Investigations of Videofluoroscopy via Machine Learning: Novel Ways for Swallowing Disorders Assessment

by

Zhenwei Zhang

M.S in Electrical Engineering

University of Pittsburgh, 2016

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

Zhenwei Zhang

It was defended on

April 27 2021

and approved by

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

Zhihong Mao Ph.D, Professor, Department of Electrical and Computer Engineering

Amro El-Jaroudi Ph.D, Associate Professor, Department of Electrical and Computer Engineering

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

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

Investigations of Videofluoroscopy via Machine Learning: Novel Ways for Swallowing Disorders Assessment

Zhenwei Zhang, PhD

University of Pittsburgh, 2021

Videofluoroscopic swallow studies are widely used in clinical and research settings to assess swallow function and to determine physiological impairments, diet recommendations, and treatment goals for people with dysphagia. It can be used to analyze biomechanical events of swallowing, to differentiate between normal and disordered swallow function. It is also important for clinicians to understand the association between various possible physiological measures and penetration-aspiration, in order to determine the boundary values for these measures that can be validated for impaired swallows. In recent years, deep learning technique have achieved tremendous success in various medical imaging applications, including, but not limited to brain studies, disease diagnosis and prevention. In this dissertation research, we attempted to further this research in two key areas. First, we evaluated the potential association between the trajectory of hyoid bone movement and the risk of airway penetration and aspiration during VFSS examination using generalized estimation equations. In addition, the model was built based on aspects of hyoid bone displacement to predict the extent of airway penetration. Second, we aimed to explore the potentials of deep learning techniques to address different dysphagia problems. These algorithms to automatically evaluate and assess VFSS dysphagia studies are highly sought after in the dysphagia clinical and scientific communities. To demonstrate the feasibility of deep learning techniques on VFSS, we computed and compared the state of art object detection networks for hyoid bone tracking algorithm, which was the first attempt to utilize deep learning techniques in the VFSS field. In physiologic measurements, scaling of images based on length of vertebrae bodies to compensate for size differences among different patients is a crucial component of the analysis. In order to detect key anatomical points needed for a routine swallowing assessment in real-time, we presented a novel two-stage convolutional neural network trained with missing annotations to localize and measure length of the vertebral bodies. Finally, we

sought to measure the amount of residue remained in vallecular area. We implemented an ensemble method with several networks to segment and calculate the residue scale.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 Motivation

1.1.1 Definition of Dysphagia

Swallowing difficulty, also called dysphagia, describes any swallowing dysfunction [51, 65, 176, 186] that causes subjective discomfort or objective difficulty in the formation or transportation of a bolus safely and completely from mouth to stomach without entering the airway [50]. It usually occurs in patients who suffer from a variety of neurological disorders (such as stroke, brain tumors, Parkinson's disease and dementia), mouth or neck cancer, throat pouches and different types of infections. It also appears in the patients who have weak muscular conditions which result in the inability to relax during the swallowing process. Dysphagia may present in many different ways, these sign and symptoms include having pain during swallowing, difficulty to swallowing, and unexpected weight loss. Patients may have to cut food into small pieces to avoid these swallowing troubles. Furthermore, patients may experience drooling, regurgitation, coughing or gagging during swallowing, which largely decreases the quality of life. Dysphagia can be classified into two main categories: oropharyngeal dysphagia [235], and esophageal dysphagia [123]. Oropharyngeal dysphagia [1] describes the swallowing problems that happen in the mouth or throat. This includes the lips, the tongues, the oral cavity, the pharynx, the airway, and the esophagus and sphincters. Patients may have the sensation of foods passing through the trachea or up the nose. The main reason of this problem is due to the weakness of muscle which make it difficult to form boluses and move it from mouth to throat. Esophageal dysphagia describes the problems in the esophagus. Patients may have the sensation of foods sticking under the throat or chest. One major concern of dysphagia is aspiration of food or liquid during the swallowing process, which may cause these boluses to pass the vocal folds in the airway and into respiratory system which leads to obstruction and pneumonia [169]. These phenomena directly endanger patients' life, and is a major cause of death in Parkinson's disease [267].

1.1.2 Incidence and Prevalence in United States

Dysphagia is a sorely neglected disorder that impacts millions of Americans every year. It can occur in all age ranges, but appears more commonly among elder patients. In the United States, approximately 4% of adults have swallowing related issues annually [28]. Investigations have estimated that around 9.44 million adults had a swallowing problem in the United States with 40% being male and 60% being female [28]. In the literature, adult dysphagia has been reported as high as 20% of the population over age 50 years [208]. The increasing age causes changes in anatomy and neural or muscular mechanisms, resulting in loss of functions that affect the swallowing process. Dysphagia appears commonly among people admitted into hospitals or nursing care facilities. It is estimated that 12%-13% of patients in short-term care hospitals and around 60% of nursing home occupants have swallowing difficulties [52]. Approximately half of Americans over 60 will suffer from dysphagia [167]. Among all the causes of dysphagia, stroke is the most commonly reported etiology (11.2%), followed by other neurological causes (7.2%) and head and neck cancer (4.9%) [28]. Dysphagia is present in above half of stroke patients and head and neck cancer radiation therapy patients [167, 191]. Dysphagia is frequent and clinically relevant in PD patients and multiple sclerosis patients. More than 80% of patients develop a swallowing impairment during the course of their disease [177], and dysphagia occurs in one third of multiple sclerosis patients, who are potentially at risk for aspiration and malnutrition during swallowing [205].

Despite swallowing problems significantly impacting workday and daily life, only a relative minority of affected patients seek professional treatments. Studies estimated that only 22.7% of patients with swallowing problems saw a professional health care professional [28]. The exact number of patients is difficult to calculate as the first step in evaluating dysphagia is to recognize the problem and some patients are not aware that they are experiencing swallowing difficulty. This phenomenon is called silent aspiration, which is poorly investigated in the related research area. Investigations have showed that about 30% of dysphagia patients aspirated silently [212]. Furthermore, many symptoms associated with swallowing difficulty are poorly recognized by nurses and physician staff who are not dysphagia clinicians. The

standard of care for these patients across institutional settings remains highly variable as each patient has different requirements for liquid thickness and food consistency [24], and also due to the unavailability of trained clinicians in underserved settings.

Although dysphagia usually occurs in elder population, it can be present in younger populations as well, particularly in infants with specific developmental and medical disorders. Causes of pediatric dysphagia can exist alone or combine with other medical conditions such as neurological disorders, prematurity, failure to thrive and brain injury [208]. The report showed that the prevalence of pediatric dysphagia is increasing due to high survival rates of children born prematurely [17]. It is estimated that feeding problems occur in 25% - 45% of typical developing children and 30% - 80% of children with developmental disorders. As children can not communicate their problems efficiently as an adults, there is a high risk that their conditions remain undiagnosed [195].

1.1.3 Impact on Hospital Resources and Cost of Treatment

Swallowing difficulty has a significant impact on health care in hospital and nursing facilities. According to the National Hospital Discharge Survey (2005 - 2006), there were about 271,983 (0.35%) hospital admissions in United States associated with dysphagia, with a total of 77 million hospitalizations during that period [8]. The costs of health care increase due to extended hospital stays, need for expensive respiratory, emergency room visits or nutritional support [138]. The report showed that only 14% of stroke patient with dysphagia required more than 7 days of hospitalization [8]. Patients with dysphagia have about a 40% increase in length of hospital stay in all age groups, with an estimated cost $ 547 million per year. The average one-year medical health care cost for patients with dysphagia post ischemic stroke was $4510 higher than patients without dysphagia [32]. It is reported that the total enteral cost for feeding supplies is over $670 million in 2003, which is about 6% of the total Medicare budget for that year. In 2007, there were over 188,000 percutaneous endoscopic gastronomy placement procedures and 68% of the procedures were for patients age 65 years or older [69]. The population over 65 years is expected to double by 2050, which will result in a dramatic impact on hospital resource and health care [219].

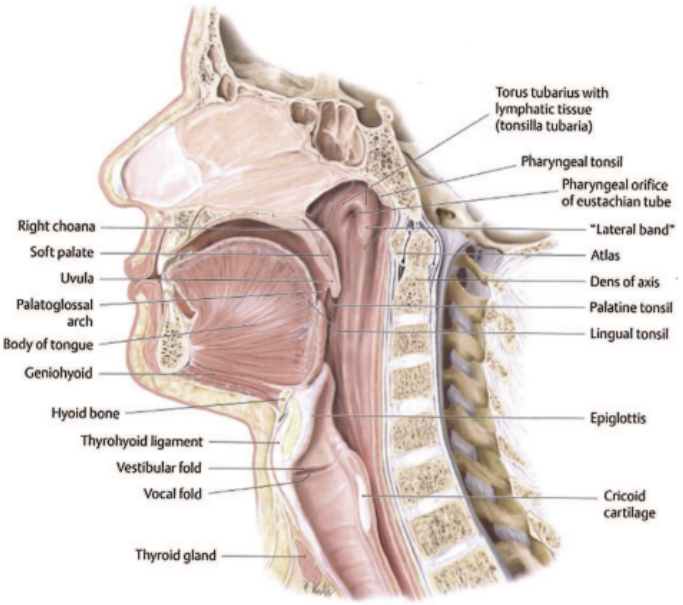


Figure 1: Midsagittal view of head and neck [231]

1.2 Anatomy and Physiology

1.2.1 Stages of Deglutition and Safe Swallowing

Swallowing is an essential motor activity needed for proper daily nutrition and hydration in humans. It is considered a complex neurological process as it involves multiple central and peripheral subsystems to accomplish the swallowing task in one to two seconds. There are few investigations on swallowing compared to other fundamental motor activities such as locomotion or mastication due to the complexity of the motor pattern associated with the greater number of muscles and cranial nerves [70]. The important structure related to swallowing activity is shown in figure 1.

Swallowing is separated into four main stages in order to describe the numerous events during a relatively short duration. These stages include the oral preparatory stage, oral transit stage, pharyngeal stage and esophageal stage [51, 65, 176, 186]. The first stage of swallowing is oral preparatory stage, which is a totally voluntary event [100]. During the first stage, food is reduced and formed into a soft consistency called bolus by being chewed, and mixed. In this phase, the formed bolus is shaped and positioned within the mid-line groove of the tongue and in spoon-like depression of the mid tongue [62]. The soft palate depresses toward the base of the tongue to seal off the oral cavity posteriorly, preventing the spillage of liquid or food into oropharynx, which is important for airway protection [70]. Then the posterior part of the tongue elevates against the soft palate and pushes downward, which keeps the bolus in the mouth and entry into the pharynx. The tongue plays an essential role in both the oral preparatory phase and the oral transit phase.

The second stage of swallowing is the oral transit phase. It starts when the tongue compresses the bolus against the palate to propel the bolus posteriorly into the pharynx [70]. In this stage, several events occur together to increase the volume of the pharynx for the bolus to pass into the oropharynx. The soft palate is elevated and the nasal cavity is sealed off from oropharynx [70]. The tongue moves upward and backward when the bolus is conveyed into the hypopharynx [100]. The upward and forward movements of hyoid bone elevates the larynx, which serves to expand the pharynx in the sagittal plane and protects the airway. The pharyngeal levators shorten length, may also increase transverse diameter of the pharynx [62].

The third stage of swallowing is pharyngeal phase, which is an involuntary independent event [39]. The pharyngeal phase is the shortest and the most complex phase in the whole swallowing process, occurring within a second [162]. The general definition of the beginning of pharyngeal phase is the moment the bolus enters the oropharynx [33]. However, the exact initiation of this phase is slightly different due to bolus characteristics and swallowing pattern [27]. For example, the pharyngeal phase begins before the head of bolus enters the pharyngeal in healthy young individuals while it starts after the bolus head enters the pharynx in older individuals. A normal pharyngeal phase includes palatal closure, bolus passage through the pharynx, glottal closure to prevent aspiration and upper esophageal sphincter (UES) opening

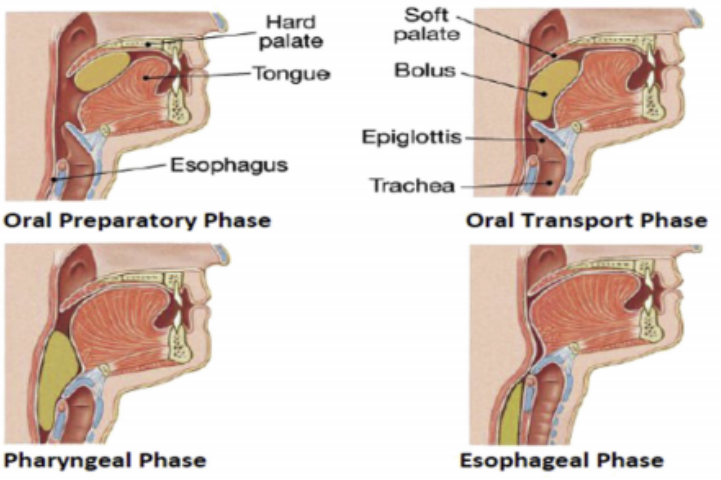


Figure 2: Four swallowing phases and associated neuromuscular action [162]

[100]. During this phase, the oropharynx is blocked by the tongue, nasopharynx is sealed by the soft palate and proximal pharyngeal wall, laryngeal vestibule is covered by epiglottis and laryngeal opening are closed by the vocal cords and arytenoids [82]. The muscles of the larynx (lateral cricoarytenoid, transverse arytenoid, and thyroarytenoid muscles) close the vocal folds and cease the aspiration. The pharyngeal constrictor contractions, creating a peristaltic wave to push the bolus to the UES. When the bolus reaches the UES, the cricopharyngeus muscle relaxes and allows it to pass into the digestive system.

The last stage is esophageal phase, which starts after the bolus passes the UES. The esophagus is a tubular structure connecting the lower part of UES to the lower esophageal sphincter which prevents regurgitation from the stomach. The esophagus relaxes as the bolus enters, the large parts of the bolus then move into the stomach due to gravity while other bolus may be transported to the stomach by peristaltic contraction [82]. During this stage, all structures move from their positions during the previous phase return to their initial position. In a healthy adult, it take about 8 to 13 second to transport a bolus from UES to the stomach [162].

1.2.2 Methods of Swallowing Assessment

Clinical non-instrumental examination, also called "bedside" examination, plays an important role in dysphagia assessment. This method aims to observe the presence of signs and symptoms of dysphagia during a patients' swallow by considering the factors such as fatigue during the meal, posture, positioning and environmental conditions. These non-instrumental methods are highly varied in design, targeted groups, and assessment domains [92]. Many of these assessments have issues like the misinterpretation of results and inconsistent use because of the lack of instruction for use and interpretation for assessment scores. In the daily clinical setting, the 3-oz water swallow test, the Toronto Bedside Swallowing Screening test, and the Standardized Swallowing Assessment are all considered to be clinically useful among other tests [260, 171]. The main procedures of non-instrumental swallowing assessment are divided into several steps [17]. The first step concerns the investigation of general conditions, which includes patient's generic data, breathing condition and functionality, nutritional situation and duration of meal, quality of phonation and speech articulation, review of medical records, and social environment. Beside these factors, neurologic diagnosis, neuropsychologic conditions and communicative level for neurologic patients are considered in the first step assessment as well. The second step of assessment is the morphodynamic evaluation, which contains the following evaluations: structural assessment of lips (kissing, opening and closing), jaw, tongue (motility, protrusion and backwards pushing), soft palate (check sufflating), larynx (elevation of the larynx) and the muscular control of the head.

Non-instrumental assessment provides many useful information for clinicians to diagnose oral dysphagia. However, this assessment can only provide limited information and accuracy, and as a result, clinicians often over-estimate the severity of pharyngeal dysphagia [67]. Furthermore, clinicians could not obtain the information about aspiration and other physiologic problems in the pharyngeal phase. Instrumental assessment is the only method which can directly observe these events. Non-instrumental evaluation may serve as a tool for determining the potential requirements for additional instrumental evaluation and specifying the clinical questions to be answered by these instrumental tests.

There are many instrumental diagnostic techniques applied for swallowing disorders, including videofluoroscopy, ultrasound, manometry, manofluorography, scintigraphy, and fiberoptic endoscopic evaluation (FEES). Other advanced techniques, including functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (MEG), and electroencephalography (EEG), investigate the relationship between brain activities and dysphagia. Clinicians use these techniques to assess swallowing physiology in patients who have the symptoms of swallowing disorders and estimate the degree of swallowing impairment.

The American Speech-Language-Hearing Association (ASHA) suggests that a clinical-instrumental examination of swallowing should reveal several physiology events [188]. These events include organic and functional alterations in involved structures, the degree of efficacy of swallowing in each stage, co-ordination between breathing and swallowing, and protection of airways during swallowing. Furthermore, ASHA also suggests to detect and quantify the penetration of boluses in the tracheal-bronchial passage.

Among the various techniques, there are two principal imaging instrumental examinations widely used in daily diagnosis and treatment of dysphagia: video fluoroscopic swallowing studies (VFSS), commonly called modified barium swallow studies, and fiberoptic endoscopic evaluation of swallowing (FEES). A complete swallowing investigation should be viewed from different planes to have a better treatment program. These planes include the coronal plane (front to back), transverse plane (upper to lower) and sagittal plane (left to right) [247]. FEES can provide transverse views of the larynx during the swallowing, while fluoroscopy can study swallowing from any of these planes. During VFSS, a patient is seated before an X-ray machine and instructed to swallow different liquids and/or food mixed with barium [215]. Typically, the swallowing assessment is carried out by a radiologist, a speech-language pathologist or another specialist [170]. By observing biomechanical functions of the numerous oropharyngeal structures such as the hyoid bone, larynx, tongue, and esophageal sphincter that lead to bolus flow such as the flow of these radio-opaque boluses through the upper aerodigestive tract, clinicians are able to observe the effects of various bolus textures, bolus volumes, and compensatory strategies on swallowing physiology [156, 255]. These biomechanical events are used by clinicians to evaluate the integrity of

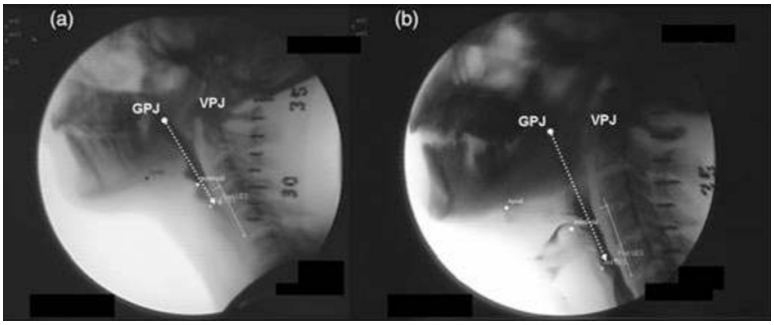


Figure 3: Examples of videofluoroscopic images in a healthy volunteer (a) and a patient showing an aspiration (b) [49]

neuromuscular function and the coordination of events in order to determine the safety and efficiency of each swallow observed during the examination. According to Logemann, there are four main purposes to using radiographic assessment [247]: (1) to measure the speed of swallowing; (2) to measure the efficiency of swallowing; (3) to examine the effectiveness of rehabilitation strategies; (4) to define the movement patterns. VFSS also detects the presence, exact time, and depth of airway laryngeal penetration and tracheal aspiration during a swallow, which assists clinicians in identifying the causes of aspirations and can identify appropriate interventions to minimize or eliminate airway compromise during swallowing. Figure 3 illustrates the equipment for VFSS investigations and images acquired from the equipment.

The primary disadvantage of VFSS is x-ray radiation. Each individual has to be verified if they have undergone other x-rays in the past year as exposure to radiation is cumulative. Patients receive excessive amount of radiation under extended or repeated use of VFSS. Therefore, investigations of patients swallowing varieties of food with several attempts of different volumes, variety and consistency is limited. According to the National Institutes of Health (NIH) Guidelines for Radiation Safety, the maximum exposure for adults over age 18 years is 5 rem/year for all diagnostic purposes [247]. These guidelines also suggest that the maximum permissible radiation to any tissue is 3 rem in a 13 week time span. The resolution of the images is related to the dose area product. Higher doses can generate a better resolution and the balance between image quality and dose exposure has to be considered.

In the case of swallowing studies, the radiation exposure is limited to 270 to 660 mrad/study. The other disadvantages include difficulty accessing these equipment and limitations from patient's conditions. Dysphagia commonly appears in patients with neurological disorders or other patients already admitted in the acute care facilities [78]. These patients may not have the physical or mental ability to complete the whole videofluoroscopy examination procedure as it requires patients to follow complex orders [25].

Beside VFSS, FEES is another popular method to evaluate swallowing disorders [188]. FEES examination has an important advantage when compared to VFSS examination: easy to implement, well tolerated, less costly and possible to implement in bed examinations. It has become the first choice in Europe when the patients are required to have an instrumental examination. FEES is performed with a fiberoptic rhinopharyngoscope to study the swallowing physiology and physiopathology of certain stages of swallowing, particularly for pharyngeal stage. Instead of using an x-ray imaging machine, the speech-language pathologist can directly observe the patients' anatomy of swallowing through a small camera [25]. FEES examinations provide the details of the relative functions of the upper airways and upper digestive tract. It also allows examiners to see smaller details as well as the colors of tissues, which can not be achieved by VFSS. FEES also provides important information for diagnosis decision making. It also can test laryngeal sensitivity by using the tip of the rhinopharyngoscope to stimulate the various pharyngeal-laryngeal areas. Furthermore, it offers the possibility to evaluate the presence, degree and type of dysphagia and it is also applied to establish the best way of feeding, advise diets and plan other diagnostic investigations [10].

Both methods have limitations that preclude identification of all components of dysphagia, however, both are considered as close to a gold standard as is available. VFSS is ubiquitous and has been considered the gold standard in the study of swallowing. It plays key roles in the diagnosis of dysphagia and revealing the presence of underlying biomechanical causes of aspiration [53, 226]. Compared to VFSS, FEES has several limitations in dysphagia diagnosis. The small camera has a limited field of view, thus it can't observe the whole swallowing process simultaneously and only a limited section of anatomy is visible at one time [112]. Clinicians can determine which parts of the patient's swallowing are not functioning

properly through VFSS and determine whether patients aspirate or not during swallowing and also how significantly they aspirate. Another limitation of FEES is that the camera can be placed in a limited area during swallow: FEES cannot address oral and esophageal stages; it is particularly applied to observe swallow physiology during the pharyngeal stage.

Though VFSS and FEES are currently the most common techniques accepted in the daily diagnosis of swallowing disorder, they are not the only screening techniques in clinical practice or research area. There are also screening methods for swallowing disorders in clinical environments such as electromyography (EMG) and cervical auscultation. EMG is a simple method to apply: electrodes are placed on patient's neck and provide electrical information about muscle activity in real-time during a swallow [259]. The ideal hypotheses is that compared to the recording from a healthy subject, the signal will significantly change in some clinical way if the muscles related to swallowing can not work well [61]. However, the limitation of EMG is due to the indirect description of swallowing, lack of knowledge about factors affecting the signal, and the anatomy and physiology of the head and neck musculature. Current techniques do not allow the separation of isolated muscle information of speech musculature, however, it may be used for between-condition or between-group comparisons. Cervical auscultation is a listening device which is used to record swallow sounds and airway sounds [31, 140]. These devices, including stethoscope, microphone, accelerometer or Doppler sonar, are placed over the patient's thyroid cartilage and examiners listens to the various sounds produced during patient's swallow. This method is mainly used to assess the relationships of swallowing sound components and physiology events during the pharyngeal phase, as well as the interaction with breathing. The goal is to use the difference of these signals between healthy swallows and unsafe swallows to help clinicians make decisions. The advantage of this method is that it is easy to perform and can be applied in any age groups. However, in the current condition, the accuracies and reproducibility of this technique are still unreliable [139]. The connections between the swallow sound components and physiological events of pharyngeal phage are still lacking of strong evidence.

1.3 Directions and Goals

Over 16 million people in the USA and over 40 million people in Europe suffer from a swallowing disorder [270]. Patients with severe dysphagia has a high morality rate, and about 40,000 people die every year in the United States because of aspiration pneumonia, which is believed to be the result from dysphagia [265]. VFSS was considered the gold standard for studies on swallowing disorders due to its wide use in the evaluation of dysphagia. This technique can make measurements more precise as clinicians can analyze the images frame-by-frame, which increase intra and inter rater reliability [111]. A standardized VFSS recording protocol was developed for the VFSS study analysis [196]; a range of software applications use a standard plane to correct head positions and magnification, several of them are allowed to mark the well-defined anatomic reference points of interest [155, 210]. The major steps of these applications include digitization, identification of reference points and anatomical points of interest.

The hyoid bone, a radiographically landmark indicating excursion of the hyolaryngeal complex (HLC), has been shown in numerous kinematic analysis to move both upward and forward, in patterns that vary slightly from person to person, then return to the starting position when muscular contractions subside [103]. The onset of hyoid bone displacement initiates the pharyngeal phase of swallowing [279]. This displacement pattern reconfigures the upper aerodigestive tract to facilitate closure of the laryngeal vestibule and in the presence of neutrally modulated relaxation of the upper esophageal sphincter (UES), applies traction to the anterior wall of the UES facilitating, opening the esophagus in order for food to be delivered into the esophagus and subsequently the stomach. Inadequate anterior hyoid displacement leads to impaired laryngeal vestibule closure and inadequate traction forces on the UES which when combined can lead to airway penetration and incomplete opening of the UES. As a result, evaluation of anterior hyoid excursion from VFSS images is considered important in evaluating the nature of the swallowing impairment and extent to which impaired excursion of the HLC contributes to airway compromise, inefficient clearance into the esophagus, post-swallow pharyngeal residue caused by separation of the bolus due to premature UES closure which can subsequently be aspirated after the swallow [170].

The hyoid bone displacement is influenced by various factors such that bolus characteristics, age, gender and etiology of dysphagia. The displacement of the hyoid bone shows the importance of penetration and aspiration evaluation [101]. However, the biomechanical analysis of the hyoid bone is still limited, further investigations associated with hyoid bone are required. On the other hand, VFSS is a powerful examination that offers possibility to reveal the anatomy during swallowing. However, VFSS examination requires experts to guide the examination. The presence of a speech pathologist is highly recommended not only to guide the radiologist performing VFSS, but also to modify the examination procedure to obtain more detailed information. Software applications help clinicians and researchers for the further study of swallowing after examination. However, in most cases, a complete biomechanical analysis is not always performed because of clinician lack of experience or unavailability of a qualified examiner. In addition, experts have to spend lots of time checking frame by frame to determine the exact time of swallowing event happened and also manually select the points of interest before they output these numeric results for the investigations. Furthermore, the analysis requires intra and inter rater reliability testing, which causes many repeated and useless workloads.

In addition to physiologic measurements, scaling of images to compensate for size differences among different patients is a crucial component of the analysis that enables each patient's swallowing function to be compared to norms that would be expected of a healthy person of the same size. For example, Seo & Molfenter developed a method of scaling images by using the distance between antero-inferior margin of the second and fourth cervical vertebral bodies, in order to correct influence from patient head movement and participant size [185, 233]. Without scaling, the distance of structural displacements can be over- or under-estimated, leading to inaccurate diagnosis. In practice, each of these landmarks is manually marked on VFSS images, not in real-time but following the examination to serve as the anatomic scalar.

A better understanding of swallowing biomechanical analysis and a faster way to analyze the VFSS images would be beneficial to the general public. Ideally, these objectively analyze should be automated and applied aside VFSS equipment during the patient's examinations. Examiners can easily get the points of interests in the real-time during examination, which

can help them make quick and accurate decisions. At the moment, we feel that some research interests related to the hyoid bone still require investigation, including the relationship between hyoid bone trajectory, penetration and aspiration in a whole swallow process, and whether we can predict the aspiration based on the trajectory information. Furthermore, machine learning techniques such as regression models has the potential to analyze these potential relations. In addition, machine learning techniques such as support vector machine, neural network, and random forest have shown huge success in other biomedical research fields [77]. The VFSS images contain huge amounts of information. Whether this image information can be applied to answer different clinical dysphagia questions is a relatively new research directions.

1.4 Dissertation Scope

In order to have the most reliable results in the computer-assisted system, it is important to manually analyze the data in the VFSS images at the first step. Each frame of images has been manually marked including location of hyoid bone, and cervical vertebrae. Work has been done with these hyoid bone data, investigating the association between features extracted from hyoid bone trajectory and the penetration and aspiration event using regression models. The penetration and aspiration were widely investigated by controlling the age, gender, bolus viscosity and head position [7, 36, 76]. However, the other factors influencing the penetration and aspiration are still limited. Through this investigation, we want to know how the trajectory influences the penetration and aspiration scale during a whole swallow process. After investigating the association between the hyoid bone and the penetration-aspiration, an important part of the investigation is to discover whether we can use trajectory information to predict the penetration and aspiration based on the results of association. A good understanding of hyoid bone trajectory for patients with different penetration and aspiration scale could potentially lead to the development of a classification method which can help in the diagnosis of dysphagia.

In addition, we want to develop an algorithm that allows the detection of key components in the VFSS images. These findings can potentially help clinicians and researchers easily get the information they want. Since machine learning techniques are hot topics in the recent computer vision field, we want to employ techniques, such as deep learning, in order to localize the hyoid bone position from a random frame, automatically segment the cervical vertebrae which can help to build a reference coordinate more efficiently. Deep learning is a powerful tool which can extract complex patterns from massive volumes of data without direct human input [189]. Beside the detection of key points in the video frames, we also plan to use neural network techniques to build an application which can automatically clip the videos to only contain the swallow. Clinicians can benefit from this application instead of manually checking frame by frame to verify the starting point and end point of each swallow. Furthermore, the machine learning techniques may participate in the score measurement, the residual score, $\%(C2 - C4)^2$ residue scale measure the amount of residue occupying pharyngeal space, which is an important scale for swallowing research and diagnosis.

1.5 Main Contributions

We hypothesize that it is possible to automatically identify important components in VFSS images acquired from patients who have undergone the VFSS instrumental evaluation. We suggest that this process can be automated and implemented with existing diagnosis techniques to help clinician make decisions more easily and accurately. To address these goals, the following key topics will be completed in this project.

- Investigate the association between trajectory features extracted from hyoid bone motion during the swallow and penetration-aspiration scale
- Investigate the possibility of using hyoid bone trajectory features and patient's personal information to predict the penetration-aspiration scale
- Develop an algorithm to automatically identify and localize the hyoid bone in the VFSS video sequence using neural network techniques

- Develop an algorithm to automatically segment the 2nd and 4th tail of cervical vertebrae in each frame for VFSS video instead of manually segmenting
- Develop an algorithm to automatically mark the residual score for the given VFSS image

1.6 Dissertation Organization

Chapter 2 explores the background of machine learning and its application in radiology imaging field to understand the principles discussed in the later chapters. Chapter 3 provides an overview of the topics covered in this manuscript. Chapter 4 discusses our attempt to discover the association between hyoid bone movement features and penetration/aspiration scale. Chapter 5 offers similar material, but presents the prediction of penetration-aspiration scale using generalized estimating equations. Chapter 6 introduced the hyoid bone tracking algorithm in VFSS using deep learning techniques. Chapter 7 and Chapter 8 present the automatic vertebrae localization and vallecular residue segmentation. Combining two networks, we achieved promising results on $\%(C2-C4)^2$ residue measurement scale estimation. Last chapter concludes our research and provide insight for possible avenues of research for future studies in this field.

2.0 Background

The majority of this chapter has been previously published in and reprinted with permission from [310]. Zhenwei Zhang and Ervin Sejdić. Radiological images and machine learning: trends, perspectives, and prospects. *Computers in biology and medicine*, 108:354–370, 2019

Radiological imaging has become indispensable in assisting medical experts in clinical diagnosis, treatments, and research studies. Technological advancements in radiology have enabled higher imaging resolutions for visualization of small structures and abnormalities in the body. As the broader use of radiologic image analysis increases the workload for radiologists, the development of intelligent computer-aided systems for automated image analysis becomes essential. The goals of computer-aided systems are to achieve faster and more accurate results in handling large volumes of radiological imaging. The rapid expansion in machine learning algorithms have make these kinds of computer-aided systems possible.

2.1 Machine Learning in Radiology

In recent years, machine learning algorithms have become very effective tools for the analysis of medical images in many radiology applications [289]. These algorithms are able to extract multiple details from medical images without an understanding of where useful information may be coded in images [228]. Computer-aided systems based on machine learning help radiologists to make informed decisions while interpreting these images [289].

2.1.1 Types of Learning

Depending on the utilization of labels in training data, machine learning algorithms can be divided into supervised learning, unsupervised learning, and semi-supervised learning. Supervised-learning is the most common form in machine learning [128]. Data is usually collected and labeled in categories, as the purpose of supervised learning is to find an appro-

priate input-output function from training data, which generalizes well against the testing data. We can compute an objective function to measure the error between the desired pattern and output score. In general, many scientific contributions focus on finding a suitable objective function with adjustable parameters. The supervised learning method is widely used in classification and regression. Unsupervised learning is used for data without corresponding label information [57]. The purpose of unsupervised learning is to discover the hidden structure or distribution of data. Algorithms discover potential patterns in the data, which differ from supervised learning. Approaches using unsupervised learning include clustering and blind signal separation techniques such as principle component analysis and independent component analysis. The idea of semi-supervised learning is somewhere between supervised learning and unsupervised-learning [178, 314]. Semi-supervised learning uses a small amount of labeled data and a large amount of unlabeled data during a training phase. One idea of semi-supervised learning is that it begins with a small set of labeled data and augments the training data size by gradually labeling unlabeled data. This method is typically utilized in cases where labeled data sets are relatively rare or difficult to acquire.

