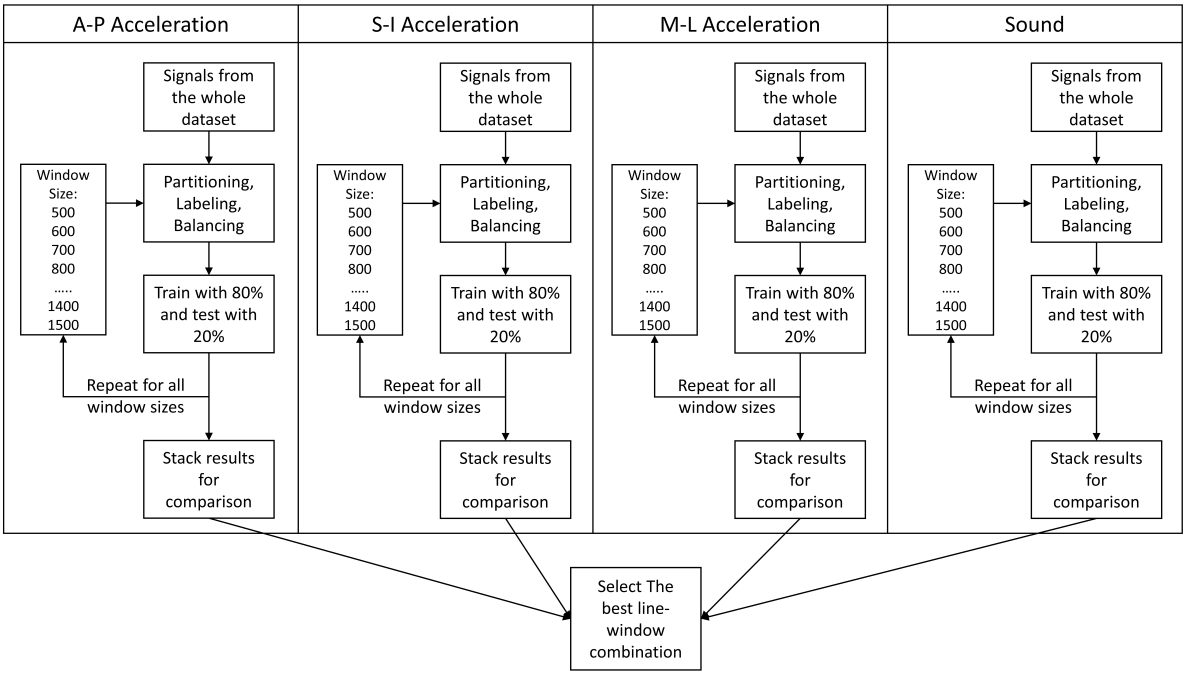


Figure 13: System's parameter selection process

All windows were labeled by comparing the start and end times to the timing of manual segmentation done by SLPs. A window is considered a part of a certain swallow if the the manually labeled swallowing segment overlaps with 50% or more of the automatically selected window size as shown in Fig. 14. The spectrogram of each window is calculated through the use of short-time Fourier transform (Eq. 4.1) with 5 non-overlapping time samples each

of $(N/5)$ length, a fixed length of 512 for the calculated Fourier transform and a Hanning window to reduce variance and leakage. This setup provided spectrograms of 257 frequency bins and we only used the magnitude of spectrogram in building the model while the phase was not of interest for this study. The magnitude of each spectrogram was unpacked into a (257×5) length vector to be used for the training process and prior training, all spectrograms were normalized to unit scale.

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]\exp^{-j\omega n} \tag{4.1}$$

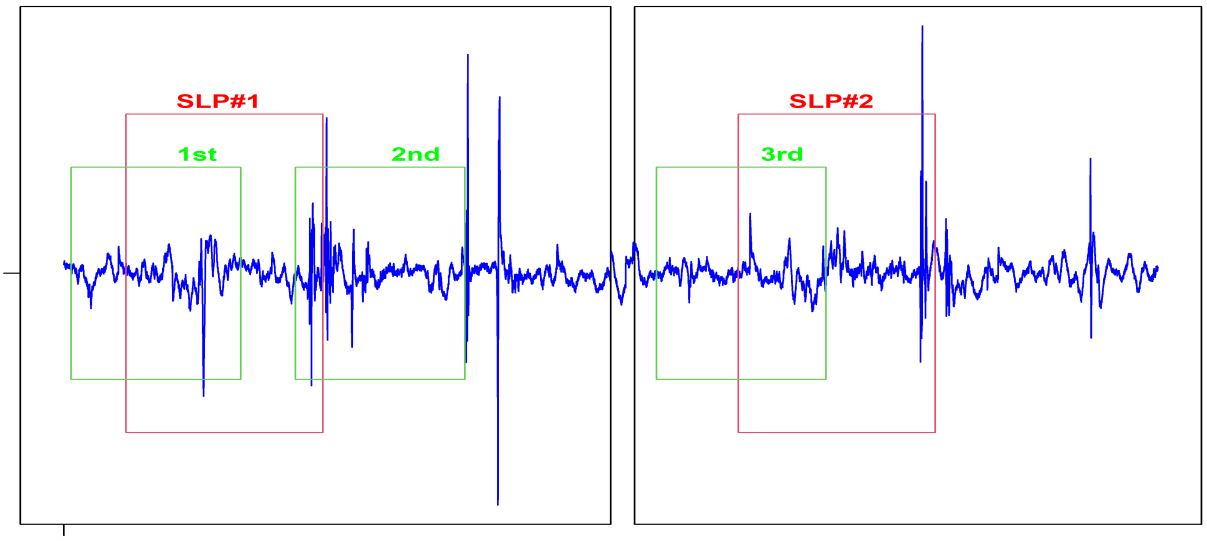


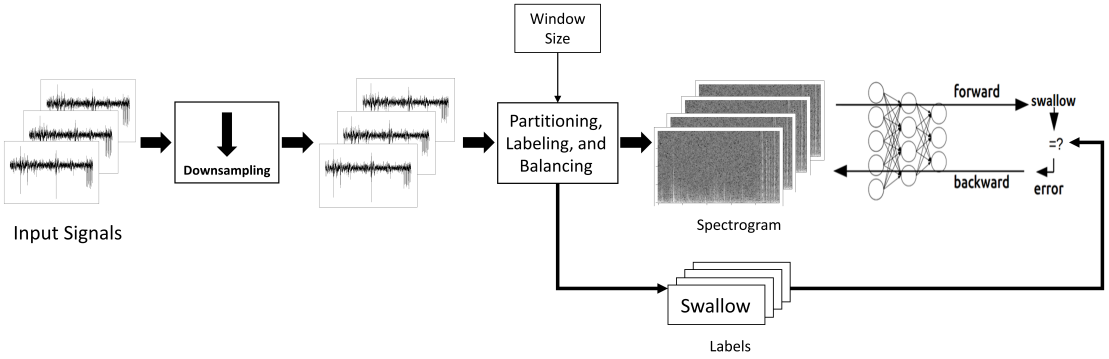
Figure 14: The labeling process of a sample swallowing sound signal. Red windows represent the swallowing segments identified by human expert SLP's. Green windows represent different positions of the sliding window. The 1^{st} and 3^{rd} positions are labeled as swallows due to large overlap

The used window sizes produced 5574 to 20121 swallowing windows and 94211 to 280043 non-swallowing windows for window sizes 1500(375 $msec$) and 500(125 $msec$) respectively. This imbalance between swallows and non-swallows comes from the fact that each recording file contains longer blank (background noise) periods than swallowing periods. As a result, the balance between both types needed to be restored for the training of the system to mitigate bias. Therefore, we used the full set of the swallowing data at each window size and randomly selected an equal group of the non-swallowing data. Single swallows were also separated in order to form a smaller dataset so that we could test the system performance

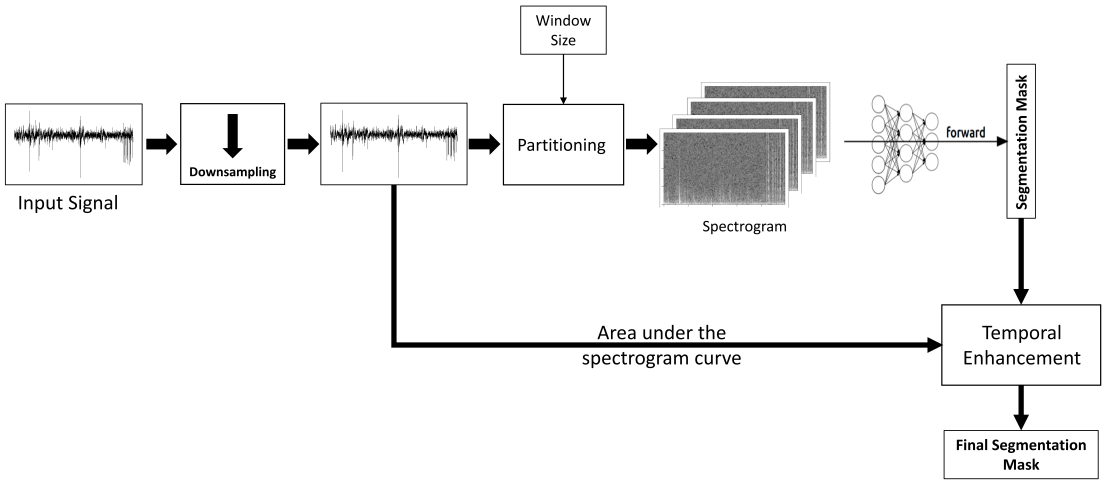
over single and other types of swallows (multiple and sequential) because the later categories are known to be more complex. The resultant datasets were randomly reordered and divided into two parts, 80% for training and 20% for testing.

A DNN was trained to create a feed forward probabilistic model of size $1285 \times 1285 \times 1$ units. The DNN was created such that the input layer is the spectrogram vector of each window and the output layer represents the synthesized probability of whether the window is a part of a swallow or not. The output layer was configured to use the biased-sigmoid as an activation function with zero bias. The DNN was trained using a 100 iterations stochastic gradient descent (SGD) [338]. In addition, the DNN was configured to use dropout free training along with full sweep iterations of SGD.

Once we got the best window size and the best performing line of swallowing signals from the previous step, we retrained and tested the system using these parameters as the block diagram shows in Fig. 15. The whole dataset was divided randomly into 10 equal subsets in terms of recordings and a holdout method is repeated 10 times by training with 9 subsets and testing with the remaining one. Furthermore, the segmentation masks generated from this step were processed in order to enhance the temporal accuracy of the detection compared to the manual segmentation. This step is intended to check the boundaries of the detected segment and add a couple of samples on each side for a better match with SLP segments. The segments added to each side are determined through inspection of the area under the spectral estimate curve (AUC) of the swallowing signal (summation across frequencies for each time sample). The whole temporal enhancement process is illustrated in the flowchart shown in Fig. 16(a). The width of the segment is determined through simple thresholding of the AUC in the area around the detected segment with a threshold calculated from statistics of the segment (min and max). Fig. 16(b) shows the AUC for a swallowing sound signal with swallowing segments annotated with rectangles. The inspection area was limited to 2 windows around the borders of each detected segment because more than this, will not be reasonable compared to the duration of swallows.



(a)



(b)

Figure 15: Flow of the training (a) and testing (b) paths of the proposed system

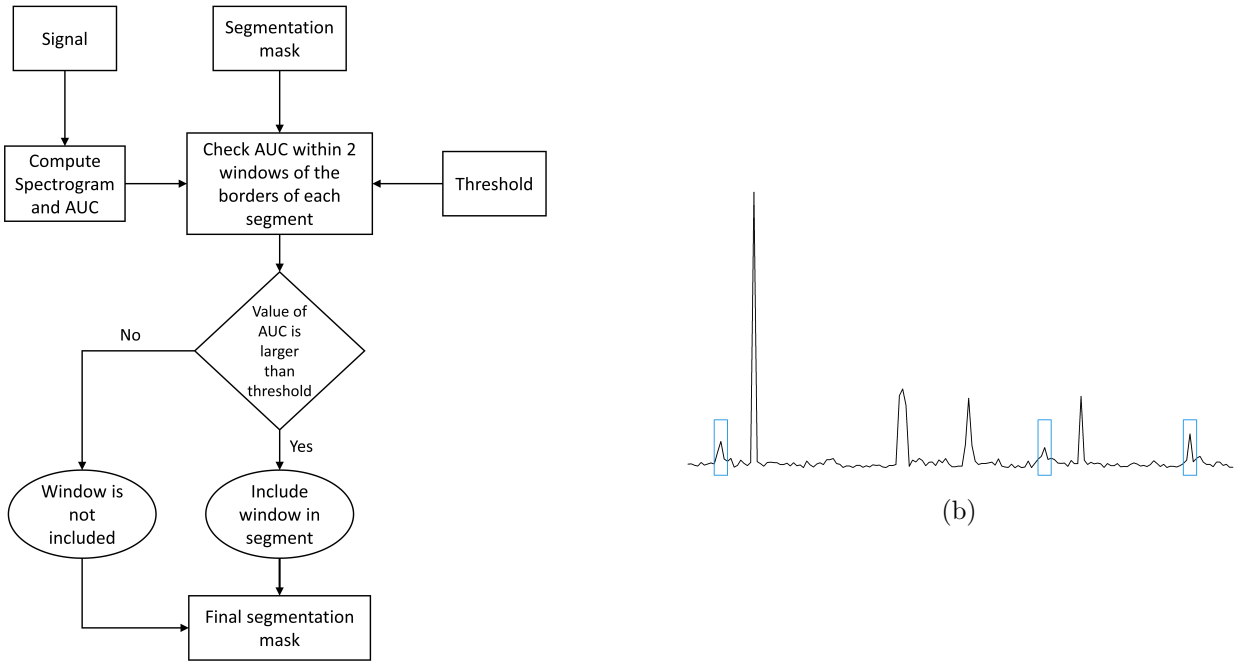


Figure 16: Temporal enhancement process: (a) shows the flowchart of the process. (b) A sample area under the spectrogram curve of a swallowing mic signal.

4.2.4 Temporal Assessment

An assessment criterion was defined to validate the results of this segmentation work against the human expert manual segmentation as shown in Fig. 17. Manual segmentation defined swallow segments as the duration between the time when the leading edge of the bolus passes the shadow cast on the x-ray image by the posterior border of the ramus of the mandible (segment onset) and the time the hyoid bone completes motion associated with swallowing related pharyngeal activity and clearance of the bolus from the video image (segment offset). When patients swallow more than once to clear a single bolus (multiple swallow), the offset was based on the time when the hyoid returns to the lowest position before the next hyoid ascending movement associated with a subsequent swallow. A swallow segment was considered correctly identified (auto-detected) if and only if there exists a

certain percentage overlap between the reference window determined by a human judge performing manual segmentation and the window produced by the proposed segmentation algorithm (as shown in Fig. 17(b)-(e)) [108]. In this study, we tested multiple overlap ratios representing two different approaches. The first approach was a fixed overlap irrespective to the segment duration and the used overlap included 2SD below the average swallow duration (431.89 $msec$) and 1SD below the average swallow duration (675.56 $msec$). The second approach was using a 90% and 95% overlap ratio of the manually measured duration for the compared segment. Otherwise the swallow was deemed to be incorrectly segmented (as shown in Fig. 17(f)-(g)). In addition to this assessment criterion, we used accuracy, specificity, and sensitivity to evaluate the overall performance of the segmentation process.

4.2.5 Segmentation Validation

The videofluoroscopy instrument was controlled by a radiologist who had a switch to stop the imaging procedure when there was no bolus administered to the patient in order to reduce the radiation dose. This pausing in the x-ray machine operation caused the collected videos to have static frames for long periods with no visual clue about the events occurring while vibratory and acoustic data continued to be recorded. These events included swallows, talking, coughing, and head motion occuring between elicited swallow events. Without the visual help of VFSS, these events cannot be labeled; hence, included in the evaluation of this segmentation procedure. However, the algorithm was applied to these areas after training to see if it would pick up any of these events. This, alongside with the presence of unexplained false positives, necessitated manual inspection and validation of the segmentation results against the videos and logs kept by research associates collecting the swallow data. A trained rater validated each event detected by the algorithm in order to identify the origin of non-swallow events as a qualitative assessment for the proposed algorithm.

4.2.6 Clinical Validation

In order to further explore the performance of the proposed segmentation framework, it was evaluated as well in a standard clinical setup during the workflow of an ongoing

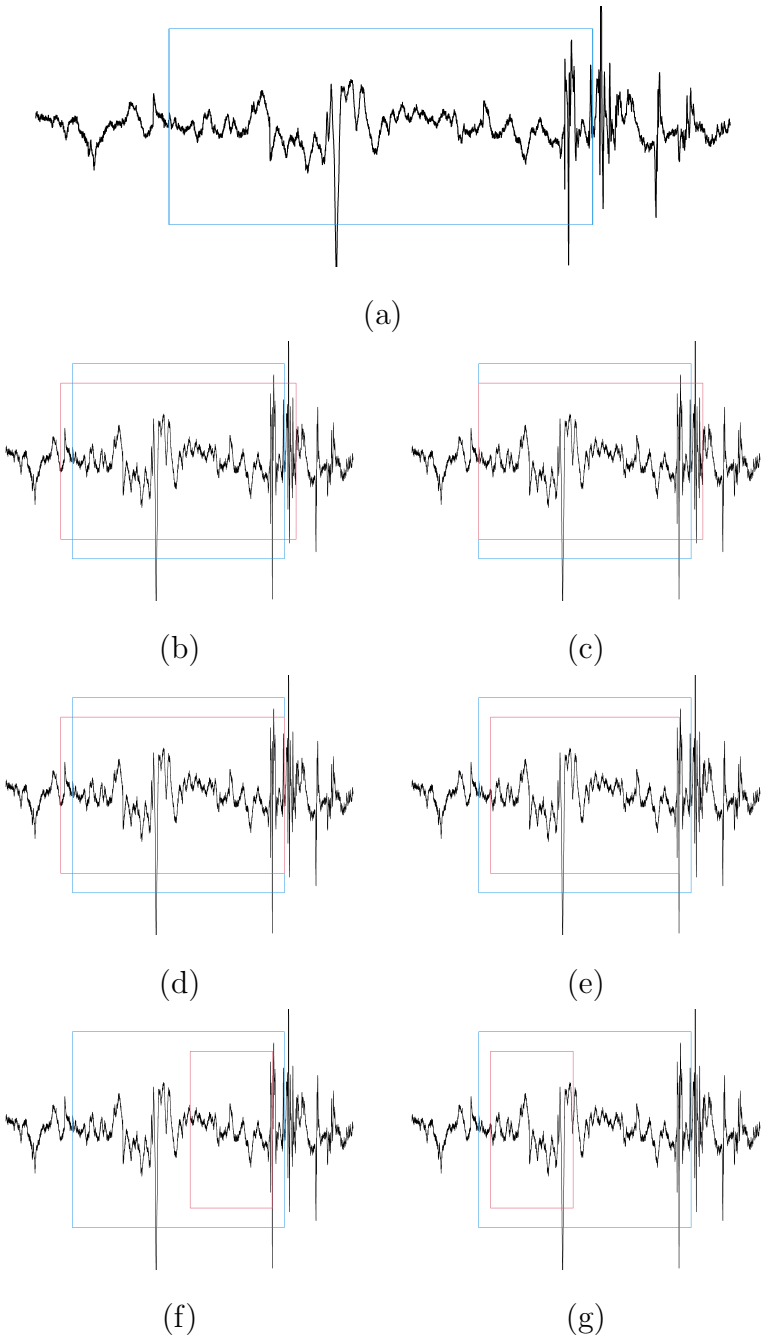


Figure 17: Possible swallow segmentation results. (a) Sample swallowing sound signal and definition of the swallowing segment (in blue). (b)-(e) Examples of correctly identified swallow segments (in red). (f)-(g) Examples of incorrectly identified swallow segments.

swallowing experiment. The experiment was performed on healthy community dwelling adults who had no history of swallowing difficulties. Twenty subjects (9 males, 11 females, age: 65.8 ± 11.4) who provided informed consents, participated in the experiment. The participants in this sample were selected randomly from a population that had no history of surgeries to the head or neck region or neurological disorders and underwent swallowing evaluation as a part of bigger study. The data collection procedure was nearly identical to the aforementioned procedure except for the bolus sizes and consistencies. Only thin liquid boluses: 3mL by spoon and unmeasured self-administered volume cup sips, were administered in a completely randomized order. A total of 76 swallows with an average duration of 1011 ± 216 msec, were used to test the proposed system for the detection of the onset and offset of pharyngeal swallows after being trained over the full 3144 swallows dataset mentioned previously. The used swallows in this validation procedure were meant to be completely unseen in order to test the robustness and generalizability of the proposed segmentation algorithm and never used in anyway in the training process. Both training and evaluation were performed using the best performing window size (800) and only the A-P acceleration.

4.3 Results

Fig. 18 shows sample spectrograms for two segments (non-swallowing and swallowing) of a randomly selected record, calculated with the same settings used in our experimental setup (STFT, 5 time samples and 257 frequency bins). The spectrogram of the AP axis acceleration in both segments is shown in the top panel while the bottom panel contains spectrograms for microphone signal (swallowing sound). This represents the typical folded input to the DNN for each of the training models described previously.

Fig. 19 shows the results of testing the DNN trained with 80% of the data for the three axes of accelerometer (A-P, S-I, and M-L) and microphone signal. At each window size, the performance of swallowing identification is shown in terms of accuracy, specificity, and sensitivity. According to Fig. 19, we can clearly see that the best results are achieved for AP

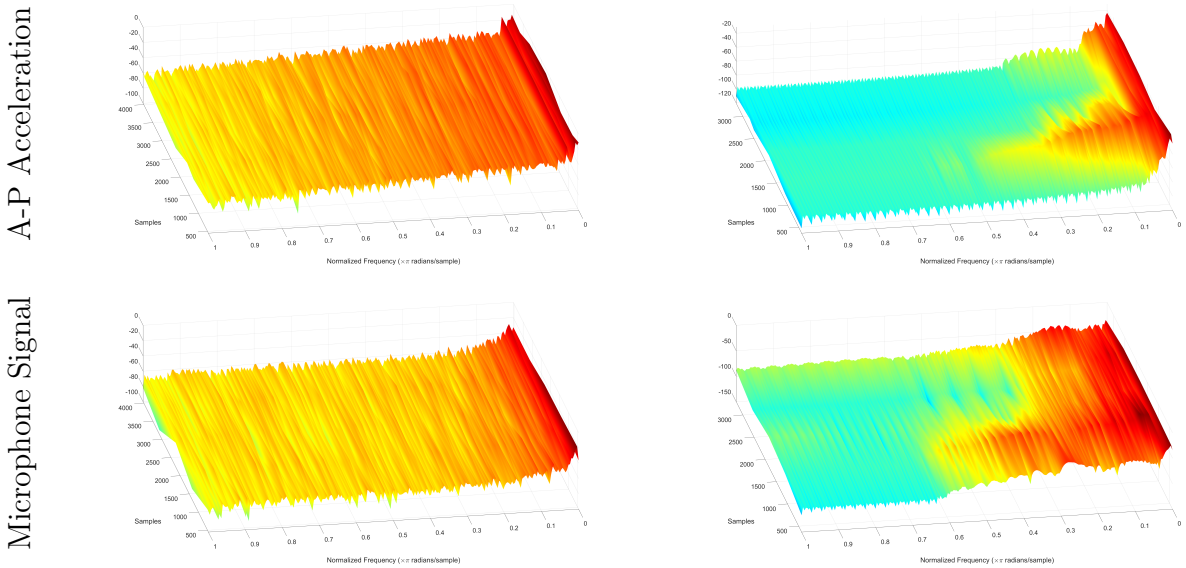


Figure 18: Spectrogram of a non-swallowing and swallowing segments for both acceleration and sound.

acceleration data at window sizes of 800 and 900 (900 and 1000 for the whole dataset). As a result, the 10-fold cross validation model was trained with A-P acceleration 10 times while excluding a randomly selected set of recordings each time for testing (without replacement). The top detection results achieved across all folds are shown in Table 8 for the two window sizes and the different overlap criteria. Ninety to 100% detection was accomplished for all four overlap ratios across single, multiple, and sequential swallows, and the precision of all four overlap ratios for single swallows was greater after post-processing. Multiple and sequential swallow detection also increased after post-processing however both the 900 and 1000 window sizes performed comparably. Overall, algorithm-based detection was most accurate using the window size of 800 (200 msec) for single swallows.

The algorithm also achieved 85.3±12.5% sensitivity and 83.8±9.5% specificity per each of the dataset records. These values were calculated over the whole dataset after removing the visually uncovered parts from records.the values came close to the anticipated results from the initial trials at Fig. 19. There may have been a slight drop in sensitivity and specificity

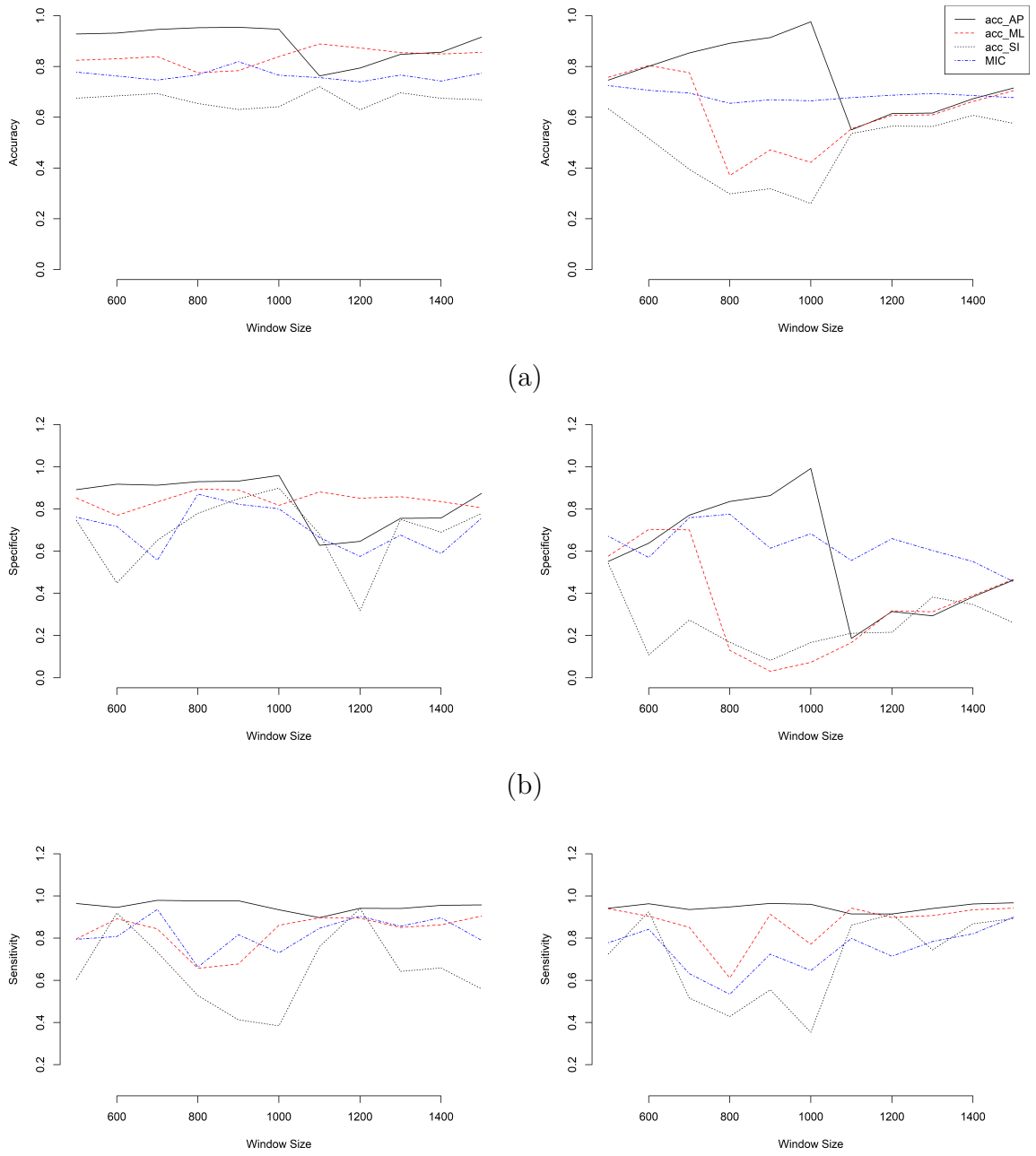


Figure 19: Quality measurements of the full run of the system. (a) Accuracy. (b) Specificity. (c) Sensitivity.