2.1.2 Feature Selection

Feature extraction and representation is a crucial step in medical image processing. With the development of modern medical techniques, higher resolution and more features have become obtainable to feed the classifiers; however, this is an obstacle for machine learning techniques in achieving an optimal solution using high dimensional features. Great interest exists in extracting and identifying reliable features from radiological images to improve classification performance [40, 127]. Several methods exist for extraction of features from medical images including region-based features, shape-based features, texture-based features, and bag-of-words features [278, 104, 296, 273, 295, 214, 174]. The performance of most image retrieval systems is dependent on the use of these features. Table 1 shows the summary of image features used in radiological image analysis. Color features are one of the most important features of images including RGB, histogram [304], color moments [198] and color coherence vector. Texture features are measured from a group of pixels, which is useful in

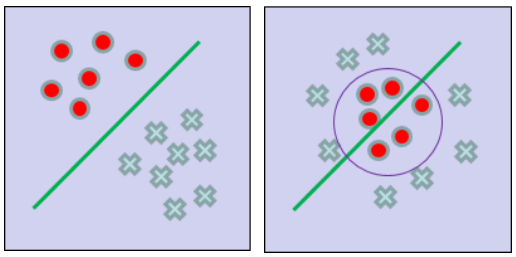
Table 1: A summary of image features used in ML systems

Features		Examples
Color	Invariant from different size and direction	Histogram [252, 87, 301]
Shape	Binary representation of images	Sphericity [301, 60]
Texture	Description of image structure, randomness, linearity, roughness, granulation, and homogeneity	Haralick's features [266, 60] Gabor features [113, 315, 134] Co-occurrence [187] Curvelet-based [59, 234] Wavelet-based [201, 165]
Local	Description of local image information using region, object of interest, corners, or edges	Local binary pattern [301] Scale invariant feature transform [15, 132, 6] Speed up robust features [283, 6]
Other	Other methods to extract image features	CNN [253]

characterization of a wide range of images. The Gabor filter is the most common method for texture extraction [273]. Scale invariant feature transform and speed up robust features algorithm are two popular methods for scale and rotation invariant feature detector and descriptor in computer vision [109]. Different types of images have significant contrast variation, thus visual features such as color, shape and texture are not enough to classify images easily. Thus high-level features are useful to overcome the intensity variations in different types of images and the extract suitable information from these images. How to select ideal features that can reflect the contents of images as useful as possible remains a challenging problem in machine learning.

2.1.3 Overview of Machine Learning Methods

Machine learning has been developing rapidly in recent years, and it is impossible to cover all recently-developed techniques in one section. In this section, we will review the most commonly used machine learning methods in radiology, such as linear models, the support vector machine, decision tree learning, ensemble classifier, as well as neural networks and deep learning. This section provides a general description of machine learning techniques that will be helpful to understand their applications in the field of radiology, as described in subsequent sections.



(a) (b)

Figure 4: Basic idea of linear classification and non-linear classification, (a) linear case (b) non linear case. The linear model uses linear functions to separate the data yet is not suitable for non-linear cases. SVM is one way to separate non-linear models using different kernel functions.

2.1.3.1 Linear Models for Regression and Classification

The purpose of regression is to predict the value from the given input features, whereas the purpose of classification is to assign input x to one of the predefined classes [29]. Commonly used linear models include linear regression, Fisher's linear discriminant (LDA), and logistic regression. The simplest linear models establish a linear relationship among input variables. Given $x_i, i = 1, 2, 3..., N$, the input feature vector, the output $y(x, \omega) = \omega_0 + \sum_{i=1}^{N} \omega_N x_N$. Logistic regression is the most basic classifier, it predicts the probability that an input x belongs to a class (class 1), versus the probability of another class (class 0). The basic idea of logistic regression is that we learn the logistic function of the form:

$$P(y = 1|x) = \frac{1}{1 + exp(-\omega^T x)} \tag{2.1}$$

where x is the input vector and ω is a weight vector for input. The logistic function is a continuous function and can turn any input from negative infinity to positive infinity into an output that is always between zero and one [58]. Fig. 4 illustrates linear and non-linear separable cases for a dataset.

2.1.3.2 Support Vector Machine

Support vector machines (SVM) are kernel-based supervised learning techniques widely used for classification and regression [29, 268]. The basic idea of SVM is to find an optimal hyperplane for linear separable patterns. It attempts to maximize the geometric margin on the training set and minimize the training error. Then, a kernel function is introduced for non-linearly separable cases by mapping original data into a new space. A two-class classification problem was used in many research cases. $x_i, i = 1, 2, ..., N$ are feature vectors of the training set X, and of corresponding class indicator $y_i \in \{-1, +1\}$. The goal of SVM is to construct a classifier in the form of:

$$y(x) = \text{sign}[\sum_{i=1}^{N_s} \lambda_i y_i K(x_i, x) + \omega_0] \tag{2.2}$$

The function $K(x_i, x)$ is called the kernel function and many pattern recognition and regression model were developed around their different mathematical properties. SVM with a linear kernel equation is computationally faster than SVM with quadratic kernel functions. SVM models using fewer but more significant features are most likely robust and less prone to overfitting [276].

2.1.3.3 Decision Tree Learning

Decision trees are one of the most popular classification approaches in machine learning [158]. The decision tree consists of a "root", "leaves", and internal nodes [222, 21, 249]. The node "root" has no incoming edges, "leaves" only have incoming edges but no outgoing edges, and the rest are internal nodes. The internal nodes use certain features to split the instance space into two or more subspaces. Each leaf represents one class. The leaf may represent the most appropriate target value or indicate the probability of the target having a certain value. Fig. 5 is an example for the decision tree model. Decision trees are capable of handling datasets that may have missing values and errors, however, this method may overfit training data and add irrelevant features. In radiological image analysis, decision trees are usually ensembled to form random forests for prediction and classification.

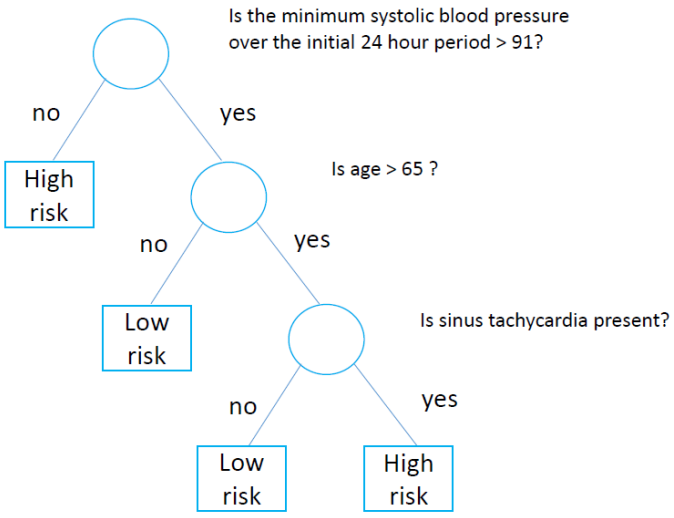


Figure 5: A medical example of decision trees. In this example, patients are classified into two classes: high risk and low risk. The features include blood pressures, age, etc. In this case, the classification tree operates similarly to a clinician's examination process.

2.1.3.4 Ensemble Learning

Ensemble learning combines multiple classifiers and applies voting algorithms to achieve a final classification, with popular ensemble approaches including boosting and bagging [26]. Fig. 6 shows the basic idea of ensemble learning. In boosting, extra weight is given to the incorrectly predicted points, a set of weak classifiers are applied to deal with data in the training phase, and the final prediction is derived from the weighted inputs resulting from the outputs of the weak classifiers. In bagging, the sub-classifier is independently constructed using a bootstrap sample of the data set and a majority voting method is taken for the final prediction [146]. The random forests are an ensemble learning method that consist of a multitude of decision trees. In standard tree construction, the node is split using the best split among all features. In a random forest, the node is split among a random subset of features. The random forest is one of the most powerful machine learning predictors used in detection, classification, and segmentation [277], particularly for brain [317, 80] and heart [229, 137] images.

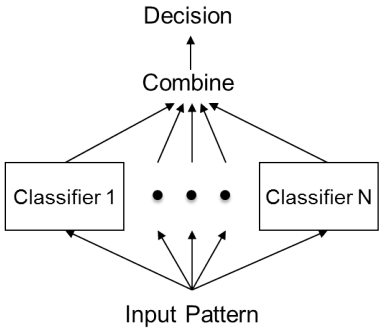


Figure 6: The concept of ensemble learning: an ensemble classifier is made up of several sub-classifiers, the final output is combined with outputs from these sub classifiers and their weights.

2.1.3.5 Neural Networks and Deep Learning

Deep learning is a relatively new paradigm in machine learning, which can learn effective features directly from the data for classification and detection purposes [238, 246]. Deep learning avoids designing specific features from the data, which is its main advantage in comparison with other machine learning methods. Some outstanding frameworks such as the restricted Boltzmann machine [227], convolutional neural networks (CNNs) [130] and sparse autoencoder have proven useful tools in many applications such as Alzheimer's disease diagnosis [261], segmentation [207], and tissue classification [54]. CNNs have a large number of parameters, which requires huge volumes of labeled training data. This requirement makes the training of CNNs from medical images difficult due to the difficulty of acquiring a database with labeled data [239]. However, several studies use CNNs to extract features for medical images and achieve good performance in classification [280, 46].

2.1.4 Evaluating Machine Learning Techniques

The goal of applying machine learning is to predict or classify diseases and produce useful results that the physician may rely on. Single training and testing on data sets may not yield a meaningful idea of the accuracy of an algorithm. Cross validation reduces the variance of accuracy scores by ensuring that each data instance is used for both training and testing

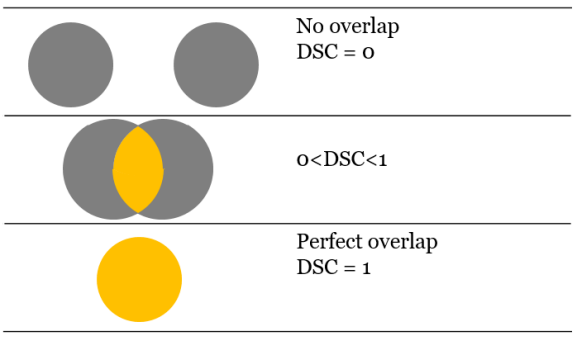


Figure 7: The dice similarity coefficient represents spatial overlap.

an equal number of times. Cross-validation method randomly splits data into k subsets and hold out each one while training on the rest.

The dice similarity coefficient is used in segmentation, as it measures the spatial overlap between two segmented target regions [318]. A and B are target regions or volumes, and the dice similarity coefficient is defined as the ratio of their intersection to the average [217]:

$$DSC(\mathrm{A,B}) = \frac{2(\mathrm{A} \cap \mathrm{B})}{\mathrm{A} + \mathrm{B}} \tag{2.3}$$

The dice similarity coefficient has a value of 0 for no overlap and 1 when perfect agreement is present. Fig. 7 illustrates the dice similarity coefficient with different overlap.

The goal of a computer-aided diagnosis system is to detect as many true positives as possible and minimize the detection of false positives. Sensitivity is defined as the proportion of patients labeled with diseases who are shown to have a disease. Specificity is the proportion of subjects labeled healthy who are tested healthy. They can be written as:

$$\text{sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2.4}$$

$$\text{specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \tag{2.5}$$

Some popular performance measures used to measure performance include the area under the receiver operating characteristic (ROC) and the top precision value. ROC curves describe

the relationship between sensitivity and specificity. The top precision value is defined as the portion of top-ranked relevant images out of all relevant database images [174, 143].

2.2 Application of Machine Learning in Radiology

2.2.1 Segmentation

The accuracy of segmentation in medical imaging applications affects diagnoses and treatments [83]. For example, MRI segmentation, such as tissue segmentation, helps to understand the progression and prognosis of diseases such Alzheimer's disease, Parkinson's disease, and multiple sclerosis. Medical image segmentation also plays an important role in many computer-aided systems and image retrieval systems [294]. Medical images contain many normal structures such as organs, bones, fat and muscles and abnormal structures such as fractures and tumors. The techniques applied in radiological image segmentation are specific to the type of body part, application, and clinical requirements [237]. In addition, the accuracy of medical segmentation suffers from several artifacts: noise, partial volume effects, bias field, insufficient resolution, anatomic variability, and complexity [182, 181, 271, 206, 289]. Table 2 is a summary of current studies in medical image segmentation.

Shape segmentation based on X-rays is seldom done due to difficulty in practice despite its usefulness. Typical X-ray segmentation includes landmark detection and shape regularization [148]. Chen et al. proposed to improve the prediction of individual landmarks by jointly estimating displacements from all patches and considering both the training data and geometric constraints on the test data [41]. They generated the shape contour using the sparse composition model for landmark position regularization.

Manual segmentation is the gold standard to determine the morphology of the brain region, however it depends on the experience of clinicians and is very time-consuming [228]. The whole-brain automatic classification methods are essential for improving the diagnosis analysis and for the reproducibility of large-scale clinical studies. Brain extraction from MRI is crucial in neuroimaging research. Kleesiek segmented the brain and non-brain tissues by

Table 2: Overview of segmentation methods for different radiological images

	image types	# images	goal	methods	Dice coefficients
[225]	MRI (T1-weighted)	12	Brain tissue	Sparse dictionary learning	0.91 (Gray matter) 0.87 (White matter)
[179]	MRI (T1-weighted, T2-weighted, fluid-attenuated inversion recovery and diffusion weighted)	36	Stroke lesion	Random forest	0.82
[22]	CT	42	Liver tumor	Random forests & supervoxels	0.93
[144]	CT	30	Liver tumor	Convolutional neural network	0.84
[152]	MRI	70	Knee	Multi-atlas context forests	0.97 (Bone) 0.81 (Cartilage)
[275]	CT	150	Multi-organ	Discriminative dictionary learning	0.90 (Liver) 0.88 (Kidney) 0.55 (Pancreas) 0.92 (Spleen)
[305]	MRI (T1-weighted, T2-weighted, diffusion-weighted)	10	Brain tissue	Deep convolutional neural networks	0.95 (Gray matter) 0.86 (White matter)
[224]	CT	82	Pancreas	Deep neural network	0.72
[84]	MRI (T1-weighted)	30	Stroke lesion	Gaussian naive Bayes classification	0.81
[240]	MRI (T2-weighted)	12	Brain lesion	Artificial neural network	0.79
[88]	MRI (T2-weighted)	66	Prostate	Sparse auto-encoder & sparse patch matching	0.88
[18]	MRI (T2-weighted)	45	Left ventricle	Convolutional neural network & stacked-auto-encoder	0.97
[117]	MRI (T1-weighted, T2-weighted)	53	Brain tumor	Convolutional neural network	0.95
[281]	CT	73	Lung texture	Convolutional restricted Boltzmann machines	0.74
[180]	MRI (T1-weighted, T2-weighted)	57	Brain segmentation	Convolutional neural network	0.86
[168]	4D-CT	22	Brain tissue	SVM	0.79 (Gray matter) 0.81 (White matter)
[90]	MRI (T1-weighted, T2-weighted)	65	Brain lesion	2 pathway convolutional neural network	0.79
[97]	CT	42	Liver tumor	Convolutional neural network	0.97

feeding data to a neural network with 7 convolutional hidden layers and one convolutional soft-max output layer [117]. They trained the network using stochastic gradient descent, the methods of which can be applied on any single image modality or combination of several modalities with varying size. Brain MRI segmentation done for lesion detection is a preliminary and important step in effective disease diagnosis and treatment. Mitra et al. proposed an automated method to segment ischemic lesions, white matter and other secondary lesions. They used Bayesian-Markov random field classification first for informative sampling of the lesion class both during the training and testing phases and then used random forest to refine the segmentation from the multimodal MRI data [179]. Maier et al. proposed an automated method to locate, segment and quantify the sub-acute ischemic stroke lesion from T1-weighted and diffusion-weighted sequence data [166], with their proposition based on the extra tree forest, which is well performed from noisy training data and robust against overfitting. However, their method can only deal with the T1-weighted and diffusion-weighted data sequences and high quality images. Griffis proposed a supervised learning method that automatically delineates stroke lesions using naive Bayes classification in single T1-weighted MRI sequence data [84]. Their approach focuses on using single scan data in order to save time and money, which detects direct lesion effects and has a better performance than manual delineation. However it showed limitations for subtle white matter lesions due to lack of image information. Si proposed a semi-automatic method to classify the pixels of brain MRI into lesioned and healthy tissues by use of an artificial neural network with gray levels and statistical features as input [240]. Their segmentation results show better performance than k-means. Yoo segmented multiple sclerosis lesions in multi-3D MR images from unsupervised features [302]. Features were extracted from T2-weighted and proton density MR images using a deep relief network and a random forest was built for the supervised classification. Roy used sparse dictionary learning that learns relevant patches from the atlas [225]. In their method, statistical priors are used to localize tissues with similar intensities. The segmentation of early-brain tissues is more difficult than adult brains due to the lower tissue contrast [142]. Multiple image modalities provide complementary information for insufficient tissue contrast [286]. Zhang et al. [305] showed that fractional anisotropy images are more powerful in distinguishing gray matter and white matter, and that T2-weighted images

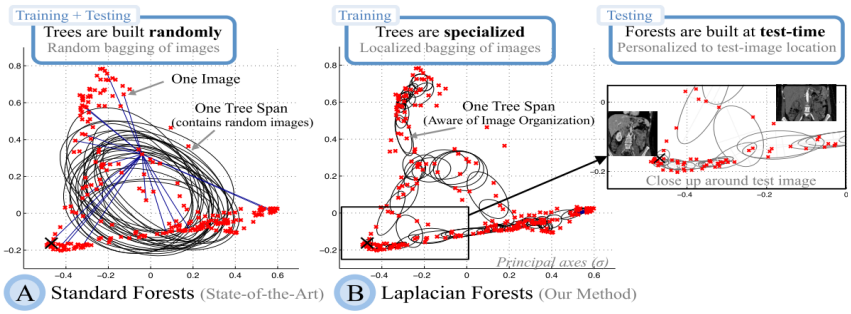


Figure 8: Instead of standard random forest, Laplacian Forests use guided bagging by creating subtrees with neighboring images on the Laplacian eigenmap. If the black cross is the test image, only neighboring trees are required for a test image [159].

have higher performance in capturing cerebrospinal fluid. They proposed a CNN method combining these multiple modality image data to improve segmentation performance.

Segmentation is also applied to identify other structures, such as organs, bones, muscles, and fractures. Lombaert et al. improved the random decision trees model by using guided bagging approaches for training images and non-uniform tree weighting during testing [159]. The use of a guided bagging strategy considers the affinities between images and, producing more related image information for tree model and have larger improvements in accuracy of kidneys segmentation, Fig. 8 compares the standard random forests and Laplacian forests. Conze proposed a semi-automatic liver tumor segmentation combining a simple linear iterative clustering super pixel algorithm and random forest, which considers the inter-dependencies among voxels [22]. The multi-phase cluster-wise features extracted in their approach are more robust for a random forest. Tong et al. proposed an automated method for multi-organ segmentation (liver, kidneys, pancreas, and spleen) using dictionary learning and a sparse coding technique [275]. The atlases selected against which to segment the images highly influence the performance of multi-based methods [5]. To deal with the high inter-subject variation in CT images, they applied a voxel-wise local atlas selection strategy to improve performance. The analysis of the knee also plays an important role in clinical assessment and surgical planning of the disease. The cartilage is typically small and the segmentation results of Haar-like operators are often unreliable in extracting context

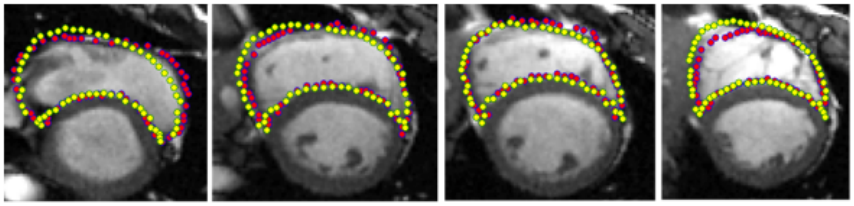


Figure 9: Sedai et al. proposed a shape regression method for right ventricle segmentation [232]. Their method more accurately segmented the right ventricle and outperformed the multi-atlas label fusion method. The yellow contour is automatic segmentation and the red contour is manual segmentation.

features. To overcome these limitations, Liu proposed a novel method using a multi-atlas context forest, which segments bones first and then cartilage [152]. They trained classifiers using appearance features and context features to align the expert segmentation of the atlases in each iteration.

Right ventricle structure segmentation in MRI is an essential task for investigating most cardiac disorders. The main challenge of this task is the large shape variation among different patients [232]. Sedai proposed a segmentation method using shape regression for the right ventricle in cardiac MRI. Their results are shown in Fig. 9. They applied gradient boosted regression to regress multidimensional right ventricle shape landmarks from image appearance, which consider correlations between landmarks. Their method minimizes the shape alignment error over training data and shows better segmentation performance than multi-atlas-label-fusion based segmentation methods.

Identification of intervertebral discs is an important process for diagnosis and operation planning of spine pathologies. An automated method to localize and segment intervertebral discs from MRI was proposed in [42]. They used unified regression and classification framework to estimate displacements for image points or voxels and achieved good results.

Image quality limits the extraction of features from the radiology images, and in many cases such as brain boundary segmentation, the data is by nature of low contrast and both resolution and partial volume effects influence the definition of boundaries [291]. Extraction of useful features from low quality images is one good way to handle this issue. Some

research focuses on different modalities to get complementary information, however it is difficult and inconvenient to apply different testing methods on patients. Multi-modality radiological approaches, including MRI, PET are acquired to provide complementary information for Alzheimer's disease diagnosis [108, 311, 274] and stroke segmentation [179]. In stroke segmentation, more information about the extent of infarcted territory and the anatomical location is found in diffusion-weighted [38], T2-weighted and fluid-attenuated inversion recovery are used to delineate the final lesion volume [293]. In brain segmentation, some approaches used multiple MRI modalities (T1-weighted, T2-weighted, fluid-attenuated inversion recovery, diffusion-weighted images) to achieve optimal performance [179, 166], however, due to time and monetary reasons, only a single anatomical scan is generally taken. In segmentation, the accuracy of the system is difficult to measure and compare, one reason being that the "ground truth" varies based on the guidelines specified for manual delineation by different experts [71]. Moreover, in medical analysis, accurate manually labeled data and high quality data is difficult and expensive to obtain [94].

2.2.2 Computer Aided Diagnosis

Promising results have been published dealing with computer aided diagnosis (CAD) systems in applications such as lesions [193, 300], epidural masses [149], fractures [297], as well as degenerative disease [272] and cancer detection. CAD is a software tool that can detect, mark, and assess potential pathologies for radiologists to help improve identification accuracy in the case of data overload and human resource limitation. The analysis, quantification, and categorization of images with these methods is an important technique, which can improve patient safety and care. Many researchers have shown promising CAD system results. However, to meet clinical requirements, the performance of systems still needs to be improved [135]. The main steps of CAD systems are the training phase and testing phase. Fisher's linear discriminant, Bayesian methods, artificial neural network, and SVM are widely used as classifiers in CAD applications [299, 298]. Table 3 summarizes some current CAD investigations with machine learning techniques.

Table 3: A summary of recent CAD studies.
AUC = area under curve; ROC = receiver operating cruve; TP = true positive rate; MAE = mean average error

	year	image type	# cases	disease	results	keywords
[202]	2014	mammography	956	Breast cancer	AUC:0.81	Combination of classifiers
[105]	2014	mammography	500	Breast cancer	AUC: 0.91	Naive Bayes classification
[276]	2014	MRI	81	Cervical cancer	Accuracy:0.69	Texture features, SVM
[263]	2015	mammography	340	Breast cancer	AUC:0.73	Texture features, SVM
[203]	2015	mammography	772	Breast cancer	AUC:0.89	Feature selection method
[288]	2015	CT	750	Lung	AUC:0.98	Structured SVM
[12]	2015	X-ray	5440	Lung	Accuracy:0.92	SVM
[81]	2015	MRI	83	Pediatric cardiomyopathy	Accuracy:0.81	Bayesian rule learning
[14]	2016	mammography	736	Breast cancer	AUC:0.82	CNN
[243]	2016	mammography	2604	Breast cancer	AUC:0.93	Adaptive wavelet neural network
[45]	2016	ultrasound	520	Breast lesions	Accuracy:0.82	Stacked denoising auto-encoder
[213]	2016	ultrasound	95	Liver lesions	Accuracy:0.87	SVM
[35]	2016	CT	104	Vertebral body fractures	TP:0.81	SVM
[66]	2016	CT	409	Wrist, radius, ulna fractures	ROC:0.89	Random forest
[121]	2017	mammography	45000	Breast cancer	AUC:0.906	CNN
[154]	2017	CT	1012	Lung cancer	Sensitivity: 0.89	ANN
[175]	2017	CT	52	Teeth	Accuracy: 0.888	CNN
[172]	2017	CT	344	Prostate cancer	ROC: 0.8	CNN
[248]	2017	X-ray	1391	Bone age	MAE: 0.8	CNN

Breast cancer is one of the most common cancers in the world. Currently, about one in ten women suffer from it, and early diagnosis and treatment of breast cancer could increase the chance of survival significantly [131]. Mammography, thermography, and ultrasound imagery are the most common techniques used to identify breast cancer [163]. Among these techniques, mammography is the best approach to detect breast cancer in its early stages and features indicating abnormalities can be extracted directly from medical images [192]. Masses and micro calcifications are two main indicators of breast cancer. The identification of benign and malignant masses is the core principle for using mammography as a means to diagnose breast cancer [106]. Perez et al. developed machine learning classifiers that combine suitable feature selection methods with different machine learning techniques [202]. The feature selection methods include chi-square discretization, information gain, one rule, relief and u-test based filter. Then, they improved their feature selection algorithm called uFilter, which ranks features in a descending manner [203]. Their method was effective for different datasets and reduced the number of employed features without decreasing the classification performances.

The SVM classifier is widely used in breast cancer diagnosis with different feature extraction methods, such as wavelet features, gray-level-co-occurrence matrix features, intensity

features, and some other texture features [264, 263]. Arevalo et al. trained an SVM model that integrated 1 and 2 layer CNN for supervised feature learning [14, 13]. Similarly, Jiao et al. trained two SVM classifiers using deep features extracted from two different layers of CNN networks [107]. The method they proposed is less time-consuming and uses less storage space. Xie et al. proposed a classification method based on SVM and extreme learning for the feature selection, and the use of extreme learning for classification reduced computational cost [295]. An automated CAD system was proposed combing the content-based image retrieval to detect masses in [105]. The main idea of their approach is to use scale invariant feature transform features to match query mammogram and exemplar masses in the database, and then uses naive Bayes classification and thresholded maps to detect masses. In their method, the computational cost is low because there is no sliding window-based scanning. A semi-supervised algorithm is proposed to deal with a large amount of unlabeled data with CNN approaches [262]. Their approaches using unlabeled data increased the overall accuracy, rather than just using labeled data.

Besides in breast cancer diagnosis, CAD is also widely studied in other diseases such as cervical cancer, lesion detection, traumatic spine and vertebral fractures. Torheim et al. used gray-level-co-occurrence matrices based textural features from dynamic contrast enhanced MRI as explanatory variables for SVM classification to predict cervical cancer [276]. Wang improved the performance of lung lesion detection from CT images by using 3D matrix patterns-based SVM with latent variables. Their study focused on detecting lung lesions that had irregular shape and low-intensity, rather than the nodules, which provides a new thought for detection of lung lesions [288]. Traumatic spine injuries are common, which are associated with neurologic deficits [11]. An accurate, rapid, and detailed injury diagnosis is important for the treatment decisions. Burns proposed a fully automated system that detects and localizes thoracic and lumbar vertebral fractures in CT images [35]. They extracted 28 features from the cortical shell as computed by an SVM classifier. This approach divided a large complex problem into small, modular pieces for fracture detection, which reduces program complexity. However, their approach relies on an essential element (Denis 'middle column') and is specific to detection of fracture discontinuities on vertebral body cortices.

There are many advantages to using machine learning techniques in CAD systems. The first advantage of machine learning is its accurate and robust performance in many radiology studies. In certain research, CAD systems have reached perfect accuracy e.g, over 99% in oral cancer detection [72], which is comparable to manual diagnosis. Moreover, CAD system can perform well and produce robust results with large amounts of data at any time and in any space. Manual diagnosis results may be affected by fatigue, reading time, and emotion on the part of the practitioner, while CAD systems perform more consistently than humans. The second advantage of applying machine learning is saving time. Many radiology analyses require experienced radiologists and are usually very time consuming. With the help of machine learning systems, the diagnosis can be finalized in a very short time. The software developed for breast cancer prediction [199] can review charts 30 times faster than humans can. With the help of machine learning, radiologists may no longer spend time on these time-consuming analyses. Another example is that the suggested approach in breast diagnosis is double reading of mammograms by two radiologists. However, the cost and workload are very high [202]. With the help of a CAD system, only one radiologist is needed instead of two, which could help to increase the survival rate among women in a cost-effective manner [19].

2.2.3 Image Retrieval

With the increased use of modern medical diagnostic techniques, there are numbers of medical images stored in hospital archives. Manual annotation and attribution of these images becomes impractical [102]. Picture archiving and communication systems have been widely introduced in many hospitals. However, most of these systems only contain textual information with limited functionalities [124]. To find similar patient cases, physicians usually look for images pathologically similar to the given image. Picture archiving and communication systems could retrieve images based on keywords, however these images may not be directly useful in helping to making clinical decisions. Systems based on semantic annotations also depend on the description of image content, which is linked to user experience and differs from experts to experts. Different from traditional image search systems, which

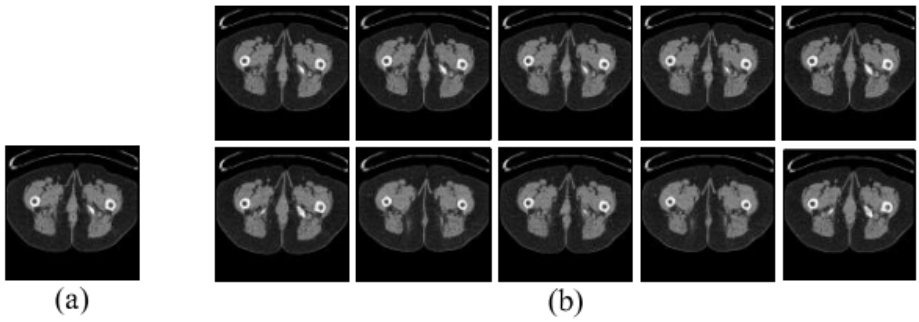


Figure 10: The method for retrieving images using Local wavelet pattern features and similarity measurement. All retrieved images are from the same category, achieving 100 % precision in this example [64].

are based on matching keywords and image tags, content-based image retrieval extracts rich contents from images and searches for other images with similar contents. Content-based image retrieval is becoming important for the medical image databases, which may potentially become efficient tools of anatomical and functional information for diagnostic, educational, and research purposes [292]. Table 4 lists current investigations on image retrieval. The main steps of content-based image retrieval are image visual extraction and use of similarity functions [95]. As an example, local wavelet pattern features are applied in Fig. 10 for different image category retrieval.

Recently, similarity or distance learning is a hot topic in the machine learning field, with traditional choices including the Euclidean distance function, x^2 square distance function, Mahalanobis distance, l_1 norm distance function [174], maximum likelihood approach [303] and Bayes ensemble [68]. Kurtz proposed an approach that includes evaluation of semantic features using hierarchical semantic-based distance and retrieves images based on semantic relations [125] Then, they extended this approach to a semantic framework that learns image description of each term using Riesz wavelets and SVM. Image based and semantic similarities using dissimilarity measurements are under consideration for use in describing the image content [126]. Their method can automatically annotate the content of radiology images, and the use of hierarchical semantic-based distance distance shows a lower computa-

Table 4: A summary of recent image retrieval research using machine learning techniques

	year	image types	# images	results	keywords
[126]	2014	CT	72	AUC:0.93	Riesz wavelets, hierarchical semantic-based distance
[75]	2015	MRI	30	Accuracy:0.88	Partial least square discriminant analysis, principal component analysis
[64]	2015	CT	EXACT09:40 TCIA: 604	Precision: 0.88	Local wavelet pattern
[37]	2015	Multimodality	ImageCLEF: 10 thousand	MAP:0.29	Deep Boltzmann machine
[284]	2015	MRI	OASIS:421	Precision: 0.48	Local binary patterns, gray-level-co-occurrence matrices
[251]	2016	X-ray & CT	ImageCLEF:5400	Accuracy:0.98	Sparse representation, online dictionary learning
[174]	2016	Multimodality	Indoor:15620, Caltech256:30670, Corel5000:5000, ImageCLEF:2785	Top precision: 0.36	Support top irrelevant machine

tional cost. Meng et al. proposed a novel similarity learning algorithm which focused on top precision and the l_2 norm, in which they considered a top precision performance measure in the loss function, which is different from traditional similarity learning that only maximizes the margin [174].

Several image retrieval systems are based on an online dictionary learning method. The main advantage of an online dictionary learning system is the computational time, as learned dictionaries are used to represent the dataset in a sparse model, which is an effective tool for representing data [211]. A method using online dictionary learning and its features extracted by multi-scale wavelet packet decomposition from different types of images is proposed in [250]. Srinivas et al. proposed a medical image classification approach using online dictionary learning with edge and patch based features to distinguish 18 categories [251].

Ahn developed a robust method to improve X-ray image classification [4]. A fusion strategy is proposed that combines domain transferred convolutional neural networks and sparse spatial pyramid classification. The combined method performs better than the single method used. Faria et al. proposed a retrieval method for brain MRI images. They captured anatomical features from T1-weighted images using least-square discriminant analysis and principal component analysis and performed a search for images between healthy controls and patients with primary progressive aphasia [75].

Besides the existing steps, semi-supervised and unsupervised learning methods were developed for image retrieval as well. An unsupervised image retrieval based on clustering method using K-SVD is proposed in [252]. The main idea of this approach is to execute iterations between grouping similar images into clusters and generating a dictionary for clusters until clusters converge. The advantage of their method is that it requires no training data for classification and is not restricted to a specific context. As labeled data is limited, Herrera proposed a semi-supervised learning method for image classification using k-nearest neighbors to expand the training data set and a random forest for final classification [94].

2.2.4 Brain Functional Studies and Neurological Diseases

The majority of brain related studies include two main steps: (1) extraction and selection of features from medical images such as MRI (2) designing a supervised classifier for the different prediction and classification stages. In brain image diagnosis, a large number of features can be extracted from brain regions related to the nature of pathological changes. However, it is challenging to design an effective classifier with these features [150, 47]. Cortical thickness [118], volume of brain structures [48], and voxel tissue probability maps around certain regions of interest [74] are popular choices for feature extraction [43]. Different MRI modalities such as T1-weighted or fluid-attenuated inversion recovery imaging contain huge amounts of information, which may be noisy and too large of inputs for a supervised classification [240]. Furthermore, not all image features are useful for the specific classification, and for the limitations of a data set, therefore using all features may influence the performance of a single global classifier. Therefore, an effective feature fusion strategy is necessary and important for neuroimaging analysis and classification.

Brain metastasis is one of the most common forms of brain tumor and application of multi-parametric MRI and PET images is a popular method to differentiate metastatic from radiation necrosis [114, 127]. Larroza et al. developed a classification model of brain metastasis and radiation necrosis in contrast-enhanced T1-weighted images. Features were extracted by texture analysis and reduced by using a linear SVM model that provided better performance for classification [127]. Ahmed applied an automated approach to detect neocortical

structural lesions, which contained five surface-based MRI features and combined them in a logistic regression [2]. To deal with imbalance issues, they used a "bagging" approach and an iterative-reweighted least squares algorithm. The base-level classifier was trained on all the minority class instances and the same size of random data from majority class instances.

Focal cortical dysplasia leads to abnormality of the structure of the cerebral cortex, which is a common cause of epilepsy [30]. Hong proposed a machine learning technique combining surface-based analysis in patients with a subtype of focal cortical dysplasia [96]. Their automated approach used features of Focal cortical dysplasia morphology and intensities, Fisher's linear discriminant was applied as a classifier to identify Focal cortical dysplasia in patients. In recent years, various machine learning methods have been designed for identification of clinical status and analysis of complex patterns in neuroimaging data [55].

Neurodegenerative diseases such as Parkinson's disease and Alzheimer's Disease are widely studied with the support of machine learning. The neurodegeneration begins before the onset of diseases; medical treatment is more effective if it is detected in early stage.

Parkinson's disease is the most common degenerative movement disorder and its diagnosis in early-stage is still a challenge [204]. Among the various forms of Parkinsonism, progressive superanuclear palsy is one of the most difficult to be identified from Parkinson's in early disease stages [79]. Salvatore et al. proposed a supervised method to classify control subjects, progressive supranuclear palsy patients, and Parkinson's disease patients with features extracted by principal component analysis from T1-weighted sequences and SVM. The accuracy of discrimination of Parkinson's disease and progressive supranuclear palsy is above 90% [228]. Fig. 11 uses color scale to express the importance of each region during classification. To improve the performance of classifying Parkinson's disease patients, Singh proposed an unsupervised feature extraction method from a T1-weighted sequence by using a Kohonen self-organizing map algorithm. With the least square SVM, the accuracy of identifying the affected area in Parkinson's disease is up to 99% [242].

Alzheimer's disease is estimated to affect around 5.4 million patients in America, and is the most common form of dementia among the elderly population [9, 272]. Alzheimer's disease leads to the loss of cognitive function and death in elderly people. Liu proposed a classification framework that works on different image modality for the classification of

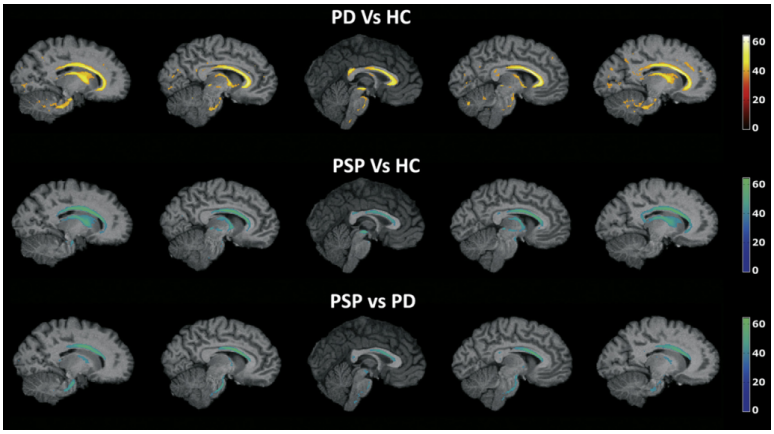


Figure 11: Salvatore et al. [228] proposed a supervised learning method to identify PD and PSP using MR images. The figures show maps of voxel-based pattern distribution of brain structural differences. The color scale express the importance of each voxel in SVM classification.

Alzheimer's disease patients [150]. Their method contains level classifiers: low-level classifiers that use different types of low-level features from patches, high-level classifiers that combine coarse-scale imaging features in each patch and outputs of low-level classifiers, as well as a final ensemble classification that combines the decisions of a high-level classifier with a weighted voting strategy. Their algorithm structure is shown in Fig. 12. Zhu et al. focused on the identification of Alzheimer's disease patients with multi-view or visual features of image data. They proposed several feature selection approaches for Alzheimer's disease classification. They integrated subspace learning into a sparse least square regression framework for multi-classification in 2014 [312]. Then, they mapped the histogram of oriented gradient features (which are diverse) onto a region of interest features (which is robust to noise), which provided complementary information for features and enhanced disease status identification performance [313]. Bron proposed a feature selection method based on the SVM significance value [34]. The significance value (p-value) serves to quantify the contribution of each feature to the SVM classifier and is used to reduce features. Chen developed a framework using patch extraction and a deep network based feature through a stacked denoising sparse autoencoder, which makes the input data points more linearly separable in SVM [43]. The diagnosis of Alzheimer's disease patients and its early stage,

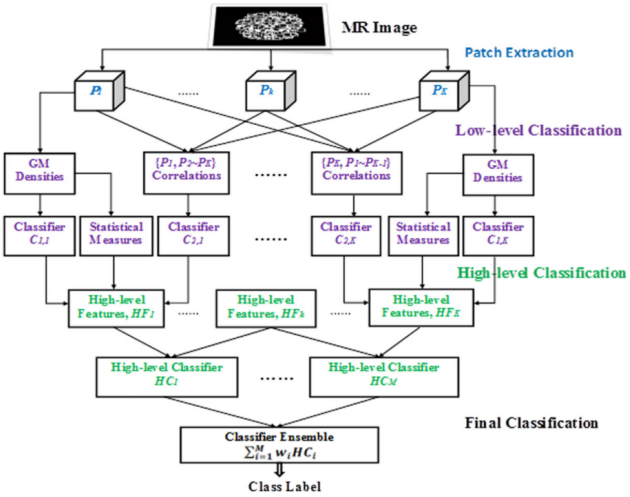


Figure 12: Flow chart of the hierarchical classification algorithm proposed in [150], the low-level classifiers are used to transform imaging and spatial-correlation features from the local patch, and the output of these low-level classifiers is integrated into high-level classifiers with coarse-scale imaging features. The final classification is achieved by ensemble outputs from high-level classifiers.

mild cognitive impairment is important for treatment. In the early stages of Alzheimer's disease, it is difficult to predict whether mild cognitive impairment subjects will progress to Alzheimer's disease in the future. Liu proposed an inherent structure guided multi-view learning method to classify Alzheimer's disease and mild cognitive impairment patients [151]. They extracted 1500 features from gray matter density and multi task feature selection was applied to reduce the dimension, followed by an ensemble classification method using multiple SVM classifiers. Some researchers are interested in mapping and reducing features by combining a lasso regression and principle analysis component. Huang proposed to use a soft-split random forest to predict clinical scores in Alzheimer's disease patients [99]. In their method, lasso regression is applied to map MRI features and then features are reduced by principle component analysis. Li combines principle component analysis, the lasso method, and a deep learning framework to extract features by fusing information from MRI and PET images and classified Alzheimer's disease/mild cognitive impairment patients by SVM [141]. A method using correlated information from different types of data was proposed in [44]. They developed a multimodal multi-label feature selection method based on a sparse

multi-label group Lasso method to capture informative features from multi-domain data and multimodal regression and classification to predict clinical scores in Alzheimer's disease patients. In [141], high accuracy results were obtained from Alzheimer's disease/healthy and mild cognitive impairment/healthy classification. However, accuracies in classifying mild cognitive impairment as converted to Alzheimer's disease are very low (57.4%), which is little higher than majority classification. Komlagan developed an ensemble learning method using gray matter for a weak classifier and selecting the most relevant sub-ensembles through sparse logistic regression [119]. They trained a global linear SVM classifier for the final classification. Combining high quality biomarkers with advanced learning methods makes results comparable to those of multi-modality methods. Tab. 5 summarizes recent research on Alzheimer's disease classification.

Besides the disease studies, some research work applied machine learning techniques to understand the brain's functional network architecture. Smyser compared the fMRI data from 50 preterm-born and 50 term-born infants using SVM [245]. Their results show that inter and intra hemispheric functional connections throughout the brain are stronger in full-term infants. Their findings might be helpful for the development of models for defining indices of brain maturation.

Table 5: Recent studies on Alzheimer's diseases. NC: normal; AD: Alzheimer's disease; pMCI: progressive mild cognitive impairment; sMCI: stable mild cognitive impairment

	year	databses	image #	image types	classification gruops	accuracy	keywords
[274]	2014	ADNI	834	MRI	AD vs. NC pMCI vs. sMCI	89% 70%	Multiple instance learning
[312]	2014	ADNI	202	MRI+PET	AD vs. MCI vs. NC AD vs. pMCI vs. sMCI vs. NC	73.35% 61.06%	Sparse discrimination feature selection
[85]	2014	ADNI	1071	MRI	AD vs. NC pMCI vs. sMCI	89% 73%	Manifold and transfer learning
[119]	2014	ADNI	814	MRI	pMCI vs. sMCI	75.6%	Gray matter grading, weak-classifier fusion
[151]	2015	ADNI	459	MRI	AD vs. NC pMCI vs. sMCI pMCI vs. NC	93.83% 80.9% 89.09%	Hierarchial fusion of features
[44]	2015	ADNI	202	PET+ MRI	pMCI vs. sMCI AD vs. NC	78.7% 91.31%	Multimodel multi-label transfer learning
[313]	2015	ADNI	830	MRI	MCI vs. NC pMCI vs. sMCI	78.07% 75.54%	HoG mapping
[113]	2016	OASIS	416	MRI	AD vs. NC	80.76%	Gabor filter
[230]	2016	ADNI	416	MRI	AD vs. NC vs. MCI	89.1 %	CNN
[161]	2016	Self-collected	67	MRI	AD vs. NC	96.77%	SVM
[16]	2016	Dartmouth College	116	MRI	AD vs. NC	97.14%	Feature ranking selection

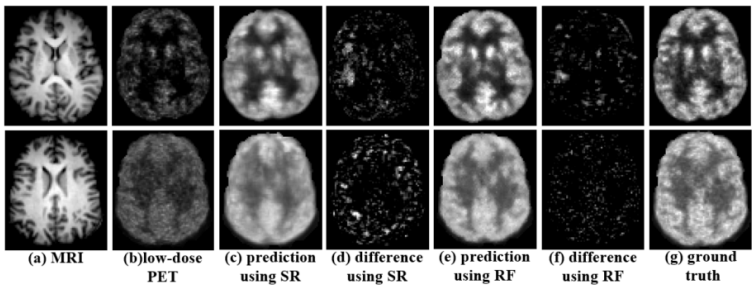


Figure 13: A new method using a regression forest based framework to predict standard-dose PET images [110]. The figures compare their new method and sparse representation method on two different subjects in the first and second rows. The new method outperforms the sparse technique in this compassion.

2.2.5 Image Registration

PET is a molecular imaging technique which is widely used in clinical cancer diagnosis. It produces 3D images, which can reflect tissue metabolic activity in the human body [221]. Low-dose PET images are widely used in clinical applications. However, the image quality is proportional to the dose injected and imaging time. Thus, a great deal of effort has been made to improve PET image quality.

Kang proposed a regression forest based approach to predict standard-dose PET images from low-dose PET and multimodal MRI images, [110], their results are shown in Fig. 13. They used a regression forest as their non-linear prediction model and features from local intensity patches of MRI data and low-dose PET. Meanwhile, Wang used a mapping-based sparse representation approach for prediction [290]. They used a graph-based distribution mapping method to reduce the patch distribution differences between MRI and low-dose PET and constructed a patch selection based dictionary learning method to predict standard-dose PET. Both methods performed better when compared with a path-based sparse model. Huynh predicted CT images from MRI data using a structured random forest instead of a classical random forest [277]. A structured random forest is an extension of a random forest, which predicts structured output instead of scalar outputs [120, 63]. Characterizing the information obtained from multiple sources improves prediction of performance.

3.0 Areas of Investigation

Our literature review of machine learning techniques has noted several important application fields related to radiology imaging.The majority of researchers focus on imaging data associated with diagnosing the brain, breast, lung. Studies of videofluorosocpic image remains undeveloped. We identified six key areas of investigation which we feel would be beneficial to the swallowing study field. The following sections present each topic, explain the reasons of importance, and explain the strategies that will be used.

3.1 Association between Hyoid Bone and Penetration / Aspiration

3.1.1 Motivation

Hyoid bone displacement influences epiglottis inversion, laryngeal elevation, and cricopharyngeal muscle opening, which is considered important for the penetration and aspiration [101]. Some work has been done to investigate the factors which influence the penetration-aspiration scale [279]. These factors are independent controlled variables, including age, swallow position, volume and viscosity. However, the investigation between penetration-aspiration and hyoid bone, one of key component in VFSS, is still limited. In order to better understand the swallowing and the factors influencing its function, we must first study hyoid bone motion which may be one of factor that has association with swallow mechanism.

3.1.2 Plan of Action

The first step is to manually collect information of hyoid motion during each swallow process. In the VFSS video sequence, experts manually checked the beginning and ending frame of each swallow. We define the moment when hyoid bone moves upward as the beginning time and the moment when the hyoid bone arrived at the lowest position as the ending time. In each frame, we marked the location of 2nd, 3rd and 4th cervical vertebrae.

Furthermore, we marked the anterior and posterior part of the hyoid bone. The features containing the clinical meanings were extracted for further investigation. What remains to be done is to apply relevant statistical analysis techniques to determine which features from hyoid bone motion show importance. Factors such as the maximum displacement of hyoid bone and the average speed of the motion is included in the analysis. Additional factors such as subject's age, gender, the viscosity of the bolus, the volume of the bolus, the head position during the swallowing is also added in the investigation as necessary.

3.2 Prediction Penetration-Aspiration Scale based on Hyoid Bone Motion

3.2.1 Motivation

The detection of penetration and aspiration is important for the daily clinical settings. Knowing whether the hyoid bone motion can be used to predict penetration and aspiration would still be beneficial to clinical examination activities. In the previous investigations, we studied important hyoid bone motion features associate with the penetration and aspiration event. In this study, we are going to determine whether the hyoid bone motion features can be applied to the prediction of penetration and aspiration event.

3.2.2 Plan of Action

Data from past studies is available and each swallow is marked by the penetration-aspiration scale. Similar to the previous studies, the factors such as patient's gender, age, head position, the volume of the bolus and the viscosity of the bolus is considered as the additional independent variables in the model. The statistical model is built based on the hyoid bone motion features and these additional variables, to investigate whether there are ways to predict the level of penetration/aspiration.

3.3 Identification and Localization of Hyoid Bone in Videofluoroscopy

3.3.1 Motivation

The hyoid bone is one of the key components that clinicians examine in the videofluoroscopy study. During the examination, researchers usually manually annotate points of interest frame by frame to have the location information. This process is very time-consuming and it causes inter and intra reliability variation among the landmarks annotated by different examiners. A detailed and quantitative system is required to help examiners make a quicker decision. Several tracking methods have been proposed in previous contributions, however, none of them can annotate automatically, and researchers have to manually annotate points of interest in the first few frames and check these points work well on the following video sequence. It is necessary to develop a method to automatically recognize the location of the hyoid bone. If such a method can be established, it could provide a great convenience for dysphagia assessment.

3.3.2 Plan of Action

We have data from 266 patients who performed several swallows during examination. The data annotation process has been done as a part of the previous study, and the annotations of the hyoid bone in the VFSS image are available for use. The goal of this study is to use object detection methods to identify and locate the position of the hyoid bone. As this is the first study working on the hyoid bone segmentation, we examine the performance of state-of-art methods which have achieved good performance in the object detection and computer vision field. The popular methods include single shot multibox detection (SSD), you just look once (YOLO), and Faster-RCNN. These methods are built based on pre-trained deep neural networks, including VGGNet, ResNet and ZFNet, which are powerful techniques extracting the features from images. For each method, the mean average precision will be calculated and then compared for different image conditions.

3.4 Automatically Annotation for Vertebrae

3.4.1 Motivation

For a long time, medical experts have had to manually measure parameters and points of interest in VFSS images. The disadvantage of this work is obvious: time-consuming, imprecise and subjective. With increase use of VFSS studies, these heavy annotation tasks require more manpower, becoming the obstacle of limited diagnosis resources. If points of interest are marked automatically by the computer-assist system, it will be easier for clinicians and experts to deal with large numbers of diagnosis daily in the clinical and research practice. Vertebrae annotation is an important process during VFSS studies. Experts usually take the 2nd and 4th vertebrae as references to measure the different key components during one swallow. For example, the line through the tail of c2 and c4 is usually set as y-axis of the new coordinate and the hyoid bone motion can be measured and adjusted based on the new coordinate, which allows further investigations. Several contributions worked on vertebrae region segmentation for the vertebrae related disease diagnosis. The work on annotating the point of vertebrae is limited and requires further investigation.

3.4.2 Plan of Action

Data has been collected and the location of vertebrae has been annotated. Several deep learning network have achieved outstanding performance for point and line detection in natural images such as the face and street scene [116, 218]. However, few of them were applied on the medical dataset as medical images have a set of challenges to overcome. Here, we showed that a novel machine learning algorithm can with high accuracy automatically detect key anatomical points needed for a routine swallowing assessment in real-time. We trained a novel two-stage convolutional neural network to localize and measure the vertebral bodies using 1436 swallowing videofluoroscopy from 265 patients. We compared the model performance on C2,C3,C4 points detection between one-stage model and two stage model. In addition, We compared the reliability between model and human raters, and showed a high reliability not only on our five-pixel errors, but also on reliability score.

3.5 Automatic Measurement of Residue Scale

3.5.1 Motivation

The vallecular residue is considered to be an indicator of OPD and is generally included in assessment. The vallecula is a bilateral space between the base of the tongue and the epiglottis. Material may be retained in these spaces due to suboptimal contact between base of tongue and posterior pharyngeal wall and reduced lingual propulsion force generation [282]. This type of residue may lead to an increase in risk of aspiration because it may enter the airway directly once protective mechanisms return to baseline after the swallow. Several methods have been developed and investigated to evaluate the post-swallow residue in VFSS. However, the major of these methods rate pharyngeal residue based on observation; $\%(C2{-}C4)^2$ measurement scale is one of two well-established, quantitative scales of vallecular [183, 200, 258]. Currently, researchers use image analysis software tools, such as ImageJ, to implement this quantitative residue measurement. As this scale is a novel measurement, no previous literatures have worked on it on the semi-automatic or automatic estimate this scale. In addition, the judgment made by researchers are time-consuming and subjective. It would be important to develop an algorithm which can measure $\%(C2 - C4)^2$ measurement scale automatically based on the frame we provided.

3.5.2 Plan of Action

We selected the qualified frames that contains the vallecular residue after the each swallow. Then the trained expert manually annotated the residue area using Matlab and calculated their corresponded $\%(C2 - C4)^2$ measurement scale. There are several popular segmentation networks used in various scenario such as medical applications, street view applications and sports. In this study, we compared the performance of four state of art networks in the field, including U-Net, ATTU-Net, SQ-Net, and SegNet. In this study, we focus on comparison of the reliability between human raters and model predictions. We also combined the our previous vertebrae landmark localization networks to estimate $\%(C2 - C4)^2$ measurement scale.

4.0 Association between Hyoid Bone and Penetration / Aspiration

The content of this chapter is currently under review with SN APPLIED SCIENCES. Zhenwei Zhang, Atsuko Kurosu, James Coyle, Subashan Perera, and Ervin Sejdić. A generalized equation approach for hyoid bone displacement and penetration-aspiration scale analysis. 2021

4.1 Motivation

We sought to investigate the motion of the hyoid bone by analyzing trajectory features during swallowing in 265 patients with dysphagia. We are looking at not only kinematic motions but also mathematical features of the displacements in order to determine whether there are relationships between characteristics of hyoid bone trajectory and a score on the penetration and aspiration (PA) scale [223]. We hypothesized that hyoid trajectory features would differentiate between normal PA scores (score of 1-2) and abnormal PA scale scores (scores of 3-8). A generalized estimate equation model was built to test our hypothesis based on trajectories extracted from VFSS images during various swallowing tasks. If these findings are confirmed, the analysis of hyoid trajectory patterns would be a useful additional component to judge the nature of penetration and aspiration in patients with dysphagia, and inform clinicians as to appropriate interventions to restore more normal HLC displacement during swallowing.

4.2 Methods

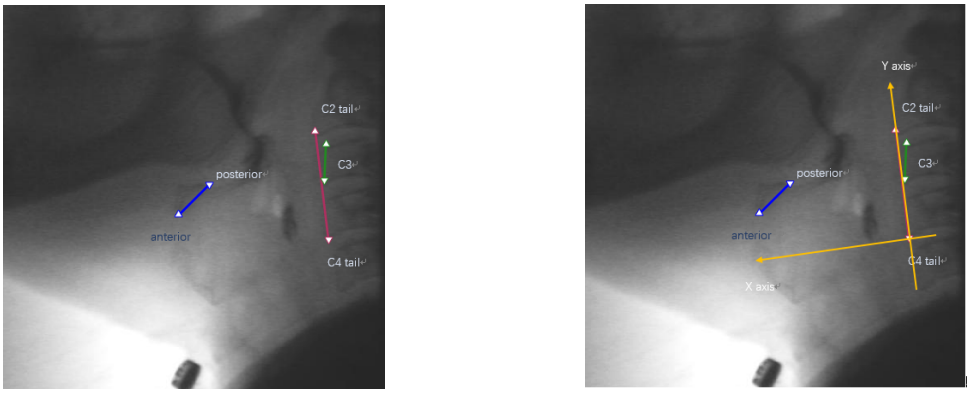
4.2.1 Data Acquisition

265 patients with clinical suspicion of dysphagia underwent videofluoroscopic examination at the Presbyterian University Hospital of the University of Pittsburgh Medical Center (Pittsburgh, Pennsylvania) in the study. The protocol for this study was approved by the Institutional Review Board at University of Pittsburgh and all participants provided informed consent. The average age of the subjects was 64.833 $\pm$ 13.56 years old, and the age range was from 19 to 94. There were 48 patients with stroke and 217 patients with non-stroke etiology. Patients with tracheostomy or anatomic disruption or abnormalities of the head and neck were excluded. Patients swallowed boluses of liquids of different consistencies and volumes as well as cookies during VFSS. The number and order of the swallow trials for each consistency and volume were determined by the examining clinician based on the patient's history and clinical evaluation observations. These liquids included thin liquid barium (Varibar Thin Liquid with < 5 cPs viscosity) and nectar-thick liquid (Varibar Nectar with about 300 cPs viscosity). 1136 swallows were evaluated in the lateral/sagittal plane with neutral head position though 252 swallows were performed in a head-neck flexion position (chin down). Patients swallowed boluses administered by a spoon in 3-5mL volumes, or self-administered boluses from a cup at a self-selected, comfortable volume.

Fluoroscopy was set at 30 pulses per second (full motion) and video images were acquired at 60 frames per second by a video card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA) and recorded to a hard drive with a LabVIEW program. The videos were made into two-dimensional digital movie clips of 720 x 1080 resolution, and in our study, we down-sampled it to 30 frames/second to eliminate duplicated frames.

4.2.2 Image Analysis

Each video sequence contained one swallow which was defined as the duration between the frame at which the head of the bolus reached the lower mandibular margin to when the tailing end (tail) of the bolus passed the upper esophageal sphincter (UES). The anterior and



(a) Markers for anterior hyoid bone, posterior hyoid bone, C2 tail, C4 tail and C3

(b) Coordinate established based on the C2-C4

Figure 14: The figures illustrates the markers for hyoid bone, C2, C3, C4 and how to establish the coordinate for hyoid bone trajectory

posterior projections of the body of the hyoid bone were marked in each video frame using MATLAB (R2015b, The MathWorks, Inc., Natick, MA, USA), as shown in Fig. 14. To create the coordinate system which normalized the motion points from different subjects of different sizes, we defined the border of anterior-inferior of the fourth cervical vertebral body as the origin, and defined the straight line connecting this origin and the anterior-inferior corner of the second cervical vertebra as the y-axis. The straight line perpendicular to the y-axis and intersecting with the origin is defined as the x-axis [184]. All distance numbers were measured as the actual distance in image pixels. In order to normalize participants of different heights to a common anatomic referent, the distance between the anterior-inferior and the anterior-superior corners of the third cervical vertebra is set as our reference scale, which we refer to as the length of C3. Then we used the length of C3 to normalize the coordinate.

Blinded to the hyoid trajectory results, the presence/degree of penetration/aspiration from the 1434 swallows were identified by using an 8-point PA scale another trained judge [223]. Among these swallows, 1129 swallows have PA scores of 1 or 2 and 304 swallows have PA scores greater or equal to 3. The mean and standard deviation of PA scores of these subjects are 2.117 ± 1.580.

To evaluate the reliability of the swallowing analysis, 100 swallow cases were utilized. Experts analyzed the same cases and their results are compared. Furthermore, the inter-rate reliability were tested as well, experts analyze the same case several months later to ensure less difference among different markers and time period.

4.2.3 Feature Extraction

We constructed six discrete series to represent the change of hyoid bone motion trajectory: the changes along the x- and y axis of the anterior/ and posterior inferior margin landmark of the body of the hyoid bone, the changes along the y-axis of the anterior and posterior margin of hyoid bone, and the distance series of anterior superior/and posterior margin of hyoid bone. The distance series was constructed from the Euclidian distance between every point and the starting points. This series shows how consecutive points move closer or farther from the reference point. The distance series can be written as:

$$D_i = \sqrt{\sum_{j=1}^{2}(X_{ij} - X_{0j})^2} \tag{4.1}$$

In our investigations, independent variables of each VFSS examination such as patient's age, bolus viscosity and size based on whether spoon or cup was used to administer the bolus, were used. Furthermore, to capture the key statistical differences between series, several features were extracted. Each of the features are described in the following subsections.

- length of the series x
- mean of the series $\bar{x}$
- number of values in x that are lower/higher than $\bar{x}$
- distance of the longest consecutive subsequence in x that is smaller/larger than $\bar{x}$
- minimum/maximum number in x
- the position at which the minimum/maximum number occurs in the series
- median of the series
- standard deviation
- duration length of series between minimum point to maximum point

• the sum over the absolute difference between subsequent time series values:

$$\sum_{i=1}^{n-1} |x_{i+1} - x_i| \tag{4.2}$$

• mean of the absolute value of consecutive changes in the series x:

$$\frac{1}{n} \sum_{i=1}^{n-1} |x_{i+1} - x_i| \tag{4.3}$$

• skewness quantifies how symmetrical the amplitude distribution is, this feature can be computed as follows:

$$\frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2]^{\frac{3}{2}}} \tag{4.4}$$

• kurtosis measures whether the distribution is peaked or flat relative to a normal distribution, it can be expressed as follows:

$$\frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2]^2} \tag{4.5}$$

4.2.4 Statistical Analysis

A generalized estimating equations (GEE) model is popularly applied for clustered data in clinical studies. It is an extension of quasi-likelihood approach [89]. The method was employed to construct a function of the feature set to match the outcome. The data are assumed to be dependent within subjects and independent between subjects. This model is quite useful with longitudinal data, which account the correlations between repeated measures on the same participant [316]. A GEE model assumes a relationship between $E(Y)$ and $Var(Y)$ rather than a specific probability distribution for Y. A GEE model provides a best guess for the variance-covariance structure $(Y_1, Y_2, ..., Y_T)$ by a linear predictor linking each marginal mean [244]. Y_{it} represents the category for each subject i, measured at different time points t. The working correlation matrix is applied to make a guess for the correlation structure among Y_t. Exchangeable correlation structure is applied here, which is a useful structure when correlations are small, which treats $Corr(Y_i s, Y_i t)$ as identical for all pairs s and t. The GEE model assumes a probability distribution for each marginal distribution and

provides reasonable estimates and standard errors. The GEE model estimates are obtained by using an iterative algorithm as there are no closed-form solutions.

In our studies, PA scores have a skewed non-normal distribution and our data consists of multiple swallows from each participant, making common statistical techniques such as (generalized) linear models and classification/regression trees not readily applicable. Therefore, we employed GEE model with low (1-2) or high (3-8) PA scores as the dichotomous dependent variable, a binomial distribution, a logit link function and an exchangeable working correlation structure to predict the probability of a high PA score. Age, swallow type, viscosity, volume/utensil, and head position were used in the model as independent variables based on face validity. In addition, we used an independent variable forward selection approach to identify a parsimonious set of trajectory variables using a criterion of p=0.05 entry into the model. Using the final model, we obtained odds ratios, and their 95% confidence intervals and statistical significance for each independent variable. Also, to assess the concordance between predicted and observed high PA scores, we created subgroups of swallows based on the predicted probability deciles, and examined the actual observed percentage of high PA score swallows within each decile. SAS® version 9.3 (SAS Institute, Inc., Cary, North Carolina) was used for all statistical analyses with GENMOD procedure for obtaining the main results.

4.3 Results

The generalized estimating equation is built to estimate the relation between various features and PA scores. Table 11 shows the clinical information of the patients and swallows. In the examination, subjects swallow different volumes of food and clinicians changed the viscosity and patient's head position in the examination. That explains the unbalanced data for the viscosity and head position. Patients usually start by swallowing the smaller thin liquid bolus with neutral head position and change to other viscosity depending on clinical need. Table 7 provides an overview of the contribution of important variables with entry criterion 0.05, based on model estimate, odd ratio, and p-value. The independent

Table 6: Clinical information of the patients and swallows. multiple(1) indicates the first swallow in the multiple swallow and multiple(2) indicates the subsequent swallows.

Age	**64.83±13.56**	**Total swallows**	**1434**
Gender		viscosity	
male	155	thin	879
female	110	nectar	405
Utensil		pudding	94
spoon	594	cookie	42
cup	832	not recorded	14
not recorded	8	Type	
Head position		single	498
neutral	1136	multiple (1)	360
chin down	252	multiple (2)	534
not recorded	46	not recorded	42

characteristics forced into the model, regardless of their p-value, are basic information data: age, swallow type, viscosity, utensil, sex, head position and swallow duration. Patients may have multiple swallows during the examination when some of bolus remains in the oral cavity or pharynx after first swallow. We indicate multiple(1) for the first swallow and multiple(2) for the following swallows. Table 7 indicates the important features with a p-value less than 0.05, providing strong contributions to the model related to the PA score. Patient and swallow condition independent variables of older age, first multiple swallow, and thin liquid viscosity, were significantly associated with higher PA scores, and the hyoid horizontal displacement independent variable was also significantly associated with PA scores .