Table 8: Detection measurements for the top two configurations

Overlap Ratio	Property	Single Swallows		Multiple & Sequential Swallows	
		800(200 $msec$)	900(225 $msec$)	900(225 $msec$)	1000(250 $msec$)
2 SD below Average	Detected Swallows	100%	98.8%	96.5%	96.4%
	Average Duration (msec)	1461.6 ± 499.5	1564.9 ± 472.9	1335.4 ± 893.8	1474.1 ± 956.1
1 SD below Average	Detected Swallows	100%	97.7%	94.5%	95.3%
	Average Duration (msec)	1504.1 ± 465.7	1599.2 ± 452.3	1382.2 ± 631.6	1495.6 ± 644.2
90%	Detected Swallows	98.3%	90.8%	93.4%	94.2%
	Average Duration (msec)	1495.2 ± 355.9	1580.6 ± 270.4	1392 ± 625.8	1475.4 ± 366.5
95%	Detected Swallows	98.3%	90.8%	93.1%	94.2%
	Average Duration (msec)	1495.2 ± 355.9	1580.6 ± 270.4	1391.6 ± 625	1475.4 ± 366.5

due to misclassification at the borders of each swallow, in addition to the unlabeled swallows treated as false positives. These values go up to more than 90% for the clean records that don't contain these pause areas and/or weren't logged to have any visually missed events. Fig. 20 shows the results of applying the segmentation algorithm on one of the clean records. It can be clearly seen that the algorithm successfully captured all the swallowing events in the signal and didn't misidentify any part of the signal including the hyoid bone motion event prior to the last swallow (Fig. 20 lower right corner). The segmentation framework also presented similar performance when tested on the swallows from the independent clinical study where 97.4% of the swallows were correctly detected when considering an overlap window of 2 SD below the average swallow duration calculated from the original dataset, 84.2% of the swallows for 1 SD below average swallow duration, and 65.8% of the swallows when considering overlap ratios of 90% or more.

More than 6500 detected segments were analyzed and validated visually against the videos and session logs for a window size of 900 and a 90% overlap criterion. The outcomes of the analysis (Table 9) show that the algorithm captured more than 94% of the swallows which is nearly a match with the results of the whole dataset in Table 8. Moreover, the rater reported that the algorithm successfully detected 353 swallows that were not captured/labeled in videos. The visually uncovered events reported in Table 9, are the segments detected by the algorithm during video pause times with no reference in session logs.

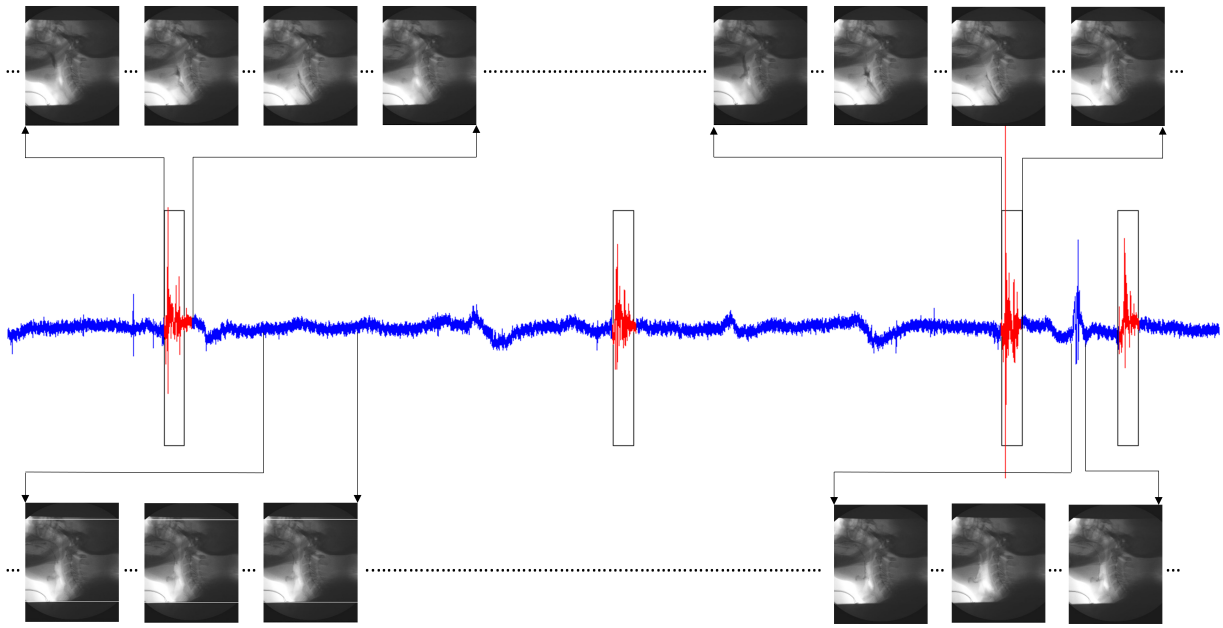


Figure 20: A clean AP acceleration record. The red segments represent swallows as labeled by SLP. Black boxes are segments detected by the algorithm. Images on each corner are simultaneous VFSS snapshots of the signal events

Table 9: Outcomes of the manual validation of automatic segmentation results

Event type	Details		Total count
Swallowing events	Detected by the algorithm	2225	2353
	Undetected by the algorithm	128	
Non-swallowing events	Reduced oral containment (Premature Spillage)	38	1275
	Hyoid bone movement	434	
	Coughing	134	
	Head and neck movement	266	
	Unexplained	403	
Visually uncovered events			2936

4.4 Discussion

The results confirmed our hypothesis that the proposed algorithm can correctly and without human intervention, detect 95% of known swallow durations in more than 90% of attempts across simple (clean swallows) and complex (non-swallow activity co-occurring with swallows) swallow events. We can clearly see from Fig. 19 that training a DNN with the spectral estimate for the raw swallowing vibrations of a single channel can produce accuracies as low as 26.1% and up to 97.6% on window level over the whole dataset. In addition, the system showed robustness in terms of true positive and true negative rates. The best performing channel was the A-P accelerometer axis with an average accuracy of 89.44% for single swallows (75.9% for the whole dataset) and superior sensitivity and specificity which is comparable to the results in [109]. The performance of other channels was close to the A-P axis, but the lowest performance was given by feeding the network with the spectrogram of the SI axis for all considered quality measurements.

The selection of proper window size highly depends on signal temporal characteristics which is obviously clear in the demonstrated results. We stated that the whole set of collected swallows is on average of 862.6 ± 277 msec. This makes the best window size to detect these swallows, located around the middle of used range (800-1000) because each swallow can be represented as integer multiples of the selected window in this range especially since we did not use any overlap between the sliding windows. This effect is most highly illustrated in the results of the A-P acceleration where the accuracy, true negative and true positive rates increase to their maximum at window size of (900-1000) and then drop sharply. They return to increase after this drop because the window size increases and approaches multiples of the effective values mentioned before. The effect is almost the same with other components of acceleration and microphone signals.

The temporal accuracy of detection was examined as well for the best two systems with window sizes of 800 and 900 for single swallows (900 and 1000 for the whole dataset) and validated against the manual segmentation by SLPs as shown in Table 8. Among the examined assessment criteria, we found that a 2 SD below average swallow duration criterion as the minimum overlap ($431.89\ msec \approx 47\%$ of average swallow duration) between

the detected and manually segmented swallows, is very low considering the duration of the examined swallows, however it gives excellent detection results. So, we tested 1 SD below the average duration (675.56 $msec \approx 73.5\%$ of average swallow duration) as well as 90% and 95% minimum overlap. The average duration of detected segments in the three criteria are close in duration and all of them are not far from the average duration of the actual segments. Moreover, the fluctuations in segment duration are considered very convenient compared to the length of segments. Therefore, all of these criteria proved to deliver excellent automated detection accuracy of swallow events without human intervention.

Encouragingly, the system has also shown promising segmentation quality when applied on completely unseen data collected from different group participants with control parameters that were not included in the main dataset under investigation. Despite these promising results, there is a little drop in the number of swallows correctly segmented considering different overlap windows when compared to the original dataset. The reason behind this drop in performance may be returned to the fact that there is actually a difference in the average swallow duration between the two dataset (a little longer in case of swallows from healthy participants) which in turn reflects on the needed window size that best represent the swallow. Another factor that may contributed into this performance drop, is the possibility that the used set of swallows contains some multiple swallows which cause the detection quality to drop when included as shown in Table 8. Nevertheless, the performance presented by the system on the new dataset suggest that it is likely to generalize to other swallowing datasets.

Evidently, the proposed algorithm achieves results better than most of the segmentation work in the literature, especially the work in [108] which achieved high segmentation accuracies. What makes our algorithm superior, is that it was validated using a wide dataset rather than a controlled limited dataset like most of the previous studies. The used dataset was at least 10 times larger than any used dataset in swallowing segmentation and covered most of the known swallowing conditions encountered in dysphagia screening which occurs in typical healthcare environments that allow for very limited control of patient position and other variables. This is important because our results, obtained in a naturalistic setting, are more externally valid than they would be had the data been collected under strict experimental controls as seen in many prior published studies. In addition, the proposed algorithm

has a better response time in testing phase that will be suitable for real time processing, if we ignored the burden of training the network. The algorithm uses massive computational resources for the training phase like any deep neural network, but this can be overcome using the newly emerging platforms with GPUs or special architectures to accelerate the process. Fusion between acceleration axes and/or swallowing sounds may also be of the future directions to be investigated for boosting the detection quality.

The start and end of each pharyngeal swallow can be roughly identified through visual and tactile inspection of hyo-laryngeal excursion and other observations of the patient swallowing. However, these methods are subjective and not reliable. Traditional cervical auscultation using a stethoscope to observe swallowing sounds, is particularly unreliable despite its commonplace use. This renders the advancements in high resolution cervical auscultation and machine learning methods demonstrated in this investigation and others, especially encouraging toward a goal of unsupervised detection of swallow events and many of their physiologic components and more timely identification of patients with dysphagia who need intervention. Adding a robust method that can automatically identify swallows is of a great clinical significance to diagnosis and rehabilitation of swallowing disorders. Such methods can detect swallows that are hard to observe in patients who have difficulty initiating oropharyngeal swallow (e.g. Parkinson's disease) or patients with weak pharyngeal swallow (e.g. medullary stroke) [59]. Future directions for this technology include developing computational deglutition methods to pre-emptively detect airway compromise (e.g. aspiration) and other clinically significant swallowing disorders at the bedside [339], facilitate behavioral treatments by providing real-time swallow biofeedback [340] , and in day-to-day management of swallowing disorders in settings that lack adequate qualified dysphagia clinical specialists.

4.5 Conclusion

In this study, a novel automatic segmentation algorithm for swallowing accelerometry and sounds was proposed, and its potential in dysphagia screening was discussed. The algorithm scans the swallowing signals through a sliding window of a specific size and each

window is classified as a swallow or non-swallow through feeding its spectral estimate into a deep neural network. Swallowing signals from 248 participants were collected for different swallowing tasks, manually labeled by experts and used to train and validate the system. The proposed algorithm yielded over 95% accuracy at the window level in addition to similar values of sensitivity and specificity. On the temporal side, the algorithm nearly did not fail in detecting any swallowing activity (2SD below average) and proved superior in detection despite high overlap ratios with accuracies that exceeded 90% for all types of swallows. Our algorithm has demonstrated the potential of convolutional deep neural network and spectral representation of swallowing signals to event detection in swallowing accelerometry.

5.0 Automatic Swallowing Segmentation using Convolutional Recurrent Neural Networks and Sensor Fusion

The content of this chapter is currently under review with the Journal of *Information Fusion*. Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdić, "Autonomous swallow segment extraction using deep learning in fused neck-sensor vibratory signals from patients with dysphagia," *Information Fusion*, submitted August 2021.

5.1 Objective

This study introduces 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, neurodegenerative diseases and those suffering from iatrogenic dysphagia due to cardiothoracic surgery. The dataset was used to validate the precision of swallowing segmentation using the proposed deep learning framework and compare its accuracy to other networks that have the potential of producing competing results in similar event detection problems.

5.2 Methods

5.2.1 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 [107]. 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) [341]. 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 fits 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 [342, 343]. 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 smooth the transient noise such as sudden head movements [106, 107].

5.2.2 Expert Manual Swallow Segmentation (Ground Truth)

VFSS streams were inspected by expert raters who are trained to perform swallow kinematic judgments, in order to identify the onset and offset of individual swallows. 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 [106, 344]. The offset is defined as the frame in which the hyoid bone returns to its resting location after completing the swallowing associated motion [106, 344]. The raters were blinded to participants' demographics and diagnoses and maintained an inter-rater and intra-rater reliability with ICC's that exceeded 0.99 during rating the swallows of the dataset. The ratings were used to label the concurrently collected swallowing vibratory signals.

5.2.3 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 [106]. 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. 21 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 $T \times N_2 \times C$.

5.2.4 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 algorithm used; however, both schemes

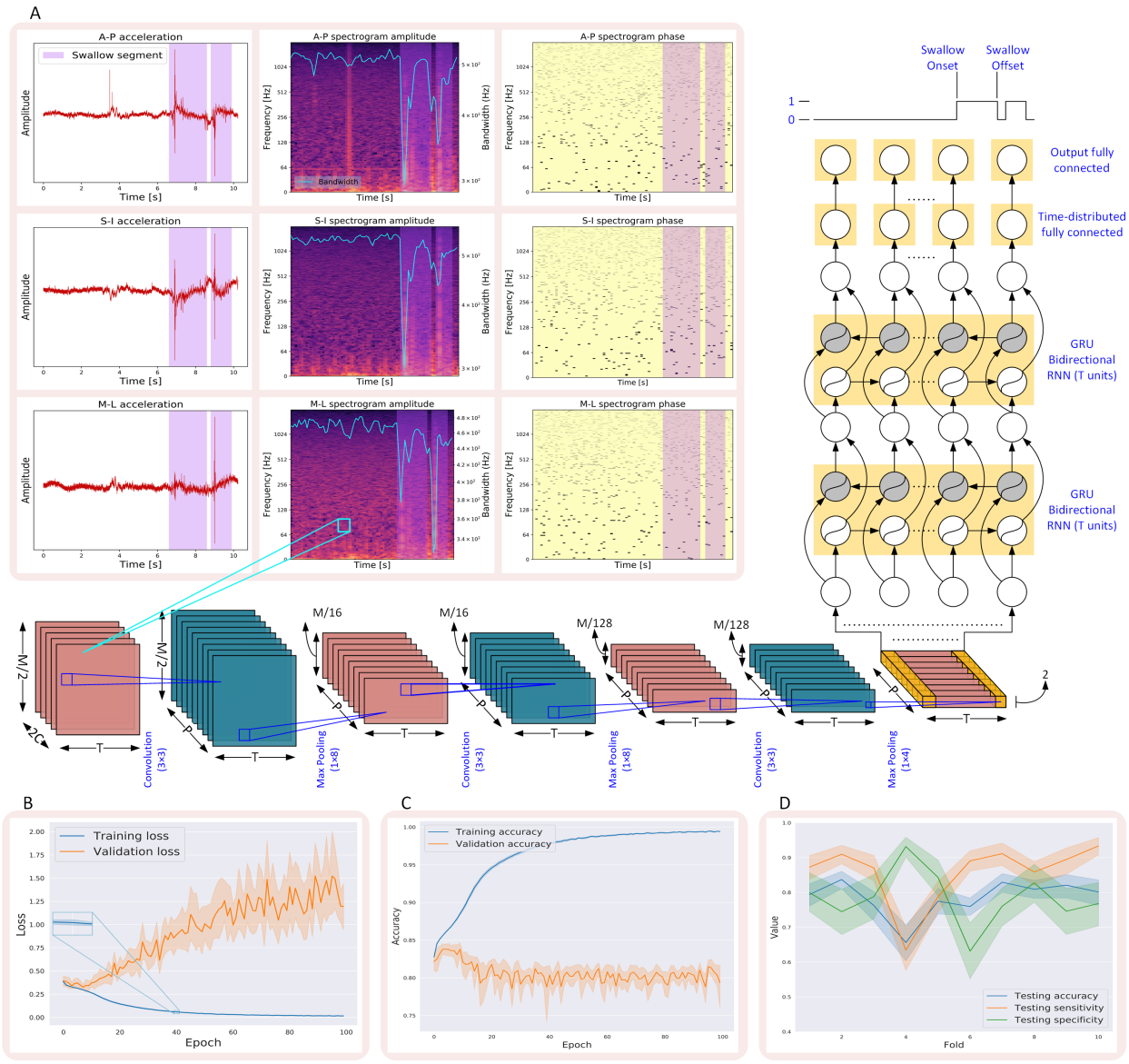


Figure 21: The architecture of the main proposed deep network. **A.** shows a typical unfolded example of the network input of acceleration signals with two swallow segments as indicated by the purple shadows in the figures. The first column represents raw acceleration signals, and the second and third 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.

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 the sequence-based models take sequence of windows as input instead of separate windows for recurrence modeling. Partitioning for the 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.

5.2.5 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 [106] 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 [345]. Fig. 22 show the architecture of the aforementioned model and its variants that are described later in text.

5.2.6 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) which is a special type of neural networks that processes sequential data one step at a time and selectively transfer information across time steps [92]. The first part of the architecture in these models is a convolutional neural network (CNN) that represents another special type of neural networks and is used to extract local features from input's time steps before passing them into the RNN to process the temporal depen-

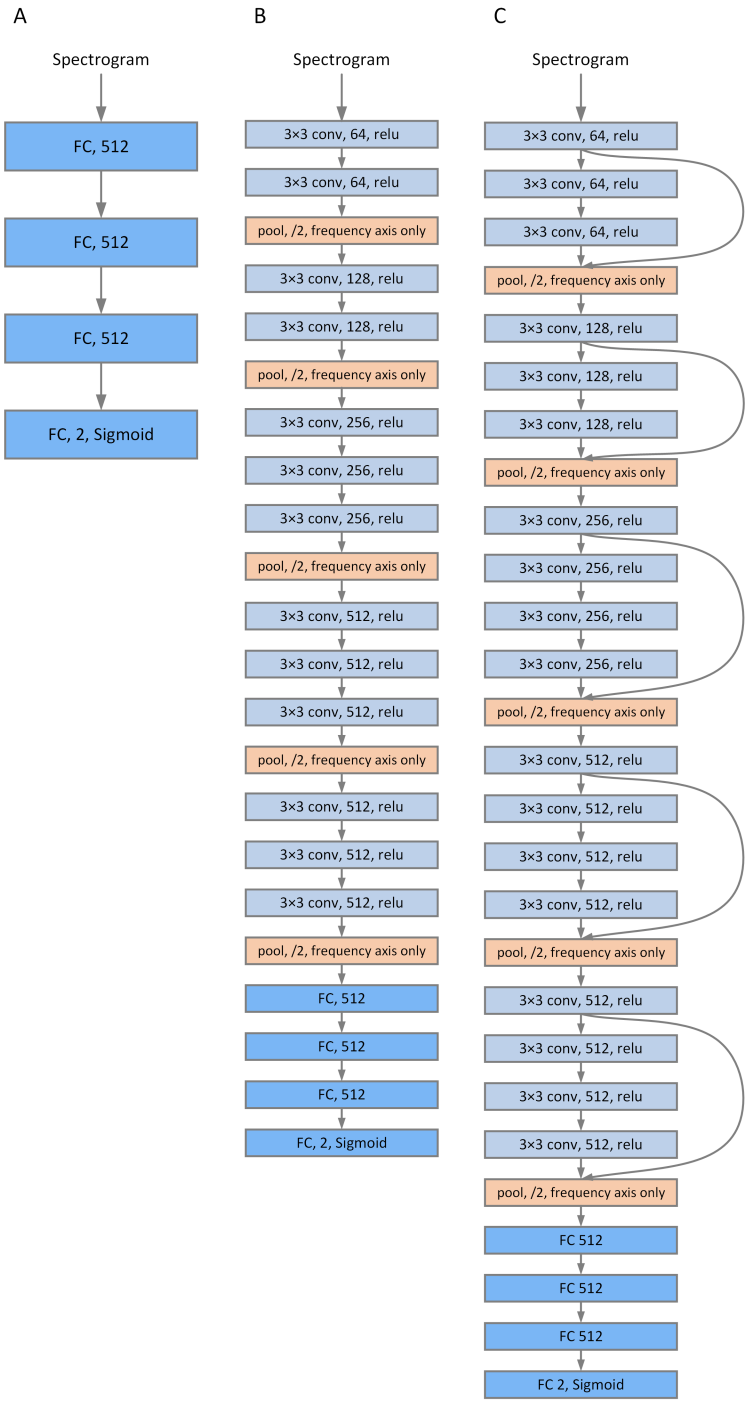


Figure 22: Layer stacking in each of the network variants. **A.** shows the network that uses only fully connected layers to process the spectrogram. **B.** shows how the VGG16 CNN layers were stacked ahead of the fully connected layers. **C.** shows how the skip connection that perform the residual learning were introduced to the VGG16 design of the network.

dencies. 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 of 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 model the time dependencies a long the sequence. RNN is 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 [92]. The last part of the model is a fully connected neural network that combines the temporal features generated by the RNN in order to generate a final segmentation sequence that represent the orientation of each window in the sequence. Fig. 21 shows one of the sequence-to-sequence architectures used in this study which takes spectrogram as input and is composed of 2D CNN.

The 2D CRNN model shown in Fig. 21 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 a 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 te time steps. The final CNN output is fed into a 2-layer GRU-based bidirectional 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 [345]. The sequence-based 10-fold cross validation scheme is used to evaluate all sequence-to-sequence-based models.

5.2.7 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 [346–348]. However, as the depth increases, the accuracy gets saturated and degrades rapidly [346]. 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 is easier to optimize than the unreferenced mapping [346]. 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. 22 demonstrates how layers are stacked to modify the simple deep fully connected network model to be more deeper (Fig. 22 B) and to use residual learning represented by the introduced skip connections (Fig. 22 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 [347]. For the residual network, we inserted skip connections into the VGG16 model which can be used 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 for dimensions matching. We used extra convolutional layers prior each identity shortcut to match dimensions (Fig. 22 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 [349]. All networks are 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 [349].

5.2.8 Performance Metrics

The main segmentation problem in this study is a binary classification task, for which the 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. 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 [106].

5.3 Results

5.3.1 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 10). 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

Table 10: Characteristics of the participating patients with suspected dysphagia

Admitting diagnosis	Included conditions	Subject-level (N, %)	Age, year (mean ± s.d.)	Female (N, %)
Neuro-degenerative 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%

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 characteristics of the collected swallows are detailed in Table 11. 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).

5.3.2 Real-Time Prediction of Swallow Segment Onset and Offset Using HRCA Signals Alone

We tested multiple deep networks to detect the onset and offset of 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

Table 11: 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

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 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 [106]. 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 [107]. 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

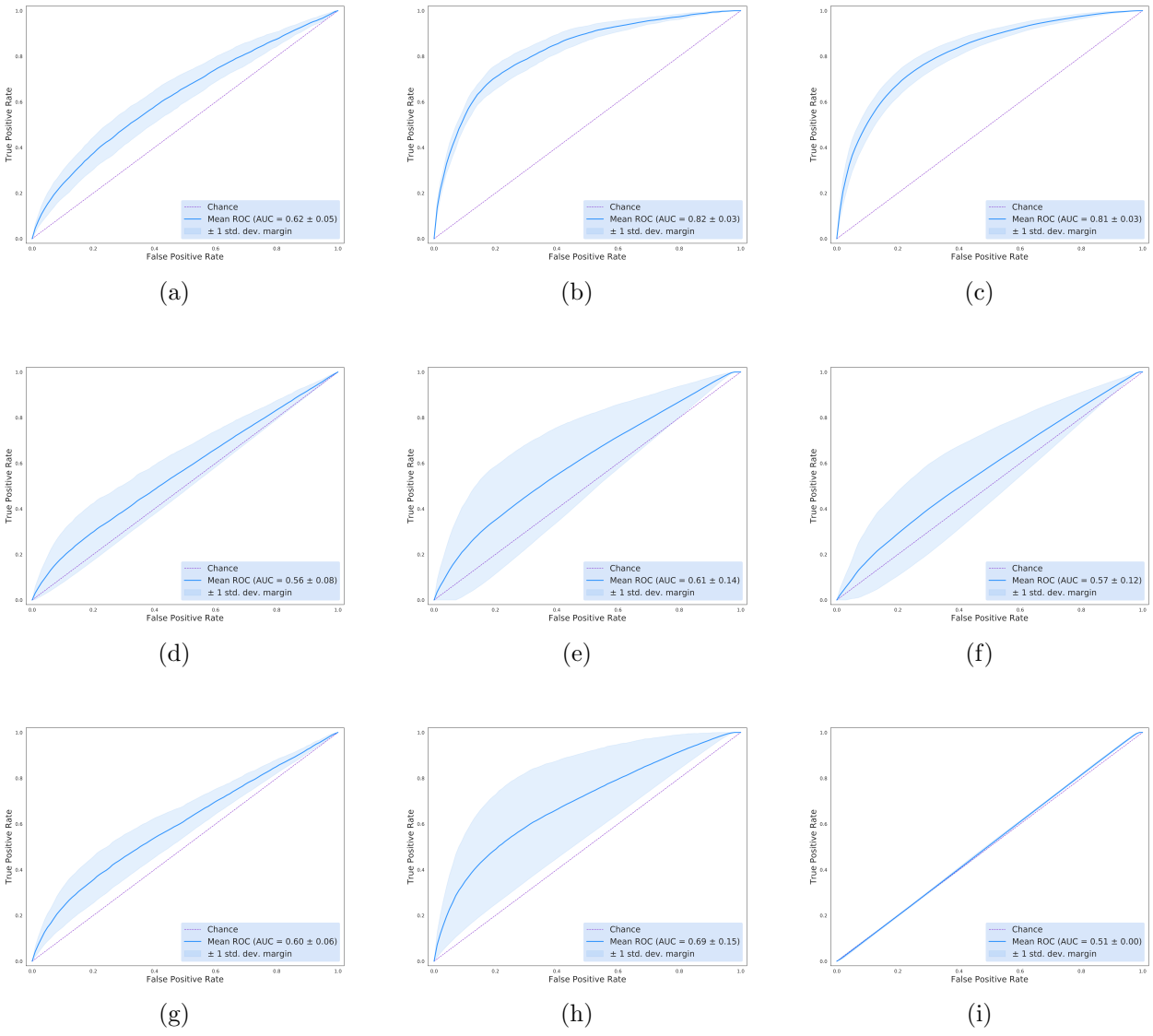


Figure 23: Receiver operating characteristic curves of the window-wise predictions of swallow segments. The nine models are (1) a 4-layer fully connected neural network with the spectral estimate as input (2) a 2D shallow CRNN with the spectral estimate as input (3) a 1D shallow CRNN with the raw signals 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.

Table 12: 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

stacking 16 weight convolutional layers (called VGG16 network [347]) before the base model layers (Fig. 22 B). 2D convolutional layers were stacked in the case of power spectral estimate inputs while 1D convolutional layer 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. 22 B) which was described elsewhere [346] 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 [106]. 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 [107]. 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

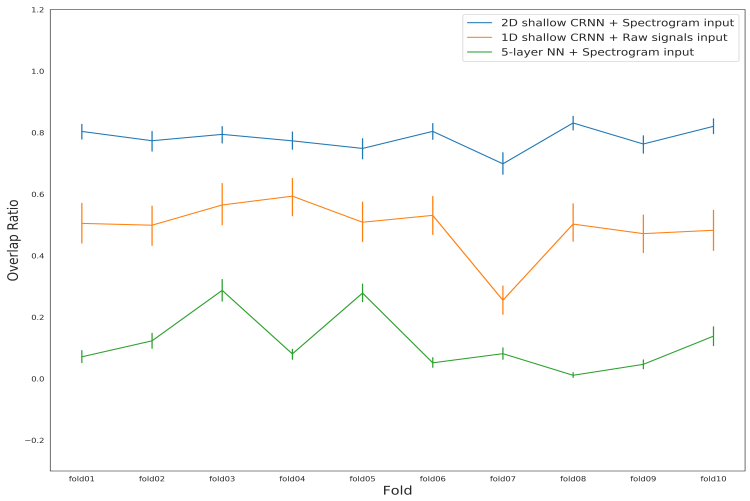


Figure 24: Average overlap ratio between detected swallow segments by the three best performing models and the reference swallow segments labeled by the gold standard across the 10 folds of the cross-validation process.

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. 23 and Table 12). Adding more layers to the CNN parts of the network did not improve the performance as can be seen in Fig. 23d-23f. On the other hand, residual learning achieved a performance that was between the base models and the VGG16 variant models (Fig. 23g-23h) except for the model that used raw signals as input (Fig. 23i).