Table 7: Generalized equation model with forward selection with 0.05 entry criterion

Parameter	Estimate	P value	Odds ratio	Odds ratio lower	Odds ratio higher
age	0.0265	0.0178	1.03	1.00	1.05
type: single	-0.4435	0.0708	0.64	0.40	1.04
type: multiple(1)	0.4545	0.0040	1.58	1.16	2.15
type: multiple(2)	0.0000	.	1.00	1.00	1.00
sex: male	0.1398	0.6998	1.15	0.57	2.34
sex: female	0.0000	.	1.00	1.00	1.00
viscosity: thin	1.2862	0.0096	3.62	1.37	9.58
viscosity: nectar	0.7049	0.1664	2.02	0.75	5.49
viscosity: pudding	-0.5334	0.3789	0.59	0.18	1.92
viscosity: cookie	0.0000	.	1.00	1.00	1.00
utensil: spoon	0.1622	0.3538	1.18	0.83	1.66
utensil: cup	0.0000	.	1.00	1.00	1.00
head position: neutral	0.0994	0.7104	1.18	0.65	1.87
head position: chin down	0.0000	.	1.00	1.00	1.00
swallow duration	-0.0004	0.9549	1.00	0.99	1.01
maximum displacement in horizontal direction	-0.0583	0.0064	0.94	0.90	0.98

4.4 Discussion

In the present study, we sought to investigate whether there is any relationship between hyoid bone displacement features and examination condition variables on airway protection as measured by the PA scale. We evaluated not only the maximal distance and velocity of the hyoid bone, but also other features extracted from the trajectory of hyoid bone. We focused the information, such as age, bolus volume, swallow type (single/multiple), and head position as the necessary variables in the GEE model, and we used forward selection to choose the important variables for the model prediction. Our results demonstrated that the hyoid bone displacement features were significantly related to PA scores. First, we will discuss the significant trajectory features related to PA scores. Then, we will discuss and compare the findings of basic variables with other contributions.

We tested the features extracted from the motion of the hyoid bone and variables such as age, bolus volume, viscosity, and head position. From the GEE model, we found that the maximum displacement of anterior-inferior hyoid bone has significant relation to the PA score: the decrease of the displacement will lead to higher PA score. Other features extracted from the hyoid bone displacement didn't show any significant association with the PA score. Our results agrees with the Steele *et al.* study that indicated that occurrence of higher PA scale scores were found in swallows with reduced anterior hyoid movement when the hyoid

displacement measurements were normalized by the C2-C4 distance [256]. On the other hand, our results do not agree with the Kim *et al.* study that reported there was no difference on the maximum anterior displacement of hyoid between aspirators and non-aspirators in patients with stroke [115]. The results do not agree with the Molfenter *et al.* study that reported no difference on hyoid displacements between aspirators and non-aspirators in patents with stroke with the anatomically normalized units [184]. Seo *et al.* also indicated there was no relationship on the actual and normalized hyoid displacements between stroke patients with and without penetration/aspiration [233]. Steele *et al.* suggested the C2-C4 vertebral distance should be used to normalize the hyoid displacement measurements in order to account for individual anatomical size differences [257]. The anatomically normalized unit C3 was used in the Kim and McCullough study, and they tracked the superior-anterior of the hyoid bone, this methodological difference may explain the discrepancy in our results and the study by Kim and McCullough. However, both the Molfenter and Steele and Seo *et al.* studies used the normalized measurements. This discrepancy may indicate the variability in the hyoid displacement measurements [184]; further investigations are needed to clarify this disagreement. It is worth noting that previous studies categorized patients who showed aspiration at least on one swallow was identified as aspirators. Instead of separating patients into two groups, i.e., either aspirators or non-aspirators, our study investigated the relationship between the hyoid and the PA scale at the swallow level. It is worthwhile to evaluate each swallow level in order to account for variability in hyoid displacement within individuals, as well as to determine whether the deployed research methods are capable of detecting relationships between hyoid movement patterns and airway protection during swallowing, since the frequency and severity of laryngeal penetration and/or aspiration are key diagnostic factors leading toward appropriate interventions to mitigate the effects of dysphagia.

Age is a significant influence on PA scores (p value < 0.05). We found that the risk of penetration increased 5% as age increases one year. This finding matches the results from several previous studies [36]. Daggett *et al.* found that the percentage of penetration and aspiration dramatically increased with healthy subjects over 50 years old [56]. Steele *et al.* [257] reported that individuals over the age of 80 years old had more risk for penetration

and aspiration. Differing from our findings, Allen *et al.* [7] reported that increasing age was not associated with more incidences of penetration. Robbins *et al.* also concluded that age was associated with higher PA scores – only one of their healthy elderly subjects aspirated while no aspiration was seen in the young and middle-aged groups.

Volume was not significantly related to PA scores (p value $>>$ 0.05) according to our results. Several contributions showed the similar findings in the bolus impact on PA scores when bolus volume was less than 20 mL. Park *et al.* investigated the relationship between the pharyngeal and bolus volume to check whether there are influences between penetration/aspiration and increased bolus volume for stroke patients [197]. They examined 10 patients with different volumes and showed that increased volume did not affect the penetration and aspiration status. Hedstrom *et al.* [93] obtained very similar results to ours in their studies with 38 patients. On the other hand, Butler *et al.* [36] studied 76 healthy older subjects and demonstrated that the bolus volume over 20 ml yielded higher risk of penetration and aspiration than the 10 ml bolus volume. However unlike our study, these studies used exact bolus volumes which do not account for variations in patient aerodigestive tract size.

Our results shows that thin liquid has the highest risk to higher PA score, followed by nectar, cookie and pudding, which matches the previous findings. Several previous studies showed that thicker bolus generally resulted in lower PA scores, both in healthy group and patient group. Rofes *et al.* studied 146 subjects with different viscosities: thin, nectar, and thick. Their results showed that PA scores were reduced when using thicker viscosities [220]. Daggett *et al.* found that the frequency of penetration was significantly less during swallows of the thick viscosities across all age groups evaluated [56]. Newman *et al.* [190] collected 33 articles related to the effect of bolus viscosity and indicated that increasing the viscosity of the bolus will lead to a safer swallow and thickening liquids are considered as the practice of choice for many clinicians to manage dysphagia. Logemann *et al.* in a study of 711 patients with dysphagia due to Parkinson's disease or dementia, similarly found that thin-liquid aspirators had nearly 50 % reduction in aspiration with the thickest of liquids they administered [157].

The head position, neutral or chin down, showed no statistical significant association to PA scores (p value >> 0.05) while there is a trend that chin down position has less risk of penetration. Swallowing in the "chin-down" position narrows the airway entrance and therefore has traditionally been considered to reduce the risk of aspiration [236, 20]. On the other hand, several contributions showed that head position were not significantly related to aspiration, which matches our findings. Shanahan *et al.* [236] investigated 30 neurologically impaired patients for different postures, discovering that half of their patients could benefit from a "chin-down" position and those who still had the aspiration issues in a "chin down" position were younger, and continued to aspirate because accumulated hypopharyngeal residue overflowed from the pyriform sinuses into the airway rather than aspirating a portion of the swallowed bolus during the swallow. The kinematics of different structures in "chin-down" (comfortable chin down position) and "chin-tuck" (strict chin down) postures were investigated in [136], where Leigh *et al.* studied the swallowing cases from 40 healthy patients and showed that there were no significant differences among postures in maximal vertical displacement. The "chin down" position had no significant effect on hyoid bone movement, while the "chin tuck" posture influenced horizontal hyoid bone movement.

In this investigation, we have several limitations that might be considered in the future study. In [136], "chin down" posture can be separated into "comfortable" and "strict", which show different mechanisms. However, our data was collected and judged by different clinicians, only "chin down" term was applied, which may result in unstandardized effects on swallowing. In [194], Okada *et al.* revealed that clinicians may have different understandings of the same posture, or single-term "chin down" to represent two different postures. More investigations related to kinematics and aspiration could be done depending on the different head and neck position. Another potential limitation is that we did not strictly measure bolus volumes that were administered to patients. In clinical practice, a spoon and a cup are two common utensils to feed the patients during VFSS, and during eating and drinking, people do not measure specific volumes when self feeding with a spoon or drinking from a cup – they self administer volumes that are comfortable. Furthermore, C3 was applied for a distance marker in our investigation while different distance markers were applied in different contributions. For example, in [136], the diameter of a coin is set as the reference

rule, the coordinate was adjusted based on the coin diameter, and then normalized to the same scale. Rules based on different normalization methods should be investigated.

4.5 Conclusion

This study employed the generalized estimating equation model to investigate the association between the hyoid bone displacement and penetration and aspiration. We have shown that the maximum displacement of the anterior-inferior hyoid bone landmark is significantly related to PA scores. Reduced maximum anterior displacement of the hyolaryngeal complex leads to higher PA score. Furthermore, age has relation to PA scores while volume, viscosity, and head position show weak associations to penetration-aspiration. These findings suggest that analysis of the trajectory of the hyoid bone could provide useful diagnostic information toward identifying patients with an elevated risk of penetration and aspiration. Further investigations based on the hyoid trajectory including other hyoid landmarks and hyoid rotational patterns should be performed to improve our understanding of the relationship between hyoid movement and risks of penetration and aspiration.

5.0 Prediction Penetration Aspiration Scale based on Hyoid Bone Motion

The majority of this chapter has been previously published in and reprinted with permission from [309]. Zhenwei Zhang, Subashan Perera, Cara Donohue, Atsuko Kurosu, Amanda S Mahoney, James L Coyle, and Ervin Sejdić. The prediction of risk of penetration–aspiration via hyoid bone displacement features. *Dysphagia*, 35(1):66–72, 2020

5.1 Motivation

In this investigation, we sought to use maximum displacement of anterior-inferior of hyoid bone in horizontal direction and other variables such as age, bolus volume, and viscosity to predict the risk of penetration and aspiration during swallowing. Our hypothesis is that generalized estimation equations model can correctly indicate whether the penetration or aspiration occurs or not based on these variables. A generalized estimate equation model was built to test our hypothesis based on trajectories extracted from VFSS images during various swallowing tasks. The model based on hyoid bone motion with good performance would be a useful additional tool to help the experts to diagnose penetration and aspiration in patients with dysphagia.

5.2 Methods

5.2.1 Data Acquisition

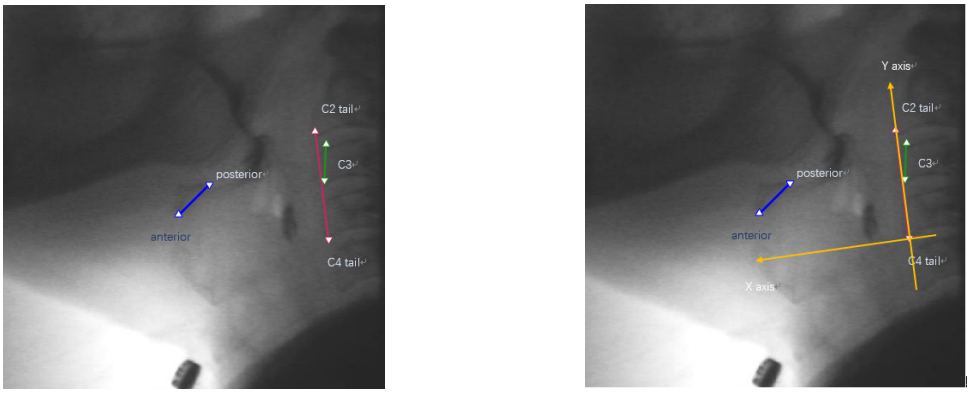
In this investigation, we considered the image data from 265 patients who underwent videofluoscopic examination at the Presbyterian University Hospital of the University of Pittsburgh Medical Center (Pittsburgh, Pennsylvania). The protocol for this study was approved by the Institutional Review Board at University of Pittsburgh and all participants

agreed and signed informed consent. The data applied in this study was obtained from the videofluoroscopic swallow study exam under the guidance of at least one speech language pathologist. 48 patients with stroke and 217 patients with non-stoke etiology participate in this study. The age range of the subjects was from 19 to 94 and the average age was 64.833 ± 13.56 years old. We exclude patients with tracheostomy or anatomic disruption or abnormalities of the head and neck in this study. Patients followed the instruction of clinicians to swallow boluses of liquids of different consistencies and volumes as well as cookies in VFSS exam. The speech pathologist determined the number and order of the swallow trials for each consistency and volume according to the patients' condition and the clinical indications. Measurements in this study were made during the swallowing of thin liquid (Varibar Thin Liquid with < 5 cPs viscosity), nectar-thick liquid (Varibar Nectar with about 300 cPs viscosity). Patients swallowed boluses in two types of utensils: a spoon in 3-5mL volumes, or a cup of a self-selected, comfortable volume. Patients primarily kept neutral head position during swallowing and depending on clinician's request, some of swallows were performed in a head-neck flexion position (chin down).

Fluoroscopy was set at 30 pulses per second (full motion) and the swallow study images were recorded on high quality at 60 frames per second by a video card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA) and later captured digitally into a hard drive with a LabVIEW computer software program. All videos were obtained from lateral view and the resolution of video clips were made into 720 x 1080. Finally, the videos were down-sampled into 30 frames/second to eliminate duplicated frames.

5.2.2 Image Analysis

The experts who are with dysphagia research experience measured the points of interest by using MATLAB (R2015b, The MathWorks, Inc., Natick, MA, USA). The onset of each video was defined as the moment when the bolus head arrives at the lower mandibula margin. The termination of each video was defined as the moment when the hyoid bone returned to its lowest position after the bolus passed the UES. We obtained over 3000 video clips which contain one swallow using this criterion. Over half of the clips has the issues of poor image



(a) Markers for anterior hyoid bone, posterior hyoid bone, C2 tail, C4 tail and C3

(b) Coordinate established based on the C2-C4

Figure 15: The landmarks for hyoid bone, C2, C3, C4 and established coordinate.

quality or interest of points obstructed by the shoulder or other medical equipment which results the tracking frame by frame impossible. The final video data set used for analysis included 1434 swallow video clips. As shown in Fig. 15, the experts tracked the following points of interest in each video frame: (1) anterior inferior corner of C2 vertebra; (2) anterior inferior corner of C4 vertebra; (3) anterior inferior corner of the hyoid bone; (4) posterior superior corner of the hyoid bone; (5) anterior inferior corner of C3 vertebra; (6) anterior superior corner of C3 vertebra. A coordinate system is created normalize the motion points from different subjects. we defined (2) as the origin, and defined the straight line connecting (2) and (1) as the y-axis. The x-axis is defined the horizontal line perpendicular to the y-axis and intersecting with (2). In order to normalize patients with different heights to a common anatomic referent, the anatomical scaling factor for displacement measure was defined as the length between (5) and (6): length of C3 vertebra. We used the actual distance in image pixels for all distance numbers. In the previous contribution, we showed that the maximum displacement of hyoid bone in horizontal direction has strong association with the penetration and aspiration. Thus, we extracted the maximum distance from the hyoid bone trajectory in horizontal direction in this investigation.

We used an 8-point PA scale [223] to identify the degree of penetration/aspiration from the 1434. Among these swallows, 1129 swallows have PA scores of 1 or 2 and 304 swallows

have PA scores greater or equal to 3. To evaluate the reliability of the swallowing analysis, 10 swallow cases were utilized. Three experts analyzed the same cases and their results are compared. Furthermore, the inter-rate reliability was tested as well, experts analyze the same case 1 month later to ensure less difference among different markers and time period.

5.2.3 Statistical Analysis

SAS ® version 9.3 (SAS Institute, Inc., Cary, North Carolina) was used for all statistical analyses with the GENMOD procedure for obtaining the main results. A dichotomous (normal; disordered) operational definition of PAS scores (1-2, and 3-8 respectively) was used for analyses, because there was a skewed distribution of PAS scores. Logistic regression models that are typically used with dichotomous data could not be used, because the independence criterion was not met due to having multiple swallows in the data set from each patient. Therefore, a GEE model [287] with a binomial distribution, a logit link function, and an exchangeable working correlation structure (which is an extension of a logistic regression model suitable for analyzing auto-correlated data) was used. Age, gender, swallow type (single/multiple 1/multiple 2), viscosity (thin/nectar/pudding/cookie), utensil (cup/spoon), head position (neutral/chin down), and swallow duration were used as forced-in independent variables based on face validity and prior knowledge of their dependence on PAS scores. In addition to these independent variables, we examined various aspects of hyoid bone displacement using a forward selection strategy with an entry criterion of $p < 0.05$. The measurement of these landmarks (superior hyoid hone and anterior hyoid bone) includes maximal displacement, maximal peak position, velocity, acceleration and duration in horizontal and vertical direction. To assess the predicted and observed disordered PAS scores, we created a contingency table based on the predicted probability deciles. The deciles were formed by sorting and separating the predicted probabilities into ten subgroups based on each patient's risk profile, from lowest to highest risk (1-10). We examined the observed percentage of disordered PAS swallows (3-8) within each decile compared to the predicted percentage according to the model. See Appendix A for the predictive model.

Table 8: Statistics and characteristics of patients involved in the investigation

Features		Frequency(%)	Features		Frequency(%)
PA	1	687(47.94%)	**Viscosity&Volume**	thin liquid by teaspoon	264(18.4%)
	2	442(30.84%)		thin liquid by cup	614(42.8%)
	3	138(9.63%)		not recorded utensil with nectar	1(0.007%)
	4	48(3.35%)		nectar by teaspoon	195(13.6%)
	5	29(2.02%)		nectar by cup sip	209(14.6%)
	6	33(2.30%)		pudding by spoon	94(6.6%)
	7	23(1.61%)		cookie	42(2.9%)
	8	33(2.30%)	**Gender**	male	155(58.49%)
Type	single	498(34.73%)		female	110(41.51%)
	multiple(1)	360(25.10%)	**Head Position**	neutral	1136(79.22%)
	multiple(2)	534(37.24%)		chin down	252(17.57%)
	not record	42(2.93%)		not record	46(3.21%)

5.3 Results

Table 8 illustrates the descriptive statistics and participant characteristics. The swallow analysis data was presented in this study for 1433 swallows from 265 distinct patients. Ninety-one swallows were excluded from the analysis due to missing information or incorrect recording. The age range of the subjects was from 19 to 94 and the average ± standard variation age was 64.8 ± 13.6 years. 1129 swallows had PA scores of 1 or 2 and 304 swallows had PA scores greater or equal to 3.

Table 7 illustrates the statistical results of focused-in clinical variables and aspects of hyoid bone displacement that met the 0.05 entry criterion for the model. Clinical variables shown in Table 7 were forced-in to the model with forward selection. Maximum anterior-horizontal hyoid bone displacement was the only aspect of hyoid bone displacement that was significantly predictive of normal versus disordered PAS scores and included in the model. Patient age was significantly predictive of normal versus disordered PAS scores, although the confidence interval included OR = 1.00. For each additional year of age, the odds of a disordered PAS score increased by 3% (OR=1.03, 95% C.I. = 1.00 1.05; p=0.0178). There was a trend toward a single swallow being less likely (36%) to have a disordered PAS score compared to multiple swallows, (OR=.64, 95% C.I. =.40-1.04; p=0.0708). Two swallows per bolus (multiple 1) was significantly more likely to have a disordered PAS score (OR=1.58, 95% C.I = 1.16 2.15; p=0.0040) than more than two swallows per bolus (multiple 2). There was strong evidence that swallows of thin liquid had a significantly greater odds

of a disordered PAS score than a cookie swallow (OR=3.62, 95% C.I. = 1.37 9.58; p=0.0096). The model predicted the risk of penetration and aspiration for each patient based on the variables included in the model. Table 3 shows the predicted probability of having a disordered PAS score in each decile compared to the observed percentage of disordered PAS scores in each decile. For instance, as shown in the table, the predicted probability for decile 1 indicates that 0-7% of the swallows will be disordered. The predictive model effectively captured patient risk profiles for this decile because 6.72% of the swallows had a disordered PAS score. Similar observations can be made for deciles 2, 4, 8, and 9. Deciles 3, 5, 6, 7, and 10 captured the increasing probability trend of penetration and aspiration, although the observed percentage of swallows with disordered PAS scores were slightly outside of the predicted ranges.

5.4 Discussion

This study found that a predictive model that included maximum anterior-horizontal hyoid bone displacement and other variables known to affect penetration and aspiration risk can reasonably predict the risk of penetration and aspiration in patients with dysphagia. While this predictive model accurately captured the increasing probability trend of penetration and aspiration risk of patients, the predicted and observed probabilities did not always match. Current clinical practice is for clinicians to assess physiological impairments of swallowing and reduced airway protection by subjectively interpreting VF images. However, one limitation of using VF as an assessment tool is that aspiration may not be observed during VF due to the time constraints of the examination to minimize radiation exposure. Creating a predictive model based on objective measurements of physiological swallowing events, such as the measurements of hyoid bone displacement that were used in this study, would allow clinicians to more accurately capture patient risk profiles of penetration and aspiration. This model could be used to improve assessment of swallow function, effectively track progress in therapy, and proactively and objectively identify physiologic markers of elevated risk of adverse events that occur secondary to dysphagia, such as aspiration pneumonia.

Table 9: Predicted probability decile cut-off and observed percentage based on the model (* actual% of swallows with disordered PA scores was within the predicted probability range based on hyoid displacement features)

Predicted Probability Decile	Predicted Percentage of High PA swallows	Number of Swallows	Actual Number (Percentage) of High PA Swallows
1	0.0 − 7.0	134	9(6.72)*
2	7.0 − 10.4	134	13(9.70) *
3	10.4 − 13.9	134	20(14.93)
4	13.9 − 16.9	135	21(15.67)*
5	16.9 − 19.7	134	21(15.56)
6	19.7 − 22.8	134	37(27.61)
7	22.8 − 25.7	134	27(20.15)
8	25.7 − 30.0	134	38(28.36)*
9	30.0 − 36.4	134	41(30.60)*
10	36.4 − 100	135	44(32.59)

5.5 Limitation

The GEE model in this study used anterior-horizontal hyoid bone displacement and other independent variables to reasonably predict penetration and aspiration risk for patients with dysphagia. However, swallowing and airway protection are complex, multifactorial processes. It is probable that the variables included in this model are not the only predictors of aspiration. One limitation of the current predictive model is that it underestimates the risk of penetration and aspiration for patients with disordered PAS scores. The predictive model will likely be improved by including other swallow kinematic measurements.

5.6 Conclusion

This research work developed a preliminary GEE model that can reasonably predict penetration and aspiration risk for patients with dysphagia. This is an important and necessary first step toward developing a more sophisticated and accurate predictive model that can be used in clinical settings. In the future, clinicians could use a predictive model based on physiological aspects of swallow function to calculate penetration and aspiration risk profiles for patients by entering patient specific information into the equation. By objectively determining patient risk profiles, clinicians could develop individualized treatment plans to prevent adverse outcomes (i.e. dehydration, malnutrition, and aspiration pneumonia) based on risk severity level, and objectively track the effectiveness of dysphagia treatment on functional patient outcome measures. Future research should examine the predictive ability of additional swallow kinematic measures on penetration and aspiration risk in patients with dysphagia. Variables such as hyoid bone velocity, initiation of the pharyngeal swallow, laryngeal elevation, laryngeal vestibular closure, UES duration, and other physiological parameters related to swallow function should be investigated. Including these kinematic events in the predictive model may increase the model's predictive value, which would further improve its clinical application.

6.0 Identification and Localization of Hyoid Bone in Videofluoroscopy

The majority of this chapter has been previously published in and reprinted with permission from [306]. Zhenwei Zhang, James L Coyle, and Ervin Sejdić. Automatic hyoid bone detection in fluoroscopic images using deep learning. *Scientific Reports*, 8(1):12310, 2018

6.1 Motivation

In the previous contributions, users had to manually mark region of interest in the first frames for the hyoid bone motion tracking. Furthermore, in their studies, the images evaluated are quite limited which cannot cover all the patients' cases. Therefore, in this study, we sought to develop a software platform that can localize the region of interest containing the hyoid bone in the alternative frames from the video with the help of objection detection method based on CNN. Our hypothesis is that the detection algorithm can detect the location of hyoid bone with high performance. State-of-art methods were applied in the investigation and we evaluated the performance of these detection algorithms by comparing the 'ground truth' manually segmented and the detected regions. Detection of hyoid bone localization accurately could help clinicians for a quicker diagnosis and to develop a fully automatic hyoid bone tracking system.

6.2 Material and Methods

6.2.1 Data Collection

In this investigation, 265 patients with swallowing difficulty underwent videofluoscopic examination at the Presbyterian University Hospital of the University of Pittsburgh Medical Center (Pittsburgh, Pennsylvania). The protocol for this study was approved by the Insti-

tutional Review Board at University of Pittsburgh and all participants provided informed consent. The age range of these subjects was from 19 to 94, and the average age of them was 64.833 ± 13.56 years old. Patients swallowed boluses of liquids of different consistencies and volumes as well as cookies during their VFSS examination. The amounts and viscosity they swallowed was judged by clinicians based on factors such as patients' history and the clinical indications. These liquids included thin liquid (Varibar Thin Liquid with < 5 cPs viscosity), nectar-thick liquid (Varibar Nectar with about 300 cPs viscosity). The position of patients during swallowing was primarily neutral head position though some swallows were performed in a head-neck flexion position. Patients swallowed boluses in a spoon which contains 3-5mL volumes, or self-administered boluses form a cup, which contains 10-20mL volumes. Fluoroscopy was set at 30 pulses per second (full motion) and video images were acquired at 60 frames per second by a video card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA) and collected into a hard drive with a LabVIEW program. The videos were two-dimensional digital movie clips of 720 x 1080 resolution, and in this investigation, we down-sampled it to 30 frames/second to eliminate duplicated frames.

6.2.2 Methods

In this investigation, our solution is to build a detection system based on the single shot multibox detector, which is one of the most popular detection algorithm in recent years. The SSD algorithm can generate high detection performance at the cost of high computational complexity. Thus, we also evaluate the performance of several other state-of-art detection methods, i.e., Faster-RCNN and YOLOv2, for the results comparison. The following paragraphs describe the approach in SSD, the data set ground truth creation and the training and testing details.

6.2.2.1 Network Architecture

Machine learning has been widely used in medical image and videos to help users to better understand the properties of these data [285]. Neural network is one of the popular type of machine learning models. The basic idea of neural network is to multiply the input

data with layers of weighted connections. Deep neural networks is a typical architecture of neural networks, which is constructed by multiple layers. Each layer implements a series of convolution operator on input, followed by a non-linear activation function, such as logistic function or rectified linear unit (Relu). Then pooling layer is applied to reduce the size of features to the following layers [129]. Popular fully convolutional networks for image tasks includes AlexNet [122], GoogleNet [269], VGG net [241] and Residual Net [91].

The SSD is a feed-forward convolutional neural network built on image classification neural network, called base network, such as VGGNet, ZFNet or ResNet [153]. Additional 8 convolutional feature layers are added after these base networks replacing the last few layers of the base networks. The size of these layers decreased progressively and used as output layers for prediction of detections at multiple resolutions. SSD integrated both higher and lower feature layers, as the lower layers contain better location information and the higher layers have more image details [160]. The images are divided into different sizes of grid sizes which is associated to default bounding boxes. The correspondence between the position of default box and the feature cell are fixed. SSD predicts the objects based on default boxes instead of predicting the bounding boxes directly. The default boxes are assigned with different scales and aspect ratios, which provide information of different object scales. The scale of each feature map is manually designed as:

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k - 1), \qquad k \in [1, m] \tag{6.1}$$

where m is the number of feature maps used for prediction. s_{min} is 0.2 and s_{max} is 0.9.

Each feature map cell corresponds to 6 default boxes, which are assigned with different aspect ratios, which are denoted as $\alpha_\gamma = \{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$. The width and height of the default box is computed as $w_k^\alpha = s_k\sqrt{\alpha_\gamma}$ and $h_k^\alpha = s_k/\sqrt{\alpha_\gamma}$. For the aspect ratio of 1, another scale $s'_k = \sqrt{s_k s_{k+1}}$ is added for the default box as well. The center of each default box is set at $(\frac{i+0.5}{|f_k|}, \frac{j+0.5}{|f_k|})$, and $|f_k|$ is the size of k-th feature map. By using these default boxes with various scales and aspect ratios from all locations of added feature maps, SSD predictions can covers different input sizes and shapes. Fig. 16 illustrates the idea of default boxes.

A set of convolutional filters are applied to the added features layers to perform the bounding box regression and category classification. For each feature layer of size $m \times n$

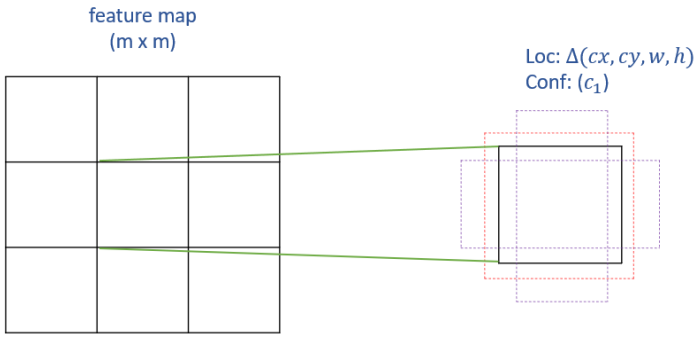


Figure 16: The idea of default boxes applied in SSD. For each default box, the offsets and confidence for categories are predicted.

with p channels, a $3 \times 3 \times p$ small kernel filter are applied to produce one value at each feature map cell, where the outputs are classification scores as well as the offsets relative to the bounding box shape.

The label of SSD include the class and the offsets from the default boxes. The default boxes is matched with ground truth if their intersection over union (IOU) is over 0.5. IOU is defined as $Area\ of\ Overlap/Area\ of\ Union$. The loss function of SSD combine a softmax loss for the confidence loss and a Smooth L1 loss for localization loss. The overall objective loss function is

$$L_{tot} = \frac{1}{N}(L_{conf} + \alpha L_{loc}) \tag{6.2}$$

where N is the number of matched default boxes and α is set to 1 by cross-validation. The SSD framework is shown in Fig. 17. For more details of the SSD network and loss function please refer to [153].

6.2.2.2 Training and Testing

We annotate the hyoid bone location (coordinate of left corner, height and width) in the frames of each videos as ground truth. These ground truth annotations for the hyoid bone locations were obtained by the experts. To evaluate the reliability of the swallowing

analysis, 10 swallow cases were utilized. Three experts analyzed the same cases and their results are compared. Furthermore, the inter-rate reliability were tested as well, experts analyze the same case 1 month later to ensure less difference among different markers and time period. The data is randomly separated based on the patients. 70% of patients were split into training data which contains 30000 frames with annotations, while 30 % of patients were split into test data which contains 18000 frames. The investigation shows that SSD provides a better overall accuracy using ResNet-101 as base network compared to VGGNet-16 [98]. Thus, we choose VGG-16 and ResNet-101 as base networks, and consider two image resolutions for inputs: 300×300 and 500×500. We compare models trained on both base networks and both resolutions inputs. The input with size 500×500 should provide the better performance as more details can be detected in higher resolution images. However, larger image size increase the computation time.

6.2.2.3 Evaluation of Accuracy

The performance of the detection module is measured by mean average precision (mAP), which which is the most commonly used evaluation method for object detection. Average precision estimated whether detected bounding boxes match the corresponding ground truth. Mean average precision is the area below the precision-recall curve, which integrates precision and recall and varies from 0 to 1. As we have just one class to classify here, mean average precision is exactly the average precision for hyoid bone class. The bounding box is labeled as true positive if IoU is greater than 0.5. Precision evaluates the fraction of true positive bounding box over all predictions and recall evaluates the fraction of the true positive detected bounding boxes among all ground truths.

6.3 Results

Tab. 10 shows results of the state-of-art published methods on our VFSS image dataset. In global, SSD method outperforms the results produced by YOLOv2 and Faster-RCNN.

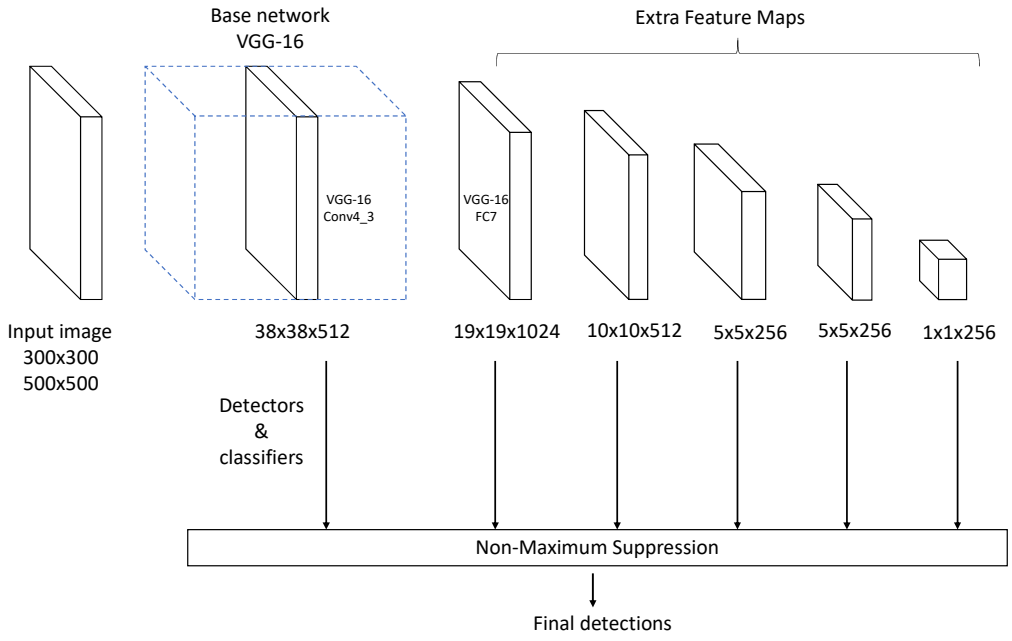


Figure 17: Architecture of Single shot multibox detector

Table 10: Comparison of mAP with different models

Model	mean average precision
YOLOv2	33.10%
Faster-RCNN + ZF	69.01%
SSD300-VGG	84.37%
SSD300-ResNet	79.25%
SSD500-VGG	89.14%
SSD500-ResNet	89.03%

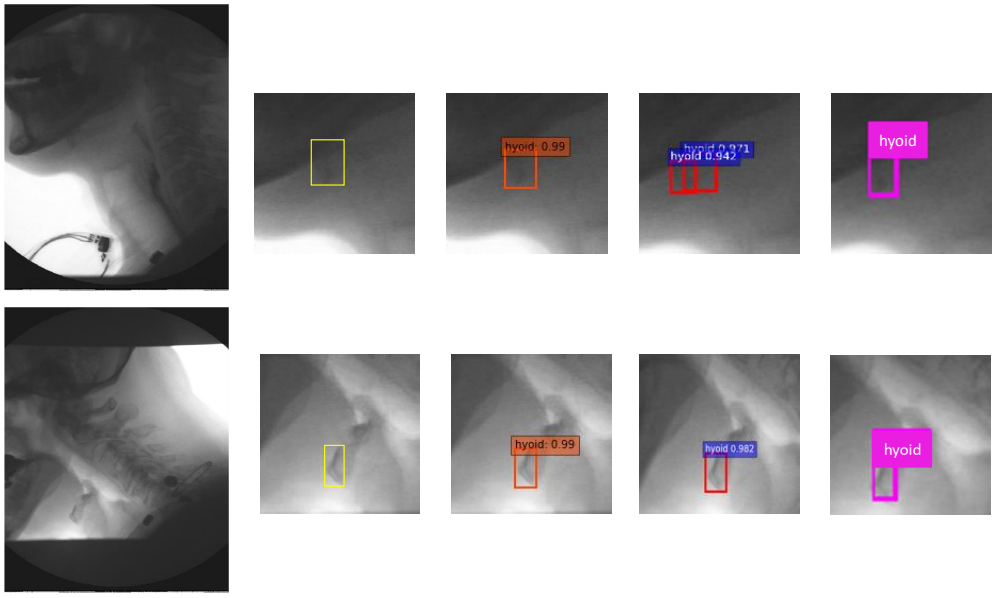


Figure 18: The identification of hyoid bone using different methods: ground truth (yellow), SSD500-VGG (orange), Faster-RCNN (red), and YOLOv2 (pink)

Among SSD method, VGGNet with input size of 500 × 500 produced the best result compared to ResNet and input size with 300 × 300. The mAP of SSD500-VGGNet is 89.14%, which is 0.11% better than using ResNet-101 as base network and 4.77% better than using smaller image input size. Fig. 17 shows the example results by manual segmentation, SSD500-VGGNet, Faster-RCNN and YOLOv2. We select two different cases as example, patient swallowed the bolus in neutral head position and chin down position. In the ground truth, the bounding box is used to locate the hyoid bone location as most of the object detection method using bounding box to locate and classify the content inside it. In the example case, all of 3 tested method showed a good result, detecting the hyoid bone location successfully. However, it can been clearly seen that Faster-RCNN method produced two regions of interest that it considers as the hyoid bone with very close confidence score.

Fig. 19 illustrate results using SSD500-VGGNet method with different hyoid bone location, under the mandible, behind the mandible, and the results with different image qualities. From these results, SSD500-VGGNet showed stable detection result, clearly finding the hyoid

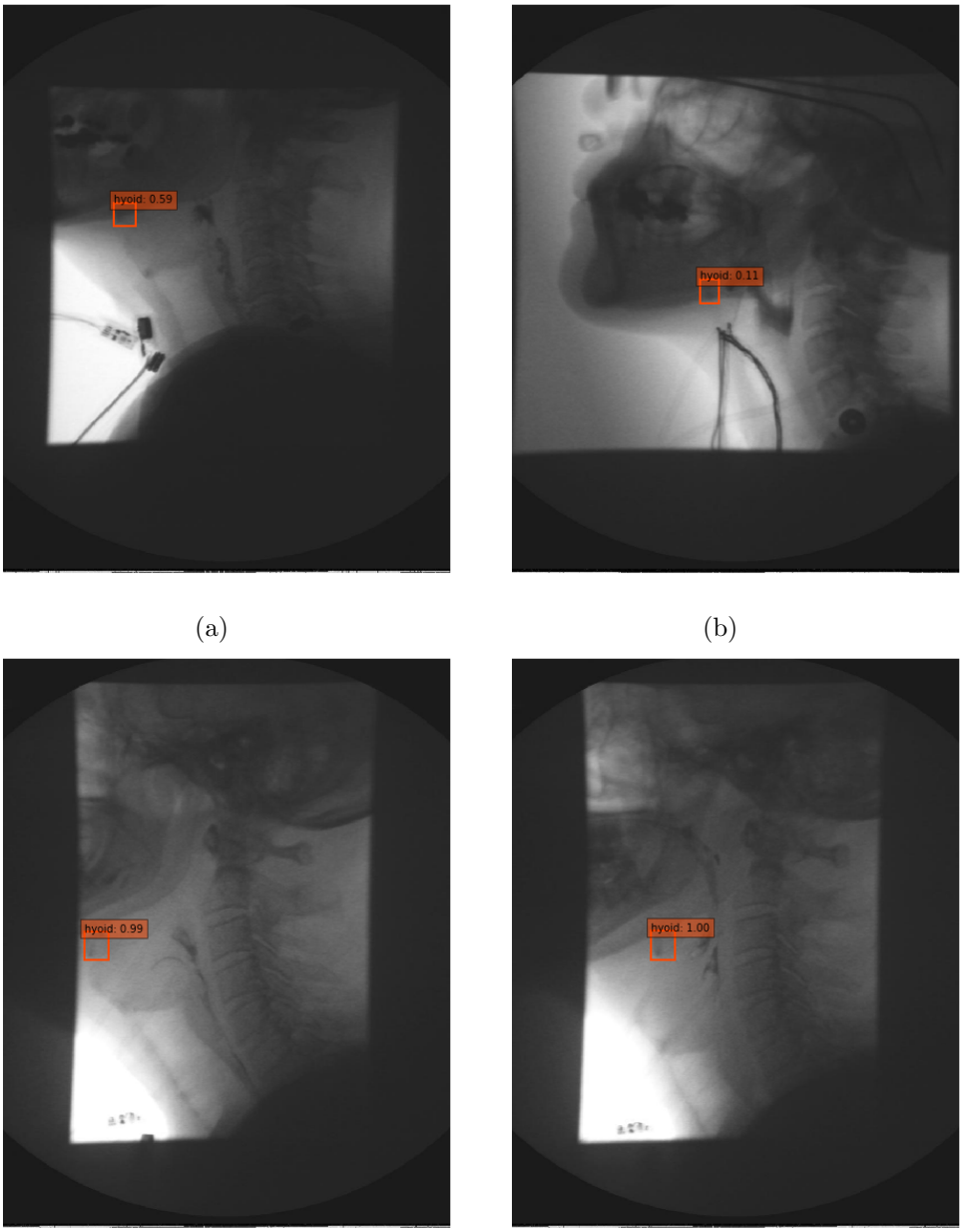
bone. The hyoid bone is hidden behind the mandible in the case (a) and (b), the algorithm detect the hyoid bone with a relative low confidence score while It perform well in the case (c) and (d) where the hyoid bone present under the mandible.

Fig. 20 shows the change of training loss function and the performance on test data during the SSD500-VGGNet model training. From these figures, we can learn how the performance of the model change during the training. The loss function has dramatically decreased in the first 10000 iteration and then the loss function only have very slight decrease in the following training iteration. On the other hand, the performance model arrived at 87 % mAP at first 10000 iteration, than it has slight variance in the further training iteration and than reached around 89%.

6.4 Discussion

In this investigation, we aimed to detect the location of the hyoid bone in the videofluoroscopic images without any human intervention. The hyoid bone is one of the key important components in the daily dysphagia assessment, whose motion is related with the severity of dysphagia and treatment effect. Manually tracking hyoid bone data from VFSS is still the golden standard accepted by experts and clinicians. However, manually segmentation and annotation is very time-consuming and unreliable. The valid hyoid bone motion data can be applied in further investigations such as statistical methods and classification based on machine learning. A quantitative and qualified computer-aided system is highly required in this field. In the dysphagia research field, limited contribution works on the hyoid bone semi-automatically tracking which requires the manually region selection by the experts in the first step. The automatic localization of hyoid bone can help researchers for a more efficient study. We are going to discuss the performance of each method and the possible reasons that may influence the results in the following paragraphs.

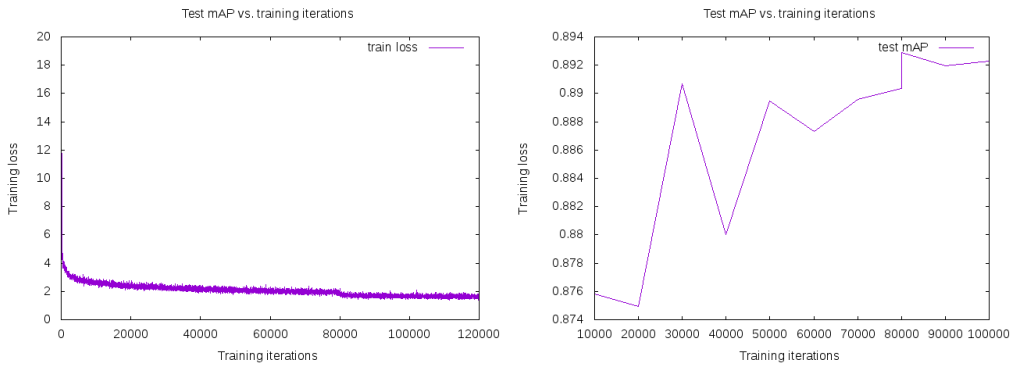
We examined the performance of different object detection methods (Faster-RCNN, YOLOv2, and SSD) on locating hyoid bone in our own VFSS images dataset. For the deep architecture, we employ the medium-size network VGGNet and the relative larger-size



(a) (b)

(c) (d)

Figure 19: Results on different image conditions using SSD500-VGGNet: (a)(b) hyoid bone hide behind mandible (c)(d) hyoid bone is slightly blurred during motion



(a) (b)

Figure 20: The influence of training iteration in the SSD500-VGG model (a) training loss vs. training iteration (b) performance on test data vs. training iteration

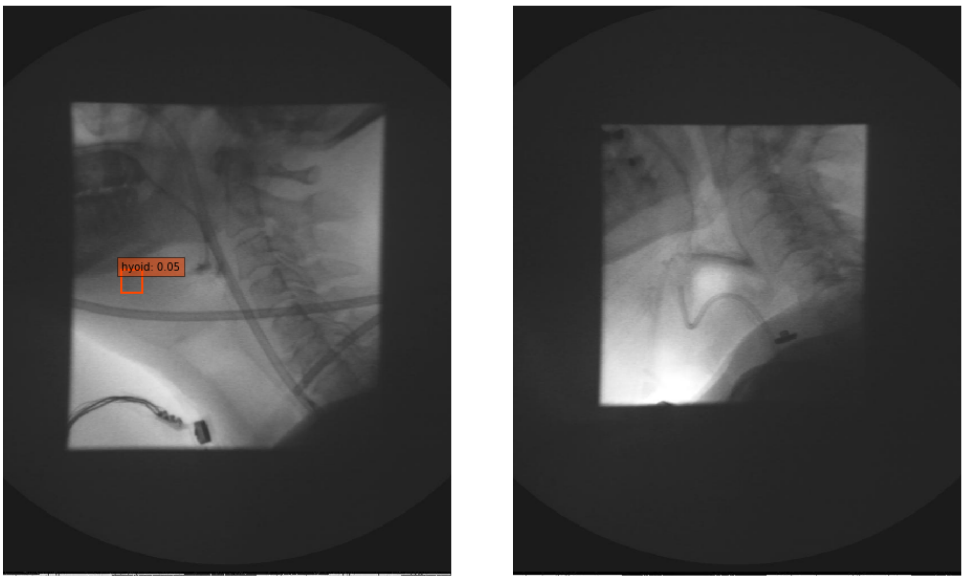
network ResNet 101 for the SSD, a small network ZFNet for Faster-RCNN. YOLOv2 is from the original Darknet model [216]. The SSD500-VGGNet achieves good results than other CNN based models, which can be considered as the most suitable method for the hyoid bone detection in the VFSS images. It is not surprising that YOLO achieve the worst performance on the VFSS data. Hyoid bone can be considered as the small objects in the images while YOLOv2 is a fast object detection method but is weak for the small object detection as it applies global features for the detection which can't get enough details of small object. SSD500 is better than SSD 300 in all settings by using ResNet-101 or VGGNet-16. The reasons might be follows. SSD resizes the input images to the fix size: SSD300 resize the images into 300×300 while SSD500 resize them into 500×500. In SSD300, as the image is resized into smaller size, each cell in the feature map can cover a relative larger area than those in SSD500. Since the hyoid bone is very small in the image, SSD300 may not learn the details of the hyoid bone, which leads to the worse performance. Furthermore, ResNet reached the similar mAP compared to VGGNet in SSD500 while it has worse performance in SSD300. This interesting observation may be explained based on the model structure. ResNet-101 is a neural network with 101 layers, while VGG-16 only has 16 layers. The similar results in SSD500 may indicate that both network provide detailed information for the added features. As ResNet is a more complex model, it may overfit for the size of input 300×300, which

leads to the lower detection results. SSD method is a powerful tool to detect the hyoid bone location, however, training SSD model with ResNet-101 and VGGNet with larger input size is very time-consuming. We implemented our algorithms on GPU NVIDIA Tesla M40, it took over one week to train the SSD500-VGG16 models and SSD500 with ResNet-101 took much longer time while Faster-RCNN took only one day as ZFNet is a small neural network.

Hyoid moves upward and forward during patient's swallow. It will move and be hidden in the mandible sometimes. As the mandible represents a dark region in the image, it is quite difficult for the clinician to find the hyoid bone directly. In general, experts have to compare frame by frame to check whether there is some changes around the mandible in order to determine the location of the hyoid bone. The result in Fig. 19 (a) and (b) shows the detection of hyoid bone. Although the confidence score is low, it still can be considered as a huge success as even the experts may not find the hyoid bone location. (c) and (d) are the examples of blurred hyoid bone. The hyoid bone may be blurred when it moves quickly between two frames, the algorithm can detect this kind of case with very high confidence score.

X-ray image varies from the quality as the clinicians always control the dose in order to let the patients receive as few dose as possible. Thus, as shown in the Fig. 19, the brightness, the contrast of each x-ray images are different. X-ray contains many useful information, but it lose many of them due to these operations. As shown in the Fig. 21, the SSD method detect the hyoid bone location with very low confidence score or totally can't detect the hyoid bone location. It is like a guess when humans locates these cases. We know the location of the hyoid bone as the pre-knowledge, and seek to find the target around the possible location and eliminate the impossible region one by one. The object detection algorithm classify the regions based on the default boxes, which is a direct way to make the decision and can't make full use of the information outside information.

The paper [98] indicated that Faster-RCNN with inception ResNet v2 has the best object detection results comparing to the other modern object detection method. Furthermore, several researches focus on the small object detection such as feature pyramid network [147], which might be the further research interest to increase the detection performance of the hyoid bone. On the clinical side, future work should investigate on automatic segmentation



(a) (b)

Figure 21: The cases which algorithm didn't detect the hyoid bone (a) the case with low confidence score (b) the case totally not detected

of hyoid bone areas and extract more useful information such as anterior and anterior part of the hyoid bone in video sequence. Moreover, since we showed that SSD detection method can solve the hyoid bone detection problem, we would also like to explore the possibility to detect other key components in the videofluoroscopy images. Given that millions of VFSS studies implemented, high-accuracy component detection can save experts considerable time during their diagnosis.

6.5 Conclusion

In this work, we have investigated the hyoid bone detection in the videofluoroscopy images using deep learning approach. We considered 1434 swallow with VFSS videos as our dataset. The hyoid bone location is manually annotated in each frame of videos. We consider each frame as the single sample, and trained 70% of frames using current state-of-

art object detection method. The SSD-500 model can track the location of the hyoid bone on each frame accurately while the hyoid bone motion information can help for physiological analysis. We believed that this proposed model has the potential to improve the diagnosis assessment of dysphagia in the near future.

7.0 Automatic Annotation of Cervical Vertebrae in Videofluoroscopy Images via Deep Learning

The content of this chapter is currently under review with Medical Image Analysis. Zhenwei Zhang, Shitong Mao, James Coyle, and Ervin Sejdić. Automatic annotation of cervical vertebrae in videofluoroscopy images via deep learning. 2021

7.1 Motivation

The purpose of this study is to demonstrate how deep learning neural networks can achieve unprecedented accuracies in anatomical landmark localization that can change the clinical assessment of dysphagia. Most importantly, our models maintain excellent performance even when validated on an independent test dataset, demonstrating its robustness and the generalizability needed for clinical settings. Specifically we present an investigation of deep learning in identifying the necessary anatomic scalar, the distance between the 2nd and 4th cervical vertebral bodies used to correct for size differences among patients, on all frames of a VFSS examination. We further sought to investigate how closely individual vertebral lengths (e.g,, C3 alone) corresponded to the longer C2-C4 segment currently used in kinematic analysis but whose most inferior landmark may not always be visible in VFSS images due to patient posture.

7.2 Methods

7.2.1 Videofluoroscopic Swallow Study Dataset

Our dataset was collected from 265 patients with swallowing difficulty and 70 healthy volunteers who underwent videofluoroscopic examination at the Presbyterian University Hos-

pital of the University of Pittsburgh Medical Center (Pittsburgh, Pennsylvania, USA). The Institutional Review Board at the University of Pittsburgh approved the protocol of this study and all participants provided informed the consent. We didn't use statistical methods to predetermine sample size or subject age range. In this preliminary feasibility study, a convenience sample was used because there are no data upon which to base power calculations for sample size. The age range of these subjects was from 19 to 94, and the average age was 64.83 $\pm$ 13.56 years old. All experiments in this data collection were performed in accordance with relevant guidelines and regulations. Participants in this study include During the VFSS examination, patients were required to swallow liquid boluses of various consistencies and volumes as well as pureed food and cookies, all containing barium. A standard data collection protocol was not followed for the patient data set. Instead, clinicians who conducted the VF modified the protocol for the administration of boluses (e.g. number of swallows, bolus consistencies, bolus volume and patient's head position) based on clinical appropriateness. The following consistencies were used in our studies: Varibar (Bracco Diagnostics, Inc.) thin liquid (<5cPs viscosity), Varibar nectar (300 cPs viscosity), Varibar pudding (5000 cPs viscosity), and Keebler Sandies Mini Simply Shortbread Cookies (Kellogg Sales Company). Patients swallowed liquid boluses administered from a spoon containing 3-5mL volumes for all consistencies, or self-administered liquid boluses from a cup containing in patient self-selected, comfortable volumes between 10-20mL. Pudding and solids were administered from a spoon.

In our investigation, the videofluoroscopy system were set at 30 pulses per second (full motion). The first dataset consisting of 265 patients was collected from 2012 to 2015 using Ultimax system (Toshiba, Tustin, CA) and the second dataset of 70 volunteers was acquired through Precision 500D system (GE Healthcare, LLC, Waukesha, WI) from 2018 to 2019. Video images were acquired at 60 frames per second by a video card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA) and recorded into a hard drive with a LabVIEW program. The first dataset was captured with 720 x 1080 resolution in real time while the second dataset was captured with 1280 x 1024. Due to poor image quality or obstruction of the fourth vertebrae by the shoulder or other medical equipment, over half of swallow videos were not ideal for marking the points and our data set included 1518 swallow video clips.

Human experts who were trained as previously described in swallow kinematic analysis identified anatomical points of interest (second vertebra and fourth vertebra) in 1518 swallow videos and annotated the landmark frame by frame in MATLAB (R2015b, The MathWorks, Inc., Natick, MA, USA). In addition, the head and tail of third vertebra were labeled on only first three frames of each subjects. Each swallow was segmented to include ll activity beginning with the frame in which the head of the bolus reached the lower mandibular margin to when the tail end of the bolus passed through the upper esophageal sphincter (UES). 10% of the videos were randomly selected for ongoing inter- and intra-rater reliability tests to maintain intraclass correlation coefficient over 0.9 to avoid judgment drift over time.

7.2.2 Image Preprocessing and Data Augmentation

The total number of frames extracted from videos with annotations is 59810 images for our dataset. As we only collected the data from 335 subjects, the head position and image condition of VFSS images from the same patient were quite similar. The problem with the data set from the limited patients is that the trained model may suffer from overfitting and would not generalize to test dataset. The data augmentation is well accepted practice to directly augment the input data to the model to increase the variety of perturbations in training data information, which more stringently trains the algorithms in detecting events during various common clinical testing conditions. In our dataset, we preprocessed the images from each patient. The augmentation methods included: random flipping half of images horizontally, rotating the images from -45 degree to 45 degree, shearing all images by -10 to 10 degrees, random cropping or padding 75% to 125% to original images, and changing the brightness of the images by multiplying 0.8 to 1.2. After data augmentation, all of augmented images still contain the C2 - C4 landmarks and the total number of the training images remains unchanged. The deep learning networks highly require computation resources, we resized the input images into 448 $\times$ 448 considering the model training time. The original landmark point is shifted with respect to the image center, and normalized by (w, h) as given by:

$$(x_i', y_i') = (\frac{x_i - 0.5w}{w}, \frac{y_i - 0.5h}{h}) \tag{7.1}$$

where (x_i, y_i) are given ground truth coordinate of landmark points and (x_i', y_i') are normalized and centered coordinates, treated as labels for networks training.

7.2.3 Overview of Model Development

Convolutional neural networks are commonly applied in medical imaging field, which can be used to discover the subtle patterns in a dataset. The main architecture tested in this study was a convolutional neural networks which used ResNet blocks followed by two convolutional layers. We implemented a two-stage networks architecture for vertebrae landmark detection. The basic idea of our two-stage network was inspired by [164]. In our design, the netwroks consist of two stages, the global detection network and the local detection network. The global stage provides the rough detection results of vertebrae locations and crops the vertebrae regions. We employed a CNN structure, which contains ResNet block, as our localization model to predict the coarse locations. ResNet block is popular architecture that makes use of the idea of 'short connection', skipping one or several layers and carrying input to the output, which allows to prevent vanishing gradient problem and fasten the training of the networks. We adopt the structure of ResNet-50 in global stage, which performs identity mapping for shortcut connections. We adjusted the last fully connected layer, which was originally designed for classification, to predict the vertebrae region.

Due to the various shape of vertebrae across the population, the global network may not capture all the variations of these difference, especially for the edge and the order of the vertebrae. To overcome the errors of local parts, we introduce the local network for the finer landmark localization, which is essential for accuracy improvement. Images are cropped via the prediction results from the global stage network, then scaled and fed into the local stage network. Similar with the global stage network, we adopt ResNet-34 structure, with the last fully connected layer adjusted to directly regress the landmark locations on the input images. The inverse transformation function is applied to map the predicted points to the original image.

Normalization is widely adopted techniques that enables more stable and faster training of deep learning models. In our study, we found that the switchable normalization showed

better performance than batch normalization layers in ResNet blocks in the training phase. Switchable normalization combines batch normalization, layer normalization and instance normalization using weight average, which allows the custom choice of normalization depending on the depth of the layer and training batch size (Fig. 22). Batch normalization was proposed and widely implemented in ResNet and similar convolutional network architecture. It reduces internal covariate shift by using mini-batch mean and variance to normalize each mini-batch of data. The normalized version of a mini-batch of inputs $\{x_1, ..., x_m\}$ is computed as follows:

$$\hat{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad \text{with} \quad \mu = \frac{1}{m}\sum_{i=1}^{m} x_i \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(xi - \mu)^2 \tag{7.2}$$

The layer normalization normalizes features within each sample, instead of normalizing across samples. The layer normalization is computed over all hidden units (H) in the same layer:

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l \quad \sigma^2 = \frac{1}{H}\sum_{i=1}^{H}(a_i^l - \mu^l)^2 \tag{7.3}$$

Similar to layer normalization, instance normalization normalizes features within channels.

The loss function was defined as Euclidean loss for landmark location prediction, which is computed from

$$loss = \frac{1}{2}\sum_{N}^{i=1}((\hat{x_i} - x_i')^2 + (\hat{y_i} - y_i')^2) \tag{7.4}$$

where $(\hat{x_i}, \hat{y_i})$ are landmark location predicted by the network. We computed the loss function on the training and validation data, and we selected the model with best loss function score on validation dataset as our final model. We fine-tune the ResNet via transfer-learning and also trained networks from the scratch. The advantages of normalization layers is to regularize the model, reduce the overfitting and improve the model performance. Normalization layers change the distribution in network weights during training.

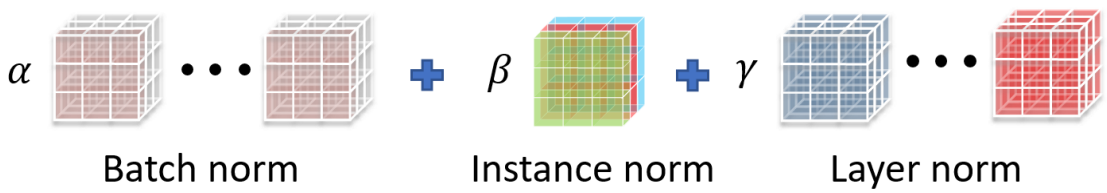


Figure 22: **Switchable normalization** Switchable normalization combines batch normalization, layer normalization and instance normalization using weighted average the means and variances. It allows networks to find the suitable ratios among three normalizations for each layer during training.

7.2.4 Training Two-Stage Network Model

In this investigation, two datasets were utilized in the model training and evaluation. The first data was collected from 265 patients using Ultimax system, with 70% of subjects for training, 30% for validation. An extra independent data collected from 70 volunteers was applied for the final testing. We ensure that no person in the training group is in the validation and test group to make it a truly independent group. In the original paper, the ResNet block utilized the batch-normalization layer. In our model, we implement and tested the residual block using switchable normalization instead of batch-normalization layer. The training curve of batch-normalization and switchable normalization is listed in supplemental files. We trained our two-level neural network models via fine-tune of original ResNet and fully trained switchable ResNet block. The model that performs best on the validation dataset is selected for testing. The switchable normalization showed slight better accuracy compared to the transfer fine-tuning using original ResNet structure. In this study, training and Testing procedures were implemented using Pytorch on the NVIDIA Tesla M40 GPU. We utilized Xavier initialization to initialize weights in the networks, and we used exponential decay learning rate starting from 0.01 and the learning rate was scale by 0.95 after each epoch. The whole-images were resize into 448 $\times$448 and the models were trained over 80 epochs on the first patient dataset with 80 % for training and 20 % for validation. Due to the limitation of C3 annotations, we trained first stage network only with C2 and C4 labels, then the second network were trained with all annotations.

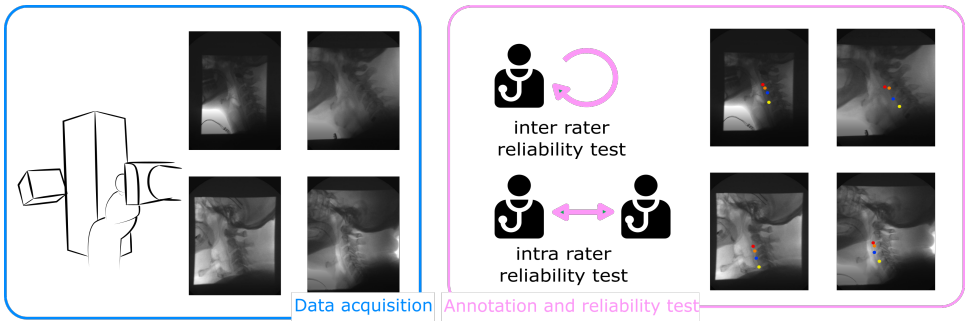


Figure 23: **Data acquisition and annotation procedure** Our dataset included annotated swallows collected from 335 subjects for the model training and evaluation. Video clips were recorded directly during VFSS examination. C2, C3, C4 vertebra locations were manually labeled by one main experienced expert during analysis. Inter-rater reliability test was implemented one month later and intra-rater reliability was tested with two other raters to ensure the accuracy of the judgment.

7.2.5 Testing and Analysis

Once the model finished the training, the evaluation of model was implemented on the testing dataset, which independent and not included in the training dataset. All parameters in the models were frozen and we predict landmark points by a forward-pass through the networks. As we rescaled and shifted the landmark points during training phase, these points should be scaled and shifted back to the original image coordinates:

$$(x_i, y_i) = (\hat{x_i'}w + 0.5w, \hat{y_i'}w + 0.5w) \tag{7.5}$$

The purpose of this study is to locate the key points of vertebrae in the videofluoroscopic images, whose information can be used as an important reference in clinical kinematic analysis. First, we evaluate the mean and standard deviation of location pixel difference between ground truth and points predicted by the models. We also evaluate the percentage of pixel difference compared to whole image size. In addition, we checked how the results were affected when various normalization layers and different input size were applied during model training. We asked three well-trained pathologists to manually label the C2, C4 landmarks points, comparing the results tolerance within human and the error range between humans

and machine predictions. Vertebrae information are used to build a coordinate for kinematic analysis in dysphagia field. To evaluate the performance of model, we calculated and compared the ratio of C2, C4 unit, and angle of C2-C4 coordinate. The ratio of C2-C4 is calculated by predicted C2-C4 length over annotated C2-C4 length. The angle of C2-C4 indicates angle between vector of predicted C2-C4 and vector of annotated C2-C4.

7.3 Results

We demonstrated an automated pipeline to measure the location, length and orientation of several cervical vertebrae in videofluoroscopic images. First, experienced raters conducted manual anatomic annotation of frame-by-frame videofluoroscopic data, which was collected from 265 subjects with suspected dysphagia and 70 healthy participants (Method). Raters annotated the location of antero-inferior corner of C2, and the anterior-superior and anterior inferior corners of the C3 and C4 vertebral bodies, as shown in Fig. 23. These measurements served as the ground truth for determining the length of this vertebral axis. Given an input image, the first step is to crop the image by removing the patient information and baffle region (black regions shown in Fig. 23) around the patient's neck region which was used to reduce the radiation during examination. Then the cropped region is scaled to a fixed size and fed into a two-stage network (Fig. 24). The convolutional networks were trained to learn features and patterns from images and mathematically describe the relationship between human annotations and the input images. After training the networks, these parameters were frozen in order to make the prediction on the validation dataset and the test dataset. The first stage network predicted the coarse location of vertebrae landmark regions and the second network finely improved the landmark regions. The model performance is evaluated by measuring mean localization distance, length ratio and angle error. Localization distance measures the actual distance in pixels between predicted landmark coordinates and the labeled landmark coordinates. Length ratio measures the ratio between predicted C3/C2-C4 length and the labeled length while the angle error measures the angles between predicted C3/C2-C4 vector and manually labeled vector. These two metrics are important parameters in the dysphagia

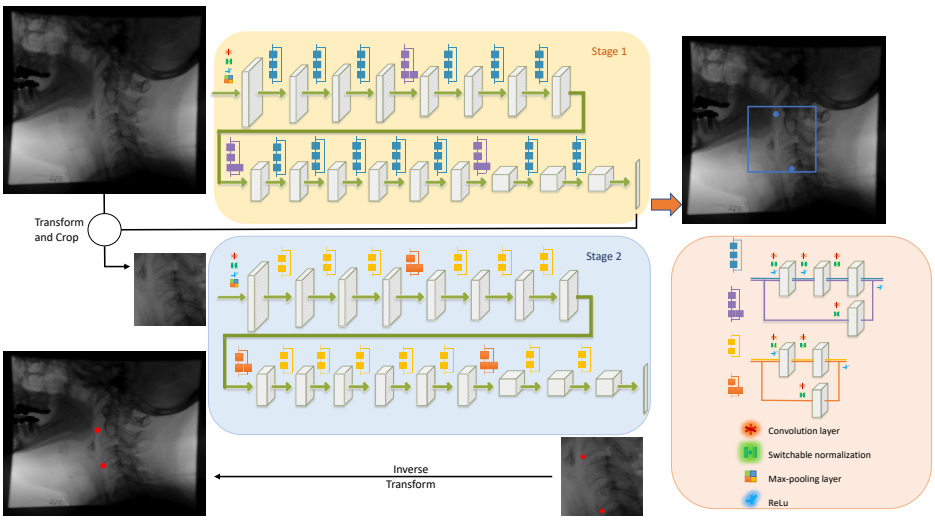


Figure 24: **The pipeline of the proposed two-stage network architecture for vertebrae landmark localization** First, a new input image is preprocessed to remove the patient information and dark regions in the videofluoroscopy image. After preprocessing process, the input image is fed into the first stage of the network to achieve the coarse detection, which allows to crop the image for finer detection. Then, the cropped image, which covers the vertebrae region, is fed into the local stage network for a better landmark localization. The output vectors from the network, which indicates the location of the vertebrae in the cropped image, are projected back to the initial image. The two-stage network consists of several ResNet blocks in each stage network. The first stage network follows the idea of ResNet50 while ResNet34 structure is implement in the second stage network. The ResNet block include several Convolutional layers, followed by normalization layers and a rectified linear unit(ReLU), then an extra identity map create a shortcut between input layer and output layer of the block. Different from the traditional ResNet block, we implemented switchable normalization layers instead of batch normalization layers, which allows to adaptively switch among various normalization techniques.

analysis, which are widely used to reduce the bias among population in decision making. Thus, we mainly focus on the accuracy of length and orientation measurement.