5.3.3 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 (**onset:** bolus passes the ramus of the mandible, **offset:** hyoid bone returns to its lowest position after clearance of the bolus tail through the upper esophageal sphincter) as it may detect 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

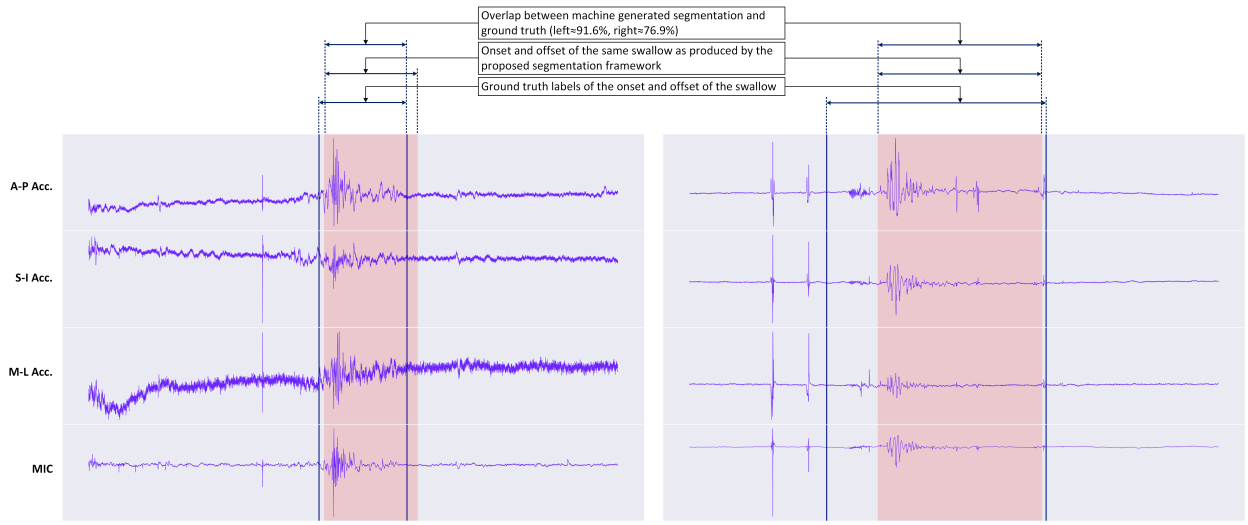


Figure 25: This figure shows two swallows from two different subjects, a male (age: 44) who developed dysphagia secondary to stroke (left panel) and a female (age:69) who developed dysphagia secondary to subdural hematoma (right panel). The onset and offset of the swallow segments are marked with dark 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.

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. 24). The indicated model consistently detected around 79% (s.d.: 11% and 95% CI: 77.8-79.6%) of the swallow segment across all folds with small variation in each fold. 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. 24. 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. 25 with an overlap with the gold standard labels of 91.6% and 76.9% (left to right).

5.4 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 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 clean signals for the evaluation process in contrast to other studies that used signals with blind segments [106]. 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 clean 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 12) 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 82% (s.d.: 3% and 95% CI: 80.7-84.1%) across the 10-folds of the entire dataset (see Fig. 23). 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 algorithm and the human labeled swallow segments. This overlap refers to the percentage of the swallow segment that was detected by the algorithm. 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.25 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 clinical importance of the proposed network is three-fold. 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 [85, 87, 107, 282, 339], the proposed automated segmentation framework mitigates the unavoidable human error in manual segmentation on which most of dysphagia characterization algorithms are reliant [344]. 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 [85, 107, 282]. 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 [350]. 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.

5.5 Conclusion

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 tools. 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.

6.0 Non-Invasive detection of Upper Esophageal Sphincter Opening

The majority of this chapter has been previously published in and reprinted with permission from [107]. © 2021 IEEE. 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, February 2021.
DOI: 10.1109/JBHI.2020.3000057

6.1 Objective

This study investigates and describes a concrete implementation that uses HRCA signals as input to estimate the exact timing of upper esophageal sphincter (UES) opening and closure, and to compare that estimate to gold-standard judgment of videofluoroscopic images. In this implementation, we use recurrent neural networks (RNNs) in association with convolutional neural networks (CNNs) in order to extract the dynamics of the swallowing vibrations from HRCA signals and use them to infer the moments when the UES first opens and re-closes during swallowing. A strength of this approach is that non-linear RNNs can perfectly fetch the complex temporal activity patterns of swallowing signals that reflect the major physiological events controlling the swallowing mechanism. We have also considered a diverse dataset that covered a variety of etiologies, bolus sizes, consistencies, and maneuvers to assure the generalization and robustness of the implementation instead of using a controlled dataset used for specific etiology diagnosis purposes which is beyond the goal of this study. In addition to that, the detected duration of UES opening (DUESO) was compared to the labeled DUESO by trained raters from VFSSs for the same swallows to calculate the temporal accuracy considering the start and end edges.

6.2 Methodology

6.2.1 Materials and Methods

Permission for this study was granted by the institutional review board of the University of Pittsburgh and all participating patients provided informed consents including consent to publish before enrollment. A total of one hundred and sixteen patients (72 males, 44 females, age: 62.7 ± 15.5) with suspected dysphagia resulting from a variety of diagnoses, underwent an oropharyngeal swallowing function evaluation by a speech language pathologist using VFSS at the University of Pittsburgh Medical Center Presbyterian Hospital (Pittsburgh, PA). Of the sample, 15 patients were diagnosed with stroke while the remaining 101 patients were diagnosed with different medical conditions unrelated to stroke.

Swallows for this study, were collected as a part of standard clinical care rather than for research purposes alone. As a result, speech language pathologists who conducted the VFSSs, had the ability to alter the evaluation protocol based on the patient's clinical manifestation of dysphagia. This included how the boluses were administered to patients (i.e. spoon, cup), the volume and viscosity/texture of each bolus of food and liquids, the number of trials, and head position during swallowing (i.e. head/neck flexion, head rotation, head neutral). The following consistencies were used during VFSSs: 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). Boluses were either self-administered by patients via a cup or a straw or administered by the clinician through the use of a spoon ($3 - 5$ mL).

This study yielded 710 swallows (132 from patients diagnosed with stroke and 578 from patients with other diagnoses) with an average duration of pharyngeal bolus transit of $869.5 \pm$ 221 msec and an average DUESO of 604.9 ± 150 msec. The collected swallows included 224 single swallows (single bolus swallowed with one swallow), 477 multiple swallows (single bolus swallowed using more than one swallow), and 9 sequential swallows (multiple boluses swallowed sequentially in a rapid manner).

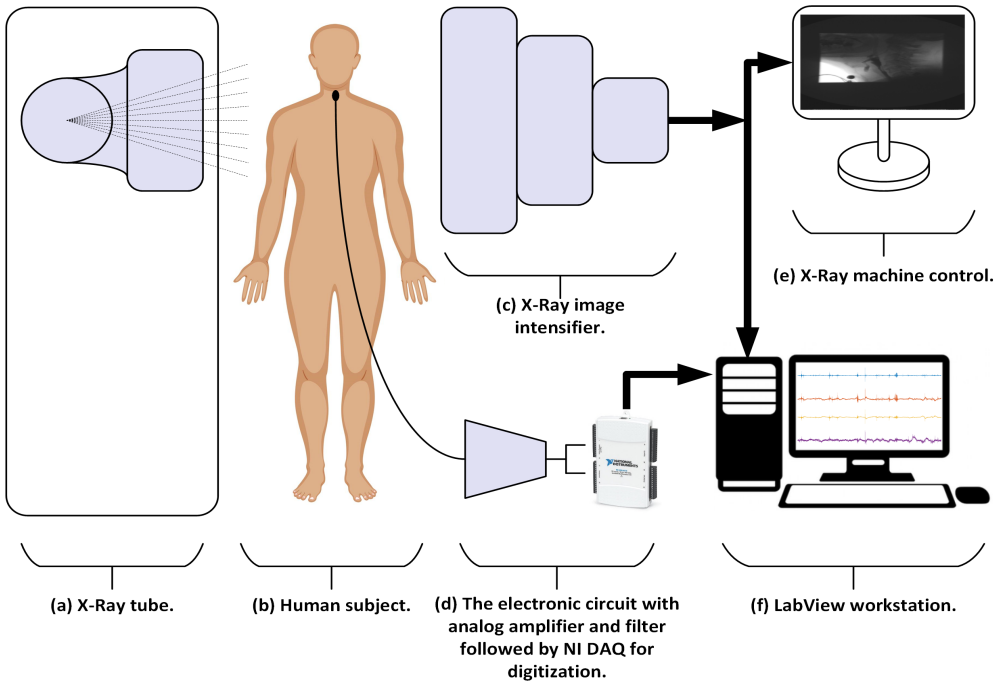


Figure 26: The experimental setup of the study. (a) An X-Ray tube that resides in a table is adjusted in a vertical position to be parallel to the swallowing path. (b) The human subject is standing or comfortably seated between the x-ray tube and the image intensifier with the HRCA sensors attached to the anterior neck. (c) The image intensifier is positioned and adjusted according to the subject height, so that the produced frames capture all of the important anatomical landmarks of the oropharyngeal swallow (jaws, pharynx, and esophagus). (d) The sensors are connected to the electronic circuit that supplies power and performs analog amplification and filtration and then to the NI DAQ for sampling. (e) The video feed is taken directly from the image intensifier to the X-Ray control workstation where clinicians and radiologists create, save , and view the exams. (f) The video feed from the image intensifier is cloned into the video capture card installed on the research workstation which is also connected to the NI DAQ and runs LabView for means of data collection and synchronization.

6.2.2 Data Acquisition

The general experimental setup is illustrated in Fig. 26. During all recording sessions, VF equipment was controlled by a radiologist and the patients were comfortably seated with the swallowing sensors attached to the anterior neck region using double sided tape. VF was conducted in the lateral plane using a Precision 500D system (GE Healthcare, LLC, Waukesha, WI) at a pulse rate of 30 pulses per second (PPS) and with the images acquired

at a frame rate of 30 frames per second (FPS) [341]. The video stream was captured and digitized using an AccuStream Express HD video card (Foresight Imaging, Chelmsford, MA) into movie clips with a resolution of 720×1080 at 60 FPS.

A tri-axial accelerometer (ADXL 327, Analog Devices, Norwood, Massachusetts) and a contact microphone (model C 411L, AKG, Vienna, Austria) were used to collect swallowing vibratory and acoustic signals. The accelerometer was mounted into a small plastic case with a concave surface that fits on neck curvature and the case was attached to the skin overlying the cricoid cartilage using a tape. The accelerometer was attached such that its main axes are aligned parallel to the cervical spine, perpendicular to the coronal plane, and parallel to the axial/transverse plane. These axes are referred to as superior-inferior (S-I), anterior-posterior (A-P), and medial-lateral (M-L) respectively. The microphone was mounted towards the right lateral side of the larynx to avoid contact noise with the accelerometer and guarantee a clear radiographic view of the upper airway. Attaching the sensors around the area of cricoid cartilage is logical given that most of the pharyngeal swallowing activity is produced by the anatomical structures present at this level and it has been reported to yield the best signal-to-noise ratio for the acquisition of swallowing signals [106, 314, 342, 343].

The accelerometer has a bandwidth of 1600 Hz after which the response falls to -3dB of the response to low frequency acceleration. In other words, the accelerometer has a low pass filter with a cut-off frequency at 1600 Hz. The contact microphone was chosen as well so that it produces a flat frequency response over the entire range of audible sounds which was proved to pass most of the frequencies encountered during swallowing [332, 342, 351]. The signals from both the accelerometer and microphone were hardware band-limited to 0.1-3000 Hz with an amplification gain of 10. The cut-off frequencies for the band-limiting filter were chosen so that most of body sway components below 0.2 Hz are suppressed and the signal components with the vast majority of energy are passed [314, 333, 351, 352]. The signals were sampled using a National Instruments 6210 DAQ at a sampling rate of 20 kHz. Both signals and video were acquired simultaneously using LabView's Signal Express (National Instruments, Austin, Texas) with a complete end-to-end synchronization.

6.2.3 VF Image Analysis

Video clips were segmented based on individual swallow events by tracking the bolus in a frame by frame manner. The onset of the pharyngeal swallow event was defined as the frame in which the head of the bolus passes the shadow of the posterior border of the ramus of the mandible and the offset as the frame in which the bolus tail passes through the UES [344], in order to capture the entire duration of pharyngeal bolus flow. Three expert judges trained in swallow kinematic judgments, identified the video frame of first UES opening and the video frame of first UES closure in the segmented videos. All raters who segmented swallowing videos and analyzed UES opening and closure established a priori intra- and inter-rater reliability with ICC's over 0.99. All raters maintained intra- and inter-rater reliability throughout measurements on 10% of swallows with ICC's over 0.99 and were blinded to participant demographics and diagnosis and any bolus condition information.

6.2.4 Signals Preprocessing

Numerous physiologic and kinematic events such as coughing and breathing occur in close temporal proximity to the pharyngeal swallow event. These events can contribute to the collected vibratory and acoustic signals [109]. As a first step to overcome confounding noise in the signals due to multi-source environmental data collection and other measurement errors, the signals accrued at a sampling rate of 20 kHz were down-sampled to 4 kHz. A more intense down-sampling could have been adopted as previous studies reported that the frequency with the maximum energy for swallowing accelerometry signals occurs below 100 Hz and the central frequency almost below 300 Hz [91, 314, 317, 353]. However, we chose down-sampling to 4 kHz so that we match twice the max frequency component present in the acceleration signals (1600 Hz). Down-sampling was performed through applying an anti-aliasing low pass filter then picking up individual samples to match the new rate.

The baseline outputs of accelerometer and microphone (produced by zero-physical input) were recorded earlier before the main data collection procedure and device noise was characterized through modified covariance auto-regressive modeling [353, 354]. The order of the auto-regressive model was 10 and it was determined using the Bayesian information

criterion [353]. The coefficients of the auto-regressive model were then used to create a finite impulse response filter (FIR) to remove the device noise from the recorded swallowing signals [353]. Afterwards, the low-frequency noise components and motion artifacts were eliminated from accelerometer signals using fourth-order least-square splines [355, 356]. Particularly, we used fourth-order splines with a number of knots equivalent to $\frac{N \times f_l}{f_s}$, where N is the data length and f_s is the sampling frequency. f_l is called the lower sampling frequency and it is proportional to the frequency associated with motion artifacts. The values for f_l were calculated and optimized in previous studies [355]. Finally, the effect of broadband noise on signals was reduced through wavelet denoising [357]. Specifically, we used tenth-order Meyer wavelets and soft thresholding. The threshold was calculated using $\sigma\sqrt{2\log N}$, where N is the number of samples and σ is the estimated standard deviation of the noise (calculated through down-sampling the wavelet coefficients) [331, 357].

6.2.5 System Design

Due to the fact that there is no specific rule of thumb to calculate the number of layers and layer sizes for a certain problem, the used architecture was fine-tuned based on an experimental approach and by following the best network configurations that achieved good results in similar problems [300, 301, 358]. Particularly, we tested multiple architecture depths that included more layers of CNN (3, 4, and 5 layers) with up to 32 filters per channel and more RNN unit sizes up to 128 as well as different RNN units (GRU and LSTM) with both Sigmoid and tanh recurrent activations. The chosen architecture was found to be the most stable among the tested configurations and the fastest to converge. In other words, it included the smallest number of parameters to be optimized while achieving a detection accuracy that doesn't sharply change when adding more layers or increasing the layer sizes. The used architecture employed also dropout between layers as well as early stopping techniques to control the network from over-fitting to the training data [359].

The longest swallow event duration in the collected dataset was around 1500 msec (90 frames of VF). The signals were divided into chunks 16.67 msec in length (equivalent to one frame in VF or 66 samples in signals). Each signal chunk is composed of 3 axes of

acceleration which makes the dimensions 66 samples $\times$ 3 channels. The chunks were fed into a 1D convolutional neural network that included two convolutional layers with a max pooling layer in between as in Fig. 27. Both convolutional layers were followed by a rectified linear unit (ReLU). The first convolutional layer applied 16 "1 $\times$ 5" filters per channel which produced 3 "62 features $\times$ 16 channels". The max pooling layer applied a window of size 2 with 2 strides and reduced the features into "31 features $\times$ 48 channels". The last convolutional layer was identical to the first one except that it used only one filter per channel which produced "27 features $\times$ 48 channels".

The complete sequence of features $x_{1:T}$ (for a full swallow) coming out of the convolutional layer was then fed into a 3-layers dynamic RNN with gated recurrent units (GRUs) as building blocks each of 64 units and a sequence of 90 time steps. The RNN computed an output sequence $\hat{y}_{1:T}$ using the following nonlinear model:

$$
\begin{aligned}
r_t^{(k)} &= \begin{cases} \sigma(W_r^{(1)}\left[h_{t-1}^{(1)}, x_t\right] + b_r^{(1)}), & \text{k=1,} \\ \sigma(W_r^{(k)}\left[h_{t-1}^{(k)}, h_t^{(k-1)}\right] + b_r^{(k)}), & \text{k=2, 3} \end{cases} \\
z_t^{(k)} &= \begin{cases} \sigma(W_z^{(1)}\left[h_{t-1}^{(1)}, x_t\right] + b_z^{(1)}), & \text{k=1,} \\ \sigma(W_z^{(k)}\left[h_{t-1}^{(k)}, h_t^{(k-1)}\right] + b_z^{(k)}), & \text{k=2, 3} \end{cases} \\
\hat{h}_t^{(k)} &= \begin{cases} tanh(W^{(1)}\left[r_t^{(1)} h_{t-1}^{(1)}, x_t\right] + b^{(1)}), & \text{k=1,} \\ tanh(W^{(k)}\left[r_t^{(k)} h_{t-1}^{(k)}, h_t^{(k-1)}\right] + b^{(k)}), & \text{k=2, 3} \end{cases} \\
h_t^{(k)} &= z_t^{(k)} \hat{h}_t^{(k)} + (1 - z_t^{(k)}) h_{t-1}^{(k)}, \qquad \text{k=1, 2, 3} \\
\hat{y}_t &= U h_t^{(3)} + c
\end{aligned}
$$

The output sequence $\hat{y}_{1:T}$ coming out of the RNN was masked (ones/zeros mask) before being fed in to the following stages to balance for the shorter swallows (less than 90 frames). Furthermore, the length of each swallow was considered in the architecture of the RNN and the same mask was used in the calculation of the cost function for the whole problem. The sequence was then fed in to 4 fully connected layers in order to fuse the temporal features from RNN into a meaningful UES opening segmentation mask. This part of the network featured 3-ReLU activated layers with 128 units and an output layer that assembled 90 units, one for each time step in the swallow as shown in Fig. 27 plus Sigmoid activation for a zeros and ones segmentation mask. Each two fully connected layers were separated by a dropout layer with a drop rate of 20%.

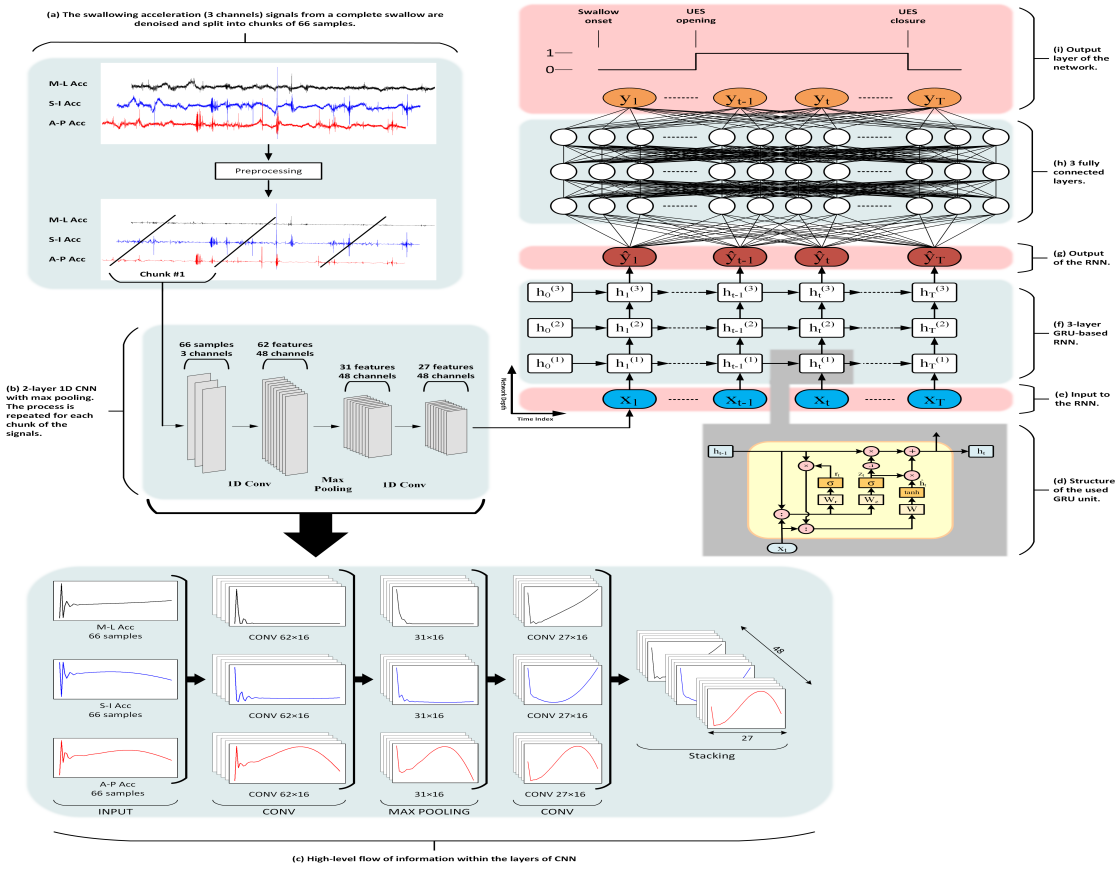


Figure 27: The architecture and data flow in the UES opening detection system. (a) This part is where the 3-channel acceleration signals from each swallow are denoised and split into equal chunks each of 66 samples (equivalent to 1 VF frame). (b) This part shows the operation of the CNN network part per data chunk. The architecture of the used 1D CNN which is comprised of two layers, the first applies 16 filters on each channel and produces 48 channels. The first CNN layer is followed by a max pooling layer and another CNN layer identical to the first except that it applies 1 filter per channel then a max pooling layer to reduce the size of the features. (c) This is an illustration for the operation of the CNN after training that shows a chunk of 3-channel acceleration pushed throw the first layer of CNN to produce 16 feature-channels per original channel. The length of chunks is shorter after this layer due to convolution on the edges of the chunks (no padding is used). (d) This is an illustration that shows the architecture of the GRU unit with the reset and update parts that help propagate states across time steps. (e) $(x_{1:T})$ is the output train from the CNN for chunks $(1:T)$ which is fed into the RNN units. (f) The architecture of the 3-layer RNN used for time sequence modeling. (g) The output sequence from the last layer of the RNN $(\hat{y}_{1:T})$ is flattened and fed into the first fully connected layer. (h) A diagram of the 3 fully connected layers (each of 128 units) used to combine the features coming out of the RNN. (i) The output layer of the network which is composed of 90 units $(y_{1:T})$ that resemble the UES opening mask.

The final cost function was defined as the mean squared error between the zero-padded ground truth $\bar{y}_{1:T}$ labeled by the expert judges and the masked output coming from the final connected layer $\hat{y}_{1:T}$ as follows:

$$MSE = \frac{1}{T}\sum_{i=1}^{T}\left[(\bar{y}_i - \hat{y}_i) \times mask_i\right]^2 \tag{6.1}$$

where $mask_i$ is the mask used to compensate for short swallows. We used the Adam optimizer to train the network due to its superiority in convergence without fine tuning for hyper-parameters [345].

6.2.6 Evaluation

The dataset was randomly divided into 10 equal subsets in terms of the number of swallows. A holdout method was repeated 10 times by training with 9 subsets and testing with the remaining one (10-fold cross validation). The results of the proposed system are in the form of a segmentation mask that tells when the UES opens and closes with respect to the start (onset) of the swallow segment as shown in Fig. 28 (b). This mask is calculated for approximately each swallow in the dataset when passed as a test sample through the trained system. In order to acquire a solid evidence about the detection quality of the system, a confusion matrix is constructed for each swallow based on the predicted segmentation mask and the reference mask as labeled by judges. The confusion matrix is then used to calculate accuracy, sensitivity, and specificity as follows:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ Sensitivity &= \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{FP+TN} \end{aligned}$$

where TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, and FN stands for False Negative. Furthermore, the difference between the actual and predicted UES opening and UES closure was measured, so that we could compare it to the human judges' tolerance reported in the literature.

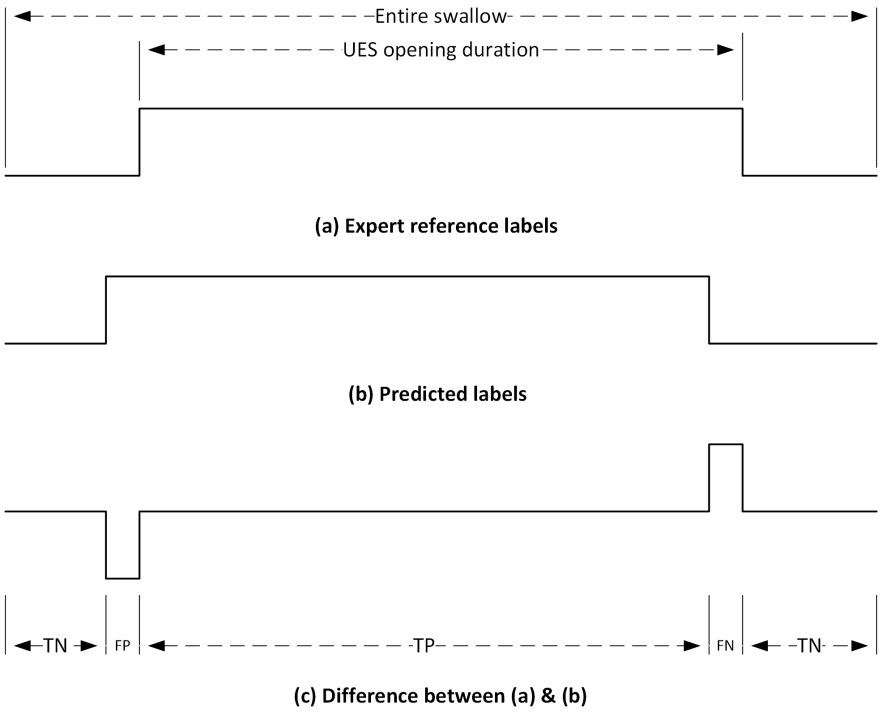


Figure 28: The evaluation procedure for each swallow. (a) The UES opening mask created from the expert manual segmentation in VF images. (b) The UES opening mask as predicted by the proposed algorithm. (c) Comparison is performed between the masks from (a) and (b) to create a confusion matrix. The confusion matrix is created in this way for each swallow included in testing. The values of accuracy, sensitivity, and specificity are calculated through this confusion matrix.

6.2.7 Clinical Validation

In order to evaluate the proposed system in a clinical environment, it was tested during the workflow of an ongoing clinical experiment performed on 15 (8 males, 7 females, age: 63.7 ± 6.2), community dwelling healthy adults who provided informed consent, and who had no reported current or prior swallowing difficulties. Participants in this validation sample also had no history of neurological disorder, surgery to the head or neck region, or chance of being pregnant based on participant's report. The experimental setup of this clinical experiment relied on the same equipment and hardware used for the collection of the main dataset as shown in Fig. 26. This included recording VF in the lateral plane using a Precision 500D system (GE Healthcare, LLC, Waukesha, WI) at a pulse rate of 30 pulses

per second (PPS) and with the images acquired at a frame rate of 30 frames per second (FPS). The video stream was captured and digitized using an AccuStream Express HD video card (Foresight Imaging, Chelmsford, MA) at 60 FPS. Swallowing vibratory and acoustic signals were acquired concurrently with VF using the same tri-axial accelerometer and microphone (ADXL 327, Analog Devices, Norwood, Massachusetts andmodel C 411L, AKG, Vienna, Austria). The sensors were attached to the same location on the anterior neck to the skin overlying the cricoid cartilage. The signals from both sensors were also band-limited between 0.1-3000 Hz and amplified with a gain of 10 then sampled at a rate of 20 kHz via an NI 6120 DAQ through LabView's Signal Express (National Instruments, Austin, Texas).