In the experiment, the model was trained using swallows from 265 consenting patient subjects, and then tested on the second dataset from 70 additional healthy volunteers, which were treated as unseen samples for the deep learning model to evaluate generalization. Notably, our second data was collected three years later and used a different videofluoroscopy machine, which can present the challenge of the invariant performance of our method on vertebrae location given different imaging resources.

In this study, the performance of our model referred to how closely the predicted vertebral locations corresponded to human judgment. An example of a continuous swallowing video captured at 30 images per second, is shown in Fig. 25(a). At each time point, the two-stage model localizes the location of C2, C3 and C4 vertebra. The images on the left show the ground truth and the frame with the largest distance error in vertical direction and the right images right images are those with largest localization error in horizontal direction. Overall, the location results from our model for one subject are reliable. Fig. 25(b) presents several location detection results on the test dataset, with orange for the ground truth, blue for the first stage results and red for our model's final results. The model was applied to the testing set, an independent dataset involving 70 subjects, and mean localization distance (MLD) achieved 4.20 ± 5.54 pixels. In order to verify the advantages of using two-stage networks, we compared the results with the model which uses ResNet50 for training. ResNet50 architecture led to a MLD at 7.44 ± 5.38 pixels. The summaries of localization distance distribution in testing the dataset compared to the human raters' annotations is shown in Fig. 26(b). As there were no established gold standard or previous experiences that could inform our methods. Regarding the acceptable localization distance tolerance, we chose 1 % of the whole image size as our criteria (i.e error less than 5 pixels range). The percentage of acceptable predicted locations via ResNet50 is 49.66% while the two-stage networks gave 87.36 %. The variability across multiple raters is unavoidable due to the limited quality of VFSS images, which is why the reliability test is deployed in routinely in research and routine clinical practice. In this study, the overall kappa ICC between two human raters and between the rater and the model both achieved over 0.9, showing that our

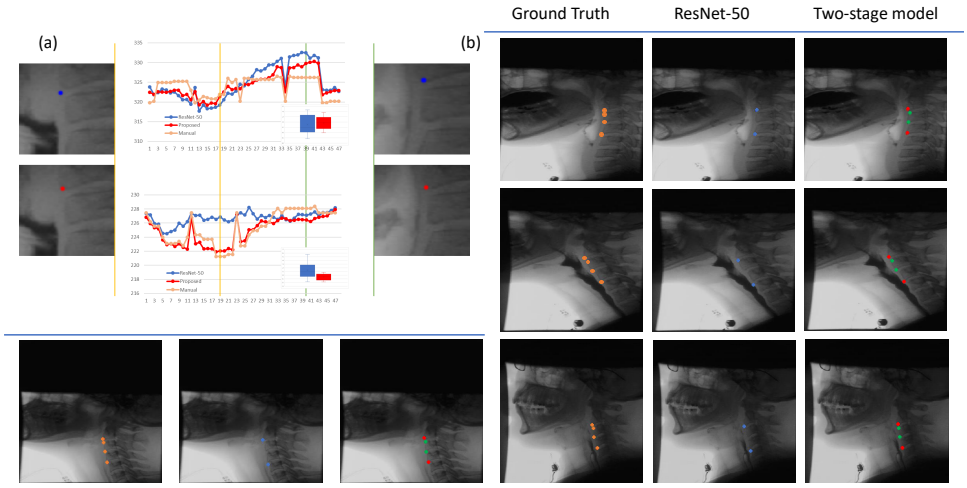


Figure 25: **Landmark localization results demonstrating the two-level model's robustness to variations among patients** (a) localization results predicted on a continuous swallowing video. Blue lines indicate the prediction from predictions, which show larger error variance comparing to red lines (the two-stage model), demonstrating the benefit of our model. Left images illustrates the largest error in y direction and the right images corresponds to the x direction. (b) Examples of the selected videofluoroscopic images with manually annotations, predictions from ResNet50 (first stage) and final prediction results. Note how the second stage achieved invariance to the scale and is able to perform localization despite head pose, vertebrae shape and lighting for different individuals.

model is comparable to human raters. Fig. 26(a) compared the model's predictions errors and one human rater judgment bias on the test data. Ninety percent of the predicted data shows comparable predictions to the second rater judgment while the model still has about 5% of results which demonstrated larger locations errors than the likely errors produced by the human rater during the manual annotation process.

Compared to the exact location of vertebrae, estimating the cervical vertebrae length and orientation is highly desired in the clinical settings as these information are usually served as patient-specific criterion referenced correction factor. In our study, we measured the length between C2-C4 and the length of C3 unit. Fig.26 (c) and (d) present the length ratio distribution and angle error distribution between estimated cervical vector and label vector respectively. The mean estimated length ratio from ResNet50 is 1.04 ± 0.09 and 45.95% of them are located in the length ratio range 0.95 to 1.05 while 93.76 % of predictions from two-stage model are located in the same range with mean estimated length ratio 0.99 ± 0.04. The mean absolute angle errors from ResNet 50 is 0.06 ± 0.05 rads and 0.03 ± 0.03 rads for our two-stage model.

To evaluate the performance of the model, we implemented 5-fold cross-validation on patient data and tested each model on healthy data as well. Table 11 presents the MLD, angle error and C2-C4 length ratio for each fold. The average of MLD is 4.07 pixels on patient group and 4.67 pixels on healthy group. The results indicate that the model generalized well on both data set while they were collected from two different video fluoroscopic machines.

7.4 Discussion

This study is the first step toward a fully automatic diagnostic image analysis system based upon computational methods, rapidly offering the vertebral scaling information that facilitates objective and accurate measurement in real time. The finding that our two-stage model could accurately and autonomously determine the anatomic scalar necessary for accurate measurements kinematic sets the stage for advancing automated analysis methods from VFSS images. The potential for speeding VFSS interpretations with automated data

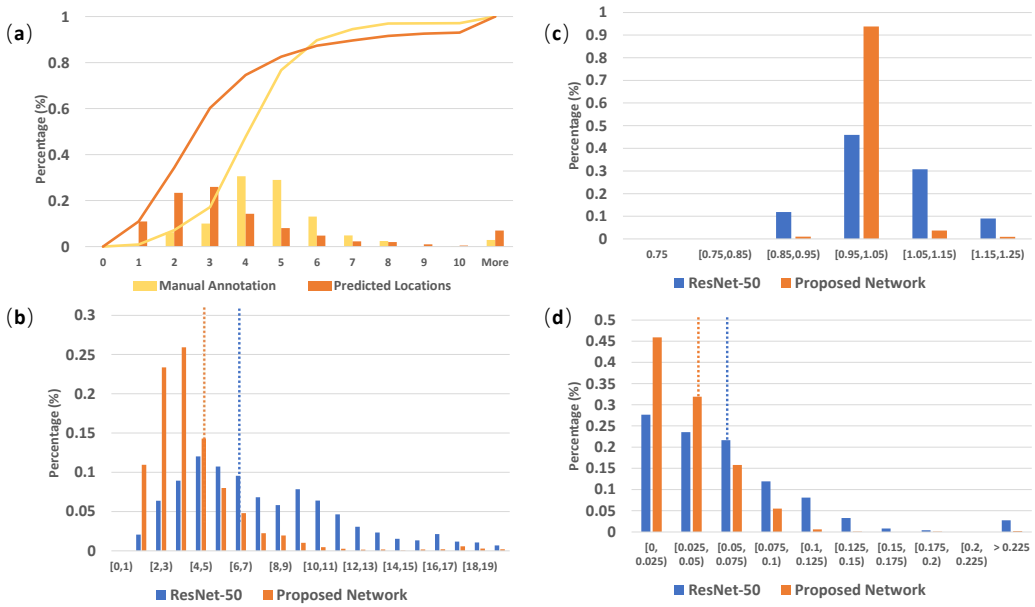


Figure 26: **Human judgment and landmark localization results**(a) The curve indicates the accumulative sum of locations distance errors. Yellow line indicates pixel distances between two human rater judgment and orange line indicates pixel distances between model prediction and one of human raters. (b) Distribution of localization distance errors between predicted and labeled annotation from first stage network and second stage network (c) Length ratio between predicted C2-C4 vector length and the manual annotation (d) Angle errors between predicted vector and manual annotation

Table 11: **Model performance with 5-fold cross validation** The performance of the model was evaluated with 5-fold cross-validation and each trained model was also tested on the healthy data set.

	Patient Data			Healthy Data		
	MLD	Angle Error	Length Ratio	MLD	Angle Error	Length Ratio
fold1	4.19 ± 4.77	0.04 ± 0.05	1.02 ± 0.04	4.14 ± 5.65	0.03 ± 0.04	1.00 ± 0.05
fold2	4.00 ± 4.26	0.03 ± 0.04	1.01 ± 0.04	4.54 ± 5.66	0.04 ± 0.04	1.00 ± 0.03
fold3	4.13 ± 4.51	0.03 ± 0.05	1.03 ± 0.04	5.37 ± 7.76	0.04 ± 0.04	1.00 ± 0.05
fold4	4.17 ± 6.64	0.02 ± 0.02	0.99 ± 0.03	4.85 ± 5.68	0.05 ± 0.04	0.99 ± 0.05
fold5	3.82 ± 4.90	0.03 ± 0.07	1.00 ± 0.03	4.49 ± 5.44	0.04 ± 0.03	1.01 ± 0.06

reduction methods while maintaining precise measurement is broad can improve the consistency of interpretations of VFSS images by providing standard measurements of swallow physiology that lower subjectivity in judgment leading to interventions for dysphagia. In current clinical setting, the importance of an anatomic scalar in VFSS measurement cannot be understated. Given the differences in the sizes of different patients and the direct association between a person's height and the dimensions of the upper aerodigestive tract [256], the ability to equalize measurements for differences in patient size provides the ability to compare results across patients of different dimensions. Moreover, real-time scaling of images provides immediate raw data for clinical interpretations which accelerates decision-making and increases efficiency of clinical workflow. In dysphagia diagnosis, the use of the vertebral scalar serves as the reference scale for linear measurements commonly used to infer about the nature of a patient's swallowing disorder (e.g., hyoid bone displacement, upper esophageal sphincter opening) that are the basis for determining appropriate treatments and judging the effects of those treatments objectively [185]. In turn, researchers investigating differences in swallow physiology in different disease states, and generation of population-based against which to compare patient function in disease states, provides for accurate determination of the magnitude of various kinematic impairments and a roadmap for determining the success or failure of treatments that restore that function.

Our two-level framework demonstrates the efficacy of using a large dataset and deep learning architectures for vertebrae landmark localization in videofluoroscopy images. Unlike previous semi-automation attempts for dysphagia keypoints [133], we conducted our model on a relatively large dataset, including over 300 subjects. Compared to other studies, we included the subjects across the adult age span varying from 19 to 94 years old and included both people with dysphagia and healthy subjects, showing the robustness of the algorithms. Additionally, our dataset not only collected single swallows, but also multiple sequential swallows and swallows in neutral and chin down head positions, all factors that are known to alter judgment of kinematic events when there is large scale motion of the patient during testing. Such diverse dataset prompted us to utilize deep learning approaches, avoiding the attempts of unstable, less powerful traditional image processing methods and classifiers.

Traditional image processing methods focuses on matching local edge and corner features. However, specific frames are rendered unmeasurable with these methods due to noisy edge and corner information in cases of patient motion during the exam, and the effect of the flowing bolus through the video field, influencing the performance of feature matching. In addition, these corner and edge features are influenced by image quality and various vertebral shape across different subjects. To overcome these limits and accurately detect the vertebrae shape with various location and edge shape, we adopted the two level framework in our study, which leverages deep learning technology and learns coarse representation from the VFSS dataset, followed by fine learning from the sub-regions to localize the keypoints on vertebrae. The coarse detection provides the approximate region of interest which contains C2, C3, C4 vertebrae information, removing the irrelevant information and also reducing the burden of computation for the second stage network. As shown in Fig. 25(b), the second stage network well improved the detection performance from the first network, which shows the importance of the usage of local network structure.

In this study, we have also demonstrated that the current framework can cope with the vertebral locations from videofluoroscopic images via two different videofluoroscopy systems and perform better than transfer learning techniques. Our framework was built based on ResNet-like structures with switchable-normalization, which is beneficial to the model generalization and stability. To compare the performance, we also trained our model using transfer learning techniques via the pre-trained network on Image-Net, a huge image database which contains various natural images. Transfer learning is a popular method that allows deep learning transferred the pre-knowledges to the new dataset, usually lower training burden and achieve better results. However, our results suggested that the usage of ResNet with switchable normalization instead of batch normalization and training the network from the scratch shows better performance to transfer learning techniques. Shown in the supplement figure, switchable normalization trained from scratch converged better than transfer learning with batch normalization . Furthermore, deeper ResNet structure proved a better accuracy.

Our study has some limitations, notably for the size of individual subjects and imaging resources. While our dataset is relatively large in the dysphagia community, it is still small compared to the popular medical imaging research on organs such as brain and lungs. The

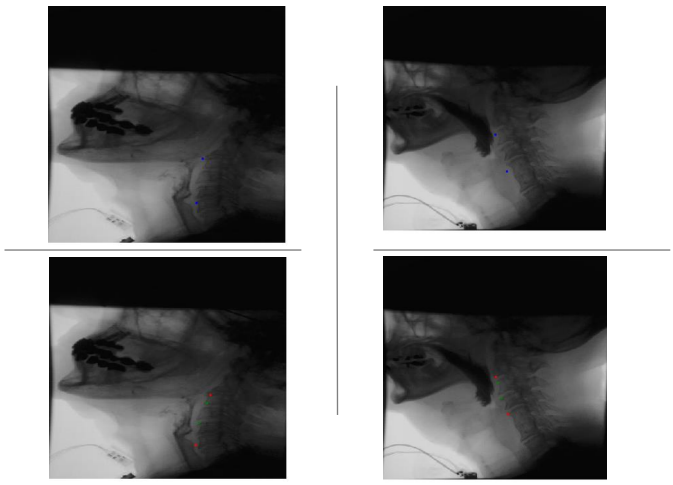


Figure 27: **Failure cases on testing dataset** Blue dots: predictions from one stage network. Green dots: C3 prediction from two stage networks. Red dots: C2, C4 tail edge detection from two statge networks. While two stage networks shows better results in numerical errors, we still can find that the landmark predictions are shifted when subjects are in a extreme posture or with an abnormal vertebra shape.

sample in this study may not be inclusive of the entire range of variety of anatomic information, which resulted in mis-localization in several cases. As shown in the figure 27, blue dots are the predictions from first stage network, and red/green dots are from second stage network. While second network improved the predictions from first network, its prediction were shifted in both case (a) and (b). In case (a), the C2 and C3 vertebrae contacted in the image due patient's head direction. The model correctly predicted the C2 tail but not other points. In case (b), the model failed to predict C4 tail due to abnormal C4 and C5 structure. The deep networks not only learned the features from the input image itself and the connection between input and output, they are able to learn the potential relationship between outputs, which might be the reason for this shifted wrong predictions. These abnormal cases such as abnormal bone shape (e.g., cervical osteophytes), postoperative anatomic disruption (e.g., anterior cervical fusion with graft or hardware), altered spinal configuration (e.g., kyphosis, excessive lordosis), or presence of feeding tubes or tracheostomies, provide direction for future research in model training that leads to better generalization of our model across more patient populations. We expect that the model performance will increase as more subjects

are included and images are collected from multiple videofluoroscopic machines. On the other hand, other techniques such as multi-stage networks and cascade network have been proposed in facial detection and pose estimation [145, 73]. These methods are not constrained by the global and local networks and use several networks to improve landmark locations step by step and may provide advantages that improve detection. However, whether these architectures can improve the performance for the VFSS detection with a larger dataset can improve the performance for the VFSS detection remains an opening question.

In the future, we would ideally extend the localization to other landmarks commonly considered in dysphagia studies (e.g., hyoid bone, arytenoid cartilages, valleculae, and epiglottis) as well as other parameters for swallow measurements. By extending our framework to study a wider range of features and providing a quantitative assessment in swallow videos, we hope that this deep learning approach is able to aid language pathologists' routine evaluation by automating some aspects of daily data analysis. This will enable clinicians to allocate their limited clinical resources on higher-level interpretations of the measurements to provide top-of-license services rather than spending valuable time performing the rote measurement necessary for these interpretations. We also hope that our framework could play an important role in research in order to develop more precise benchmarks for separating disordered from typical function that aids clinical interpretations, and in characterizing the properties of dysphagia in various disease states.

7.5 Conclusion

In this research, we introduced a deep learning neural network-based method for anatomic landmarks localization in videofluoroscopic images. We showed that our two-stage framework are are able to accurately estimate the length and angle of cervical vertebrae. We believe that deep learning approach will lead to automation of kinematic analysis that could speed up time to diagnosis and treatment.

8.0 Deep Learning-based Auto-Segmentation and Evaluation of Vallecular Residue in Videofluoroscopy

8.1 Motivation

Pharyngeal residue is widely considered as an indicator of swallow impairment in videofluoroscopic studies. The volume of residue indicates potential risk for penetration-aspiration on subsequent swallows while the accuracy and measurement guidelines of residue volume estimation varies significantly among human judges and facilities. Here, we present a machine learning algorithm that can efficiently identify the residue area remaining in vallecula and provide a normalized residue score to support clinical decision making. Here, we demonstrate how machine learning techniques can contribute to OPD assessment methods by using the strategies based on deep convolutional neural networks that achieves promising accuracy on vallecular residue measurement. These measurements are intended for use by speech-language pathologists (SLP) to help quantify certain aspects of VFSS interpretation. Most importantly, our models maintain the good performance when validated on a test dataset which is comparable to the manual labeling from experienced SLPs.

8.2 Methods

8.2.1 Videofluoroscopic Dataset Collection

Our dataset was collected from patients with swallowing difficulty who underwent videofluoroscopic examination at the Presbyterian University Hospital of the University of Pittsburgh Medical Center (Pittsburgh, Pennsylvania, USA). The Institutional Review Board at the University of Pittsburgh approved the protocol of this study and all participants were informed and signed the consent. The data collection in this experiments as performance under the relevant guidelines and regulations. The subjects are required to swallow barium liquid

during VFSS examination. These liquid contains various consistencies and volumes which was decided based on clinical hypotheses and patient's clinical presentation and symptoms of dysphagia. Subjects swallowed 3-5ml liquid bolus from a spoon, or self-selected comfortable volumes for one swallow from a cup containing 10-20 ml liquid. The following consistencies were used in our VFSS studies: E-Z-EM Canada, Inc. Varibar thin (Bracco Diagnostics, Inc.) (<5cPs viscosity), Varibar nectar (300 cPs viscosity), Varibar pudding (5000 cPs viscosity), and Keebler Sandies Mini Simply Shortbread Cookies (Kellogg Sales Company).

In this study, a LabVIEW program recorded data acquired by a video card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA) from X-ray machine (Ultimax system, Toshiba, Tustin, CA). VFSS was collected at 30 pulses per second and video clips were recorded at 60 frames per second. The VFSS videos were recorded with 720 x 1080 resolution in real time. Human experts were trained to determine the exact frame that contains the post-swallow vallecular residue and annotated the residue area remained in vallecular using segmentation tool in MATLAB (The MathWorks, Inc., Natick, MA, USA). The segmentation of vallecular residue follows the guideline described in [200]. The final annotations include 185 post-swallow cases. 10% of images were randomly selected for inter/intra-rater reliability test with intraclass correlation coefficient to avoid judgment drift.

8.2.2 $\%(C2-4)^2$ Measure Scale for Valleculae Residue

The presence of valleculae residue is an important indicators to understand the associated risk of penetration and aspiration in the subsequent swallows. Steele et al. proposed and recommended a pixel-based quantitative measurement called $\%(C2-C4)^2$ for valid, reliable and precise pharyngeal residue measurement [254] to estimate the residue severity. Their results showed that the risk of penetration and aspiration is extremely higher when $\%(C2-C4)^2 > 3\%$. In this study, we follow the findings from Steele's group, using 3% as cut-points of $\%(C2-4)$ measurement scales to evaluate our model performance. This scalar calculated the residue in C2-C4 units. An example of evaluation of vallecular residue and its $\%(c2-4)$ measurements is shown in Fig 31.

8.2.3 Annotation Principles and Quality Control

Human experts trained in swallow kinematic analysis were split into three groups for our data annotations. In each swallow video, one single frame was picked up based on the following rules: 1) post-swallow valleculae residue exists after the swallow. 2) frame was picked up when the hyoid bone returned to its rest position and valleculae is open to the largest space. The reliability test was perform among three experts with maximum of three frames differences in the video between two selected frames. The 2nd and 4nd vertebrae was manually marked by another three well-trained annotators with interclass correlation coefficient greater than 0.9. Residue area were labeled by two SLPs. In this study, we randomly selected 10 % of the frames and conducted both inter and intra reliability test for three types of annotations to ensure the robustness and high accuracy of manual labeling.

8.2.4 Dataset Augmentation Principles

In this investigation, the residue frame was selected from the swallow videos by experts. The total number of frames extracted from our swallow video dataset with residue annotations is 172 images. Due to the limitation of dataset size, the trained model may suffer a lot from overfitting and poorly generalization on a test dataset. Thus, data augmentation, a well accepted practice, was implemented to augment the variety of training image data to reduce the overfitting. We preprocessed the selected frames as follows: random flipping half of frames horizontally, randomly cropping the images containing the residue area, rotating the images from 45 degree to 45 degree, and changing the image brightness by multiplying 0.8 to 1.2. To reduce computational burdens, we resized our residue images into 224 x 224 in model training and deployment.

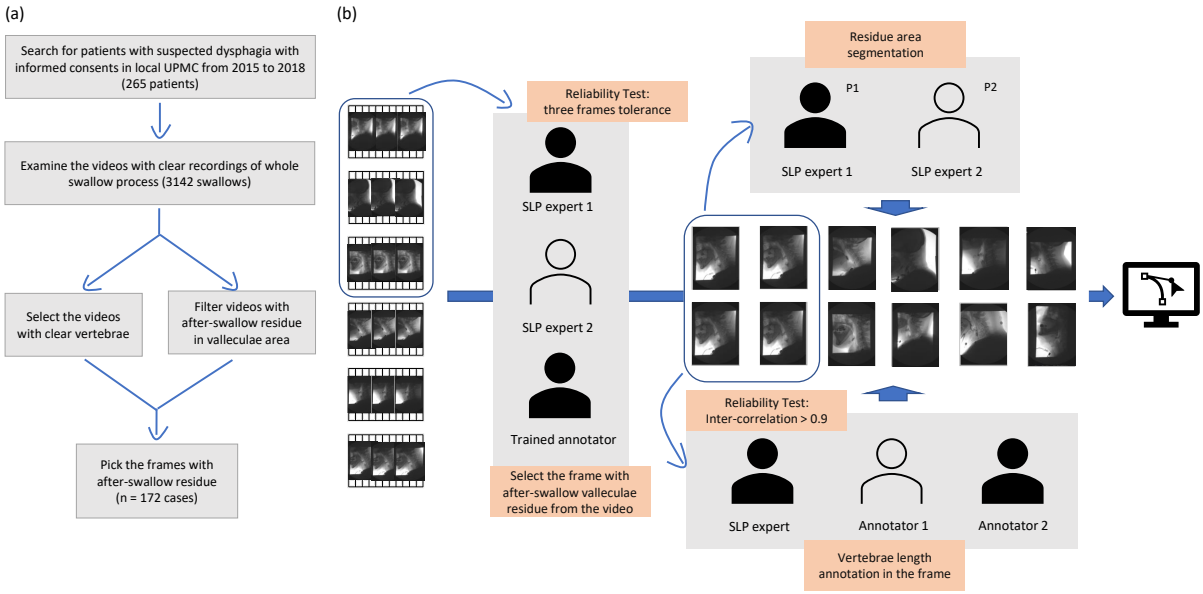


Figure 28: **Flowchart of data collection, selection and annotation** (a) Flowchart demonstrates data collection and selection. VFSS were conducted on 265 subjects suspected of dysphagia. Each subject followed instructions to swallow various consistencies and volumes of liquids within the context of routine clinical care. Video recordings that were included in analysis included clear imaging of an entire, single swallow, as defined as the bolus crossing the ramus of the mandible until the hyoid bone returned to rest. Frames were selected by ensuring that the second and fourth vertebrae were visible and the primary bolus was no longer present in the video. 172 swallow cases met the inclusion criteria for this study. (b) This flowchart presents the annotation procedures and a priori reliability testing before the model training. The residue frame was picked from the videos when vallecula was open to its largest space. These frames were selected by a trained expert in these methods, with a second rater completing reliability of these selections for 10% of the sample. Another set of experts marked vertebral locations, using the same reliability procedures. In this study, 10% of the videos and images are randomly selected for inter and intra reliability tests for all three annotation procedures to ensure the robustness and credence for those assessments.

8.3 Overview of Deep Convolutional Network

8.3.1 Motivation for Transfer learning

As the scale of radiology examinations and studies continue to grow, more and more data are being generated. Transfer learning has become an important step in radiology imaging applications. The basic concept behind transfer learning in deep learning techniques is taking the knowledge acquired from one particular domain and applying them to another specific task. Studies shows that small data regime benefits from the transfer learning techniques largely on deep architecture sizes [209]. A popular method of transfer learning is to take an existing architecture with pretrained weights on a well-known large dataset and then fine-tuning the model on the new medical imaging data. Prominent investigations have used this methodology by training architectures like ResNet, DenseNet, VGGNet on X-rays [86], mammography [173], and MRI studies [3]. In this investigation, the total amount of our dataset is far smaller than the large dataset such as COCO, and ImageNet. Considering the nature of our images: videofluoroscopic images are radiological images, which may not hold similar features as natural images from ImageNet, we chose radiological data stored on The Cancer Imaging Archive (TCIA) [23] instead to pretrain our models.

8.3.2 Segmentation Networks

Several convolutional network architectures were trained (UNet, ATT UNet, SQNet, and SegNet). Beyond these, we also implement the ensemble strategies with the idea of super-pixel on trained networks. Principles of trained networks are outlined in appendix 2. The transfer learning strategy with the initial trained weights on TCIA dataset was used for four model architectures. All the weights were further fine-tuned completely on our residue image dataset. The purpose of these models training is to segment the residue area in the videofluoroscopic frame and return the shape and pixel-based area surface to the clinicians. Four models were fine-tuned separately on the binary pixel-level classification task with two classes: residue area or not. We have selected the frame that contained the after-swallow residue and the clean swallow frames without remained residue were not selected as it's

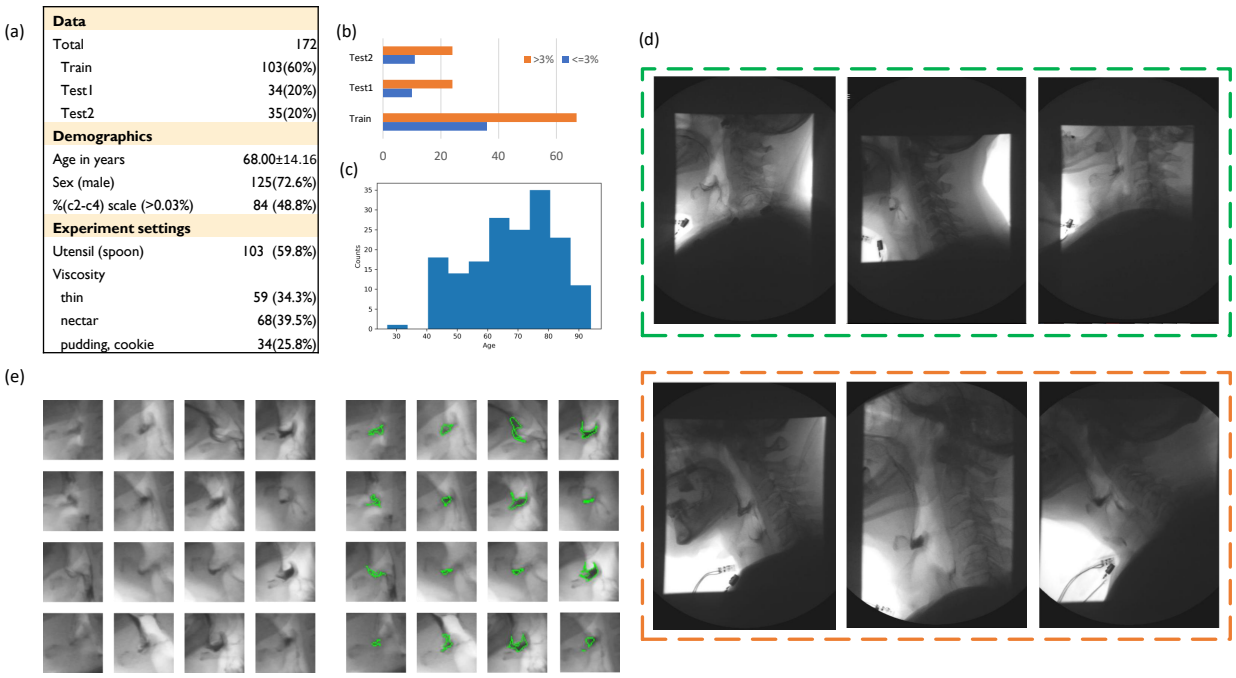


Figure 29: **Composition of dataset** (a) Total dataset consists of 172 swallow cases that were deeply annotated with regard to the residue area and vertebrae locations. We split the data into train (103), test 1 (34) in training and test 2 (33) for independent testing (b) Composition of data cohorts to which $\%(C2 - C4)$ measure scale are available. (c) Histogram of age associated to 172 selected swallows. (d) Examples of frame with the post-swallow residue, green indicates $\%(C2 - C4)$ measure scale less than 0.03% and red indicates the scale greater than 0.03% (higher risk of post-swallow penetration and aspiration). (e) Image crops containing the residue area and the manual annotation of region of interest.

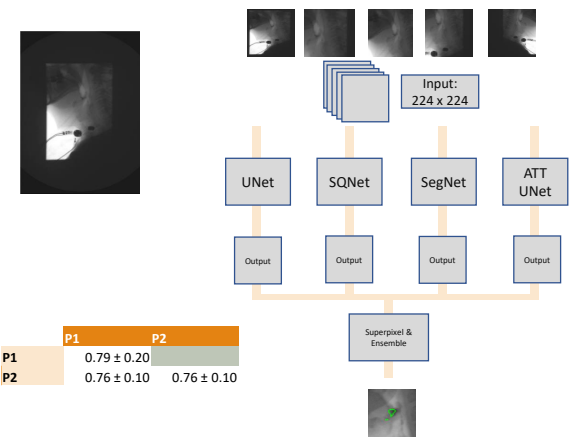
obvious for clinicians to conclude during the swallow examination and it also facilitates the training process. As mentioned in previous sections, data augmentation techniques were implemented in our four model training stage, which achieved slightly better accuracy empirically. Then the ensemble strategy was applied on output from four models. We segmented the image into super-pixels using simple linear iterative clustering method, then implemented the majority voting strategy on outputs of four models and superpixel methods for the final segmentation. We trained our models using combination of dice loss and focal loss function. The model weights were optimized by gradient descent with momentum.

After training, the performance was evaluated and reported using validation dataset (test1) and an extra test dataset (test2). No modifications were made to our models when evaluating on our test dataset. We derived receiver operator and recall curve at several threshold of probability prediction. We also derived dice coefficient and segment accuracy at 0.5 as probability cut-off of each model. Dice coefficient is defined as 2 x overlap area divided by the total number of pixel area in both segmentation. Segment accuracy is defined as the true positive percentage of the residue segmentation.

8.3.3 Development of Residue Grading Algorithm

Our goal is to provide residue details to clinicians in an efficient way. $\%(C2-C4)$ measurement scale is considers as an important scale in residue estimation. The residue pixel area and the length of vertebrae are required to compute the score of this scale. In our previous study, we developed two-stage network to localize the key points of vertebrae, then estimate the length between C2 and C4 as reference. $\%(C2-C4)$ measurement scale is calculated by the outcome of our segment networks and the vertebrae localization networks. We selected 0.03 % as our cut-off to distinguish safe swallows and swallows having higher risk of residue penetration/aspiration. We computed the receiver operator and recall curve at several threshold for segmentation network. The sensitivity and specificity is also derived by using manual labeled C2-C4 vertebrae length and predicted vertebrae length.

(a)



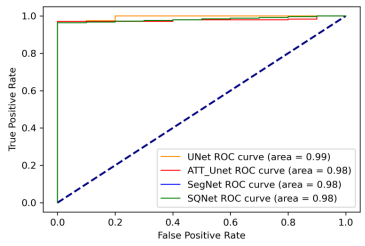
(b)

	P1	P2
P1	0.79 ± 0.20	
P2	0.76 ± 0.10	0.76 ± 0.10

(c)

	Test1		Test2	
	Dice	Segment Accuracy	Dice	Segment Accuracy
UNet	0.62 ± 0.24	0.63 ± 0.29	0.65 ± 0.21	0.63 ± 0.24
SegNet	0.66 ± 0.20	0.67 ± 0.24	0.67 ± 0.19	0.65 ± 0.22
SQNet	0.66 ± 0.20	0.67 ± 0.24	0.68 ± 0.14	0.71 ± 0.17
ATT_UNet	0.68 ± 0.20		0.67 ± 0.23	0.63 ± 0.23
Ensemble	0.68 ± 0.20	0.68 ± 0.24	0.72 ± 0.17	0.70 ± 0.19

(d)



(e)

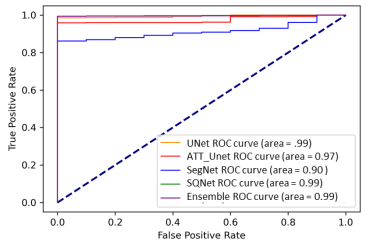


Figure 30: **Flowchart of segmentation networks and their performance** (a) Computational pipeline of segmentation networks. A given whole image is preprocessed and resized to fix sizes. Four CNNs based architectures predict the probability of residue area for each image. The final predictions are aggregated into a single prediction at each pixel by taking majority regions from super-pixel output. (b) The dice coefficient test on residue segmentation between two human expert raters (inter and intra test). (c) Table of segmentation performance from our architectures. (d) ROC curves of our models on test 1 data and test 2 data.

8.3.4 Environment

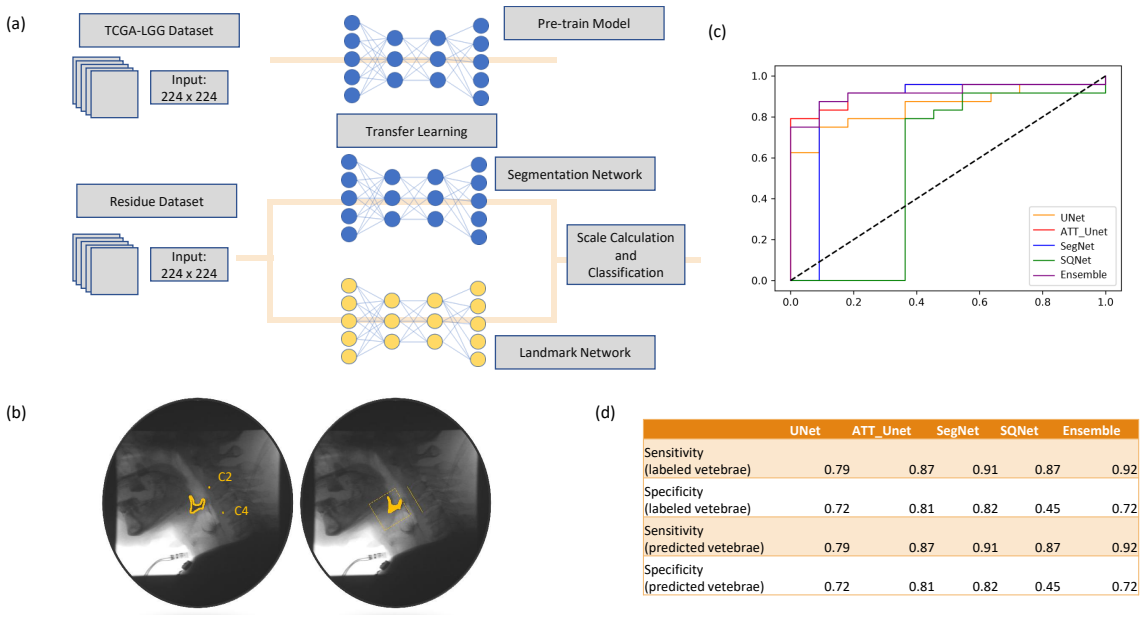
All the deep learning models were written, trained and deployed using Pytorch libraries in Python 3.6 environment. Our trainings and testings were implemented on a single workstation with Nvidia M60 card (24Gb).

8.3.5 Patient Characteristics

Fig 30(a) summarizes the clinical characteristics of each cohort. There are 172 swallows frames in total which qualified for our post-swallow residue selection rules. This study included subjects from 19 to 94 with mean age 68 ± 14.16. The age distribution of subjects is illustrated in Fig. 29(b). Subjects were instructed to swallow barium of various consistencies during the examination. The majority of swallows were administered with a spoon (5ml) and the others were taken by cup (20ml). The viscosities included thin liquid (59), nectar-thickened liquid (68), pudding and cookie (34). Among all swallows, 84 cases showed higher $\%(C2 - C4)^2$ residue measurement scale ($> 0.03\%$). To optimize the clinical utility of our algorithms, we included all swallow cases in our swallow studies, including various surgical procedures and clinical diagnoses (e.g., stroke, brain injury, neurodegenerative diseases).

8.3.6 Overview of Data

The flowchart of data selection is described in Fig. 28(a). Our full dataset of VFSS recordings were collected from 265 suspected dysphagia subjects from 2012 to 2018, which contains 3142 clear swallow recordings. The calculation of $\%(C2 - C4)$ residue measurement scale requires clear annotation of vertebrae and residue area; therefore, only frames with clear C2 to C4 vertebrae edges without obstacles were selected. As previously described, a group of experts determined frame with post-swallow residue in vallecular area. This process resulted in 172 cases with clear C2 to C4 vertebrae and measurable residue. To ensure the reliability of the frame selection and data annotations, several reliability testing measures were completed during the annotation process, as shown in Fig. 28(b). More details of reliability process and their corresponding scores are listed in the Methods section. Fig



(d)

	UNet	ATT_Unet	SegNet	SQNet	Ensemble
Sensitivity (labeled vetebrae)	0.79	0.87	0.91	0.87	0.92
Specificity (labeled vetebrae)	0.72	0.81	0.82	0.45	0.72
Sensitivity (predicted vetebrae)	0.79	0.87	0.91	0.87	0.92
Specificity (predicted vetebrae)	0.72	0.81	0.82	0.45	0.72

Figure 31: **Flowchart of $\%(C2-C4)$ measure scale prediction algorithms and their performance** (a) Flowchart of over all system, the residue frame was passed into segmentation network for residue part segmentation and landmark network for vertebrae localization, then the outputs were calculated for $\%(C2-C4)$ measure scale. (b) An example image showing present vallecular residue after a complete swallow. [left] Residue area and tail edges of 2nd and 4th cervical spine was rated by experts. [right] The length of the 2nd to 4th cervical spine (C2-C4) was measured in pixels. Then, the vallecular residue area was calculated as a percentage of the squared C2-C4 reference scalar. (c) Performance of models on $\%(C2-C4)$ measure scale (d)(e) Comparison table of models between manually labeled vertebrae locations and predicted ones.

29(d) presented several example frames with two different scales, which is challenging for clinicians to reliably judge the amount of retained residue visually. We then split the data into train, test1 and test2 cohorts. Fig. 29(b) shows the distribution of residue scale in each cohort.

8.3.7 Segmentation Performance using Various Networks

We first trained each segmentation network via transfer learning on training and validation (test 1) dataset, then tested their performance on test 2 dataset. We computed the dice coefficient and segment accuracy for the models and compared to the expert raters as reference standards. First, we examined the bias between human raters in this segmentation tasks before the model implementation. Fig. 30(b) summarized the bias between two human raters to demonstrate the accuracy of segmentation networks. The dice coefficient metric was conducted between two raters (inter) and raters themselves (intra) on randomly selected 20 residue frames. The inter-rater dice score within raters was 0.76 ± 0.10 and intra-rater reliability score was 0.79 ± 0.2 and 0.76 ± 0.10. Of note, the inter score was calculated based on the first initial segmentation results from two raters, and two raters segmented the same frames 2 weeks later for the intra score test.

Fig. 30 demonstrated that our ensemble methods achieved highly accurate residue segmentation results compared to expert clinicians. Fig. 30(d) shows the ROC curves for each segmentation networks against reference standard. Fig. 30(c) summarized dice coefficient and segment accuracy for each network on test 1 and test 2 dataset. Although ATT UNet outperformed the other networks on test 1 dataset, but we observed that four networks have a wide variability in performance on test 2 dataset, reflecting a poor generalizability. Compared to using independent architecture, the ensemble strategy show a more stable generalization, which used the four deep networks and superpixel segmentation to create an ensemble model based on the vote count and average of probability outcomes.

8.3.8 Residue Scale Classification

To explore the performance of our model in residue estimation, we calculated $\%(C2-C4)^2$ residue measurement with vertebral length and segmented vallecular residue area. As demonstrated in Fig. 31(b), the residue scale is estimated by the area of residue and square of C2-C4 vertebrae length. Fig. 31(a) presented a full flow-chart that how we arrived $\%(C2-C4)$ residue measurement scale. The residue area was predicted by the segmentation networks we mentioned in previous section, and we predicted the vertebrae edge point location in the frame using a localization network from our previous study. Fig. 31(c) shows the ROC curves for residue scale classification with each segmentation network. We chose 0.03% as cut-off value for residue scale , Fig. 31(d) presented the sensitivity and specificity for this classification. To show the performance of localization network, we present both results with vertebrae length from manual annotation and network predictions.

8.4 Discussion

This investigation provides a novel input to the application of machine learning algorithms for OPD assessment, in particular, for highly reliable segmentation of vallecular residue area and grading of $\%(C2-C4)$ residue measurement scale. Our findings show that our complex deep learning models achieved promising results on a small residue dataset by introducing the transfer learning techniques. On the independent dataset, our final deep learning ensemble classifier achieved 94 % of accuracy compared to human raters on residue area segmentation (0.72 vs 0.76 in dice coefficient performance). Moreover, it achieved a high sensitivity in residue scale estimation while maintaining high specificity.

Furthermore, this study demonstrated clinically meaningful results in measuring the area of the vallecula. Steele et al. firstly introduced $\%(c2-c4)$ vallecular residue scale measurement and compared the measurement accuracy against Normalized Residue Ratio Scale for vallecular residue estimation and demonstrated higher accuracy for vallecula residue ratings [258]. Intraclass coefficient reliability is widely accepted in swallowing research, as

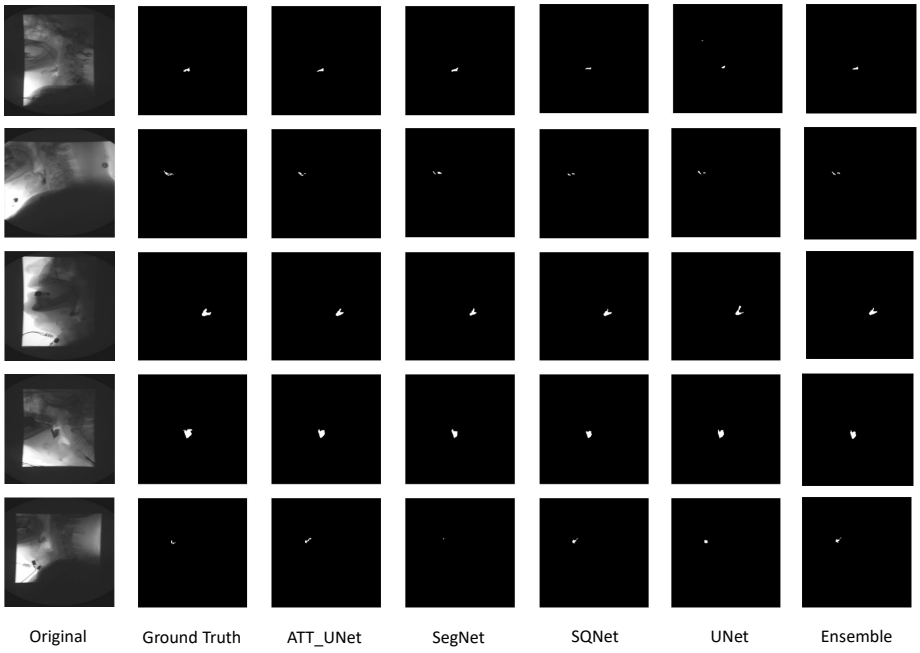


Figure 32: **An example of segmentation outputs from our tested models** From left to right: original image, human annotation, ATT Unet, SegNet, SQNet, UNet, and ensemble method.

demonstrated by its inclusion in similar measurements such as the Modified Barium Swallow Impairment Profile, hyoid bone point tracking, residue scale ratings, and upper esophageal sphincter diameter. Similar in Steele's study, they conducted intraclass correlation for reliability on pixel-based residue scale measurement. For other methods of measurement, such as manual segmentation in biomedical imaging, indicated dice coefficient or segment accuracy score is the preferred modality. Here, we provided an insight to the dice coefficient results when human raters were trained and confirmed with high intraclass correlation score. In this study, our experts were trained with high intraclass correlation score ($\geq$0.9) while the overlap in pixel-level didn't reach the same level of reliability score. We suggest more discussions on this finding, a high score criteria may not correspond to another evaluation criteria, should we introduce other evaluation criteria from a new field to ensure a better reliability.

The evaluation of post-swallow residue in OPD assessment is currently constrained by interpretation turnaround time. In addition, the related training is challenging to implement with respect tot reliability and agreement, and variants between institutions. Steele et al. claimed that one of five measurements of vallecular residue suffered from overestimate and underestimate. In contrast, the method we proposed here is only limited by computational resources and related cost considerations. Although the exact cut-off of residue measurement has not been fully established in clinical practice, our technique demonstrated good performance on both segmentation areas and one of most widely accepted thresholds of residue scale, showing robustness of our models to different practical choice. A strength of our methods is the implementation of several different strategies for the improvement of residue area segmentation. Although this strategy necessitates additional computational power increments for the hardware, they provided a valuable output to the segmentation accuracy. As shown in the Methods and Results sections, four networks were trained in the same parameter settings. However their performance was not robust on two datasets. As shown in figure 32, although the residue is very small area, and the models show different output specifically in edge and tiny connection points. Our ensemble method combines the four networks and super-pixels, an unsupervised segmentation method, and decreases the variance among outputs, which shows a better generalization compared to single network. We compared both manually-labeled and model-predicted C2-C4 length in the $\%(c2 - c4)^2$

residue scale. These showed that our landmark network in a previous study had perfectly commensurate performance also compared to the human raters. The ensemble method that showed best overall performance also showed the highest sensitivity rate compared to other methods, while its specificity was lower than other networks. It remains unclear whether sensitivity or specificity plays a more clinically significant role in residue scale diagnosis. Additional clinical application is needed to strengthen and refine these methods.

Despite supportive results, our study had some limitations. First, the number of patients and the residue cases was limited. A large swallowing dataset was collected with 3142 swallows; however, only 172 swallows met the inclusion criteria for measurable vallecular residue. Furthermore, patients with OPD of heterogeneous etiology may have confounding contributors to post-swallow residue. While the usage of transfer learning techniques contributes to the high performance of the model, a larger dataset is highly expected to enrich the image features and facilitate improved generalization of the models. Another limitation in these early stages is that the residue frames were determined by human experts first then the model was used on those frames. In addition, the images used to train the CNN are selected with known post-swallow residue areas, which is not always the case in a clinical scenario. Supplementary technique for residue frames selection is a part of necessary future directions where the ultimate goal is a fully automated assessment system. Finally, the focus of this approach was comparing the performance of our classifiers against experienced clinicians in our group. It is unknown whether our model is affected by the aforementioned judgment bias. More training and testing are required including annotations from experts with different background to reduce bias, improve the model performance, and promote generalization.

8.5 Conclusion

In this study, we developed precise deep learning based models for the analysis of VFSS images in patients with suspected OPD, which has the potential to reach commensurate accuracy to expert human judges. Our approach provides clinically-relevant information regarding vallecular residue, which enables SLPs making appropriate recommendations based

on quantifiable data that may be compared and analyzed systematically. Further data collection and validation in real-world clinical diagnostic work flow is anticipated as a future direction.

9.0 Conclusion and Future Directions

9.1 Conclusion

Oropharyngeal dysphagia poses serious health risks to people who suffer from stroke, head and neck cancer, older adults with multiple medical conditions, prematurely born infants and children with neurological, airway and developmental disorders. Despite current VFSS providing real-time visualization of swallowing during consumption of various consistencies, the kinematic analysis such as hyoid bone movement and penetration-aspiration are still under investigated. In addition, attempts to quantify and measure these kinematic parameters and events of swallowing using an computer-assisted system have been limited. As a result, developing an automatic segmentation and annotation system with high accuracy, but also can be deployed easily in the clinical practice has the potential to further research studies and daily diagnosis. The research described in this manuscript attempted to advance the use of machine learning and deep learning techniques in kinematic analysis and automated assessment. While some contributions have been made on this topic, pass research mainly included semi-automatic methods, which requires the human intervention or verification. The in-depth investigation and increasing usage of machine learning and deep learning techniques leads to the automation of image-based processing including segmentation, annotation and classification. In this manuscript, we attempted to use generalized equation to investigate kinematic features such hyoid bone movement features as input and penetration-aspiration scale as output, that revealed the significant association between maximum hyoid bone displacement and penetration/aspiration. Besides the statistical analyze, we also sought to provide methods to detect, segment and classify the kinematic parameters from VFSS images using state of art network architectures.

Ultimately, we achieved our stated goals successfully. We were able to track the hyoid bone in VFSS with high accuracy using object detection networks. We were also able to localize the vertebrae edge points (C2, C3, C4) with missing annotations in training data. Both of methods shows high reliability compared to human raters. Furthermore, we seg-

mented the vallecular residue and estimated the $\%(C2-C4)$ measurement scale with high sensitivities and specificity. Specially, the $\%(C2-C4)$ measurement scale consists of residue area and vertebrae edge points detection, which utilized the previous vertebrae networks, proving the robustness of models in kinematic analysis. All of these research justify out key points: that the deep learning techniques can play an important role in automation of dysphagia assessment like other computer-assisted medical imaging applications.

9.2 Future Directions

There are several directions that would be meaningful to investigate in future studies. First, one of the most challenging and important topics in swallowing study is the time-sink in determining kinematic biomarkers and when four swallow phases start and end. Similar to our previous work such as vertebrae detection and hyoid bone tracking, experts tend to go through each frame from the video clips to determine each swallow phase which needs the tracking of the movement of bolus and several kinematic biomarkers. For example, the onset movement of hyoid bone leads to the start of pharyngeal phase and the closing of upper esophageal sphincter indicates the end of pharyngeal phase and start of esophageal phase. It may bee worth investigating various algorithms and their understandings on these events such as swallowing phase detection, opening/closing of UES. In this work, the relationship between hyoid bone movement and penetration/aspiration has been investigated and an algorithm was presented for hyoid bone tracking in video clips. It is worth investigating using different kinds of deep learning algorithms to automatically classify the severity of PA scales in swallowing video clips.

Second, it would also be useful to refine the methods in this work in future studies. The algorithms presented in this work have shown prominent performance and demonstrated the potential validity. However, there is still gap between what has been achieved here and a clinical assessment and environment. The model requires more evaluation and validation across various facilities and clinical settings. Besides the validation purpose, it's also interesting to investigating the model robustness when the training data are noisy/incorrect labels due to

different level of expertise from various annotators. Furthermore, the data and labels may be highly imbalanced. For example, in our dataset, only 1 % of swallow videos are diagnosed with aspiration (PA scale > 6). How to deal with imbalanced data and insufficient data is an important direction for further model validation.

Finally, our ultimate goal is to develop an fully automatic assessment system. The results presented in this work are only a small portion of this system. It's critical to consider the computational limitations and robustness of the whole system when these methods are merged together. Several models with different architectures were investigated in this work, being able to build one or reduced model that can achieve all these tasks would provide great benefits to the study of swallowing assessment.

Bibliography

[1] Diseases and Conditions: Dysphagia. http://www.mayoclinic.org/diseases-conditions/dysphagia/basics/causes/con-20033444. Accessed: 2014-10-15.

[2] Bilal Ahmed, Carla E Brodley, Karen E Blackmon, Ruben Kuzniecky, Gilad Barash, Chad Carlson, Brian T Quinn, Werner Doyle, Jacqueline French, Orrin Devinsky, and Thomas Thesen. Cortical feature analysis and machine learning improves detection of "MRI-negative" focal cortical dysplasia. *Epilepsy & Behavior*, 48:21–28, 2015.

[3] Kaoutar B Ahmed, Lawrence O Hall, Dmitry B Goldgof, Renhao Liu, and Robert A Gatenby. Fine-tuning convolutional deep features for mri based brain tumor classification. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101342E. International Society for Optics and Photonics, 2017.

[4] Euijoon Ahn, Ashnil Kumar, Jinman Kim, Changyang Li, Dagan Feng, Michael Fulham, Nuclear Medicine, Royal Prince, and Alfred Hospital. X-ray image classification using domain transferred convolutional neural networks and local sparse spatial pyramid. In *2016 IEEE 13th International Symposium on Biomedical Imaging*, pages 855–858, 2016.

[5] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–738, 2009.

[6] Mohammed Alkhawlani and Mohammed Elmogy. Content-based image retrieval using local features descriptors and bag-of-visual words. *International Journal of Advanced Computer Science and Applications*, 6(9):212–219, 2015.

[7] Jacqui E Allen, Cheryl J White, Rebecca J Leonard, and Peter C Belafsky. Prevalence of penetration and aspiration on videofluoroscopy in normal individuals without dysphagia. *Otolaryngology–Head and Neck Surgery*, 142(2):208–213, 2010.

[8] Kenneth W Altman, Gou-Pei Yu, and Steven D Schaefer. Consequence of dysphagia in the hospitalized patient: impact on prognosis and hospital resources. *Archives of Otolaryngology–Head & Neck Surgery*, 136(8):784–789, 2010.

[9] Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):88, 2015.

[10] American Speech Language Hearing Association. Role of the speech-language pathologist in the performance and interpretation of endoscopic evaluation of swallowing: Guidelines. 2004.

[11] S Anderson, M H Biros, and R F Reardon. Delayed diagnosis of thoracolumbar fractures in multiple-trauma patients. *Academic Emergency Medicine*, 3(9):832–839, 1996.

[12] Sameer Antani. Automated detection of lung diseases in chest X-Rays. *US National Library of Medicine*, 2015.

[13] John Arevalo, Fabio A. González, Raúl Ramos-Pollán, Jose L. Oliveira, and Miguel Angel Guevara Lopez. Convolutional neural networks for mammography mass lesion classification. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 797–800, 2015.

[14] John Arevalo, Fabio A. González, Raúl Ramos-Pollán, Jose L. Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 127:248–257, 2016.

[15] Jacinto Arias, Jesus Martínez-Gómez, Jose A. Gámez, Alba G. Seco de Herrera, and Henning Müller. Medical image modality classification using discrete Bayesian networks. *Computer Vision and Image Understanding*, 151:61–71, 2016.

[16] R Armananzas, M Iglesias, D A Morales, and L Alonso-Nanclares. Voxel-based diagnosis of Alzheimer's disease using classifier ensembles. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–7, 2016.

[17] Joan C Arvedson. Assessment of pediatric dysphagia and feeding disorders: clinical and instrumental approaches. *Developmental Disabilities Research Reviews*, 14(2):118–127, 2008.

[18] MR Avendi, Arash Kheradvar, and Hamid Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical Image Analysis*, 30:108–119, 2016.

[19] Turgay Ayer, Mehmet Us Ayvaci, Ze Xiu Liu, Oguzhan Alagoz, and Elizabeth S Burnside. Computer-aided diagnostic models in breast cancer screening. *Imaging in Medicine*, 2(3):313–323, 2010.

[20] T Ayuse, S Ishitobi, S Kurata, E Sakamoto, I Okayasu, and K Oi. Effect of reclining and chin-tuck position on the coordination between respiration and swallowing. *Journal of Oral Rehabilitation*, 33(6):402–408, 2006.

[21] Ahmad Taher Azar and Shereen M. El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7-8):2387–2403, 2013.

[22] Snehashis Roy B, Aaron Carass, Jerry L Prince, and Dzung L Pham. Semi-automatic liver tumor segmentation in dynamic contrast-enhanced CT scans using random forests and supervoxels. In *International Workshop on Machine Learning in Medical Imaging*, pages 212—-219, 2015.

[23] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017.

[24] S R Barczi, P A Sullivan, and J Robbins. How should dysphagia care of older adults differ? Establishing optimal practice patterns. *Seminars in Speech and Language*, 21(4):347–361, 2000.

[25] Robert W Bastian. The videoendoscopic swallowing study: an alternative and partner to the videofluoroscopic swallowing study. *Dysphagia*, 8(4):359–367, 1993.

[26] Eric Bauer, Ron Kohavi, Philip Chan, Salvatore Stolfo, and David Wolpert. An empirical comparison of voting classification algorithms: bagging, Boosting, and variants. *Machine Learning*, 36(August):105–139, 1999.

[27] Tara G Bautista, Qi-Jian Sun, and Paul M Pilowsky. The generation of pharyngeal phase of swallow and its coordination with breathing: interaction between the swallow and respiratory central pattern generators. *Progress in Brain Research*, 212:253, 2014.

[28] Neil Bhattacharyya. The prevalence of dysphagia among adults in the united states. *Otolaryngology–Head and Neck Surgery*, 151(5):765–769, 2014.

[29] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[30] Ingmar Blümcke, Maria Thom, Eleonora Aronica, Dawna D. Armstrong, Harry V. Vinters, Andre Palmini, Thomas S. Jacques, Giuliano Avanzini, A. James Barkovich, Giorgio Battaglia, Albert Becker, Carlos Cepeda, Fernando Cendes, Nadia Colombo, Peter Crino, J. Helen Cross, Olivier Delalande, François Dubeau, John Duncan, Renzo Guerrini, Philippe Kahane, Gary Mathern, Imad Najm, Çiğdem ÖZKARA, Charles Raybaud, Alfonso Represa, Steven N. Roper, Noriko Salamon, Andreas Schulze-Bonhage, Laura Tassi, Annamaria Vezzani, and Roberto Spreafico. The clinicopathologic spectrum of focal cortical dysplasias: a consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. *Epilepsia*, 52(1):158–174, 2011.

[31] Geovana de Paula Bolzan, Mara Keli Christmann, Luana Cristina Berwig, Cintia Conceição Costa, and Renata Mancopes Rocha. Contribution of the cervical auscultation in clinical assessment of the oropharyngeal dysphagia. *Revista CEFAC*, 15(2):455–465, 2013.

[32] Heather Shaw Bonilha, Annie N Simpson, Charles Ellis, Patrick Mauldin, Bonnie Martin-Harris, and Kit Simpson. The one-year attributable cost of post-stroke dysphagia. *Dysphagia*, 29(5):545–552, 2014.

[33] James F Bosma. Deglutition: pharyngeal stage. *Physiological Reviews*, 37(3):275–300, 1957.

[34] E. Bron, M. Smits, J. Van Swieten, W. Niessen, and S. Klein. Feature selection based on SVM significance maps for classification of dementia. In *International Workshop on Machine Learning in Medical Imaging*, pages 272–279, 2014.

[35] Joseph E Burns, Jianhua Yao, Hector Muñoz, and Ronald M Summers. Automated detection, localization, and classification of traumatic vertebral body fractures in the thoracic and lumbar spine at CT. *Radiology*, 278(1):64–73, 2016.

[36] Susan G Butler, Andrew Stuart, Xiaoyan Leng, Catherine Rees, Jeff Williamson, and Stephen B Kritchevsky. Factors influencing aspiration during swallowing in healthy older adults. *The Laryngoscope*, 120(11):2147–2152, 2010.

[37] Yu Cao, Shawn Steffey, He Jianbiao, Degui Xiao, Cui Tao, Ping Chen, and Henning Müller. Medical image retrieval: a multimodal approach. *Cancer Informatics*, 13:125–136, 2015.

[38] Leeanne M Carey, Rüdiger J Seitz, Mark Parsons, Christopher Levi, Shawna Farquharson, Jacques-Donald Tournier, Susan Palmer, and Alan Connelly. Beyond the lesion: neuroimaging foundations for post-stroke recovery. *Future Neurol*, 8(5):507–527, 2013.

[39] Rachel Aguiar Cassiani, Carla Manfredi Santos, Luana Casari Parreira, and Roberto Oliveira Dantas. The relationship between the oral and pharyngeal phases of swallowing. *Clinics*, 66(8):1385–1388, 2011.

[40] Samuel T. Chao, Manmeet S. Ahluwalia, Gene H. Barnett, Glen H J Stevens, Erin S. Murphy, Abigail L. Stockham, Kevin Shiue, and John H. Suh. Challenges with the diagnosis and treatment of cerebral radiation necrosis. *International Journal of Radiation Oncology Biology Physics*, 87(3):449–457, 2013.

[41] C. Chen, W. Xie, J. Franke, P. A. Grutzner, L. P. Nolte, and G. Zheng. Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. *Medical Image Analysis*, 18(3):487–499, 2014.

[42] Cheng Chen, Daniel Belavy, and Guoyan Zheng. 3D intervertebral disc localization and segmentation from MR images by data-driven regression and classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 50–58. Springer, 2014.

[43] Yani Chen, Bibo Shi, Charles D. Smith, and Jundong Liu. Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 304–312, 2015.

[44] Bo Cheng, Mingxia Liu, and Daoqiang Zhang. Multimodal multi-label transfer learning for early diagnosis of Alzheimer's disease. In *International Workshop on Machine Learning in Medical Imaging*, pages 238–245. Springer, 2015.

[45] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-Aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. *Scientific Reports*, 6, 2016.

[46] Sungbin Choi. X-ray image body part clustering using deep convolutional neural network. *ImageCLEF 2015 Medical Clustering Task*, pages 6–8, 2015.

[47] Carlton Chu, Ai Ling Hsu, Kun Hsien Chou, Peter Bandettini, and ChingPo Lin. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1):59–70, 2012.

[48] M. Chupin, A. Hammers, R. S N Liu, O. Colliot, J. Burdett, E. Bardinet, J. S. Duncan, L. Garnero, and L. Lemieux. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation. *NeuroImage*, 46(3):749–761, 2009.

[49] P Clavé, M De Kraa, V Arreola, M Girvent, Ricard Farre, E Palomera, and M Serra-Prat. The effect of bolus viscosity on swallowing function in neurogenic dysphagia. *Alimentary Pharmacology & Therapeutics*, 24(9):1385–1394, 2006.

[50] P. Clavé, R Terré, M de Kraa, and M Serra. Approaching oropharyngeal dysphagia. *Revista Espanola de Enfermedades Digestivas*, 96(2):119–131, 2004.

[51] Pere Clavé and Reza Shaker. Dysphagia: current reality and scope of the problem. *Nature Reviews Gastroenterology and Hepatology*, 12(5):259–270, 2015.

[52] I. J. Cook and P. J. Kahrilas. AGA technical review on management of oropharyngeal dysphagia. *Gastroenterology*, 116(2):455–478, 1999.

[53] James L. Coyle, Lori A. Davis, Caryn Easterling, Darlene E. Graner, Susan Langmore, Steven B. Leder, Maureen A. Lefton-Greif, Paula Leslie, Jeri A. Logemann, Linda Mackay, Bonnie Martin-Harris, Joseph T. Murray, Barbara Sonies, and Catriona M. Steele. Oropharyngeal dysphagia assessment and treatment efficacy: setting the record straight (Response to Campbell-Taylor). *Journal of the American Medical Directors Association*, 10(1):62–66, 2009.

[54] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–410, 2013.

[55] Rémi Cuingnet, Emilie Gerardin, Jér{\ˆo}me Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, and Alzheimer's Disease Neuroimaging Initiative. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, 2011.

[56] Alicia Daggett, Jeri Logemann, Alfred Rademaker, and Barbara Pauloski. Laryngeal penetration during deglutition in normal subjects of various ages. *Dysphagia*, 21(4):270–274, 2006.

[57] Peter Dayan. Unsupervised learning. *The MIT Encyclopedia of the Cognitive Sciences*, pages 1–7, 2009.