The participants in this clinical experiment were community dwelling adults without report of current or prior swallowing difficulties. Therefore, only ten thin liquid boluses (5 at 3mL by spoon, 5 unmeasured self-selected volume cup sips) administered in a randomized order in order to limit x-ray radiation exposure. For all spoon presentations, participants were instructed by the researcher to "Hold the liquid in your mouth and wait until I tell you to swallow it." Liquid bolus presentations by cup varied in volume by participant, because participants were instructed by the researcher to "Take a comfortable sip of liquid and swallow it whenever you're ready." Fifty swallows, selected randomly from this independent clinical experiment, were used to test the system for UES opening detection after being trained over the full 710 swallows dataset.

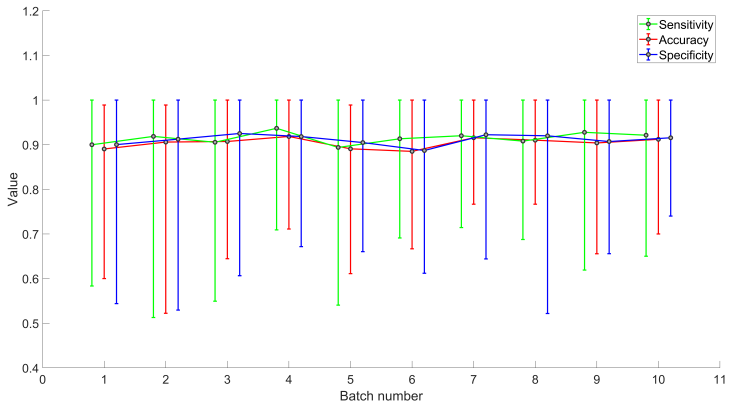
6.3 Results

A chunk of 3D acceleration (3×133) was first preprocessed to achieve denoising and artifact removal as shown in Fig. 27. After preprocessing, the filtered acceleration segments were fed into the convolutional network (CNN) part of the system as in the snapshot shown in the lower part of Fig. 27. The snapshot represents a sample feature map across the CNN that shows the evolution of inputs (low-level features) into high level features at the final layer of the CNN. The later helps identify more complex features in the input signals and promote distinctive traits while the insignificant features disappear.

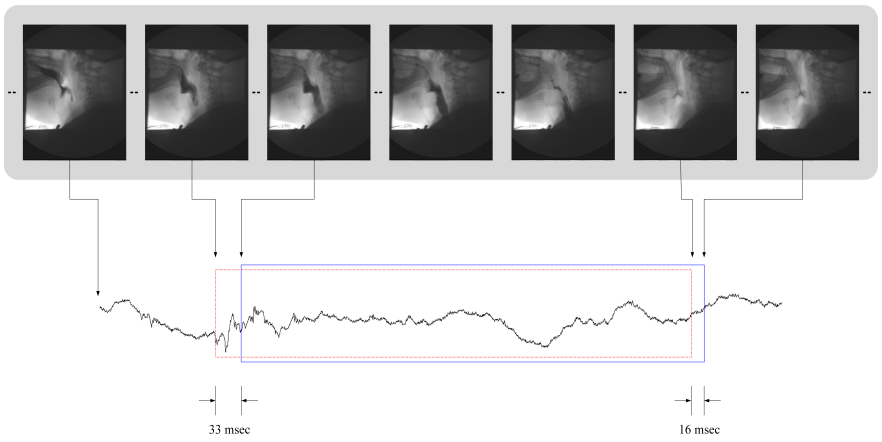
Table 13: Summary of the performance measurements that the proposed system achieved for both the main patient and the independent clinical datasets.

	Main dataset	Independent dataset
Average Accuracy	0.9093	0.8880
Average sensitivity	0.9145	0.8559
Average specificity	0.9119	0.9356
% of swallows with UES opening error $<$ 3 VF frames	82.6	84
% of swallows with UES opening error $<$ 4 VF frames	90	88
% of swallows with UES closure error $<$ 3 VF frames	72.3	66
% of swallows with UES closure error $<$ 4 VF frames	80	74

Fig. 29 (a) shows the performance of the proposed system across the 10-folds of the whole set of swallows. The values presented, represent the distribution of sensitivity, accuracy, and specificity in each fold. Each vertical line has 3 main points that represent the min average and maximum respectively from bottom up. The average accuracy of all folds across the whole dataset was 0.9039 with 0.9145 sensitivity and 0.9119 specificity. Fig. 29 (b) depicts a comparison between DUESO detection from the proposed system against the manual labeling by experts through the use of VF. On average, the network detected UES opening 33 msec earlier and closure 16 msec earlier than true opening and closure as measured by swallow kinematic analysis. The outcome of the algorithm for the whole set of swallows, was calculated and compared to the VF based labels and the differences are shown through the histograms in Fig. 30 (a-b) and Table 13. The comparison shows that for 82.6% of the swallows, the opening of UES was detected within a 100 msec ($\approx$ 3 frames at 30 FPS) of the human ratings, and within a 133 msec ($\approx$ 4 frames at 30 FPS) for 90% of the swallows (Fig. 30 (a)). Likewise, the network accurately detected UES closure within a 100 msec ($\approx$ 3 frames at 30 FPS) for 72.3% of the swallows and within a 133 msec ($\approx$ 4 frames at 30 FPS) for more than 80% of the swallows (Fig. 30 (b)). The accepted tolerance for human frame selection $\approx \pm$ 2.48 frames at 30 FPS [344].



(a)



(b)

Figure 29: Distribution of per swallow based performance measurements in each testing batch of the 10-fold cross validation process and a sample visual of the detection in one of the swallows. A sample of figures showing the timing difference between the automatically detected DUESO by our algorithm and the actual DUESO observed from VF (in frames) for both opening and closure. (a) Distribution for accuracy, sensitivity, and specificity in each batch (min, average, and max). (b) shows a sample full swallow with both the predicted (in red) and the actual DUESO (in blue) marked on the A-P acceleration component and video frames.

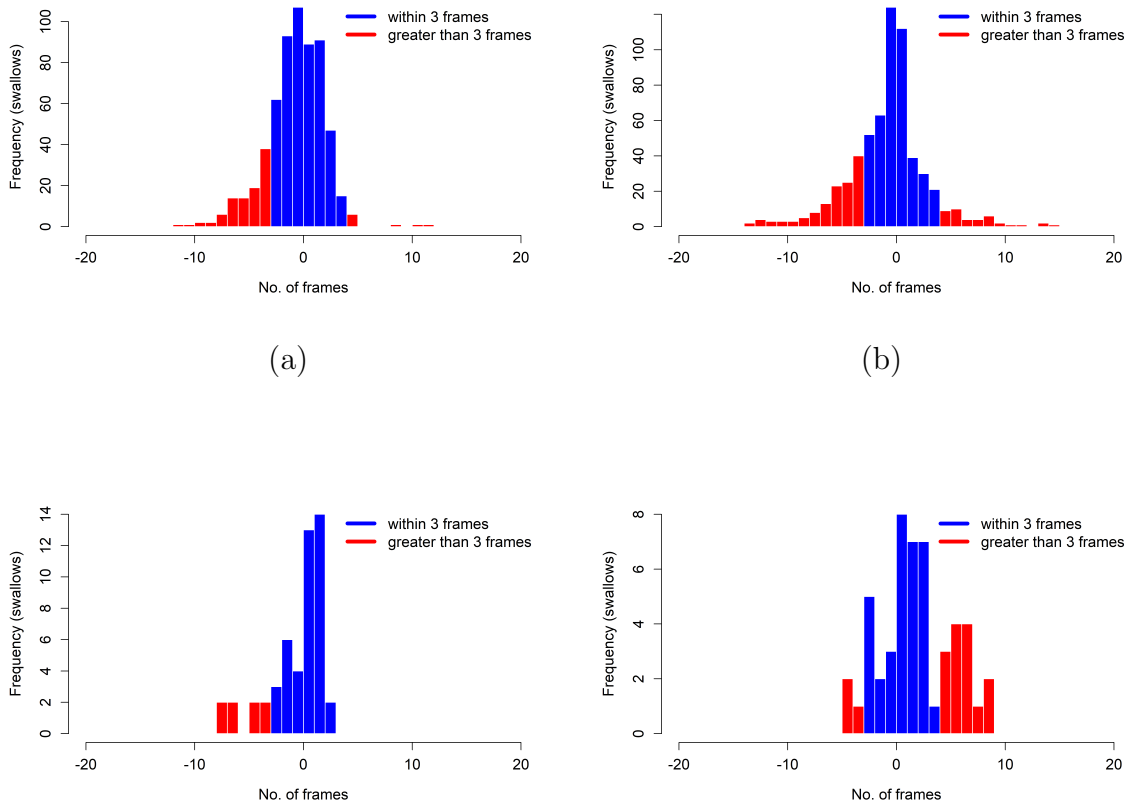


Figure 30: The timing difference between the automatically detected DUESO by the proposed system and the actual DUESO observed from VF (in frames) for both opening and closure in the whole dataset and the clinically independent data. The differences between the detected opening frame and the opening frame marked by the judges are highlighted in (a) for the 10 folds within the original dataset and in (c) for the clinically independent data. The differences between the detected closure frame and the closure frame marked by the judges are highlighted in (b) for the 10 folds within the original dataset and in (d) for the clinically independent data. The Positive values indicate that the actual UES opening and closure preceded the predicted UES opening and closure.

The system also presented similar results when tested using the swallows from the independent clinical experiment as in Table 13. for the 50 swallows, the system achieved an average per swallow accuracy of 0.8880, an average per swallow sensitivity of 0.8559, and an average per swallow specificity of 0.9356. Fig. 30 (c-d) show histograms for the difference between the automatic detection and the reference manual labeling of the DUESO in terms of opening and closure frames. The results showed that UES opening and closure were detected within a 100 msec tolerance in around 84% and 66% of the swallows in the independent test set respectively.

6.4 Discussion

The main purpose of this study was to test the feasibility of HRCA in detecting the exact timing of UES opening and closure during swallowing using non-invasive neck-attached sensors independent of VFSS images and to compare the accuracy to human ratings of the DUESO. We have established the fact that UES opening can be best visualized using VF which is clinically impractical due to the delivered radiation doses and unavailability outside clinical care settings. We have also demonstrated the critical rule that UES plays during swallowing and how monitoring its opening and closure will help identify the risks leading to unsafe swallowing. As a necessary part of the optimal goal to create a non-invasive swallowing monitoring system, UES opening/closure detection should help patients with brainstem parts, responsible for swallowing regulation, damaged and/or surgically removed to rehabilitate and relearn how to swallow. These patients will have a consistent feedback to tell if they are correctly performing swallowing compensation maneuvers in which they are taught to improve the hyolaryngeal excursion which would in turn reflect on UES duration/diameter and airway protection in order to maintain a safe function.

Prior studies have only addressed indicators and changes in HRCA signal features at the UES opening and closure moments or during the passage of the bolus through UES, but non of them offered a direct way to detect the DUESO during swallowing. Some of these studies reported the presence of localized maxima of some HRCA signal features at UES opening

and closure times [90, 360]. One study also observed changes in the acoustic component of HRCA signals while the bolus passed through the UES [361]. Although these studies were essential for establishing the association between UES opening and HRCA signals, they were just descriptive analyses about the patterns in signal features at certain points of time when physiological events occurred. Therefore, in this study we aimed to explore a more advanced predictive profile to detect the DUESO from HRCA signal through considering the time dependency along the swallowing segment. As such we have demonstrated the system's feasibility on detecting DUESO without VFSS image verification.

One major disadvantage of human ratings is the subjectivity which creates an inter-rater tolerance of 82 msec ($\approx \pm 2.48$ frames at 30 FPS) as reported for measuring swallowing kinematic events [344]. Human ratings of swallow kinematic events can also drift over time and necessitates that raters maintain ongoing intra and inter-reliability over time to maintain an appropriate error tolerance. Having an automated system that is capable of rating the swallowing kinematic events with a comparable human rater accuracy and impregnable to changes over time, is advantageous for swallowing analysis when imaging technology is unavailable, not feasible, or otherwise impractical for evaluating swallowing physiology. Based on the results, we can clearly see that the proposed system accurately detected up to 93.6% of the actual DUESO with low rates of false positives and negatives occurring only at the borders of DUESO as shown in Fig. 29 (b). These results were also achieved regardless of gender, age, or diagnosis of the subjects which assures the wide applicability of the system.

The system also showed robust performance when applied to a completely independent set of swallows that were collected from a different group of participants with different conditions and never seen in the training dataset. In terms of global measurements, the system achieved a close testing accuracy compared to the validation done through the folds of the original dataset (0.888 vs. 0.9035) and the same for sensitivity and specificity. It didn't come short either on the side of temporal properties of the DUESO, where it captured the UES opening and closure within a 100 msec tolerance in most of the swallows in the independent test set. This confirms that the high quality of DUESO detection can be carried over to completely unseen data and assures a high degree of generalization in the proposed system.

It is important to bear in mind that the accuracy of any physiological event detector cannot be judged only through comparison with human ratings which are subject to error too. The sub-events occurring during or after the detected event and their importance to the whole physiological process, control the limits to which the system can be considered accurate because one doesn't want to detect an event with 50 msec accuracy to look for another sub-event that happens within 10 msec of the original event. Previous studies have shown that the important UES events happen slightly after the initial UES opening [79]. For example, in general, entry of the bolus head into the sphincter defines UES opening; however, in 20% of swallows, air precedes entry of the bolus by 30-60 msec [79]. Maximal values of A-P UES diameter were found also to be reached after 70-170 msec of UES opening, depending on the bolus size and other factors [79]. So, it could be argued that a delayed detection of UES opening is not completely inaccurate if it happens within 100 msec ($\approx$ 3 frames at 30 FPS) after the actual opening. Conversely, anatomic abnormalities leading to reduced DUESO (e.g. cricopharyngeal bar, Zenker diverticulum, hypopharyngeal lesions) would be completely undetectable without imaging leading to the need for further research to determine if HRCA can classify patterns of DUESO that indicate the need for imaging to rule out an anatomic diagnosis reducing DUESO.

In Summary, this study along with others, demonstrates advancements in HRCA signal processing and provides substantial evidence that HRCA signals predominantly reflect the patterns in DUESO and combined with our overall growing research portfolio, swallowing physiological activity. These advancements show the capability of HRCA to provide insight into diagnostic physiological aspects of swallow function and push towards the development of more accessible tools for dysphagia screening within clinical settings. Future research directions for this study include enhancing the detection quality of DUESO while reducing the error between the predicted and actual DUESO and investigating whether characteristic differences in HRCA signal signatures may reflect underlying anatomic or other etiologic explanations warranting investigation with imaging. This point is crucial in that some causes of dysphagia are indeed anatomically based, however in situations in which such diagnoses are suspected and imaging is not available immediately, HRCA certainly shows promise toward providing interim information that can guide management.

6.5 Conclusion

In this study, we proposed an ambitious deep architecture for the temporal identification of the DUESO during swallows by using HRCA signals. Swallows from 116 patients were collected under a standard clinical procedure for different swallowing tasks and materials. 3D acceleration signals of full length swallows, were denoised and fed into a network composed of a two-layer CNN, a 3-layer GRU-based RNN, and 3 fully connected layers to generate the temporal mask marking the time of UES opening and closure during swallows. The proposed system yielded an average accuracy of more than 90% of the swallow width and more than 91% of the DUESO width (sensitivity) with a low false positive rate. Moreover, the system showed nearly identical performance when used on an independent testing set from an ongoing clinical trial. Our results have provided substantial evidence that HRCA signals combined with a deep network architecture can be used to demarcate important physiological events that occur during swallowing.

7.0 Upper Esophageal Sphincter Opening Maximal Distension Detection and Localization

Part of this chapter (the UES distension measurement protocol) has been previously published in and reprinted with permission from [362]. © 2021 IOP Publishing. K. Shu, J. L. Coyle, S. Perera, Y. Khalifa, A. Sabry, and E. Sejdić, "Anterior-posterior distension of maximal upper esophageal sphincter opening is correlated with high-resolution cervical auscultation signal features," *Physiological Measurements*, February 2021.
DOI: 10.1088/1361-6579/abe7cb

7.1 Objective

In this study, we investigate the possibility of using HRCA signals to non-invasively measure the upper esophageal sphincter (UES) opening maximal distension during swallowing. The multi-channel HRCA signals are fed into a hybrid convolutional recurrent neural network that employs attention to focus on the part of the signals where the UES is actually open. This algorithm along with the UES opening detection algorithm can give a complete picture about the efficiency and duration of the UES opening during swallowing which can be extremely useful to clinicians to determine a lot of factors regarding the swallowing condition such as the possibility of residue formation and/or penetration/aspiration possibility.

7.2 Methods

7.2.1 Study Design and Clinical Protocol

This study was approved by the institutional review board of the University of Pittsburgh. All participating subjects provided informed written consents prior enrollment including consent to publish. We collected data from one hundred and thirty three patients (93 males, 40 females, age: 64.3 ± 13.2) with a variety diagnoses. The patients enrolled in this study while undergoing an oropharyngeal swallowing function evaluation using VFSS at the University of Pittsburgh Medical Center Presbyterian Hospital (Pittsburgh, PA, USA). Of the patients enrolled in this study, 37 had stroke while the other 96 patients were admitted due to other medical conditions unrelated to stroke such as neurodegenerative diseases and lung transplant.

This study was conducted as a part of a standard clinical procedure rather than a dedicated research controlled protocol. As a result, the swallowing assessment protocol was altered according to the patient's status and condition and this included the size of the boluses, their consistencies, the way they were administered to the patient and the head position. The administered boluses included the following consistencies: 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). The boluses were either administered by the speech language pathologist conducting the experiment or self-administered by the patient. Four hundred and thirty four swallows (203 from stroke-diagnosed patients and 230 from patients with other non-stroke conditions) were collected and analyzed in this study.

7.2.2 Data Acquisition

The experimental setup of this study is similar to the setup described elsewhere [107]. Subjects were comfortably seated and VFSS was conducted in the lateral plane using a Precision 500D system (GE Healthcare, LLC, Waukesha, WI) at a pulse rate of 30 pulses per second [341]. The VFSS feed from the x-ray machine was connected to the data acquisition

workstation through an AccuStream Express HD video card (Foresight Imaging, Chelmsford, MA) that digitized the video feed at a resolution of 720 × 1080 and a sampling rate of 60 frame per second (FPS). Swallowing vibrations were collected simultaneously with VFSS through tri-axial accelerometer (ADXL 327, Analog Devices, Norwood, Massachusetts) that was attached to the skin overlying the cricoid cartilage using an adhesive tape [106]. The accelerometer's axes were aligned so that they pick vibrations in the anterior-posterior (A-P), superior-inferior (S-I), and medial-lateral (M-L) directions. The signals were fed into the same acquisition workstation as the VFSS feed through a 6120 DAQ (National Instruments, Austin, Texas) and digitized in a rate of 20 kHz. The collection og both streams from the VFSS and the accelerometer was synchronized using LabView (National Instruments, Austin, Texas). The collected accelerometer signals were later downsampled to 4 kHz to smooth out transient noise an measurement errors [107].

7.2.3 VFSS Image Analysis and UES Distension Expert Measurement

VFSS videos were segmented into individual swallow segments by tracking the bolus to determine the onset and offset of pharyngeal swallowing. The onset of the swallow was defined as the frame in which the bolus head passed the ramus of the mandible, and the offset of the swallow was defined as the frame in which the hyoid bone returned to its lowest rest position after clearance of the bolus tail through the UES [106]. The UES opening and closure were also determined by expert judges trained to perform swallow kinematics measurements in the individual segmented videos of each swallow. All judges who performed swallow segmentation and UES opening rating in VFSS established a priori intra- and interrater reliability with ICC's over 0.99. They also maintained similar reliability ICC's throughout measurements on 10% of swallows plus they were blinded to all swallow information and subject's diagnosis to avoid bias.

To measure the maximal UES opening A-P distension in a swallow, the frame in which the maximal displacement of the hyoid bone in the pharyngeal phase of swallowing ,was selected. The UES maximal distension usually happens at, shortly before or shortly after the frame of the maximal hyoid bone displacement, so we measured the UES distension

at the frame of the maximal hyoid bone displacement, 2-3 frames before and 2-3 frames after (5-7 frames in total) and the maximal A-P distension is calculated with the measured frames [79, 362, 363]. The maximal UES opening A-P distension was measured in each of the selected frames using a protocol and a software which were both developed in our lab [362]. The protocol was as follows:

1. In order to standardize judgments regarding the location of the superior and inferior limits of the height of the UES, we used the height of the third cervical vertebral body. The region of the proximal esophagus considered the UES has been quantified in manometric studies as coursing 1:3 cm inferiorly from the base of the plane of the true vocal folds (Cook et al. 1989). The height of the third cervical vertebra ranges from $1:11-1:14$ cm in adult females and $1:24-1:37$ cm in adult males based on midsagittal x-ray measurements [364]. Therefore for each selected frame, a yellow line was drawn from the anterior superior edge to the anterior inferior edge of third cervical vertebral body (C3) in Fig. 31a.
2. Next, another red line was drawn between the anterior inferior edge of the second cervical vertebrae (C2) and the anterior inferior edge of the fourth cervical vertebrae (C4) to provide a vertical axis (C2-C4) that enables the algorithms to subtract larger scale head/neck movements from the measurements [365](figure 31b). The length of the C2-C4 segment was also used as an anatomical scalar representing each subject's height.
3. The yellow line drawn in step 1 was dragged and anchored to the superior border of the posterior tracheal air column (indicating the range of UES height to limit judgments of UES opening) as shown in figure 31c.
4. A long blue line that is perpendicular to the C2-C4 segment was drawn and used as a referent axis to ensure alignment of the vertical and horizontal axes of measurement to participant position rather than to an arbitrary x-y coordinate system based on strict vertical and horizontal geometric axes of zero and 90 degrees. This line was then dragged superiorly and inferiorly between two ends of the dragged C3 segment to the location of maximal anterior-posterior distance of the UES opening (figure 31d).

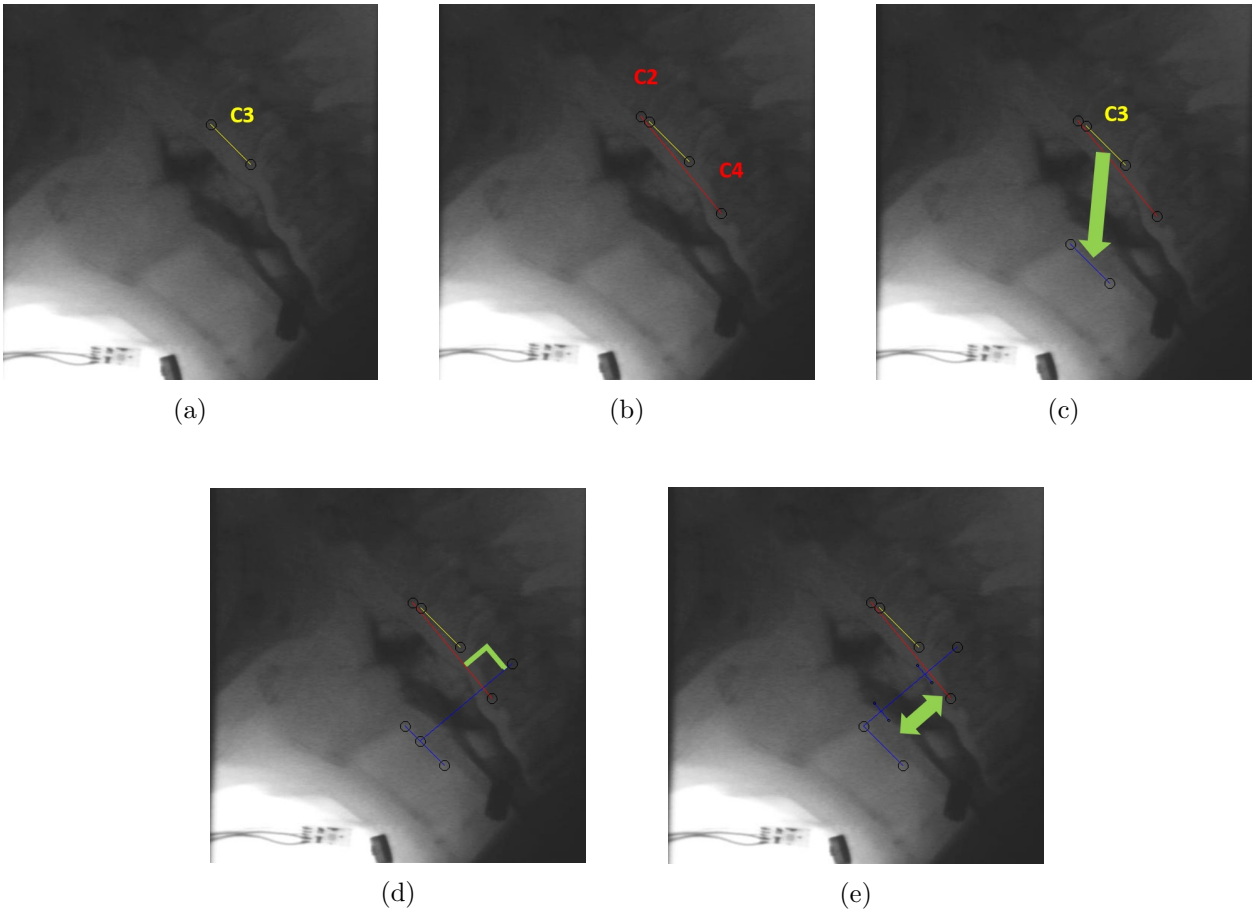
5. Then, the anterior and posterior points of UES opening on the perpendicular line were marked by short blue line segments respectively on the line segment drawn in step 6 (represented by two short parallel blue lines in (figure 31e).
6. The coordinates of the measured length of maximal UES opening (UES opening anterior end X and Y, UES opening posterior end X and Y) were returned in the output of the application.

The measured UES opening maximal A-P distension value was then divided by the length of the C2C4 segment in order standardize and compensate for the height of each patient. It was reasonable to assume that the distension value would be different among patients due to the differences in physical body characteristics (i.e. height and weight). The C2C4 segment length represents a part of the vertebral column which changes with the patient's height, so we used this as a standardization procedure for the UES opening maximal A-P distension value as followed in multiple studies [86, 366].

7.2.4 Signal Preprocessing

The signal preprocessing techniques performed in this study for the HRCA vibratory signals, followed the same techniques performed in a previous study that investigated using HRCA for automatic detection the UES opening and closure times [107]. The first step as mentioned previously, is to downsample the HRCA signals from the original sampling frequency 20kHz to 4kHz which helps minimize the transient noise such as sudden head movement other errors occuring during the measurement process.

Baseline noise was removed from the HRCA signals through recording the zero-input response of the sensors so that it can be used for building a finite impulse response filter (FIR) using the modified covariance auto-regressive modeling (of 10^{th} order) [353, 354]. The FIR filter was then used for device noise removal from each of the signals components. Afterwards, the motion artifacts and low-frequency noise were removed from the HRCA vibratory signals using fourth-order least-square splines [355, 356]. Finally, HRCA signals were denoised from the broadband noise using wavelet denoising [357]. Tenth-order Meyer wavelets with soft thresholding were performed for the denoising process.



(a) (b) (c)

(d) (e)

Figure 31: Illustration of steps for measuring the AP distension of maximal UES opening using the newly developed UES AP distension drawing application: (a) The length of anterior edge of C3 is indicated by the yellow line segment; (b) The anterior inferior edge of C2 and the anterior inferior edge of C4 were connected by the red line segment; (c) The C3 length was dragged, following the green arrow, to the position of blue line segment with the upper ends anchored to the superior border of tracheal air column; (d) The longer blue line segment perpendicular to the C2-C4 axis was positioned with its left ends sliding on the dragged C3 segment; (e) When the reference line (longer blue line segment) was adjusted to across the largest width of UES opening, two short blue line segments were placed on the extremities of UES. The length of UES opening is measured between the two short segments represented by the bidirectional green arrow.

7.2.5 Design of The Deep Prediction Model

The design of the network implemented in this study, was fine-tuned based on an experimental approach and following the best practices that achieved high performance in similar

problems [107, 300, 301]. We used a smilar network design as the network used for the detection of UES opening duration in HRCA signals through adopting a hybrid convolutional recurrent neural network that work on the raw HRCA vibrational signals directly [107]. In this study, we added some changes to the network originally implemented in [107] so that we take into account the prior knowledge about the UES opening duration in which the HRCA signals are strongly correlated to the values of the UES opening maximal A-P distension rather than the full duration of the entire swallow [362]. Therefore we added attention mechanisms that was built and trained using a a zero/one mask that resembles the UES opening duration as labeled by expert judges as shown in the lower middle part of Fig. 32.

The general network architecture was comprised of a 1D convolutional neural network that included two convolutional layers with a max pooling layer in between. Both convolutional layers were followed by a rectified linear unit (ReLU). The first convolutional layer applied 16 "1×5" filters per channel. The max pooling layer consisted of a window of size 2 with 2 strides. The last convolutional layer was identical to the first layer except except for using only one filter per channel. The longest swallow segment in the collected data lasted around 1500 msec (90 frames of VFSS @60FPS), so the signals of each swallow were divided into smaller chunks 16.67 msec in length ($\equiv$ 1 frame in VFSS or 66 samples in signals).Each chunk from the signals consisted of 3 channels of HRCA acceleration signals which made the dimensions 66 samples $\times$ 3 channels.

The attention mechanism is composed of two separate identical networks as shown in the center of Fig. 32. Each of the networks was composed of two layers, the first had a size of 2048 units and the second contained a number of units that matched the output of the layer to which the output attention mask was to be applied. The layer that generated a mask for the CNN output sequence included 90×1296 units, and the laer that generated a mask for the RNN output sequence included 90×64 units. The attention-highlighted output of the CNN, $x_{1:T}$, was fed into the RNN which was composed of 90 GRUs each of 64 units. The output sequence from the RNN was highlighted using the attention mask and fed into the next part that included the fully connected network (the middle right part of Fig. 32). The attention-highlighted output sequence of the RNN ($y_{1:T}$) was fed into 4 fully connected layers in order to fuse the temporal features from RNN into the UES opening

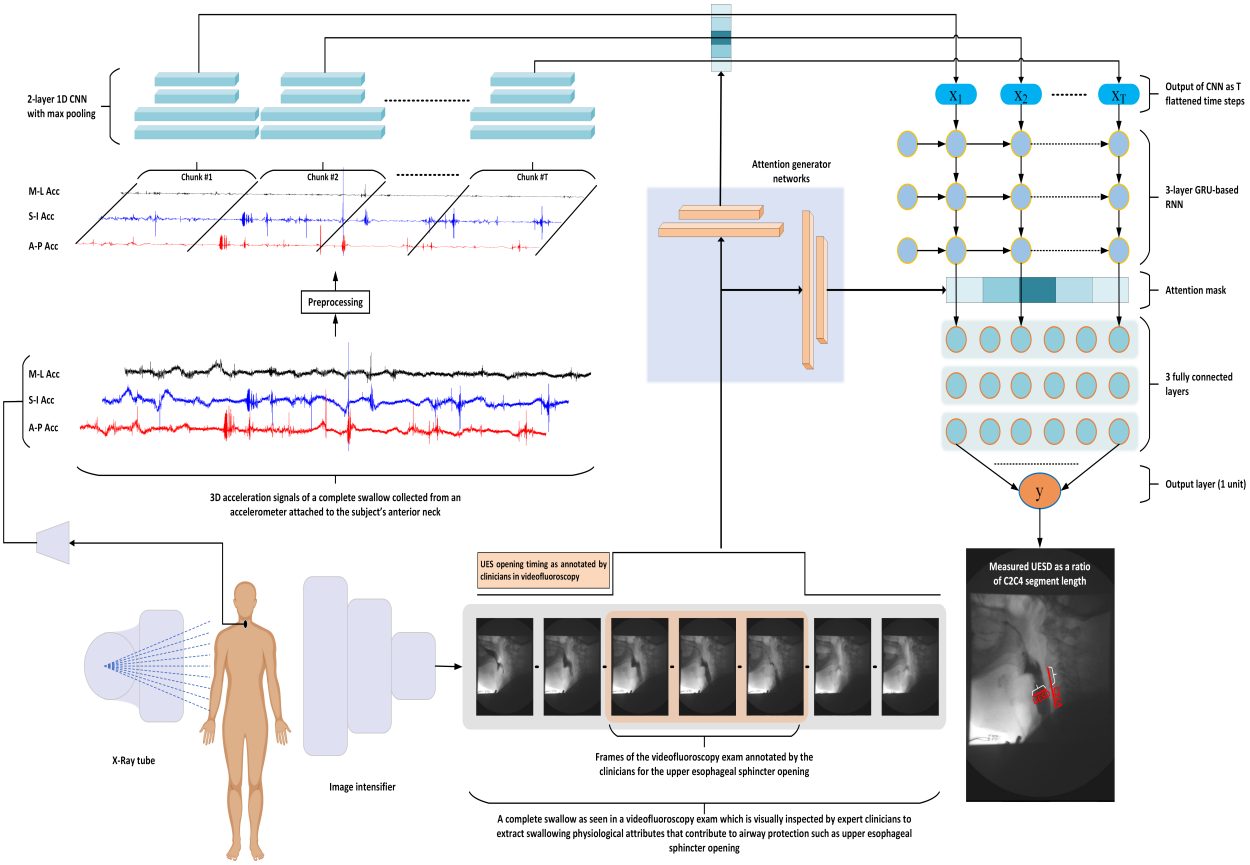
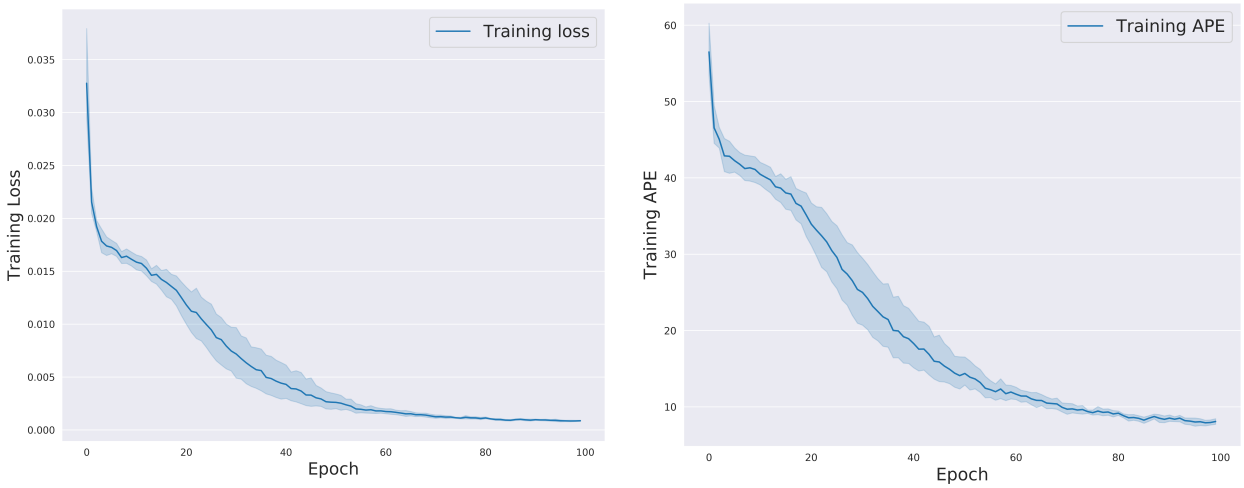


Figure 32: The architecture and data flow in the UES opening maximal distension prediction system. The lower left corner shows part of the experimental setup of the study where HRCA signals and VFSS are collected simultaneously. The 3-channel HRCA acceleration signals from each swallow are denoised and split into equal chunks each of 66 samples (equivalent to 1 VF frame). The architecture of the used 1D CNN which is comprised of two layers, the first applies 16 filters on each channel and produces 48 channels. In the middle part, the attention generator networks are shown. The attention networks (two fully connected layers) take the UES opening mask as input so that it generates the attention masks for the CNN output and the RNN output. $x_{1:T}$ is the output train from the CNN for chunks $(1:T)$ after being masked by the generated attention and fed into the RNN units. Each unit in the RNN was built based on the gated recurrent unit design (GRU). The architecture of the 3-layer RNN used for time sequence modeling is shown in the upper right corner. The output sequence from the last layer of the RNN ($\hat{y}_{1:T}$) is flattened and masked by the attention and fed into the first fully connected layer. (h) A diagram of the 3 fully connected layers (each of 128 units) used to combine the features coming out of the RNN. (i) The output layer is composed of 1 unit (y) that resembles the UES opening maximal A-P distension prediction as a ratio of the C2C4 segment length.

maximal A-P distension prediction. The first 3 layers were ReLU activated with 128 units and the output layer resembled only one unit with Sigmoid activation that generated the distension prediction value. Each two fully connected layers were separated by a dropout layer with a drop rate of 20%.

In this study, we employed the final cost function as the mean squared error between the ground truth values of the UES opening maximal A-P distension ratio to the C2C4 segment length and the predictions generated by the aforementioned network. We used the Adam optimizer to train the network due to its superiority in convergence without fine tuning for hyper-parameters [345].



(a) (b)

Figure 33: This figure includes the plots that show the progress of the MSE loss function and the APE over the epochs of training the proposed UES opening distension prediction network. (a) represents the MSE loss function over the 100 training epochs across the 10 folds. (b) represents the APE over the 100 training epochs across the 10 folds.

7.2.6 Evaluation

The swallows were randomly divided into 10 equal subsets and a holdout method was used 10 times to train the network with 9 subsets and test with the remaining subset (also known as 10-fold cross validation). The output from the proposed system is in the form of

a ratio that represents the normalized UES opening A-P maximal distension with respect to the C2C4 segment length. This ratio wasn't reported to be more than one in the same cohort before [362]. The predicted C2C4-normalized UES opening A-P maximal distension was compared to the ground truth using the absolute percentage error (APE) which is defined as follows:

$$APE = \frac{|Prediction - Groung\ Truth| \times 100}{Groung\ Truth}$$

7.3 Results

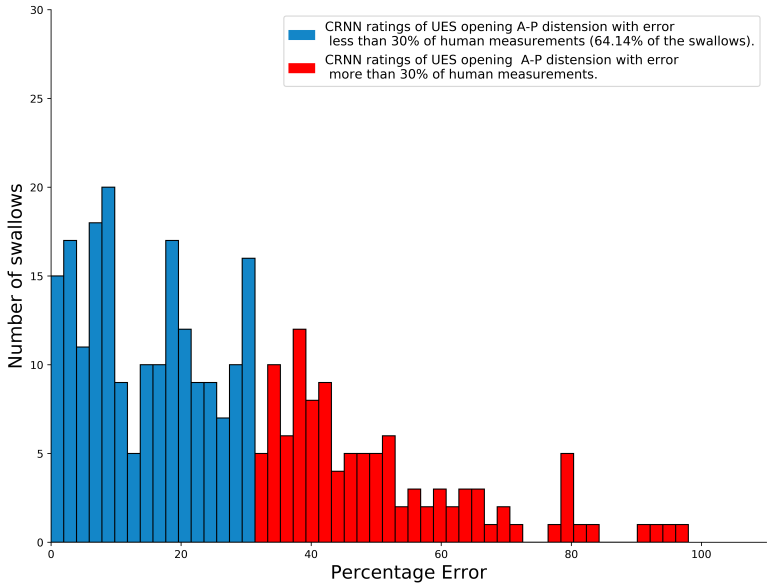


Figure 34: The APE fo each swallow in the dataset when used within the validation samples. The blue bars represent swallows that got the UES opening distension predicted with an APE of 30% or less when compared to the ground truth labeled by human experts. The swallows with predictions of APE 30% or less were around 64.14% of th entire dataset. The red bars represent swallows with an APE more than 30%.

A series of chunks of denoised multi-channel HRCA signals (each of size: 3× 66) that represent a complete swallow, were fed into the convolutional neural network as shown in Fig. 32. Simultaneously, a zeros/ones mask that represents the UES opening duration, was

fed into the fully connected network of the attention generation in order to push the network to focus on certain features that exist with the UES opening duration which were proven to be most associated with the UES maximal distension when compared to the features calculated from the entire swallow. Attention is applied in two levels, the first happens after the last layer of CNN and the second happens after the last layer of the RNN. The attention-highlighted output was then fed into a fully connected network that translated the temporally attention-highlighted features into a normalized UES opening maximal A-P distension prediction. The network was trained over 100 epochs and the evolution of both the loss function (MSE) and the absolute percentage error (APE) during training over the 100 epochs is shown in Fig. 33. As we can see in the graphs for the MSE and APE, the network seems to be training well and learning the patterns that reside within the used dataset. This can be verified by the achieved APE over the validation sets, where the network produced the normalized UES distension predictions with a mean APE of 27.24 ± 21.1.

Fig. 34 shows the performance of the proposed UES distension prediction network over individual swallows in the dataset when included in the validation set as a testing sample. The results show that the prediction network predicted the C2C4 normalized UES opening maximal A-P distension with an absolute error of 30% or less for around 64.14% of the swallows in the dataset and with an absolute error of 50% or less for around 86.84% of the swallows in the dataset. Fig. 35 shows a sample swallow that was presented to our proposed system for UES distension prediction. We can see how a prediction with 22% error (reduction) look like when compared to the ground truth measured distension. The ground truth for this swallow measured approximately 0.45 of the C2C4 segment length and the predicted segment measured approximately 0.35 of the C2C4 segment length.

Previous results that included labeling similar anatomical landmarks for the swallowing process in x-ray images of VFSS suffered from variability between human raters in rating/labeling the landmark for the same swallow. For instance, a previous study that investigated the hyoid bone displacement measured such variability between human raters [85]. In that study, the hyoid bone was marked in a similar way as we did with UES A-P distension, through two anchors placed at the anterior-inferior and posterior-superior corners of the hyoid which represent a line but for the purpose of that study, the two anchor points

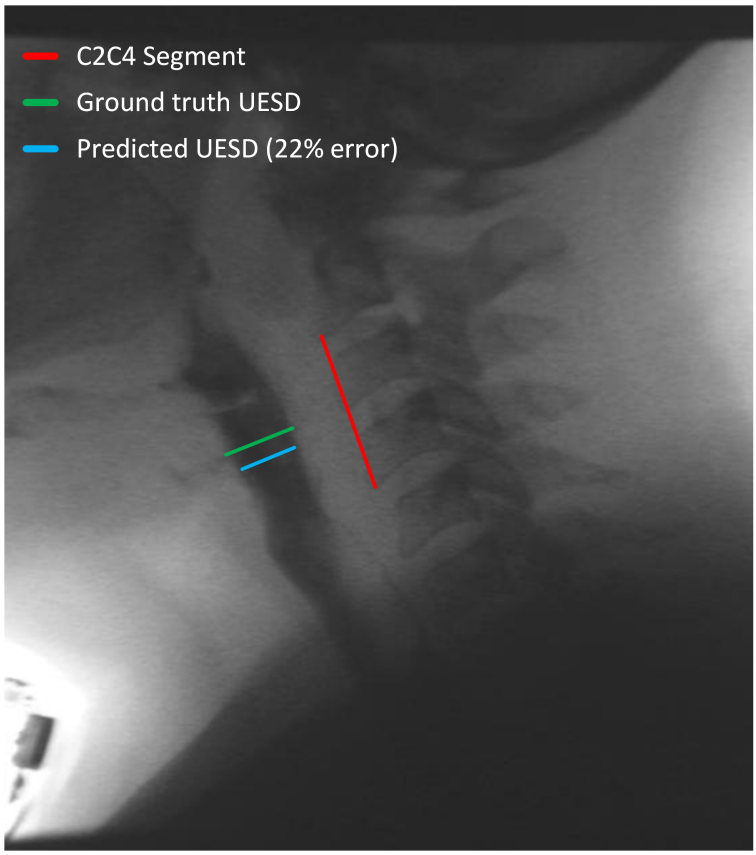


Figure 35: A sample prediction of the C2C4 normalized UES opening maximal A-P distension for one of the swallows by the proposed system. The ground truth for this swallow is shown as the green line and it measured 0.45 of the C2C4 length. The light blue segment represent the predicted distension by the network which measured 0.35 of the C2C4 length. The absolute error between the ground truth and the predicted segments is 22% of the ground truth value.

were used to construct a bounding boc that surround the body of the hyoid. The same set of swallows was presented to a group of human raters and the agreement between raters for the same swallows was measured. The overlap between the bounding boxes marked by the different raters for the same swallows never exceeded 79.09% of the hyoid bone body [85].

7.4 Discussion

The primary goal of this study was determine the feasibility of using HRCA vibratory signals as an input for a deep learning architecture to non-invasively predict the upper esophageal sphincter opening maximal anterior-posterior distension as one of the important kinematic events directly tied with healthy swallowing. We presented a a hybrid deep neural network model that used convolutional neural networks, recurrent neural networks plus attention mechanisms to extract the local features from the raw HRCA vibratory signals and temporally correlate and adjust such features to accurately predict the value of the UES maximal A-P distension. The results show that HRCA combined with deep learning can fairly accurately predict the C2C4 normalized UES opening maximal A-P distension when compared to the ground truth distension labeled by expert human judges.

Inspiration for the deep learning architecture employed in this study was taken from multiple previous studies that investigated the correlation between HRCA signals and the opening mechanisms of the UES in addition to the value of the UES maximal A-P distension itself [107, 281, 362]. In summary, these studies presented multiple outcomes that helped design the used architecture in this study. The first significant outcome was that HRCA signals are highly correlated with the UES opening duration and can be actually used with deep learning to predict the exact timing when the UES opens and closes during swallowing [107, 281]. The deep learning model presented by these studies was the basis for our work in this one with the UES distension because of the strong performance achieved for the UES opening duration. The second outcome from the literature based on which we fine tuned our model, was that the correlation between the HRCA signals features and the UES maximal A-P distension is the strongest within the duration of UES opening but not the entire swallow which guided us to use attention mechanisms to focus on just the important features [362].

Based on our results, the proposed network predicts the C2C4 normalized UES distension with a error percentage of 30% or less for more than half of the swallows in the dataset (64.14%) and less than 50% for 86.84% of the swallows in the dataset. As mentioned before, the error rates achieved in this study are comparable to the common error rates between

humans for similar measurements such as the hyoid bone labeling [85]. This enhances the clinical significance of the results of our proposed prediction system that it performs well given that even trained human raters don't ever completely agree on the same measurement. This also would have probably affected the ground truth for our study because our data were labeled by multiple human raters which suggests subjectivity in judgments. A possible explanation for such disagreement in ratings between human judges is that the resolution and quality x-ray images in VFSS are not usually high and differ from a machine to another which makes it extremely difficult to reach a very high precision in visually determining the certain pixel that marks the boundary of the UES or any other anatomical structure such as hyoid bone. A few pixels change in each anchor point could lead to a larger change in the orientation and length of the measured segment. Therefore, given the variability and errors in human measurements, we think that the performance of our network can be considered acceptable; however, we also expect that the performance can be enhanced by using a larger dataset with more labeled swallows.

A possible other way that may contribute to enhancing the performance of such prediction task, is using multi-task learning to jointly train a prediction framework to predict both the UES opening and closure moments (thus the opening duration) and the maximal A-P distension simultaneously. This will be extremely helpful due to the construction of a shared model that uses shared representations to quickly learn the common features between the downstream prediction tasks. We believe that such a shared prediction model could reduce overfitting and increase data efficiency due to the common ideas between the two prediction tasks.

Clinically, there are numerous possibilities for the use of non-invasively estimating the UES distension which all lead to a efficient diagnosis and rehabilitation of swallowing problems. For instance, it can be used as biofeedback tool for swallowing rehabilitation where the pateints ware instructed to perform certain head positions or other maneuvers during swallowing that can lead to better prolonged UES opening and thus better swallowing and less probability of forming residue which can cause aspiration. Having a utility that can tell the patient whether the maneuver or the rehabilitation procedure they are performing, is working or not or if they are performing the maneuver in the right way, could be of a great

help to a wide group of patients suffering from swallowing problems. It will also reduce the cost of dysphagia management by limiting the need for advanced diagnostic imaging studies such as VFSS. Further, non-invasive estimation of UES distension can be used to determine the typical swallowing patterns that deviate from normal/healthy swallowing by defining the acceptable range for an important kinematic such as UES distension. Furthermore, it can be used to judge the deterioration of swallowing function in certain patient populations such as patients with neurodegenerative diseases.

7.5 Conclusion

In conclusion, this study aimed to propose a new method to non-invasively estimate the upper esophageal sphincter opening maximal A-P distension during swallowing from HRCA signals. On the way to that, we developed a measurement protocol for the UES maximal A-P distension in VFSS so that we can use it as the ground truth for our study in which VFSS was simultaneously collected with HRCA signals. We employed a hybrid deep neural network that uses CNNs, RNNs and attention mechanisms to perform the prediction from the raw HRCA signals. The results revealed that HRCA combined with deep learning models can provide a fairly accurate estimate of the UES maximal A-P distension during swallowing when compared to the ground truth distension measured by trained judges in VFSS. This, along with other studies investigating the correlations between HRCA signals and swallowing kinematics, provides an evidence that HRCA combined with advanced signals processing techniques has the power to provide non-invasive, time-efficient and low cost diagnostic value in dysphagia comparable to advanced diagnostic exams.

8.0 Non-Invasive Detection of Unsafe Airway Protection

8.1 Objective

In this study, we investigate the ability of creating a screening algorithm that uses HRCA signals to detect aspiration on three levels that differentiate between normal swallow, swallow with penetration, and swallow with aspiration. The three catgeories are determined based on the 8-point penetration-aspiration scale value given to the swallow by clinicians in VFSS. Such algorithm can expedite the swallow screening process and make it more objective and immune against human error. We employ fusion between the multi-channel HRCA signals, power spectral estimation, convolutional neural networks, and residual learning to develop a deep leatning network that performs the classification of each swallow as one of the three categories mentioned earlier.

8.2 Methods

8.2.1 Study Design and Clinical Protocol

This study was approved by the institutional review board at the University of Pittsburgh where all experiments were conducted. All subjects who participated in this study provided informed written consents ahead of joining. The study was conducted at the University of Pittsburgh Medical Center Presbyterian Hospital. To achieve accurate aspiration detection in HRCA signals, this study included the collection of simultaneous VFSS and HRCA signals for the algorithm development purpose and for having a ground truth to which we can refer. The experimental setup of this study has been described in detail elsewhere [107]. In this study, clinicians followed a standard swallowing clinical evaluation procedure rather than for research purposes and determined the consistencies of the administered materials and based the status of each subject [106]. The administered consistencies 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 in the lateral plane using a Precision 500D system (GE Healthcare, LLC, Waukesha, WI) at a pulse rate of 30 pulses per second (PPS) as recommended in the literature [341]. The video stream was sampled 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). HRCA signals were collected through a 3D accelerometer (ADXL 327, Analog Devices, Norwood, Massachusetts) mounted using a small curved plastic case. The accelerometer complex was attached to the skin overlying the cricoid cartilage with an adhesive tape. This specific location was recommended for vibration signals of high signal to noise ratio [342, 343]. The accelerometer was aligned so that it picks vibratory signals in the anterior-posterior (A-P), superior-inferior (S-I), and medial-lateral (M-L) directions. The signals from the accelerometer were digitized using a 6210 DAQ (National Instruments, Austin, Texas) at a sampling rate of 20 kHz and then down-sampled to 4 kHz to reduce transient noise and measurement errors [106]. The streams from the accelerometer and the x-ray machine were synced into the acquisition workstation through LabView (National Instruments, Austin, Texas).

8.2.2 VFSS Video Analysis

VFSS videos were inspected by expert personnel to extract the onset and offset of each swallow. The onset of the swallow was defined as the frame in which the bolus head passed the ramus of the mandible, and the offset of the swallow was defined as the frame in which the hyoid bone returned to its lowest rest position after clearance of the bolus tail through the UES [106]. HRCA signals were segmented based on these timings identified by the experts in videos due to complete synchronization between videos and signals. Trained, expert judges then determined the PAS for each swallow. All judges established and maintained inter- and intra-rater reliability of ICCs more than 0.9. The PA scores were divided into three categories: safe (1-2), penetration (3-5) and aspiration (6-8).

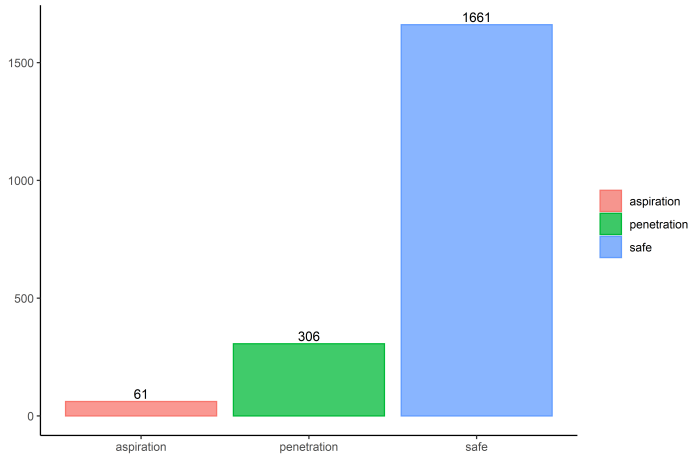


Figure 36: The distribution of the three PAS categories over the dataset. The number of the swallows in each category is represented by the height of the bars and is also written on the top.

8.2.3 Study Data Characteristics

In this study we collected 2028 swallows from 191 patients (116 males and 75 females) who were referred to VFSS for swallow function evaluation as part of their clinical care. The mean age of the patients was 63 (standard deviation, s.d.= 14.3). The patients were referred to swallow function evaluation due to suspected dysphagia secondary to multiple conditions that they originally had. These conditions included stroke, neurodegenerative diseases, lung transplant, lung lobectomy, heart disease, and head/neck surgeries. The distribution of the PAS ratings in the dataset can be seen in Fig. 36.

8.2.4 System Design

A deep neural network was implemented to process the spectrograms of the HRCA signals of each swallow and produce the PAS category to which the swallow belongs. The network design was based on ResNet architecture so that we can get the benefit of residual learning in order to increase the depth of the network and its capacity. The main idea behind residual learning is using skip or identity connection between layers to skip one or more intermediate layers thus adding more layers shouldn't degrade the network performance when compared to

its shallower counterparts [346]. The ResNet version that we used in this study was ResNet18 which includes 18 layers in addition to the residual connections and pooling layers. We just replaced the last two layers (the average pooling and the fully connected layer) with a 2D average pooling over the frequency and time domains followed by a convolutional layer that generates the classification output as seen in Fig. 37 C. We initialized the training of our classifier on a version of ResNet18 that was pretrained on ImageNet dataset [367].

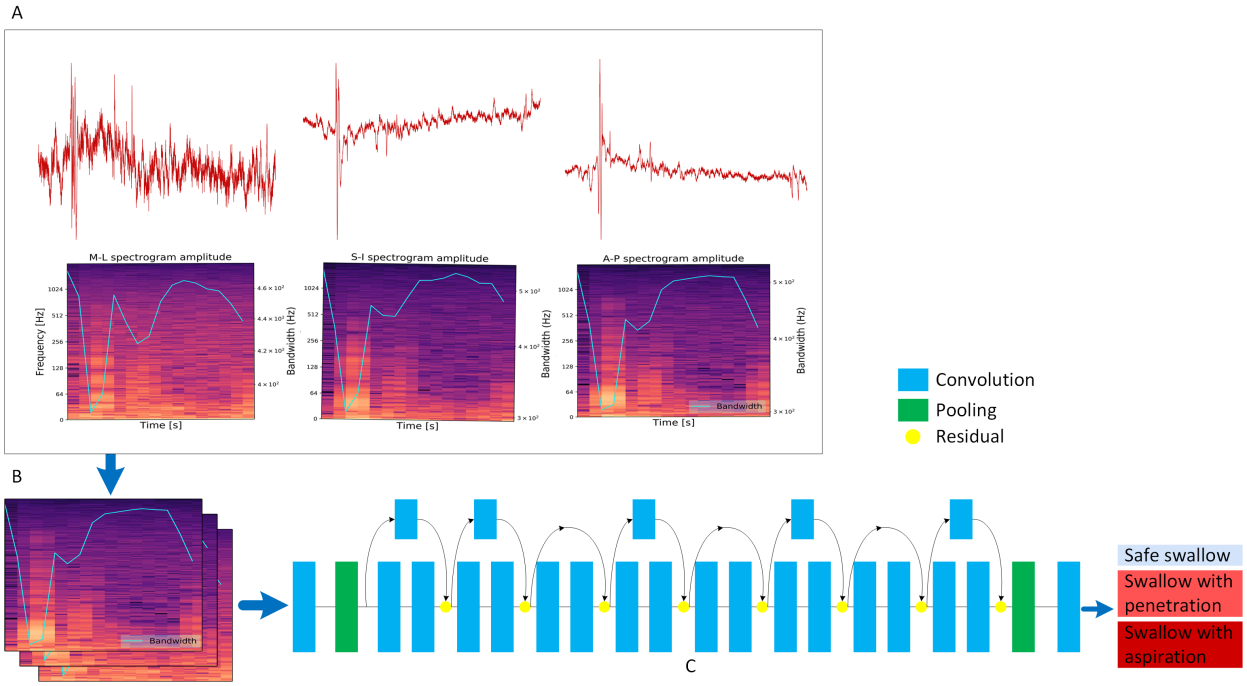


Figure 37: The architecture of the main proposed deep network. **A.** shows a typical unfolded example of the network input of acceleration signals of a swallow with PAS of 2. The first row represents raw acceleration signals, and the second row represent the spectrogram for each of the acceleration axes. **B.** represents the folded spectrogram from the three acceleration channels prepared as a 3-channel image to be fed into the convolutional residual network. **C.** represents the modified architecture of ResNet18 that was employed in this study.

The input of the network was only the 3-channel HRCA acceleration. The spectrogram was calculated from the raw signals using an M-point discrete Fourier transform ($M = 512$) using a 256-point Hanning window and overlap of 50%. We chose the spectrogram as input for our study here because it was proven effective in multiple occasions with swallow segment extraction [106]. The spectrogram of the three channels was then combined as a 3-channel image from which the MelSpectrogram was calculated and fed into the network.

8.2.5 Performance Evaluation

In this study, we are addressing a typical multi-class classification problem with three classes that represent the PAS categories. The generic classification accuracy was used as an indicator of the overall performance of the system. However, due to having three classes we calculated also the precision and recall for each of the three classes against the others to show the system sensitivity for each of the classes. At last, we used F1 score calculated for each of the the three classes against the others to show how weel the system performs given that the three classes are in imbalance regarding the number of instances in each class in the dataset. The mathematical representation of the used performance metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{8.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{8.3}$$

$$F_1\ score = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{8.4}$$

where TP represents the number of True Positives, TN represents the number of True Negatives, FP represents the number of False Positives, and FN represents the number of False Negatives. The dataset was split into 5 folds and a 5-fold cross validation scheme was applied. When splitting the data into 5 folds, we took into account that the PAS categories are not balanced and performed the process as a stratified 5 fold splitting through preserving the percentage of classes in each fold.

8.3 Results

To realize the differences between safe and unsafe swallows and what exactly happens during aspiration, Fig. 38 depicts how aspiration looks like in VFSS and the noticeable differences in HRCA signals that result from the abnormal vibration patterns originating

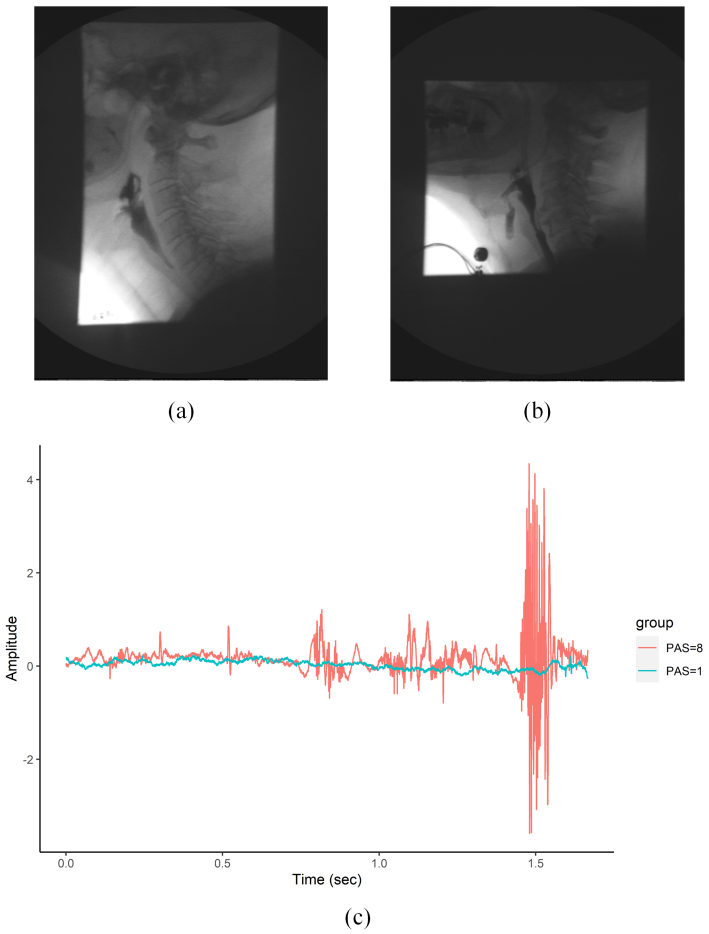


Figure 38: This figure shows two VFSS snapshots from two different swallows and the anterior-posterior acceleration signals for both swallows. (a) represents a snapshot from a safe swallow with PAS=1. (b) represents a snapshot from a swallow with silent aspiration and PAS=8. (c) shows the two acceleration signals corresponding to the previously mentioned swallows.

from aspiration. Fig. 38 (a) is a perfectly safe swallow with a PAS of 1 and Fig. 38 (b) is a swallow with aspiration that was rendered silent due to the absence of patient's attempts to eject the material out of the airway and thus there were no overt signs and the swallow was rated with a PAS of 8. Fig. 38 (c) shows the differences that occur in the raw acceleration signals picked from the anterior-posterior direction during the entire swallow for both of the swallows that are displayed in Fig. 38 (a) & (b).

In this study, we used a slightly modified ResNet18 model to classify HRCA signals from swallows into one of three categories of PAS as a way to develop a non-invasive and

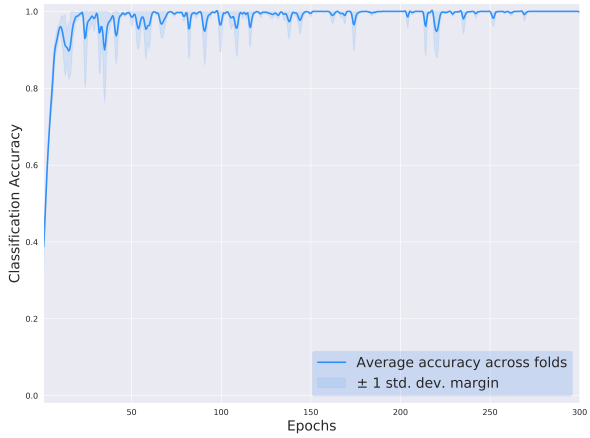


Figure 39: This figure shows the average classification accuracy across the 5 folds as it developed when testing after training the model each epoch.

reliable screening tool for swallowing. The training of the deep model relied on the frequency representation of the multi-channel HRCA signals of the full swallow as input and produced the category as one of the three categories: safe, penetration, or aspiration. The model was validated through a 5-fold cross validation scheme and the average accuracy across the 5-folds can be seen in Fig. 39 as it developed with the model training. We can see that the accuracy reaches 99% or almost 100%.

As mentioned earlier, due to the multi-class nature of our classification problem and to get more insight about the classification performance and sensitivity of the model for each of the three classes, we calculated the performance metrics including the precision, recall, and F_1 score for each of the three categories against the others as done in binary classification problems across the 5 folds of validation. This shows us how sensitive our model is when it comes to penetration and aspiration categories, especially that the number of swallows in these two categories is limited in the dataset used to evaluate the prediction model. Fig. 40 shows the precision, recall and F_1 score for each of the categories. We can see that the precision and recall are high (almost 100%) for the dominant category (safe) which is reasonable given the huge number of swallows in this category compared to the other two less dominant categories of PAS in the dataset. For the category of penetration, the precision and

recall values were around 90% and so is the F_1 score value. For the category of aspiration, the values of the three performance metrics were fluctuating around 61% with a maximum of 63%.

Although the algorithm's sensitivity to the aspiration category was around 60%, it exceeds the performance and sensitivity achieved on the same dataset in previous studies [87, 368]. The previous studies used ordinary machine learning techniques for classification such SVM, K-means, and even a feed forward neural network [368]. This shows that using a model of higher capacity can benefit the classification performance for such an imbalanced problem. Our method in this study also took into consideration the penetration category which was never considered before in any of the studies that investigated swallowing screening using vibratory signals. Our results also show that the sensitivity to the penetration category is high and reaches to 90% which a good indicator that such a category is separable and is worth considering as a separate category on the contrary to what was done before.

8.4 Discussion

In this study, we acquired HRCA signals simultaneously with VFSS from adults suspected with dysphagia in order to build a deep learning system that uses only the HRCA signals to differentiate between safe swallowing, swallows with penetration and swallows with aspiration. The used deep model used 3 PAS categories to indicate the status of airway way protection. The first category was PAS of 1-2 which indicates safe swallowing, the second was PAS of 3-6 which indicates penetration and the third category was PAS of 7-8 which indicates aspiration. The trained model achieved an average accuracy of 99% across a 5-fold cross validation and average precision values of 100%, 90% and 61% for each of the three categories of PAS, safe, penetration and aspiration respectively and close values for recall and F_1 score as shown in Fig. 40. The achieved results using the trained model in this study, are superior compared to those reported for different screening protocols where the patients are being observed for aspiration overt signs while they swallow and in which false positive rates can reach as high as 72% [78]. Another disadvantage for screening swallow

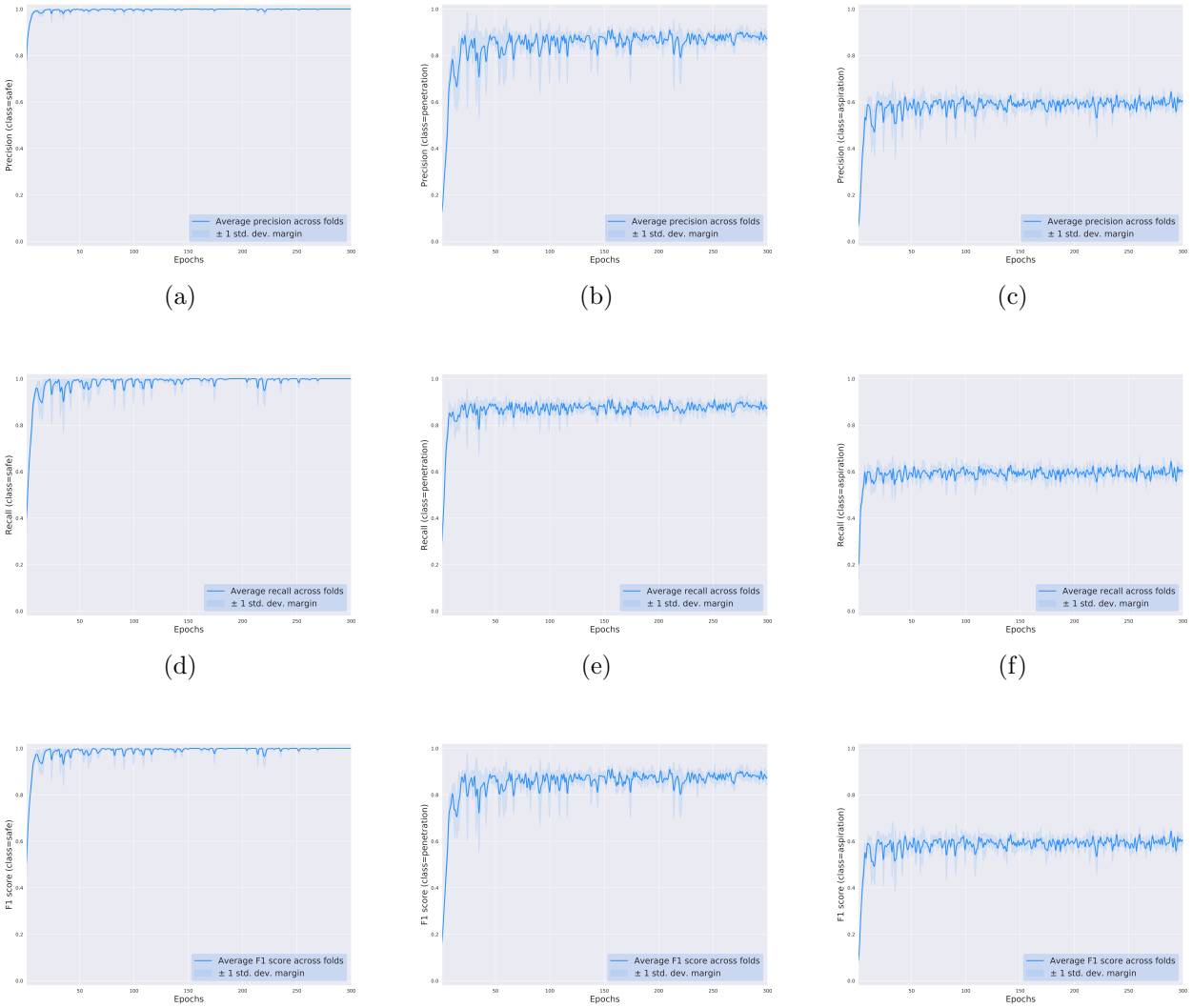


Figure 40: This figure includes the plots of the performance metrics for each one of the three PAS categories. The shown metrics are in order from the top row to the third, precision, recall and F_1 score respectively. The first column represents the plots for the safe category in PAS (the most dominant). The second column represents the plots for the penetration category of the PAS. The third column represents the plots for the aspiration category in the PAS. All of the plots are given as the average across the 5 folds of validation as evaluated after training the model each epoch.

screening protocols compared to the HRCA-based deep model, is that they look only for the overt signs of aspiration only like coughing which means that they can't detect penetration if present. Other factors that probably contributed to the higher performance of the deep model here in this study, include the fact that the swallowed materials were barium liquids or barium-mixed materials and not water as done in regular swallow screening and that the categories were based on accurate human labeling of penetration/aspiration in VFSS not just observing the swallowing pattern visually looking for aspiration overt signs.

Previous studies that investigated the use swallowing vibratory signals in prediction of aspiration, used only two categories for screening , safe and unsafe swallowing. A broad margin of cases with material penetration down till the vocal folds is hidden between safe and unsafe swallowing which might have contributed to the low sensitivity values [87, 368]. In addition, the previous studies trained only models of low capacity to perform the classification and used features that might not have been significantly different and/or inseparable between the safe and unsafe swallowing categories. In this study, we used a high capacity deep neural network which can automatically learn abstract feature representations based on the frequency domain characteristics of the multi-channel HRCA signals which have been proven extremely useful in swallowing segments extraction and denoising [106].

The importance of the results reported in this study, is that it proves that HRCA signals have the ability to identify unsafe swallowing to the sparsity of penetration and aspiration which opens many possibilities for using such algorithm in swallowing screening. It's important to place the results of this study in context compared to the screening protocols usually used in swallowing. Our study is based on barium-based swallows which have different viscosity compared to water which is used in the screening procedures which might have affected the results. This is one of the limitations of this study because if we want to deploy such system for swallowing screening, then it has to be running on water swallows not barium. So, this something that should be further investigated to test the ability of such deep models and the ability of HRCA signals to pick the different patterns between safe swallows and swallows with penetration or aspiration.

8.5 Conclusion

In conclusion, this work showed that deep learning-based architectures could be used to automatically identify the category of swallows in terms of being safe, with penetration or with aspiration using only HRCA signals as input. The combined use of CNNs and residual learning can achieve good accuracy due to the ability to increase the model capacity and depth. HRCA-based systems continue to show their ability to play a vital role in dysphagia and swallowing function assessment which can play a huge role in expediting dysphagia diagnosis and optimizing the use of clinical resources.

9.0 Conclusions and Future Work

In this research, we showed the potential of HRCA signals and deep learning to non-invasively provide diagnostic insights about swallowing physiological and pathological processes, an ability only provided through advanced diagnostic exams such as videofluoroscopy and fiberoptic endoscopic evaluation of the swallowing. This can play a vital role in expediting clinical decision making and dysphagia rehabilitation through creating widely accessible and cheap tools that provide the same diagnostic value as the currently utilized tools. In the following sections, we list the possible adaptations that we think are the optimal future directions of the work presented in this dissertation and were left out due to lack of time.

9.1 A Non-Invasive Swallowing Analysis Integrated Toolkit

The research in dissertation focused on individual tasks that can be taken as indicators about the swallowing conditions; however, the optimal use of such algorithms is to combine all of them into an integrated toolkit that can be used to give a complete picture of the swallowing condition of a subject. I would like to think that the next necessary milestone of this work is to combine the algorithms of swallow segment extraction, upper esophageal sphincter opening detection, laryngeal vestibule closure detection and airway protection detection into a single platform that takes real time HRCA signals and gives a complete report about the patient's swallowing given all the factors extracted by the aforementioned algorithms. Such a toolkit can be of a huge help to clinicians to optimize the clinical resources available and expedite the diagnosis of swallowing problems. In addition, this can be used as a biofeedback tool that gives a real-time evaluation about swallowing when the patients are actually swallowing so that they can perform the rehabilitation procedures in a better way at home independently. I can imagine how comfortable this would be to patients who are always nervous about eating because they suffer from swallowing disorders.

9.2 The potential Deployment of Multi-Task Learning Techniques

In this work, we focused on developing individual algorithms to solve each of the problems that we investigated. Although we achieved great results with single task algorithms; making use of multi-task learning should give us more opportunities to achieve even better results given that most of the learned tasks are strongly correlated as parts of the swallowing problem. Multi-task learning is currently in heavy use to solve multiple problems simultaneously such as anomaly detection and direction of arrival estimation in sound signals. I would like to think that such a step should be extremely beneficial to swallowing signal analysis.

9.3 The Potential Deployment of Unsupervised Learning Techniques

Most of the models investigated in this work, were trained in a supervised learning fashion in which we used labeled data as the ground truth for training. However, the labeling process is an extremely time wasting process and requires highly trained personnel to perform especially in the clinical data. In addition to that, the human error factor plays a huge role in biasing the results of such systems. I think that the next step for this research is to use unsupervised or partially unsupervised learning techniques such as self-supervised learning to build the models in order to reduce the amount of labeled data needed for supervised training. Self-supervised learning just requires strong augmentation techniques so that it can learn precise representations in an unsupervised fashion from unlabeled data. I can imagine that this can lead to saving huge amounts of the time wasted on data labeling and increase the productivity and efficiency of clinical research.

Bibliography

[1] A. J. Miller, "The neurobiology of swallowing and dysphagia," *Developmental Disabilities Research Reviews*, vol. 14, no. 2, pp. 77–86, Jul. 2008.

[2] N. Bhattacharyya, "The prevalence of dysphagia among adults in the United States," *Otolaryngology and Head and Neck Surgery*, vol. 151, no. 5, pp. 765–769, Nov. 2014.

[3] P. Clave, R. Terre, M. de Kraa, and M. Serra, "Approaching oropharyngeal dysphagia," *Revista Española de Enfermedades Digestivas*, vol. 96, no. 2, pp. 119–131, Feb. 2004.

[4] P. Clave and R. Shaker, "Dysphagia: Current reality and scope of the problem," *Nature Reviews: Gastroenterology and Hepatology*, vol. 12, no. 5, pp. 259–270, May 2015.

[5] D. M. Trate, H. P. Parkman, and R. S. Fisher, "Dysphagia: Evaluation, diagnosis, and treatment," *Primary Care: Clinics in Office Practice*, vol. 23, no. 3, pp. 417–432, Sep. 1996.

[6] H. Siebens, E. Trupe, A. Siebens, F. Cook, S. Anshen, R. Hanauer, and G. Oster, "Correlates and consequences of eating dependency in institutionalized elderly," *Journal of the American Geriatrics Society*, vol. 34, no. 3, pp. 192–198, Mar. 1986.

[7] B. Martin-Harris and B. Jones, "The videofluorographic swallowing study," *Physical Medicine and Rehabilitation Clinics of North America*, vol. 19, no. 4, pp. 769–785, Nov. 2008.

[8] R. Ishida, J. B. Palmer, and K. M. Hiiemae, "Hyoid motion during swallowing: Factors affecting forward and upward displacement," *Dysphagia*, vol. 17, no. 4, pp. 262–272, 2002.

[9] S. R. Barczi, P. A. Sullivan, and J. A. Robbins, "How should dysphagia care of older adults differ? Establishing optimal practice patterns," *Seminars in Speech and Language*, vol. 21, no. 4, pp. 347–364, Nov. 2000.

[10] J. A. Martin, B. E. Hamilton, P. D. Sutton, S. J. Ventura, F. Menacker, and M. L. Munson, "Births: Final data for 2003," *National Vital Statistics Reports*, vol. 54, no. 2, pp. 1–116, Sep. 2005.

[11] B. E. Hamilton, A. M. Minino, J. A. Martin, K. D. Kochanek, D. M. Strobino, and B. Guyer, "Annual summary of vital statistics: 2005," *Pediatrics*, vol. 119, no. 2, pp. 345–360, Feb. 2007.

[12] K. A. Burklow, A. N. Phelps, J. R. Schultz, K. McConnell, and C. Rudolph, "Classifying complex pediatric feeding disorders," *Journal of Pediatric Gastroenterology and Nutrition*, vol. 27, no. 2, pp. 143–147, Aug. 1998.

[13] N. Marlow, "Neurocognitive outcome after very preterm birth," *Archives of Disease in Childhood*, vol. 89, no. 3, pp. F224–F228, May 2004.

[14] L. A. Newman, C. Keckley, M. C. Petersen, and A. Hamner, "Swallowing function and medical diagnoses in infants suspected of Dysphagia," *Pediatrics*, vol. 108, no. 6, p. E106, Dec. 2001.

[15] P. Y. Ancel, F. Livinec, B. Larroque, S. Marret, C. Arnaud, V. Pierrat, M. Dehan, S. N'Guyen, B. Escande, A. Burguet, G. Thiriez, J. C. Picaud, M. Andre, G. Breart, M. Kaminski, and E. S. Group, "Cerebral palsy among very preterm children in relation to gestational age and neonatal ultrasound abnormalities: the EPIPAGE cohort study," *Pediatrics*, vol. 117, no. 3, pp. 828–835, Mar. 2006.

[16] M. Serra-Prat, G. Hinojosa, D. Lopez, M. Juan, E. Fabre, D. S. Voss, M. Calvo, V. Marta, L. Ribo, E. Palomera, V. Arreola, and P. Clave, "Prevalence of oropharyngeal dysphagia and impaired safety and efficacy of swallow in independently living older persons," *Journal of the American Geriatrics Society*, vol. 59, no. 1, pp. 186–187, Jan. 2011.

[17] A. Lee, Y. Y. Sitoh, P. K. Lieu, S. Y. Phua, and J. J. Chin, "Swallowing impairment and feeding dependency in the hospitalised elderly," *Annals of the Academy of Medicine, Singapore*, vol. 28, no. 3, pp. 371–376, May 1999.

[18] M. Cabre, M. Serra-Prat, L. Force, J. Almirall, E. Palomera, and P. Clave, "Oropharyngeal dysphagia is a risk factor for readmission for pneumonia in the very elderly persons: observational prospective study," *Journals of Gerontology. Series A: Biological Sciences and Medical Sciences*, vol. 69, no. 3, pp. 330–337, Mar. 2014.

[19] M. Cabre, M. Serra-Prat, E. Palomera, J. Almirall, R. Pallares, and P. Clave, "Prevalence and prognostic implications of dysphagia in elderly patients with pneumonia," *Age and Ageing*, vol. 39, no. 1, pp. 39–45, Jan. 2010.

[20] J. Almirall, M. Cabre, and P. Clave, "Complications of oropharyngeal dysphagia: aspiration pneumonia," *Nestle Nutrition Institute Workshop Series*, vol. 72, pp. 67–76, Oct. 2012.

[21] D. Nogueira and E. Reis, "Swallowing disorders in nursing home residents: How can the problem be explained?" *Clinical Interventions in Aging*, vol. 8, pp. 221–227, Feb. 2013.

[22] M. Rosemary, F. Norine, B. Sanjit, D. Nicholas, S. Mark, and T. Robert, "Dysphagia after stroke," *Stroke*, vol. 36, no. 12, pp. 2756–2763, Dec. 2005.

[23] J. G. Kalf, B. J. de Swart, B. R. Bloem, and M. Munneke, "Prevalence of oropharyngeal dysphagia in Parkinson's disease: a meta-analysis," *Parkinsonism and Related Disorders*, vol. 18, no. 4, pp. 311–315, May 2012.

[24] S. E. Langmore, R. K. Olney, C. Lomen-Hoerth, and B. L. Miller, "Dysphagia in patients with frontotemporal lobar dementia," *Archives of Neurology*, vol. 64, no. 1, pp. 58–62, Jan. 2007.

[25] J. Horner, M. J. Alberts, D. V. Dawson, and G. M. Cook, "Swallowing in Alzheimer's disease," *Alzheimer Disease and Associated Disorders*, vol. 8, no. 3, pp. 177–189, 1994.

[26] M. K. Suh, H. Kim, and D. L. Na, "Dysphagia in patients with dementia: Alzheimer versus vascular," *Alzheimer Disease and Associated Disorders*, vol. 23, no. 2, pp. 178–184, Apr. 2009.

[27] P. Calcagno, G. Ruoppolo, M. G. Grasso, M. De Vincentiis, and S. Paolucci, "Dysphagia in multiple sclerosis - prevalence and prognostic factors," *Acta Neurologica Scandinavica*, vol. 105, no. 1, pp. 40–3, Jan. 2002.

[28] A. Chen and C. G. Garrett, "Otolaryngologic presentations of amyotrophic lateralsclerosis," *Otolaryngology and Head and Neck Surgery*, vol. 132, no. 3, pp. 500–504, Mar. 2005.

[29] G. Ruoppolo, I. Schettino, V. Frasca, E. Giacomelli, L. Prosperini, C. Cambieri, R. Roma, A. Greco, P. Mancini, M. De Vincentiis, V. Silani, and M. Inghilleri, "Dysphagia in amyotrophic lateral sclerosis: prevalence and clinical findings," *Acta Neurologica Scandinavica*, vol. 128, no. 6, pp. 397–401, Dec. 2013.

[30] P. Garcia-Peris, L. Paron, C. Velasco, C. de la Cuerda, M. Camblor, I. Breton, H. Herencia, J. Verdaguer, C. Navarro, and P. Clave, "Long-term prevalence of oropharyngeal dysphagia in head and neck cancer patients: Impact on quality of life," *Clinical Nutrition*, vol. 26, no. 6, pp. 710–717, Dec. 2007.

[31] J. Galli, V. Valenza, L. D'Alatri, F. Reale, A. S. Gajate, S. Di Girolamo, and G. Paludetti, "Postoperative dysphagia versus neurogenic dysphagia: scintigraphic assessment," *Annals of Otology, Rhinology and Laryngology*, vol. 112, no. 1, pp. 20–8, Jan. 2003.

[32] P. D. Utsinger, D. Resnick, and R. Shapiro, "Diffuse skeletal abnormalities in Forestier disease," *Archives of Internal Medicine*, vol. 136, no. 7, pp. 763–768, Jul. 1976.

[33] D. Resnick, S. R. Shaul, and J. M. Robins, "Diffuse idiopathic skeletal hyperostosis (DISH): Forestier's disease with extraspinal manifestations," *Radiology*, vol. 115, no. 3, pp. 513–524, Jun. 1975.

[34] J. Almirall, L. Rofes, M. Serra-Prat, R. Icart, E. Palomera, V. Arreola, and P. Clave, "Oropharyngeal dysphagia is a risk factor for community-acquired pneumonia in the elderly," *European Respiratory Journal*, vol. 41, no. 4, pp. 923–928, Apr. 2013.

[35] M. Serra-Prat, M. Palomera, C. Gomez, D. Sar-Shalom, A. Saiz, J. G. Montoya, M. Navajas, E. Palomera, and P. Clave, "Oropharyngeal dysphagia as a risk factor for malnutrition and lower respiratory tract infection in independently living older persons: a population-based prospective study," *Age and Ageing*, vol. 41, no. 3, pp. 376–381, May 2012.

[36] J. A. Y. Cichero and K. W. Altman, "Definition, prevalence and burden of oropharyngeal dysphagia: A serious problem among older adults worldwide and the impact on prognosis and hospital resources," in *Stepping Stones to Living Well with Dysphagia*, vol. 72, May 2012, pp. 1–11.

[37] O. Ekberg, S. Hamdy, V. Woisard, A. Wuttge-Hannig, and P. Ortega, "Social and psychological burden of dysphagia: Its impact on diagnosis and treatment," *Dysphagia*, vol. 17, no. 2, pp. 139–146, 2002.

[38] S. Carrion, M. Cabre, R. Monteis, M. Roca, E. Palomera, M. Serra-Prat, L. Rofes, and P. Clave, "Oropharyngeal dysphagia is a prevalent risk factor for malnutrition in a cohort of older patients admitted with an acute disease to a general hospital," *Clinical Nutrition*, vol. 34, no. 3, pp. 436–442, Jun. 2015.

[39] L. Rofes, V. Arreola, M. Romea, E. Palomera, J. Almirall, M. Cabre, M. Serra-Prat, and P. Clave, "Pathophysiology of oropharyngeal dysphagia in the frail elderly," *Neurogastroenterology and Motility*, vol. 22, no. 8, pp. 851–858, Aug. 2010.

[40] P. E. Marik and D. Kaplan, "Aspiration pneumonia and dysphagia in the elderly," *Chest*, vol. 124, no. 1, pp. 328–336, Jul. 2003.

[41] 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.

[42] L. C. Lin, S. C. Wu, H. S. Chen, T. G. Wang, and M. Y. Chen, "Prevalence of impaired swallowing in institutionalized older people in taiwan," *Journal of the American Geriatrics Society*, vol. 50, no. 6, pp. 1118–1123, Jun. 2002.

[43] A. Jean, "Brain stem control of swallowing: Neuronal network and cellular mechanisms," *Physiological Reviews*, vol. 81, no. 2, pp. 929–969, Apr. 2001.

[44] J. Walton and P. Silva, "Physiology of swallowing," *Surgery (Oxford)*, vol. 36, no. 10, pp. 529–534, Oct. 2018.

[45] Blausen.com staff, "Medical gallery of Blausen Medical 2014," *WikiJournal of Medicine*, vol. 1, no. 2, Aug. 2014.

[46] P. Clave, M. de Kraa, V. Arreola, M. Girvent, R. Farre, E. Palomera, and M. Serra-Prat, "The effect of bolus viscosity on swallowing function in neurogenic dysphagia," *Alimentary Pharmacology and Therapeutics*, vol. 24, no. 9, pp. 1385–1394, Nov. 2006.

[47] K. Matsuo and J. B. Palmer, "Anatomy and physiology of feeding and swallowing: normal and abnormal," *Physical Medicine and Rehabilitation Clinics of North America*, vol. 19, no. 4, pp. 691–707, Nov. 2008.

[48] ——, "Coordination of mastication, swallowing and breathing," *Japanese Dental Science Review*, vol. 45, no. 1, pp. 31–40, May 2009.

[49] K. Kohyama, L. Mioche, and P. Bourdio3, "Influence of age and dental status on chewing behaviour studied by EMG recordings during consumption of various food samples," *Gerodontology*, vol. 20, no. 1, pp. 15–23, Jul. 2003.

[50] J. Robbins, R. E. Gangnon, S. M. Theis, S. A. Kays, A. L. Hewitt, and J. A. Hind, "The effects of lingual exercise on swallowing in older adults," *Journal of the American Geriatrics Society*, vol. 53, no. 9, pp. 1483–1489, Sep. 2005.

[51] L. Rofes, V. Arreola, J. Almirall, M. Cabré, L. Campins, P. García-Peris, R. Speyer, and P. Clavé, "Diagnosis and management of oropharyngeal dysphagia and its nutritional and respiratory complications in the elderly," *Gastroenterology Research and Practice*, vol. 2011, Aug. 2010.

[52] J. E. Aviv, "Effects of aging on sensitivity of the pharyngeal and supraglottic areas," *American Journal of Medicine*, vol. 103, no. 5A, pp. 74S–76S, Nov. 1997.

[53] J. E. Aviv, J. H. Martin, R. L. Sacco, D. Zagar, B. Diamond, M. S. Keen, and A. Blitzer, "Supraglottic and pharyngeal sensory abnormalities in stroke patients with dysphagia," *Annals of Otology, Rhinology and Laryngology*, vol. 105, no. 2, pp. 92–97, Feb. 1996.

[54] B. T. Massey, "Physiology of oral cavity, pharynx and upper esophageal sphincter," *GI Motility online*, May 2006.

[55] I. J. Cook, M. Gabb, V. Panagopoulos, G. G. Jamieson, W. J. Dodds, J. Dent, and D. J. Shearman, "Pharyngeal (Zenker's) diverticulum is a disorder of upper esophageal sphincter opening," *Gastroenterology*, vol. 103, no. 4, pp. 1229–1235, Oct. 1992.

[56] I. J. Cook, "Clinical disorders of the upper esophageal sphincter," *GI Motility online*, May 2006.

[57] G. N. Ali, K. L. Wallace, R. Schwartz, D. J. DeCarle, A. S. Zagami, and I. J. Cook, "Mechanisms of oral-pharyngeal dysphagia in patients with Parkinson's disease," *Gastroenterology*, vol. 110, no. 2, pp. 383–392, Feb. 1996.

[58] R. B. Williams, K. L. Wallace, G. N. Ali, and I. J. Cook, "Biomechanics of failed deglutitive upper esophageal sphincter relaxation in neurogenic dysphagia," *American*

Journal of Physiology: Gastrointestinal and Liver Physiology, vol. 283, no. 1, pp. G16–G26, Jul. 2002.

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

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

[61] J. A. Logemann and G. L. Shelley, "Should treatment for pharyngeal swallowing disorders begin before instrumental assessment is completed?" *ASHA*, vol. 38, no. 4, pp. 14–15, 1996.

[62] J. A. Cichero, S. Heaton, and L. Bassett, "Triaging dysphagia: Nurse screening for dysphagia in an acute hospital," *Journal of Clinical Nursing*, vol. 18, no. 11, pp. 1649–1659, Jun. 2009.

[63] D. M. Suiter, S. B. Leder, and D. E. Karas, "The 3-ounce (90-cc) water swallow challenge: a screening test for children with suspected oropharyngeal dysphagia," *Otolaryngology and Head and Neck Surgery*, vol. 140, no. 2, pp. 187–190, Feb. 2009.

[64] R. Martino, F. Silver, R. Teasell, M. Bayley, G. Nicholson, D. L. Streiner, and N. E. Diamant, "The Toronto bedside swallowing screening test (TOR-BSST): Development and validation of a dysphagia screening tool for patients with stroke," *Stroke*, vol. 40, no. 2, pp. 555–561, Feb. 2009.

[65] J. Edmiaston, L. T. Connor, L. Loehr, and A. Nassief, "Validation of a dysphagia screening tool in acute stroke patients," *American Journal of Critical Care*, vol. 19, no. 4, pp. 357–364, Jul. 2010.

[66] N. Antonios, G. Carnaby-Mann, M. Crary, L. Miller, H. Hubbard, K. Hood, R. Sambandam, A. Xavier, and S. Silliman, "Analysis of a physician tool for evaluating dysphagia on an inpatient stroke unit: the modified Mann Assessment of Swallowing Ability," *Journal of Stroke and Cerebrovascular Diseases*, vol. 19, no. 1, pp. 49–57, Jan. 2010.

[67] R. Ding, C. R. Larson, J. A. Logemann, and A. W. Rademaker, "Surface electromyographic and electroglottographic studies in normal subjects under two swallow conditions: normal and during the Mendelsohn manuever," *Dysphagia*, vol. 17, no. 1, pp. 1–12, 2002.

[68] C. Ertekin, I. Aydogdu, N. Yuceyar, S. Tarlaci, N. Kiylioglu, M. Pehlivan, and G. Celebi, "Electrodiagnostic methods for neurogenic dysphagia," *Electroencephalography and Clinical Neurophysiology*, vol. 109, no. 4, pp. 331–340, Aug. 1998.

[69] P. M. Zenner, D. S. Losinski, and R. H. Mills, "Using cervical auscultation in the clinical dysphagia examination in long-term care," *Dysphagia*, vol. 10, no. 1, pp. 27–31, 1995.

[70] P. Leslie, M. J. Drinnan, P. Finn, G. A. Ford, and J. A. Wilson, "Reliability and validity of cervical auscultation: A controlled comparison using videofluoroscopy," *Dysphagia*, vol. 19, no. 4, pp. 231–240, 2004.

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

[72] M. G. Rugiu, "Role of videofluoroscopy in evaluation of neurologic dysphagia," *Acta Otorhinolaryngologica Italica*, vol. 27, no. 6, pp. 306–316, Dec. 2007.

[73] M. Rashid, "Case 1: Diagnosing difficult deglutition," *Paediatrics and Child Health*, vol. 14, no. 7, pp. 453–454, Sep. 2009.

[74] R. W. Bastian, "The videoendoscopic swallowing study: An alternative and partner to the videofluoroscopic swallowing study," *Dysphagia*, vol. 8, no. 4, pp. 359–367, 1993.

[75] S. E. Langmore, K. Schatz, and N. Olsen, "Fiberoptic endoscopic examination of swallowing safety: A new procedure," *Dysphagia*, vol. 2, no. 4, pp. 216–219, Dec. 1988.

[76] A. M. Kelly, P. Leslie, T. Beale, C. Payten, and M. J. Drinnan, "Fibreoptic endoscopic evaluation of swallowing and videofluoroscopy: does examination type influence perception of pharyngeal residue severity?" *Clinical Otolaryngology*, vol. 31, no. 5, pp. 425–432, Oct. 2006.

[77] J. C. Rosenbek, J. A. Robbins, E. B. Roecker, J. L. Coyle, and J. L. Wood, "A penetration-aspiration scale," *Dysphagia*, vol. 11, no. 2, pp. 93–98, 1996.

[78] C. M. Steele, E. Sejdić, and T. Chau, "Noninvasive detection of thin-liquid aspiration using dual-axis swallowing accelerometry," *Dysphagia*, vol. 28, no. 1, pp. 105–112, Mar. 2013.

[79] I. J. Cook, W. J. Dodds, R. O. Dantas, B. Massey, M. K. Kern, I. M. Lang, J. G. Brasseur, and W. J. Hogan, "Opening mechanisms of the human upper esophageal sphincter," *American Journal of Physiology*, vol. 257, no. 5 Pt 1, pp. G748–G759, Nov. 1989.

[80] I. J. Cook, W. J. Dodds, R. O. Dantas, M. K. Kern, B. T. Massey, R. Shaker, and W. J. Hogan, "Timing of videofluoroscopic, manometric events, and bolus transit during the oral and pharyngeal phases of swallowing," *Dysphagia*, vol. 4, no. 1, pp. 8–15, Mar. 1989.

[81] A. Daggett, J. Logemann, A. Rademaker, and B. Pauloski, "Laryngeal penetration during deglutition in normal subjects of various ages," *Dysphagia*, vol. 21, no. 4, pp. 270–274, Oct. 2006.

[82] Y. Kim, G. H. McCullough, and C. W. Asp, "Temporal measurements of pharyngeal swallowing in normal populations," *Dysphagia*, vol. 20, no. 4, pp. 290–296, 2005.

[83] R. D. Gross, C. W. Atwood, S. B. Ross, K. A. Eichhorn, J. W. Olszewski, and P. J. Doyle, "The coordination of breathing and swallowing in Parkinson's disease," *Dysphagia*, vol. 23, no. 2, pp. 136–145, Jun. 2008.

[84] E. Sejdić, G. A. Malandraki, and J. L. Coyle, "Computational deglutition: Using signal- and image-processing methods to understand swallowing and associated disorders," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 138–146, Jan. 2019.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] D. C. Zoratto, T. Chau, and C. M. Steele, "Hyolaryngeal excursion as the physiological source of swallowing accelerometry signals," *Physiological Measurement*, vol. 31, no. 6, pp. 843–855, Jun. 2010.

[90] 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.

[91] 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.

[92] 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.

[93] L. Glass, "Synchronization and rhythmic processes in physiology," *Nature*, vol. 410, no. 6825, pp. 277–284, Mar. 2001.

[94] R. M. Rangayyan and N. P. Reddy, "Biomedical signal analysis: A case-study approach," *Annals of Biomedical Engineering*, vol. 30, no. 7, pp. 983–983, 2002.

[95] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 579–590, May 2013.

[96] J. Kim, M. Kim, I. Won, S. Yang, K. Lee, and W. Huh, "A biomedical signal segmentation algorithm for event detection based on slope tracing," in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* IEEE, Sep. 2009, pp. 1889–1892.

[97] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G. Z. Yang, "Big data for health," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[98] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68–80, May 2017.

[99] D. P. Mandic, D. Obradovic, A. Kuh, T. Adali, U. Trutschell, M. Golz, P. De Wilde, J. Barria, A. Constantinides, J. Chambers, W. Duch, J. Kacprzyk, E. Oja, and S. Zadrożny, "Data Fusion for Modern Engineering Applications: An Overview," Berlin, Heidelberg, Sep. 2005.

[100] D. Mandic, M. Golz, A. Kuh, D. Obradovic, and T. Tanaka, *Signal Processing Techniques for Knowledge Extraction and Information Fusion.* Springer US, Apr. 2008.

[101] L. Huiying, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," in *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4. IEEE, Oct. 1997, pp. 1630–1633.

[102] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, Mar. 1985.

[103] V. Srinivasan, C. Eswaran, and N. Sriraam, "Approximate entropy-based epileptic EEG detection using artificial neural networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 3, pp. 288–295, May 2007.

[104] N. Kannathal, M. L. Choo, U. R. Acharya, and P. K. Sadasivan, "Entropies for detection of epilepsy in EEG," *Computer Methods and Programs in Biomedicine*, vol. 80, no. 3, pp. 187–94, Dec. 2005.

[105] A. Schlogl, F. Lee, H. Bischof, and G. Pfurtscheller, "Characterization of four-class motor imagery EEG data for the BCI-competition 2005," *Journal of Neural Engineering*, vol. 2, no. 4, pp. L14–L22, Dec. 2005.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, May 2015.

[111] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[112] L. R. Rabiner, "A Tutorial on hidden Markov-models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[113] P. J. Werbos, "Backpropagation through time - what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

[114] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[115] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *CoRR*, vol. abs/1410.5401, Oct. 2014.

[116] A. Cohen, "Hidden Markov models in biomedical signal processing," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 3. IEEE, Nov. 1998, pp. 1145–1150.

[117] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[118] D. Jurafsky and J. H. Martin, *Speech and language processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., Jan. 2009.

[119] L. E. Baum and J. A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology," *Bulletin of the American Mathematical Society*, vol. 73, no. 3, pp. 360–363, 1967.

[120] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds," *Pacific Journal of Mathematics*, vol. 27, no. 2, pp. 211–227, 1968.

[121] L. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process," in *Proceedings of the 3rd Symposium on Inequalities*, vol. 3, Jan. 1972, pp. 1–8.

[122] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 39, no. 1, pp. 1–38, 1977.

[123] L. A. Liporace, "Maximum-likelihood estimation for multivariate observations of Markov sources," *IEEE Transactions on Information Theory*, vol. 28, no. 5, pp. 729–734, Sep. 1982.

[124] B. H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov-chains," *AT&T Technical Journal*, vol. 64, no. 6, pp. 1235–1249, Jul. 1985.

[125] Levinson, S, and M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of markov chains," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 307–309, Mar. 1986.

[126] G. Safont, A. Salazar, L. Vergara, E. Gómez, and V. Villanueva, "Multichannel dynamic modeling of non-Gaussian mixtures," *Pattern Recognition*, vol. 93, pp. 312–323, Sep. 2019.

[127] A. Salazar, L. Vergara, and R. Miralles, "On including sequential dependence in ICA mixture models," *Signal Processing*, vol. 90, no. 7, pp. 2314–2318, Jul. 2010.

[128] A. Graves, "Supervised sequence labelling," in *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 5–13.

[129] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179 – 211, Apr. 1990.

[130] M. I. Jordan, "Attractor dynamics and parallelism in a connectionist sequential machine," in *Artificial Neural Networks*, J. Diederich, Ed. Piscataway, NJ, USA: IEEE Press, 1990, pp. 112–127.

[131] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," German National Research Center for Information Technology, Tech. Rep., Jan. 2001.

[132] Y. Khalifa, Z. Zhang, and E. Sejdić, "Sparse recovery of time-frequency representations via recurrent neural networks," in *Proceedings of the 22nd International Conference on Digital Signal Processing*. ACM, Aug. 2017, pp. 1–5.

[133] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Neural Network Models of Cognition*, ser. Advances in Psychology. North-Holland, 1997, vol. 121, pp. 471–495.

[134] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, Jun. 2013, pp. III–1310–III–1318.

[135] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[136] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, May 2009.

[137] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. PMLR, May 2010, pp. 249–256.

[138] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Jun. 2014, pp. 1724–1734.

[139] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proceedings of the 9th International Conference on Artificial Neural Networks*, vol. 2. IEEE, Sep. 1999, pp. 850–855.

[140] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[141] P. Schwab, G. C. Scebba, J. Zhang, M. Delai, and W. Karlen, "Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks," in *Computing in Cardiology*, vol. 44, Sep. 2017, pp. 1–4.

[142] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, "Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation," *arXiv preprint arXiv:1506.07452*, Jun. 2015.

[143] A. Kadish, A. E. Buxton, H. Kennedy, B. P. Knight, J. W. Mason, C. Schuger, C. Tracy, W. L. Winters, A. W. Boone, M. Elnicki, J. W. Hirshfeld, B. H. Lorell, G. Rodgers, and H. H. Weitz, "ACC/AHA clinical competence statement on electrocardiography and ambulatory electrocardiography," *Journal of the American College of Cardiology*, vol. 38, pp. 3169–3178, Dec. 2001.

[144] M. H Crawford, S. Bernstein, P. Deedwania, J. Dimarco, K. J Ferrick, A. Garson, L. Green, H. Leon Greene, M. Silka, P. H Stone, C. Tracy, and R. Gibbons, "ACC/AHA guidelines for ambulatory electrocardiography," *Journal of the American College of Cardiology*, vol. 34, pp. 912–948, Aug. 1999.

[145] R. V. Andreao, B. Dorizzi, and J. Boudy, "ECG signal analysis through hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 8, pp. 1541–1549, Aug. 2006.

[146] K. S. Sayed, A. F. Khalaf, and Y. M. Kadah, "Arrhythmia classification based on novel distance series transform of phase space trajectories," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2015, pp. 5195–5198.

[147] W. Gersch, P. Lilly, and E. Dong, "PVC detection by the heart-beat interval data—Markov chain approach," *Computers and Biomedical Research*, vol. 8, no. 4, pp. 370 – 378, Aug. 1975.

[148] D. A. Coast, R. M. Stern, G. G. Cano, and S. A. Briller, "An approach to cardiac arrhythmia analysis using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 9, pp. 826–836, Sep. 1990.

[149] R. E. Hermes, D. B. Geselowitz, and G. Oliver, "Development, distribution, and use of the American Heart Association database for ventricular arrhythmia detector evaluation," *Computers in Cardiology*, pp. 263–266, Jan. 1980.

[150] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in *Computers in Cardiology*, Sep. 1997, pp. 673–676.

[151] F. Sandberg, M. Stridh, and L. Sornmo, "Frequency tracking of atrial fibrillation using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2 Pt 1, pp. 502–511, Feb. 2008.

[152] J. Oliveira, C. Sousa, and M. T. Coimbra, "Coupled hidden Markov model for automatic ECG and PCG segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2017, pp. 1023–1027.

[153] E. D. Übeyli, "Combining recurrent neural networks with eigenvector methods for classification of ECG beats," *Digital Signal Processing*, vol. 19, no. 2, pp. 320–329, Mar. 2009.

[154] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, Jun. 2000.

[155] C. Zhang, G. Wang, J. Zhao, P. Gao, J. Lin, and H. Yang, "Patient-specific ECG classification based on recurrent neural networks and clustering technique," in *Proceedings of the 13th International Conference on Biomedical Engineering*, Feb. 2017, pp. 63–67.

[156] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, May 2001.

[157] Z. Xiong, M. K. Stiles, and J. Zhao, "Robust ECG signal classification for detection of atrial fibrillation using a novel neural network," in *Computing in Cardiology*, vol. 44, Sep. 2017, pp. 1–4.

[158] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[159] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for electrocardiogram classification," in *Computing in Cardiology*, Sep. 2017, pp. 1–4.

[160] 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.

[161] Y. Chang, S. Wu, L. Tseng, H. Chao, and C. Ko, "AF detection by exploiting the spectral and temporal characteristics of ECG signals with the LSTM model," in *Computing in Cardiology*, vol. 45, Sep. 2018, pp. 1–4.

[162] S. Petrutiu, A. V. Sahakian, and S. Swiryn, "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans," *EP Europace*, vol. 9, no. 7, pp. 466–470, Jul. 2007.

[163] A. Taddei, G. Distante, M. Emdin, P. Pisani, G. B. Moody, C. Zeelenberg, and C. Marchesi, "The European ST-T database: standard for evaluating systems for

the analysis of ST-T changes in ambulatory electrocardiography," *European Heart Journal*, vol. 13, no. 9, pp. 1164–1172, Sep. 1992.

[164] F. M. Nolle, F. K. Badura, J. M. Catlett, R. W. Bowser, and M. H. Sketch, "CREIGARD, a new concept in computerized arrhythmia monitoring systems," *Computers in Cardiology*, vol. 13, pp. 515–518, Dec. 1987.

[165] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet," *Biomedizinische Technik/Biomedical Engineering*, vol. 40, no. s1, pp. 317–318, 1995.

[166] H. W. Lui and K. L. Chow, "Multiclass classification of myocardial infarction with convolutional and recurrent neural networks for portable ECG devices," *Informatics in Medicine Unlocked*, vol. 13, pp. 26–33, Jan. 2018.

[167] G. D. Clifford, C. Liu, B. Moody, L. H. Lehman, I. Silva, Q. Li, A. E. Johnson, and R. G. Mark, "AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017," in *Computing in Cardiology*, Sep. 2017, pp. 1–4.

[168] S. Singh, S. K. Pandey, U. Pawar, and R. R. Janghel, "Classification of ECG Arrhythmia using Recurrent Neural Networks," *Procedia Computer Science*, vol. 132, pp. 1290–1297, Jan. 2018.

[169] D. L. Schomer and F. L. Da Silva, *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*, 6th ed. Lippincott Williams and Wilkins, Nov. 2012.

[170] D. P. Subha, P. K. Joseph, U. R. Acharya, and C. M. Lim, "EEG signal analysis: A survey," *Journal of Medical Systems*, vol. 34, no. 2, pp. 195–212, Apr. 2010.

[171] S. H. Sheldon, R. Ferber, and M. H. Kryger, *Principles and practice of pediatric sleep medicine*, 1st ed. Elsevier Health Sciences, Mar. 2005.

[172] D. Y. Kang, P. N. DeYoung, A. Malhotra, R. L. Owens, and T. P. Coleman, "A state space and density estimation framework for sleep staging in obstructive sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1201–1212, Jun. 2018.

[173] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. D. Clifford, "A review of signals used in sleep analysis," *Physiological Measurement*, vol. 35, no. 1, pp. R1–R57, Dec. 2013.

[174] A. Flexerand, G. Dorffner, P. Sykacekand, and I. Rezek, "An automatic, continuous and probabilistic sleep stager based on a hidden markov model," *Applied Artificial Intelligence*, vol. 16, no. 3, pp. 199–207, Mar. 2002.

[175] A. Flexer, G. Gruber, and G. Dorffner, "A reliable probabilistic sleep stager based on a single EEG signal," *Artificial Intelligence in Medicine*, vol. 33, no. 3, pp. 199–207, Mar. 2005.

[176] L. G. Doroshenkov, V. A. Konyshev, and S. V. Selishchev, "Classification of human sleep stages based on EEG processing using hidden Markov models," *Biomedical Engineering*, vol. 41, no. 1, pp. 25–28, Jan. 2007.

[177] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.

[178] M. T. Bianchi, N. A. Eiseman, S. S. Cash, J. Mietus, C. K. Peng, and R. J. Thomas, "Probabilistic sleep architecture models in patients with and without sleep apnea," *Journal of Sleep Research*, vol. 21, no. 3, pp. 330–341, Jun. 2012.

[179] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl, "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, Dec. 1997.

[180] S. T. Pan, C. E. Kuo, J. H. Zeng, and S. F. Liang, "A transition-constrained discrete hidden Markov model for automatic sleep staging," *Biomedical Engineering Online*, vol. 11, no. 1, p. 52, Aug. 2012.

[181] F. Yaghouby and S. Sunderam, "Quasi-supervised scoring of human sleep in polysomnograms using augmented input variables," *Computers in Biology and Medicine*, vol. 59, pp. 54–63, Apr. 2015.

[182] J. A. Onton, D. Y. Kang, and T. P. Coleman, "Visualization of whole-night sleep EEG from 2-channel mobile recording device reveals distinct deep sleep stages with differential electrodermal activity," *Frontiers in Human Neuroscience*, vol. 10, p. 605, Nov. 2016.

[183] P. R. Davidson, R. D. Jones, and M. T. R. Peiris, "Detecting behavioral microsleeps using EEG and LSTM recurrent neural networks," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, Jan. 2005, pp. 5754–5757.

[184] Y. L. Hsu, Y. T. Yang, J. S. Wang, and C. Y. Hsu, "Automatic sleep stage recurrent neural classifier using energy features of EEG signals," *Neurocomputing*, vol. 104, pp. 105–114, Mar. 2013.

[185] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[186] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research," *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, Dec. 2014.

[187] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. B. Westover, M. T. Bianchi, and J. Sun, "SLEEPNET: Automated Sleep Staging System via Deep Learning," *arXiv preprint arXiv:1707.08262*, Jul. 2017.

[188] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. D. Vos, "Automatic sleep stage classification using single-channel EEG: Learning sequential features with attention-based recurrent neural networks," in *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Jul. 2018, pp. 1452–1455.

[189] E. Bresch, U. Großekathöfer, and G. Garcia-Molina, "Recurrent Deep Neural Networks for Real-Time Sleep Stage Classification From Single Channel EEG," *Frontiers in Computational Neuroscience*, vol. 12, p. 85, Oct. 2018.

[190] G. Klosh, B. Kemp, T. Penzel, A. Schlogl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeithofer, B. Saletu, W. M. Herrmann, S. L. Himanen, D. Kunz, M. J. Barbanoj, J. Roschke, A. Varri, and G. Dorffner, "The SIESTA project polygraphic and clinical database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 51–57, May 2001.

[191] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, Mar. 2019.

[192] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Computers in Biology and Medicine*, vol. 106, pp. 71–81, Mar. 2019.

[193] C. Sun, J. Fan, C. Chen, W. Li, and W. Chen, "A Two-Stage Neural Network for Sleep Stage Classification Based on Feature Learning, Sequence Learning, and Data Augmentation," *IEEE Access*, vol. 7, pp. 109 386–109 397, Aug. 2019.

[194] C. on Epidemiology and Prognosis and I. L. A. Epilepsy, "Guidelines for epidemiologic studies on epilepsy," *Epilepsia*, vol. 34, no. 4, pp. 592–596, Jul. 1993.

[195] W. W. Lytton, "Computer modelling of epilepsy," *Nature Reviews: Neuroscience*, vol. 9, no. 8, pp. 626–637, Aug. 2008.

[196] M. H. Abdullah, J. M. Abdullah, and M. Z. Abdullah, "Seizure detection by means of hidden Markov model and stationary wavelet transform of electroencephalograph signals," in *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, Jan. 2012, pp. 62–65.

[197] O. Smart and M. Chen, "Semi-automated patient-specific scalp EEG seizure detection with unsupervised machine learning," in *Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, Aug. 2015, pp. 1–7.

[198] S. Santaniello, D. L. Sherman, M. A. Mirski, N. V. Thakor, and S. V. Sarma, "A Bayesian framework for analyzing iEEG data from a rat model of epilepsy," in *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2011, pp. 1435–1438.

[199] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz, "On the predictability of epileptic seizures," *Clinical Neurophysiology*, vol. 116, no. 3, pp. 569–587, Mar. 2005.

[200] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, "Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic," *Physica D-Nonlinear Phenomena*, vol. 194, no. 3-4, pp. 357–368, Jul. 2004.

[201] P. E. McSharry, L. A. Smith, and L. Tarassenko, "Prediction of epileptic seizures: Are nonlinear methods relevant?" *Nature Medicine*, vol. 9, no. 3, pp. 241–242, Mar. 2003.

[202] Y. C. Lai, M. A. Harrison, M. G. Frei, and I. Osorio, "Controlled test for predictive power of Lyapunov exponents: their inability to predict epileptic seizures," *Chaos*, vol. 14, no. 3, pp. 630–642, Sep. 2004.

[203] M. Winterhalder, T. Maiwald, H. U. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage, "The seizure prediction characteristic: A general framework to assess and compare seizure prediction methods," *Epilepsy and Behavior*, vol. 4, no. 3, pp. 318–325, Jun. 2003.

[204] S. Wong, A. B. Gardner, A. M. Krieger, and B. Litt, "A stochastic framework for evaluating seizure prediction algorithms using hidden Markov models," *Journal of Neurophysiology*, vol. 97, no. 3, pp. 2525–2532, Mar. 2007.

[205] A. B. Gardner, A. M. Krieger, G. Vachtsevanos, and B. Litt, "One-class novelty detection for seizure analysis from intracranial EEG," *Journal of Machine Learning Research*, vol. 7, no. Jun, pp. 1025–1044, Jun. 2006.

[206] B. Direito, C. Teixeira, B. Ribeiro, M. Castelo-Branco, F. Sales, and A. Dourado, "Modeling epileptic brain states using EEG spectral analysis and topographic mapping," *Journal of Neuroscience Methods*, vol. 210, no. 2, pp. 220–229, Sep. 2012.

[207] M. Ihle, H. Feldwisch-Drentrup, C. A. Teixeira, A. Witon, B. Schelter, J. Timmer, and A. Schulze-Bonhage, "EPILEPSIAE - a European epilepsy database," *Computer Methods and Programs in Biomedicine*, vol. 106, no. 3, pp. 127–138, Jun. 2012.

[208] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," {PhD} {Thesis}, Massachusetts Institute of Technology, Sep. 2009.

[209] A. Petrosian, D. Prokhorov, R. Homan, R. Dasheiff, and D. Wunsch, "Recurrent neural network based prediction of epileptic seizures in intra- and extracranial EEG," *Neurocomputing*, vol. 30, no. 1-4, pp. 201–218, Jan. 2000.

[210] N. F. Güler, E. D. Übeyli, and n. Güler, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification," *Expert Systems with Applications*, vol. 29, no. 3, pp. 506–514, Oct. 2005.

[211] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state," *Physical Review. E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 64, no. 6, p. 061907, Dec. 2001.

[212] S. P. Kumar, N. Sriraam, and P. G. Benakop, "Automated detection of epileptic seizures using wavelet entropy feature with recurrent neural network classifier," in *Proceedings of the IEEE Region 10 International Conference*, Nov. 2008, pp. 1–5.

[213] G. R. Minasyan, J. B. Chatten, M. J. Chatten, and R. N. Harner, "Patient-specific early seizure detection from scalp EEG," *Journal of Clinical Neurophysiology*, vol. 27, no. 3, pp. 163–178, Jun. 2010.

[214] M. A. Naderi and H. Mahdavi-Nasab, "Analysis and classification of EEG signals using spectral analysis and recurrent neural networks," in *Proceedings of the 17th Iranian Conference of Biomedical Engineering*, Nov. 2010, pp. 1–4.

[215] L. Vidyaratne, A. Glandon, M. Alam, and K. M. Iftekharuddin, "Deep recurrent neural network for seizure detection," in *Proceedings of the IEEE International Joint Conference on Neural Networks*. IEEE, Jul. 2016, pp. 1202–1207.

[216] S. S. Talathi, "Deep Recurrent Neural Networks for seizure detection and early seizure detection systems," *arXiv preprint arXiv:1706.03283*, Jun. 2017.

[217] M. Golmohammadi, S. Ziyabari, V. Shah, E. Von Weltin, C. Campbell, I. Obeid, and J. Picone, "Gated recurrent networks for seizure detection," in *Proceedings of the 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec. 2017, pp. 1–5.

[218] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," *Frontiers in Neuroscience*, vol. 10, May 2016.

[219] M. Golmohammadi, V. Shah, S. Lopez, S. Ziyabari, S. Yang, J. Camaratta, I. Obeid, and J. Picone, "The TUH EEG seizure corpus," in *Proceedings of the American Clinical Neurophysiology Society Annual Meeting*, Feb. 2017, p. 1.

[220] S. Raghu, N. Sriraam, and G. P. Kumar, "Classification of epileptic seizures using wavelet packet log energy and norm entropies with recurrent Elman neural network classifier," *Cognitive Neurodynamics*, vol. 11, no. 1, pp. 51–66, Feb. 2017.

[221] 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.

[222] 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.

[223] R. Hussein, H. Palangi, R. K. Ward, and Z. J. Wang, "Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals," *Clinical Neurophysiology*, vol. 130, no. 1, pp. 25–37, Jan. 2019.

[224] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[225] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A brain-computer interface using electrocorticographic signals in humans," *Journal of Neural Engineering*, vol. 1, no. 2, pp. 63–71, Jun. 2004.

[226] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 140–154, Oct. 2011.

[227] K. Sayed, M. Kamel, M. Alhaddad, H. M. Malibary, and Y. M. Kadah, "Characterization of phase space trajectories for Brain-Computer Interface," *Biomedical Signal Processing and Control*, vol. 38, pp. 55–66, Sep. 2017.

[228] ——, "Extracting phase space morphological features for electroencephalogram-based brain-computer interface," *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 4, pp. 771–774, Aug. 2017.

[229] E. Donchin, K. M. Spencer, and R. Wijesinghe, "The mental prosthesis: Assessing the speed of a P300-based brain-computer interface," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 174–179, Jun. 2000.

[230] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 9, no. 3, pp. 283–288, Sep. 2001.

[231] B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller, "Hidden Markov models for online classification of single trial EEG data," *Pattern Recognition Letters*, vol. 22, no. 12, pp. 1299–1309, Oct. 2001.

[232] G. Pfurtscheller, C. Neuper, G. R. Muller, B. Obermaier, G. Krausz, A. Schlogl, R. Scherer, B. Graimann, C. Keinrath, D. Skliris, M. Wortz, G. Supp, and C. Schrank, "Graz-BCI: State of the art and clinical applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 177–180, Jun. 2003.

[233] S. Solhjoo, A. M. Nasrabadi, and M. R. H. Golpayegani, "Classification of chaotic signals using HMM classifiers: EEG-based mental task classification," in *Proceedings of the 13th European Signal Processing Conference*, Sep. 2005, pp. 1–4.

[234] G. Pfurtscheller and A. Schlögl, "Dataset III: Motor imagery," Tech. Rep., 2003.

[235] H. Suk and S. Lee, "Two-layer hidden Markov models for multi-class motor imagery classification," in *Proceedings of the 1st Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging*, Aug. 2010, pp. 5–8.

[236] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Dataset IIa: Graz dataset A," Tech. Rep., Jul. 2008.

[237] W. Speier, C. Arnold, J. Lu, A. Deshpande, and N. Pouratian, "Integrating language information with a hidden Markov model to improve communication rate in the P300 speller," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 3, pp. 678–684, May 2014.

[238] A. Erfanian and B. Mahmoudi, "Real-time ocular artifact suppression using recurrent neural network for electro-encephalogram based brain-computer interface," *Medical and Biological Engineering and Computing*, vol. 43, no. 2, pp. 296–305, Mar. 2005.

[239] E. M. Forney and C. W. Anderson, "Classification of EEG during imagined mental tasks by forecasting with Elman recurrent neural networks," in *Proceedings of the IEEE International Joint Conference on Neural Networks.* IEEE, Aug. 2011, pp. 2749–2755.

[240] D. Balderas, A. Molina, and P. Ponce, "Alternative classification techniques for brain-computer interfaces for smart sensor manufacturing environments," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 680–685, 2015.

[241] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain-computer communication: Motivation, aim, and impact of exploring a virtual apartment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 4, pp. 473–482, Dec. 2007.

[242] R. Maddula, J. Stivers, M. Mousavi, S. Ravindran, and V. de Sa, "Deep recurrent convolutional neural networks for classifying P300 BCI signals," in *Proceedings of the 7th Graz Brain-Computer Interface Conference*, Sep. 2017.

[243] J. Stivers and V. de Sa, "Spelling in parallel: Towards a rapid, spatially independent BCI," in *Proceedings of the 7th Graz Brain-Computer Interface Conference*, Sep. 2017.

[244] J. Thomas, T. Maszczyk, N. Sinha, T. Kluge, and J. Dauwels, "Deep learning-based classification for brain-computer interfaces," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2017, pp. 234–239.

[245] V. P. Oikonomou, G. Liaros, K. Georgiadis, E. Chatzilari, K. Adam, S. Nikolopoulos, and I. Kompatsiaris, "Comparative evaluation of state-of-the-art algorithms for SSVEP-based BCIs," *arXiv preprint arXiv:1602.00904*, Feb. 2016.

[246] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp. 6809–6817.

[247] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. H. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[248] T. Hosman, M. Vilela, D. Milstein, J. N. Kelemen, D. M. Brandman, L. R. Hochberg, and J. D. Simeral, "BCI decoder performance comparison of an LSTM recurrent neural network and a Kalman filter in retrospective simulation," in *Proceedings of the 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, Mar. 2019, pp. 1066–1071.

[249] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–171, Jul. 2006.

[250] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making Sense of Spatio-Temporal Preserving Representations for EEG-Based Human Intention Recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3033–3044, Jul. 2020.

[251] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, Jun. 2004.

[252] S. Tortora, S. Ghidoni, C. Chisari, S. Micera, and F. Artoni, "Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network," *Journal of Neural Engineering*, vol. 17, no. 4, p. 046011, Jul. 2020.

[253] M. J. Zwarts and D. F. Stegeman, "Multichannel surface EMG: Basic aspects and clinical utility," *Muscle and Nerve*, vol. 28, no. 1, pp. 1–17, Jul. 2003.

[254] J. Y. Hogrel, "Clinical applications of surface electromyography in neuromuscular disorders," *Neurophysiologie Clinique*, vol. 35, no. 2-3, pp. 59–71, Jul. 2005.

[255] M. A. Oskoei and H. S. Hu, "Myoelectric control systems-A survey," *Biomedical Signal Processing and Control*, vol. 2, no. 4, pp. 275–294, Oct. 2007.

[256] K. Englehart, B. Hudgins, and P. A. Parker, "A wavelet-based continuous classification scheme for multifunction myoelectric control," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 3, pp. 302–311, Mar. 2001.

[257] A. D. Chan and K. B. Englehart, "Continuous myoelectric control for powered prostheses using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 1, pp. 121–124, Jan. 2005.

[258] K. Englehart, B. Hudgins, and A. D. C. Chan, "Continuous multifunction myoelectric control using pattern recognition," *Technology and disability*, vol. 15, no. 2, pp. 95–103, Aug. 2003.

[259] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Q. Wang, and J. H. Yang, "A framework for hand gesture recognition based on accelerometer and EMG sensors," *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans*, vol. 41, no. 6, pp. 1064–1076, Nov. 2011.

[260] K. R. Wheeler, M. H. Chang, and K. H. Knuth, "Gesture-based control and EMG decomposition," *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, vol. 36, no. 4, pp. 503–514, Jul. 2006.

[261] J. Monsifrot, E. Le Carpentier, Y. Aoustin, and D. Farina, "Sequential decoding of intramuscular EMG signals via estimation of a Markov model," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 5, pp. 1030–1040, Sep. 2014.

[262] K. S. Lee, "EMG-based speech recognition using hidden markov models with global control variables," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 930–940, Mar. 2008.

[263] A. D. Chan, K. Englehart, B. Hudgins, and D. F. Lovely, "Hidden Markov model classification of myoelectric signals in speech," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 5, pp. 143–146, Sep. 2002.

[264] ——, "Myo-electric signals to augment speech recognition," *Medical and Biological Engineering and Computing*, vol. 39, no. 4, pp. 500–504, Jul. 2001.

[265] Z. Li, M. Hayashibe, C. Fattal, and D. Guiraud, "Muscle fatigue tracking with evoked EMG via recurrent neural network: Toward personalized neuroprosthetics," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 38–46, May 2014.

[266] P. Xia, J. Hu, and Y. Peng, "EMG-based estimation of limb movement using deep learning with recurrent convolutional neural networks," *Artificial Organs*, vol. 42, no. 5, pp. E67–E77, May 2018.

[267] F. Quivira, T. Koike-Akino, Y. Wang, and D. Erdogmus, "Translating sEMG signals to continuous hand poses using recurrent neural networks," in *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics*, Mar. 2018, pp. 166–169.

[268] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, Aug. 2013.

[269] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, and W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PloS One*, vol. 13, no. 10, p. e0206049, Oct. 2018.

[270] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A. G. Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Muller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Scientific data*, vol. 1, p. 140053, Dec. 2014.

[271] A. Samadani, "Gated Recurrent Neural Networks for EMG-Based Hand Gesture Classification. A Comparative Study," in *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 1–4.

[272] M. Atzori, A. Gijsberts, S. Heynen, A.-G. M. Hager, O. Deriaz, P. van der Smagt, C. Castellini, B. Caputo, and H. Müller, "Building the Ninapro database: A resource for the biorobotics community," in *Proceedings of the 2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, Jun. 2012, pp. 1258–1265.

[273] M. Simão, P. Neto, and O. Gibaru, "EMG-based online classification of gestures with recurrent neural networks," *Pattern Recognition Letters*, vol. 128, pp. 45–51, Dec. 2019.

[274] ——, "UC2018 DualMyo Hand Gesture Dataset," Jul. 2018.

[275] S. Pizzolato, L. Tagliapietra, M. Cognolato, M. Reggiani, H. Müller, and M. Atzori, "Comparison of six electromyography acquisition setups on hand movement classification tasks," *PloS One*, vol. 12, no. 10, p. e0186132, Oct. 2017.

[276] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiological Measurement*, vol. 31, no. 4, pp. 513–529, Apr. 2010.

[277] A. D. Ricke, R. J. Povinelli, and M. T. Johnson, "Automatic segmentation of heart sound signals using hidden markov models," in *Computers in Cardiology*, vol. 32, Sep. 2005, pp. 953–956.

[278] P. Sedighian, A. W. Subudhi, F. Scalzo, and S. Asgari, "Pediatric heart sound segmentation using Hidden Markov Model," in *Proceedings of the 36th Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society, Aug. 2014, pp. 5490–5493.

[279] C. S. Lima and D. Barbosa, "Automatic segmentation of the second cardiac sound by using wavelets and hidden Markov models," in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2008, pp. 334–337.

[280] P. B. Shull, W. Jirattigalachote, M. A. Hunt, M. R. Cutkosky, and S. L. Delp, "Quantified self and human movement: a review on the clinical impact of wearable sensing and feedback for gait analysis and intervention," *Gait and Posture*, vol. 40, no. 1, pp. 11–9, May 2014.

[281] 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.

[282] 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.

[283] C. Nickel, C. Busch, S. Rangarajan, and M. Möbius, "Using hidden Markov models for accelerometer-based biometric gait recognition," in *Proceedings of the IEEE 7th International Colloquium on Signal Processing and its Applications*, Mar. 2011, pp. 58–63.

[284] A. Mannini and A. M. Sabatini, "A hidden Markov model-based technique for gait segmentation using a foot-mounted gyroscope," in *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2011, pp. 4369–4373.

[285] C. Nickel and C. Busch, "Classifying Accelerometer Data via Hidden Markov Models to Authenticate People by the Way They Walk," *IEEE Aerospace and Electronic Systems Magazine*, vol. 28, no. 10, pp. 29–35, Oct. 2013.

[286] G. Panahandeh, N. Mohammadiha, A. Leijon, and P. Handel, "Continuous hidden Markov model for pedestrian activity classification and gait analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1073–1083, May 2013.

[287] M. Inoue, S. Inoue, and T. Nishida, "Deep recurrent neural network for mobile human activity recognition with high throughput," *Artificial Life and Robotics*, vol. 23, no. 2, pp. 173–185, Jun. 2018.

[288] A. Lisowska, G. Wheeler, V. Ceballos Inza, and I. Poole, "An evaluation of supervised, novelty-based and hybrid approaches to fall detection using silmee accelerome-

ter data," in *Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2015, pp. 402–408.

[289] T. Theodoridis, V. Solachidis, N. Vretos, and P. Daras, "Human fall detection from acceleration measurements using a recurrent neural network," in *Precision Medicine Powered by pHealth and Connected Health*, N. Maglaveras, I. Chouvarda, and P. de Carvalho, Eds. Springer Singapore, Nov. 2017, pp. 145–149.

[290] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[291] T. Glasmachers, "Limits of End-to-End Learning," *arXiv preprint arXiv:1704.08305*, Apr. 2017.

[292] G. Cheron, J.-P. Draye, M. Bourgeios, and G. Libert, "A dynamic neural network identification of electromyography and arm trajectory relationship during complex movements," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 5, pp. 552–558, May 1996.

[293] G. Cheron, F. Leurs, A. Bengoetxea, J. P. Draye, M. Destrée, and B. Dan, "A dynamic recurrent neural network for multiple muscles electromyographic mapping to elevation angles of the lower limb in human locomotion," *Journal of Neuroscience Methods*, vol. 129, no. 2, pp. 95–104, Oct. 2003.

[294] S. Chauhan and L. Vig, "Anomaly detection in ECG time signals via deep long short-term memory networks," in *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics*, Oct. 2015, pp. 1–7.

[295] V. G. Sujadevi, K. P. Soman, and R. Vinayakumar, "Real-time detection of atrial fibrillation from short time single lead ECG traces using recurrent neural networks," in *Intelligent Systems Technologies and Applications*, ser. Advances in Intelligent Systems and Computing, S. M. Thampi, S. Mitra, J. Mukhopadhyay, K.-C. Li, A. P. James, and S. Berretti, Eds. Cham: Springer International Publishing, Sep. 2017, pp. 212–221.

[296] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of the 2010 IEEE International Symposium on Circuits and Systems*, May 2010, pp. 253–256.

[297] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten Digit Recognition with a Back-Propagation Network," in *Proceedings of the 3rd Conference on Neural Information Processing Systems*, D. S. Touretzky, Ed. Morgan-Kaufmann, Nov. 1990, pp. 396–404.

[298] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 433–445, Mar. 2011.

[299] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, Mar. 2016.

[300] 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.

[301] J. H. Tan, Y. Hagiwara, W. Pang, I. Lim, S. L. Oh, M. Adam, R. S. Tan, M. Chen, and U. R. Acharya, "Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals," *Computers in Biology and Medicine*, vol. 94, pp. 19–26, Mar. 2018.

[302] Z. Xiong, M. P. Nash, E. Cheng, V. V. Fedorov, M. K. Stiles, and J. Zhao, "ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network," *Physiological Measurement*, vol. 39, no. 9, p. 094006, Sep. 2018.

[303] Y. M. Saidutta, J. Zou, and F. Fekri, "Increasing the learning Capacity of BCI Systems via CNN-HMM models," in *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Jul. 2018, pp. 1–4.

[304] Z.-R. Wang, J. Du, W.-C. Wang, J.-F. Zhai, and J.-S. Hu, "A comprehensive study of hybrid neural network hidden Markov model for offline handwritten Chinese text recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 21, no. 4, pp. 241–251, Dec. 2018.

[305] Z.-R. Wang, J. Du, and J.-M. Wang, "Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition," *Pattern Recognition*, vol. 100, p. 107102, Apr. 2020.

[306] N. C. Dvornek, D. Yang, P. Ventola, and J. S. Duncan, "Learning generalizable recurrent neural networks from small task-fMRI datasets," in *Proceedings of the 21st Conference on Medical Image Computing and Computer Assisted Intervention*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Springer International Publishing, Sep. 2018, pp. 329–337.

[307] S. J. Pan and Q. A. Yang, "A Survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[308] A. Isin and S. Ozdalili, "Cardiac arrhythmia detection using deep learning," *Procedia Computer Science*, vol. 120, pp. 268 – 275, 2017.

[309] C. Wei, Y. Lin, Y. Wang, T. Jung, N. Bigdely-Shamlo, and C. Lin, "Selective transfer learning for EEG-based drowsiness detection," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2015, pp. 3229–3232.

[310] Y.-Q. Zhang, W.-L. Zheng, and B.-L. Lu, "Transfer components between subjects for EEG-based driving fatigue detection," in *Proceedings of the 29th Conference on Neural Information Processing Systems*, S. Arik, T. Huang, W. K. Lai, and Q. Liu, Eds. Springer International Publishing, Nov. 2015, pp. 61–68.

[311] U. Côté-Allard, C. L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, K. Glette, F. Laviolette, and B. Gosselin, "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760–771, Apr. 2019.

[312] 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.

[313] N. P. Reddy, R. Thomas, E. P. Canilang, and J. Casterline, "Toward classification of dysphagic patients using biomechanical measurements," *Journal of Rehabilitation Research and Development*, vol. 31, no. 4, pp. 335–344, Nov. 1994.

[314] J. Lee, C. M. Steele, and T. Chau, "Time and time-frequency characterization of dual-axis swallowing accelerometry signals," *Physiological Measurement*, vol. 29, no. 9, pp. 1105–1120, Sep. 2008.

[315] N. P. Reddy, E. P. Canilang, J. Casterline, M. B. Rane, A. M. Joshi, R. Thomas, and R. Candadai, "Noninvasive acceleration measurements to characterize the pharyngeal phase of swallowing," *Journal of Biomedical Engineering*, vol. 13, no. 5, pp. 379–383, Sep. 1991.

[316] N. P. Reddy, A. Katakam, V. Gupta, R. Unnikrishnan, J. Narayanan, and E. P. Canilang, "Measurements of acceleration during videofluorographic evaluation of dysphagic patients," *Medical Engineering and Physics*, vol. 22, no. 6, pp. 405–412, Jul. 2000.

[317] 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.

[318] F. Hanna, S. M. Molfenter, R. E. Cliffe, T. Chau, and C. M. Steele, "Anthropometric and demographic correlates of dual-axis swallowing accelerometry signal characteristics: a canonical correlation analysis," *Dysphagia*, vol. 25, no. 2, pp. 94–103, Jun. 2010.

[319] 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.

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

[321] S. Singh and S. Hamdy, "The upper oesophageal sphincter," *Neurogastroenterology and Motility*, vol. 17 Suppl 1, no. Suppl. 1, pp. 3–12, Jun. 2005.

[322] N. K. Ahuja and W. W. Chan, "Assessing upper esophageal sphincter function in clinical practice: a primer," *Current Gastroenterology Reports*, vol. 18, no. 2, p. 7, Feb. 2016.

[323] P. J. Kahrilas, W. J. Dodds, J. Dent, J. A. Logemann, and R. Shaker, "Upper esophageal sphincter function during deglutition," *Gastroenterology*, vol. 95, no. 1, pp. 52–62, Jul. 1988.

[324] Y. Kim, T. Park, E. Oommen, and G. McCullough, "Upper esophageal sphincter opening during swallow in stroke survivors," *American Journal of Physical Medicine and Rehabilitation*, vol. 94, no. 9, pp. 734–739, Sep. 2015.

[325] A. L. Merati, "In-office evaluation of swallowing: FEES, pharyngeal squeeze maneuver, and FEESST," *Otolaryngologic Clinics of North America*, vol. 46, no. 1, pp. 31–9, Feb. 2013.

[326] S. G. Butler, L. Markley, B. Sanders, and A. Stuart, "Reliability of the penetration aspiration scale with flexible endoscopic evaluation of swallowing," *Annals of Otology, Rhinology and Laryngology*, vol. 124, no. 6, pp. 480–483, Jun. 2015.

[327] A. M. Kelly, M. J. Drinnan, and P. Leslie, "Assessing penetration and aspiration: how do videofluoroscopy and fiberoptic endoscopic evaluation of swallowing compare?" *Laryngoscope*, vol. 117, no. 10, pp. 1723–1727, Oct. 2007.

[328] 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.

[329] J. W. Lee, D. R. Randall, L. M. Evangelista, M. A. Kuhn, and P. C. Belafsky, "Subjective assessment of videofluoroscopic swallow studies," *Otolaryngology and Head and Neck Surgery*, vol. 156, no. 5, pp. 901–905, May 2017.

[330] J. Robbins, J. Coyle, J. Rosenbek, E. Roecker, and J. Wood, "Differentiation of normal and abnormal airway protection during swallowing using the penetration-aspiration scale," *Dysphagia*, vol. 14, no. 4, pp. 228–232, 1999.

[331] 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.

[332] K. Takahashi, M. E. Groher, and K. Michi, "Methodology for detecting swallowing sounds," *Dysphagia*, vol. 9, no. 1, pp. 54–62, 1994.

[333] J. Lee, E. Sejdić, C. M. Steele, and T. Chau, "Effects of liquid stimuli on dual-axis swallowing accelerometry signals in a healthy population," *Biomedical Engineering Online*, vol. 9, no. 1, p. 7, Feb. 2010.

[334] S. Hamlet, D. G. Penney, and J. Formolo, "Stethoscope acoustics and cervical auscultation of swallowing," *Dysphagia*, vol. 9, no. 1, pp. 63–68, 1994.

[335] J. A. Cichero and B. E. Murdoch, "Detection of swallowing sounds: methodology revisited," *Dysphagia*, vol. 17, no. 1, pp. 40–9, 2002.

[336] F. M. Bandi and J. R. Russell, "Microstructure noise, realized variance, and optimal sampling," *The Review of Economic Studies*, vol. 75, no. 2, pp. 339–369, 2008.

[337] B. C. Sonies, L. J. Parent, K. Morrish, and B. J. Baum, "Durational aspects of the oral-pharyngeal phase of swallow in normal adults," *Dysphagia*, vol. 3, no. 1, pp. 1–10, Mar. 1988.

[338] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, Aug. 2015, pp. 429–436.

[339] 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.

[340] N. P. Reddy, D. L. Simcox, V. Gupta, G. E. Motta, J. Coppenger, A. Das, and O. Buch, "Biofeedback therapy using accelerometry for treating dysphagic patients with poor laryngeal elevation: Case studies," *Journal of Rehabilitation Research and Development*, vol. 37, no. 3, pp. 361–372, May 2000.

[341] 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.

[342] 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.

[343] 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.

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

[345] 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.

[346] 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.

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

[348] 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.

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

[350] 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.

[351] J. A. Cichero and B. E. Murdoch, "Acoustic signature of the normal swallow: characterization by age, gender, and bolus volume," *Annals of Otology, Rhinology and Laryngology*, vol. 111, no. 7 Pt 1, pp. 623–632, Jul. 2002.

[352] A. El-Jaroudi, M. S. Redfern, L. F. Chaparro, and J. M. Furman, "The application of time-frequency methods to the analysis of postural sway," *Proceedings of the IEEE*, vol. 84, no. 9, pp. 1312–1318, Sep. 1996.

[353] 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.

[354] 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.

[355] 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.

[356] ——, "The effects of head movement on dual-axis cervical accelerometry signals," *BMC Research Notes*, vol. 3, p. 269, Oct. 2010.

[357] ——, "A procedure for denoising dual-axis swallowing accelerometry signals," *Physiological Measurement*, vol. 31, no. 1, pp. N1–N9, Jan. 2010.

[358] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[359] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, Jun. 2014.

[360] A. L. Perlman, X. He, J. Barkmeier, and E. Van Leer, "Bolus location associated with videofluoroscopic and respirodeglutometric events," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 1, pp. 21–33, Feb. 2005.

[361] S. Moriniere, M. Boiron, D. Alison, P. Makris, and P. Beutter, "Origin of the sound components during pharyngeal swallowing in normal subjects," *Dysphagia*, vol. 23, no. 3, pp. 267–273, Sep. 2008.

[362] K. Shu, J. L. Coyle, S. Perera, Y. Khalifa, A. Sabry, and E. Sejdić, "Anterior-posterior distension of maximal upper esophageal sphincter opening is correlated with high-resolution cervical auscultation signal features," *Physiological Measurement*, Feb. 2021.

[363] P. Jacob, P. J. Kahrilas, J. A. Logemann, V. Shah, and T. Ha, "Upper esophageal sphincter opening and modulation during swallowing," *Gastroenterology*, vol. 97, no. 6, pp. 1469–1478, Dec. 1989.

[364] P. R. Katz, H. M. Reynolds, D. R. Foust, and J. K. Baum, "Mid-sagittal dimensions of cervical vertebral bodies," *American Journal of Physical Anthropology*, vol. 43, no. 3, pp. 319–326, Nov. 1975.

[365] S. M. Molfenter and C. M. Steele, "Use of an anatomical scalar to control for sex-based size differences in measures of hyoid excursion during swallowing," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 3, pp. 768–778, Jun. 2014.

[366] Q. He, S. Perera, Y. Khalifa, Z. Zhang, A. S. Mahoney, A. Sabry, C. Donohue, J. L. Coyle, and E. Sejdić, "The association of high resolution cervical auscultation signal features with hyoid bone displacement during swallowing," *IEEE Transactions on*

Neural Systems and Rehabilitation Engineering, vol. 27, no. 9, pp. 1810–1816, Sep. 2019.

[367] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, Aug. 2009, pp. 248–255.

[368] K. Shu, S. Mao, J. L. Coyle, and E. Sejdić, "Improving Non-invasive Aspiration Detection with Auxiliary Classifier Wasserstein Generative Adversarial Networks," *IEEE Journal of Biomedical and Health Informatics*, Aug. 2021.