[58] Kan Deng. *OMEGA : On-line memory-based general purpose system classifier*. PhD thesis, Carnegie Mellon University, 1998.

[59] Sami Dhahbi, Walid Barhoumi, and Ezzeddine Zagrouba. Breast cancer diagnosis in digitized mammograms using curvelet moments. *Computers in Biology and Medicine*, 64:79–90, 2015.

[60] Ashis Kumar Dhara, Sudipta Mukhopadhyay, Anirvan Dutta, Mandeep Garg, and Niranjan Khandelwal. A combination of shape and texture features for classification of pulmonary nodules in lung CT images. *Journal of Digital Imaging*, 29(4):466–475, 2016.

[61] Ruiying Ding, Charles R Larson, Jeri A Logemann, and Alfred W Rademaker. Surface electromyographic and electroglottographic studies in normal subjects under two swallow conditions: normal and during the mendelsohn manuever. *Dysphagia*, 17(1):1–12, 2002.

[62] Wylie J Dodds, Edward T Stewart, and Jeri A Logemann. Physiology and radiology of the normal oral and pharyngeal phases of swallowing. *American Journal of Roentgenology*, 154(5):953–963, 1990.

[63] Piotr Dollar and C. Lawrence Zitnick. Structured forests for fast edge detection. In *IEEE International Conference on Computer Vision*, pages 1841–1848, 2013.

[64] Shiv Ram Dubey, Satish Kumar Singh, and Rajat Kumar Singh. Local wavelet pattern: a new feature descriptor for image retrieval in medical CT databases. *IEEE Transactions on Image Processing*, 24(12):5892–5903, 2015.

[65] Joshua M. Dudik, James L. Coyle, Amro El-Jaroudi, Mingui Sun, and Ervin Sejdić. A matched dual-tree wavelet denoising for tri-axial swallowing vibrations. *Biomedical Signal Processing and Control*, 27:112–121, 2016.

[66] Raja Ebsim, Jawad Naqvi, and Tim Cootes. Detection of wrist fractures in x-ray images. In *Workshop on Clinical Image-Based Procedures*, pages 1–8. Springer, 2016.

[67] Jeff Edmiaston, Lisa Tabor Connor, Lynda Loehr, and Abdullah Nassief. Validation of a dysphagia screening tool in acute stroke patients. *American Journal of Critical Care*, 19(4):357–364, 2010.

[68] Tobias Emrich, Franz Graf, Hans Peter Kriegel, Matthias Schubert, and Marisa Thoma. Similarity estimation using Bayes ensembles. In *International Conference on Scientific and Statistical Database Management*, pages 537–554, 2010.

[69] Enteral Nutrition ASPEN Public Policy. Disease-related malnutrition and enteral nutrition therapy: a significant problem with a cost-effective solution. *Nutrition in Clinical Practice*, 25(5):548–554, 2010.

[70] Cumhur Ertekin and Ibrahim Aydogdu. Neurophysiology of swallowing. *Clinical Neurophysiology*, 114(12):2226–2244, 2003.

[71] Simon F. Eskildsen, Pierrick Coupé, Vladimir Fonov, José V. Manjón, Kelvin K. Leung, Nicolas Guizard, Shafik N. Wassef, Lasse Riis Østergaard, and D. Louis Collins. BEaST: brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362–2373, 2012.

[72] Konstantinos P. Exarchos, Yorgos Goletsis, and Dimitrios I. Fotiadis. Multiparametric decision support system for the prediction of oral cancer reoccurrence. *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1127–1134, 2012.

[73] Haoqiang Fan and Erjin Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016.

[74] Yong Fan, Dinggang Shen, Ruben C Gur, Raquel E Gur, and Christos Davatzikos. COMPARE: classication of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105, 2007.

[75] Andreia V. Faria, Kenichi Oishi, Shoko Yoshida, Argye Hillis, Michael I. Miller, and Susumu Mori. Content-based image retrieval for brain MRI: an image-searching engine and population-based analysis to utilize past clinical data for future diagnosis. *NeuroImage: Clinical*, 7:367–376, 2015.

[76] A Forster, N Samaras, G Gold, and D Samaras. Oropharyngeal dysphagia in older adults: a review. *European Geriatric Medicine*, 2(6):356–362, 2011.

[77] Kenneth R Foster, Robert Koprowski, and Joseph D Skufca. Machine learning, medical diagnosis, and biomedical engineering research-commentary. *Biomedical Engineering online*, 13(1):94, 2014.

[78] Bernard R Garon, Tess Sierzant, and Charles Ormiston. Silent aspiration: results of 2,000 video fluoroscopic evaluations. *Journal of Neuroscience Nursing*, 41(4):178–185, 2009.

[79] Dj Gelb, E Oliver, and S Gilman. Diagnostic criteria for Parkinson disease. *Arch Neurol*, 56(4):368–376, 1999.

[80] Ezequiel Geremia, Olivier Clatz, Bjoern H. Menze, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.

[81] Vanathi Gopalakrishnan, Prahlad G Menon, and Shobhit Madan. cMRI-BED: A novel informatics framework for cardiac MRI biomarker extraction and discovery applied to pediatric cardiomyopathy classification. *Biomedical Engineering online*, 14(2):S7, 2015.

[82] Raj K Goyal and Hiroshi Mashimo. Physiology of oral, pharyngeal, and esophageal motility. *GI Motility online*, 2006.

[83] Leo Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[84] Joseph C. Griffis, Jane B. Allendorfer, and Jerzy P. Szaflarski. Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. *Journal of Neuroscience Methods*, 257:97–108, 2016.

[85] R Guerrero, C Ledig, and D Rueckert. Manifold alignment and transfer learning for classification of Alzheimer's disease. In *International Workshop on Machine Learning in Medical Imaging*, pages 77–84, 2014.

[86] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge

Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[87] Rohith Reddy Gundreddy, Maxine Tan, Yuchen Qiu, Samuel Cheng, Hong Liu, and Bin Zheng. Assessment of performance and reproducibility of applying a content-based image retrieval scheme for classification of breast lesions. *Medical physics*, 42(7):4241–4249, 2015.

[88] Yanrong Guo, Yaozong Gao, and Dinggang Shen. Deformable mr prostate segmentation via deep feature learning and sparse patch matching. *IEEE Transactions on Medical Imaging*, 35(4):1077–1089, 2016.

[89] James A Hanley, Abdissa Negassa, and Janet E Forrester. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*, 157(4):364–375, 2003.

[90] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[91] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[92] Dani-Ella Heckathorn, Renée Speyer, Jessica Taylor, and Reinie Cordier. Systematic review: non-instrumental swallowing and feeding assessments in pediatrics. *Dysphagia*, 31(1):1–23, 2016.

[93] Johanna Hedström, Lisa Tuomi, Mats Andersson, Hans Dotevall, Hanna Osbeck, and Caterina Finizia. Within-bolus variability of the penetration-aspiration scale across two subsequent swallows in patients with head and neck cancer. *Dysphagia*, pages 1–8, 2017.

[94] Alba G Seco De Herrera, Dimitrios Markonis, Ranveer Joyseeree, Roger Schaer, and Antonio Foncubierta-rodr. Semi – supervised learning for image modality classification. *Multimodal Retrieval in the Medical Domain*, pages 85–98, 2015.

[95] Steven C H Hoi, Wei Liu, Michael R. Lyu, and Ma Wei-Ying. Learning distance metrics with contextual constraints for image retrieval. In *IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition, volume 2, pages 2072–2078, 2006.

[96] Seok Jun Hong, Hosung Kim, Dewi Schrader, Neda Bernasconi, Boris C. Bernhardt, and Andrea Bernasconi. Automated detection of cortical dysplasia type II in MRI-negative epilepsy. *Neurology*, 83(1):48–55, 2014.

[97] Peijun Hu, Fa Wu, Jialin Peng, Ping Liang, and Dexing Kong. Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. *Physics in Medicine and Biology*, 61(24):8676, 2016.

[98] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv preprint arXiv:1611.10012*, 2016.

[99] Lei Huang, Yaozong Gao, Yan Jin, Kim-Han Thung, and Dinggang Shen. Soft-split sparse regression based random forest for predicting future clinical scores of Alzheimer's disease. *International Workshop on Machine Learning in Medical Imaging*, pages 194–202, 2015.

[100] Christopher V Hughes, Bruce J Baum, Philip C Fox, Yitzhak Marmary, Chih-Ko Yeh, and Barbara C Sonies. Oral-pharyngeal dysphagia: a common sequela of salivary gland dysfunction. *Dysphagia*, 1(4):173–177, 1987.

[101] Wooseop Hwang, Seunghee Ha, and Sujin Hwang. Displacement of the hyoid bone among normal, aspirated, and penetrated swallows in post-stroke patients with dysphagia. *Commun Sci Disord*, 16(3):372–387, 2011.

[102] HIU Sadiq Jaafar Ibrahim and A Mukhtar. Content based image retrieval in mammograms: a survey. *International Journal of Engineering Science*, 4638, 2016.

[103] Ryo Ishida, Jeffrey B. Palmer, and Karen M. Hiiemae. Hyoid motion during swallowing: Factors affecting forward and upward displacement. *Dysphagia*, 17(4):262–272, 2002.

[104] Md Monirul Islam, Dengsheng Zhang, and Guojun Lu. A geometric method to compute directionality features for texture images. In *IEEE International Conference on Multimedia and Expo*, number 3, pages 1521–1524, 2008.

[105] Menglin Jiang, Shaoting Zhang, and DimitrisN. Metaxas. Detection of mammographic masses by content-based image retrieval. In *International Workshop on Machine Learning in Medical Imaging*, pages 33–41, 2014.

[106] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi. Improving breast cancer diagnosis with computer-aided diagnosis. *Academic Radiology*, 6(1):22–33, 1999.

[107] Zhicheng Jiao, Xinbo Gao, Ying Wang, and Jie Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 197:1–11, 2016.

[108] Yan Jin, Yonggang Shi, Liang Zhan, Boris A. Gutman, Greig I. de Zubicaray, Katie L. McMahon, Margaret J. Wright, Arthur W. Toga, and Paul M. Thompson. Automatic clustering of white matter fibers in brain diffusion MRI with an application to genetics. *NeuroImage*, 100:75–90, 2014.

[109] Luo Juan and O Gwun. A comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing*, 3(4):143–152, 2009.

[110] Jiayin Kang, Yaozong Gao, Feng Shi, David S. Lalush, Weili Lin, and Dinggang Shen. Prediction of standard-dose PET image by low-dose PET and MRI images. *Medical Physics*, 42(9):5301–5309, 2015.

[111] Patrick M Kellen, Darci L Becker, Joseph M Reinhardt, and Douglas J Van Daele. Computer-assisted assessment of hyoid bone motion from videofluoroscopic swallow studies. *Dysphagia*, 25(4):298–306, 2010.

[112] AM Kelly, P Leslie, T Beale, C Payten, and MJ Drinnan. Fibreoptic endoscopic evaluation of swallowing and videofluoroscopy: does examination type influence perception of pharyngeal residue severity? *Clinical Otolaryngology*, 31(5):425–432, 2006.

[113] Prateek Keserwani, V. S. Chandrasekhar Pammi, Om Prakash, Ashish Khare, and Moongu Jeon. Classification of Alzheimer disease using gabor texture feature of hippocampus region. *International Journal of Image, Graphics and Signal Processing*, 8(6):13–20, 2016.

[114] Philipp Kickingereder, Franziska Dorn, Tobias Blau, Matthias Schmidt, Martin Kocher, Norbert Galldiks, and Maximilian I Ruge. Differentiation of local tumor recurrence from radiation-induced changes after stereotactic radiosurgery for treatment

of brain metastasis: case report and review of the literature. *Radiation Oncology*, 8(1):52, 2013.

[115] Youngsun Kim and Gary H McCullough. Maximal hyoid excursion in poststroke patients. *Dysphagia*, 25(1):20–25, 2010.

[116] Masatoshi Kimura, Takayoshi Yamashita, Yuji Yamauchi, and Hironobu Fujiyoshi. Facial point detection based on a convolutional neural network with optimal mini-batch procedure. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2860–2864. IEEE, 2015.

[117] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

[118] Stefan Kloppel, Cynthia M. Stonnington, Carlton Chu, Bogdan Draganski, Rachael I. Scahill, Jonathan D. Rohrer, Nick C. Fox, Clifford R. Jack, John Ashburner, and Richard S J Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008.

[119] M. Komlagan, V.-T. Ta, X. Pan, J.-P. Domenger, D.L. Collins, and P. Coupé. Anatomically constrained weak classifier fusion for early detection of Alzheimer's disease. In *International Workshop on Machine Learning in Medical Imaging*, pages 141–148, 2014.

[120] Peter Kontschieder, Samuel Rota Bulò, Horst Bischof, and Marcello Pelillo. Structured class-labels in random forests for semantic image labelling. In *IEEE International Conference on Computer Vision*, pages 2190–2197, 2011.

[121] Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017.

[122] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[123] Danielle Kruger. Assessing esophageal dysphagia. *Journal of the American Academy of Physician Assistants*, 27(5):23–30, 2014.

[124] R Senthil Kumar and M Senthilmurugan. Content-based image retrieval system in medical applications. *International Journal of Engineering Research and Technology*, 2(3), 2013.

[125] Camille Kurtz, Christopher F. Beaulieu, Sandy Napel, and Daniel L. Rubin. A hierarchical knowledge-based approach for retrieving similar medical images described with semantic annotations. *Journal of Biomedical Informatics*, 49:227–244, 2014.

[126] Camille Kurtz, Adrien Depeursinge, Sandy Napel, Christopher F. Beaulieu, and Daniel L. Rubin. On combining image-based and ontological semantic dissimilarities for medical image retrieval applications. *Medical Image Analysis*, 18(7):1082–1100, 2014.

[127] Andrés Larroza, David Moratal, Alexandra Paredes-Sánchez, Emilio Soria-Olivas, María L Chust, Leoncio A Arribas, and Estanislao Arana. Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in mri. *Journal of Magnetic Resonance Imaging*, 42(5):1362–1368, 2015.

[128] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[129] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[130] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, pages 253–256, 2010.

[131] Carol H. Lee, D. David Dershaw, Daniel Kopans, Phil Evans, Barbara Monsees, Debra Monticciolo, R. James Brenner, Lawrence Bassett, Wendie Berg, Stephen Feig, Edward Hendrick, Ellen Mendelson, Carl D'Orsi, Edward Sickles, and Linda Warren Burhenne. Breast cancer screening with imaging: recommendations from the society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. *Journal of the American College of Radiology*, 7(1):18–27, 2010.

[132] Dong-Hoon Lee, Do-Wan Lee, and Bong-Soo Han. Possibility study of scale invariant feature transform (SIFT) algorithm application to spine magnetic resonance imaging. *Plos One*, 11(4):e0153043, 2016.

[133] Jun Chang Lee, Kyoung Won Nam, Dong Pyo Jang, Nam Jong Paik, Ju Seok Ryu, and In Young Kim. A supporting platform for semi-automatic hyoid bone tracking and parameter extraction from videofluoroscopic images for the diagnosis of dysphagia patients. *Dysphagia*, 32(2):315–326, 2017.

[134] Wen Li Lee, Koyin Chang, and Kai Sheng Hsieh. Unsupervised segmentation of lung fields in chest radiographs using multiresolution fractal feature vector and deformable models. *Medical and Biological Engineering and Computing*, 54(9):1409–1422, 2016.

[135] Constance D Lehman, Robert D Wellman, Diana S M Buist, Karla Kerlikowske, Anna N A Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine*, 175(11):1–10, 2015.

[136] Ja-Ho Leigh, Byung-Mo Oh, Han Gil Seo, Goo Joo Lee, Yusun Min, Keewon Kim, Jung Chan Lee, and Tai Ryoon Han. Influence of the chin-down and chin-tuck maneuver on the swallowing kinematics of healthy adults. *Dysphagia*, 30(1):89–98, 2015.

[137] Victor Lempitsky, Victor Lempitsky, Michael Verhoek, Michael Verhoek, J Alison Noble, J Alison Noble, Andrew Blake, and Andrew Blake. Random forest classication for automatic delineation of myocardium in real-time 3D echocardiography. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 447–456, 2009.

[138] Paula Leslie, Paul N Carding, and Janet A Wilson. Investigation and management of chronic dysphagia. *British Medical Journal*, 326(7386):433, 2003.

[139] Paula Leslie, Michael J Drinnan, Paul Finn, Gary A Ford, and Janet A Wilson. Reliability and validity of cervical auscultation: a controlled comparison using video fluoroscopy. *Dysphagia*, 19(4):231–240, 2004.

[140] Paula Leslie, Michael J Drinnan, Ivan Zammit-Maempel, James L Coyle, Gary A Ford, and Janet A Wilson. Cervical auscultation synchronized with images from endoscopy swallow evaluations. *Dysphagia*, 22(4):290–298, 2007.

[141] F Li, L Tran, K-H Thung, S Ji, D Shen, and J Li. Robust deep learning for improved classification of AD / MCI patients. In *International Workshop on Machine Learning in Medical Imaging*, pages 240–247, 2014.

[142] Gang Li, Li Wang, Feng Shi, Weili Lin, and Dinggang Shen. Multi-atlas based simultaneous labeling of longitudinal dynamic cortical surfaces in infants. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 58–65, 2013.

[143] Nan Li, R Jin, and ZH Zhou. Top rank optimization in linear time. *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

[144] Wen Li, Fucang Jia, and Qingmao Hu. Automatic segmentation of liver tumor in CT images with deep convolutional neural networks. *Journal of Computer and Communications*, 3(11):146, 2015.

[145] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[146] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(December):18–22, 2002.

[147] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *arXiv preprint arXiv:1612.03144*, 2016.

[148] C Lindner, S Thiagarajah, J M Wilkinson, The Consortium, G A Wallis, and T F Cootes. Fully automatic segmentation of the proximal femur using random forest regression voting. *Medical Image Analysis*, 32(8):1462–1472, 2013.

[149] Jiamin Liu, Sanket Pattanaik, Jianhua Yao, Evrim Turkbey, Weidong Zhang, Xiao Zhang, and Ronald M. Summers. Computer aided detection of epidural masses on computed tomography scans. *Computerized Medical Imaging and Graphics*, 38(7):606–612, 2014.

[150] Manhua Liu, Daoqiang Zhang, and Dinggang Shen. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. *Human Brain Mapping*, 35(4):1305–1319, 2014.

[151] Mingxia Liu, Daoqiang Zhang, and Dinggang Shen B. Inherent structure-guided multi-view learning for Alzheimer's disease and mild cognitive impairment classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 296–303, 2015.

[152] Qin Liu, Qian Wang, Lichi Zhang, Yaozong Gao, and Dinggang Shen. Multi-atlas context forests for knee MR image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 186–193, 2015.

[153] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[154] Xinglong Liu, Fei Hou, Hong Qin, and Aimin Hao. A cade system for nodule detection in thoracic ct images based on artificial neural network. *Science China Information Sciences*, 60(7):072106, 2017.

[155] JA Logemann, PJ Kahrilas, J Begelman, WJ Dodds, and BR Pauloski. Interactive computer program for biomechanical analysis of videoradiographic studies of swallowing. *American Journal of Roentgenology*, 153(2):277–280, 1989.

[156] Jeri A Logemann. Behavioral management for oropharyngeal dysphagia. *Folia Phoniatrica et Logopaedica*, 51(4-5):199–212, 1999.

[157] Jeri A Logemann, Gary Gensler, JoAnne Robbins, Anne S Lindblad, Diane Brandt, Jacqueline A Hind, Steven Kosek, Karen Dikeman, Marta Kazandjian, Gary D Gramigna, et al. A randomized study of three interventions for aspiration of thin liquids in patients with dementia or parkinson's disease. *Journal of Speech, Language, and Hearing Research*, 51(1):173–183, 2008.

[158] Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

[159] Herve Lombaert, Darko Zikic, Antonio Criminisi, and Nicholas Ayache. Laplacian forests: semantic image segmentation by guided bagging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504, 2014.

[160] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[161] Zhuqing Long, Bin Jing, Huagang Yan, Jianxin Dong, Han Liu, Xiao Mo, Ying Han, and Haiyun Li. A support vector machine based method to identify mild cognitive

impairment with multi-level characteristics of magnetic resonance imaging. *Neuroscience*, 331:169–176, 2016.

[162] Nikki Lucci, Cole McConnell, and Chuck Biddle. Understanding normal and abnormal swallowing: Patient safety considerations for the perianesthetic nurse. *Journal of PeriAnesthesia Nursing*, 2017.

[163] Shu-Ting Luo and Bor-Wen Cheng. Diagnosing breast masses in digital mammography using feature selection and ensemble methods. *Journal of Medical Systems*, 36(2):569–577, 2012.

[164] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3317–3326, 2017.

[165] Hiram Madero Orozco, Osslan Osiris Vergara Villegas, Vianey Guadalupe Cruz Sánchez, Humberto De Jesús Ochoa Domínguez, and Manuel De Jesús Nandayapa Alfaro. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *BioMedical Engineering OnLine*, 14(1):9, 2015.

[166] Oskar Maier, Matthias Wilms, Janina von der Gablentz, Ulrike M. Krämer, Thomas F. Münte, and Heinz Handels. Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. *Journal of Neuroscience Methods*, 240:89–100, 2015.

[167] Giselle Mann, Graeme J Hankey, and David Cameron. Swallowing disorders following acute stroke: prevalence and diagnostic accuracy. *Cerebrovascular Diseases*, 10(5):380–386, 2000.

[168] Rashindra Manniesing, T Marcel, H Oei, Luuk J Oostveen, Jaime Melendez, Ewoud J Smit, Bram Platel, Clara I Sánchez, J Frederick, A Meijer, et al. White matter and gray matter segmentation in 4D computed tomography. *Scientific Reports*, 7:1, 2017.

[169] Paul E Marik. Aspiration pneumonitis and aspiration pneumonia. *New England Journal of Medicine*, 344(9):665–671, 2001.

[170] Bonnie Martin-Harris and Bronwyn Jones. The videofluorographic swallowing study. *Physical Medicine and Rehabilitation Clinics of North America*, 19(4):769–785, 2008.

[171] Rosemary Martino, Frank Silver, Robert Teasell, Mark Bayley, Gordon Nicholson, David L Streiner, and Nicholas E Diamant. The toronto bedside swallowing screening test (tor-bsst). *Stroke*, 40(2):555–561, 2009.

[172] Alireza Mehrtasha, Alireza Sedghic, Mohsen Ghafooriana, Mehdi Taghipoura, Clare M Tempanya, Tina Kapura, Parvin Mousavic, Purang Abolmaesumib, and Andriy Fedorova. Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. In *SPIE Medical Imaging*, pages 101342A–101342A. International Society for Optics and Photonics, 2017.

[173] Kayla Mendel, Hui Li, Deepa Sheth, and Maryellen Giger. Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography. *Academic radiology*, 26(6):735–743, 2019.

[174] Jiandong Meng, Yan Jiang, Xiaoliang Xu, and Irfani Prianandа. Support top irrelevant machine: learning similarity measures to maximize top precision for image retrieval. *Neural Computing and Applications*, pages 1–10, 2016.

[175] Yuma Miki, Chisako Muramatsu, Tatsuro Hayashi, Xiangrong Zhou, Takeshi Hara, Akitoshi Katsumata, and Hiroshi Fujita. Classification of teeth in cone-beam CT using deep convolutional neural network. *Computers in Biology and Medicine*, 80:24–29, 2017.

[176] Arthur J Miller. The neurobiology of swallowing and dysphagia. *Developmental Disabilities Research Reviews*, 14(2):77–86, 2008.

[177] Nick Miller, Liesl Allcock, AJ Hildreth, Diana Jones, Emma Noble, and DJ Burn. Swallowing problems in parkinson disease: frequency and clinical correlates. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(9):1047–1049, 2009.

[178] Tom Mitchell and Avrim Blum. Combining labeled and unlabeled data with co-training. In *11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[179] Jhimli Mitra, Pierrick Bourgeat, Jurgen Fripp, Soumya Ghose, Stephen Rose, Olivier Salvado, Alan Connelly, Bruce Campbell, Susan Palmer, Gagan Sharma, Soren Christensen, and Leeanne Carey. Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *NeuroImage*, 98:324–335, 2014.

[180] Pim Moeskops, Max A Viergever, Adriënne M Mendrik, Linda S de Vries, Manon JNL Benders, and Ivana Išgum. Automatic segmentation of mr brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261, 2016.

[181] J Mohan, V Krishnaveni, and Yanhui Guo. A survey on the magnetic resonance image denoising methods. *Biomedical Signal Processing and Control*, 9(September):56–69, 2014.

[182] Mohamed Mokhtar, Bendib Hayet, and Farida Merouani. Automatic segmentation of brain MRI through stationary wavelet transform and random forests. *Pattern Analysis and Applications*, 18(4):829–843, 2015.

[183] Sonja M Molfenter and Catriona M Steele. The relationship between residue and aspiration on the subsequent swallow: an application of the normalized residue ratio scale. *Dysphagia*, 28(4):494–500, 2013.

[184] Sonja M Molfenter and Catriona M Steele. Kinematic and temporal factors associated with penetration–aspiration in swallowing liquids. *Dysphagia*, 29(2):269–276, 2014.

[185] Sonja M Molfenter and Catriona 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*, 57(3):768–778, 2014.

[186] Faezeh Movahedi, Atsuko Kurosu, James L Coyle, Subashan Perera, and Ervin Sejdić. Anatomical directional dissimilarities in tri-axial swallowing accelerometry signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5):447–458, 2017.

[187] Subrahmanyam Murala and Q. M. Jonathan Wu. Local ternary co-occurrence patterns: A new feature descriptor for MRI and CT image retrieval. *Neurocomputing*, 119:399–412, 2013.

[188] A Nacci, F Ursino, R La Vela, F Matteucci, V Mallardi, and B Fattori. Fiberoptic endoscopic evaluation of swallowing (FEES): proposal for informed consent. *Acta Otorhinolaryngologica Italica*, 28(4):206, 2008.

[189] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1, 2015.

[190] Roger Newman, Natalia Vilardell, Pere Clavé, and Renée Speyer. Effect of bolus viscosity on the safety and efficacy of swallowing and the kinematics of the swallow response in patients with oropharyngeal dysphagia: white paper by the European Society for Swallowing Disorders (ESSD). *Dysphagia*, 31(2):232–249, 2016.

[191] Nam P Nguyen, Cheryl Frank, Candace C Moltz, Paul Vos, Herbert J Smith, Prabhakar V Bhamidipati, Ulf Karlsson, Phuc D Nguyen, Alan Alfieri, Ly M Nguyen, et al. Aspiration rate following chemoradiation for head and neck cancer: an underreported occurrence. *Radiotherapy and Oncology*, 80(3):302–306, 2006.

[192] R Nithya and B Santhi. Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *International Journal of Computer Applications*, 28(6):21–25, 2011.

[193] Stacy D O'Connor, Jianhua Yao, and Ronald M Summers. Lytic metastases in thoracolumbar spine: computer-aided detection at CT–preliminary study. *Radiology*, 242(3):811–816, 2007.

[194] Sumiko Okada, Eiichi Saitoh, Jeffrey B Palmer, Koichiro Matsuo, Michio Yokoyama, Ritsuko Shigeta, and Mikoto Baba. What is the chin-down posture? A questionnaire survey of speech language pathologists in Japan and the United States. *Dysphagia*, 22(3):204–209, 2007.

[195] Susan R Orenstein, Fariba Izadnia, and Seema Khan. Gastroesophageal reflux disease in children. *Gastroenterology Clinics of North America*, 28(4):947–969, 1999.

[196] Jeffrey B Palmer, Keith V Kuhlemeier, Donna C Tippett, and C Lynch. A protocol for the videofluorographic swallowing study. *Dysphagia*, 8(3):209–214, 1993.

[197] Jin-Woo Park, Gyu-Jeong Sim, Dong-Chan Yang, Kyoung-Hwan Lee, Ji-Hea Chang, Ki-Yeun Nam, Ho-Jun Lee, and Bum-Sun Kwon. Increased bolus volume effect on delayed pharyngeal swallowing response in post-stroke oropharyngeal dysphagia: A pilot study. *Annals of Rehabilitation Medicine*, 40(6):1018–1023, 2016.

[198] Greg Pass and R. Zabih. Histogram refinement for content-based image retrieval. In *IEEE Workshop on Applications of Computer Vision*, pages 96–102, 1996.

[199] Tejal A. Patel, Mamta Puppala, Richard O. Ogunti, Joe E. Ensor, Tiancheng He, Jitesh B. Shewale, Donna P. Ankerst, Virginia G. Kaklamani, Angel A. Rodriguez, Stephen T. C. Wong, and Jenny C. Chang. Correlating mammographic and pathologic

findings in clinical decision support using natural language processing and data mining methods. *Cancer*, pages 1–8, 2016.

[200] William G Pearson, Sonja M Molfenter, Zachary M Smith, and Catriona M Steele. Image-based measurement of post-swallow residue: the normalized residue ratio scale. *Dysphagia*, 28(2):167–177, 2013.

[201] Danilo Cesar Pereira, Rodrigo Pereira Ramos, and Marcelo Zanchetta do Nascimento. Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer Methods and Programs in Biomedicine*, 114(1):88–101, 2014.

[202] Noel Pérez, Miguel Angel Guevara, Augusto Silva, Isabel Ramos, and Joana Loureiro. Improving the performance of machine learning classifiers for breast cancer diagnosis based on feature selection. In *Federated Conference on Computer Science and Information Systems*, volume 2, pages 209–217., 2014.

[203] Noel Pérez Pérez, Miguel A. Guevara López, Augusto Silva, and Isabel Ramos. Improving the Mann-Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artificial Intelligence in Medicine*, 63(1):19–31, 2015.

[204] M Politis. Neuroimaging in Parkinson disease: from research setting to clinical practice. *Nature Reviews Neurology*, 10(12):708–722, 2014.

[205] Marziyeh Poorjavad, Fatemeh Derakhshandeh, Masoud Etemadifar, Bahram Soleymani, Alireza Minagar, and Amir-Hadi Maghzi. Oropharyngeal dysphagia in multiple sclerosis. *Multiple Sclerosis Journal*, 16(3):362–365, 2010.

[206] Jonathan D Power, Anish Mitra, Timothy O Laumann, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84:320–341, 2014.

[207] Adhish Prasoon, Kersten Petersen, Christian Igel, Fran{\c{c}}ois Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–253, 2013.

[208] Jane E Prasse and George E Kikano. An overview of pediatric dysphagia. *Clinical Pediatrics*, 48(3):247–251, 2009.

[209] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.

[210] Mary C Ragland, Taeok Park, Gary McCullough, and Youngsun Kim. The speed of the hyoid excursion in normal swallowing. *Clinical Archives of Communication Disorders*, 1(1):30–35, 2016.

[211] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, number 1, pages 3501–3508, 2010.

[212] Deborah Ramsey, David Smithard, and Lalit Kalra. Silent aspiration: what do we know? *Dysphagia*, 20(3):218–225, 2005.

[213] Anju Rani, Deepti Mittal, et al. Detection and classification of focal liver lesions using support vector machine classifiers. *Journal of Biomedical Engineering and Medical Imaging*, 3(1):21, 2016.

[214] Rozita Rastghalam and Hossein Pourghassem. Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images. *Pattern Recognition*, 51:176–186, 2016.

[215] Narender P Reddy, Aparna Katakam, Vineet Gupta, Rajeev Unnikrishnan, Janardhan Narayanan, and Enrique P Canilang. Measurements of acceleration during videofluorographic evaluation of dysphagic patients. *Medical Engineering and Physics*, 22(6):405–412, 2000.

[216] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[217] Valerie K. Reed, Wendy A. Woodward, Lifei Zhang, Eric A. Strom, George H. Perkins, Welela Tereffe, Julia L. Oh, T. Kuan Yu, Isabelle Bedrosian, Gary J. Whitman, Thomas A. Buchholz, and Lei Dong. Automatic segmentation of whole breast using atlas approach and deformable image registration. *International Journal of Radiation Oncology Biology Physics*, 73(5):1493–1500, 2009.

[218] Christoph Reinders, Hanno Ackermann, Michael Ying Yang, and Bodo Rosenhahn. Learning a fully convolutional network for object recognition using very few data. *arXiv preprint arXiv:1709.05910*, 2017.

[219] Dylan F Roden and Kenneth W Altman. Causes of dysphagia among different age groups: a systematic review of the literature. *Otolaryngologic Clinics of North America*, 46(6):965–987, 2013.

[220] L Rofes, V Arreola, R Mukherjee, J Swanson, and P Clavé. The effects of a xanthan gum-based thickener on the swallowing function of patients with dysphagia. *Alimentary Pharmacology and Therapeutics*, 39(10):1169–1179, 2014.

[221] Eric M Rohren, Timothy G Turkington, and R Edward Coleman. Clinical applications of PET in oncology. *Radiology*, 231:305–332, 2004.

[222] Lior Rokach and Oded Maimon. Classification Trees. In *Data Mining and Knowledge Discovery Handbook*, pages 149–174. 2010.

[223] John C Rosenbek, Jo Anne Robbins, Ellen B Roecker, Jame L Coyle, and Jennifer L Wood. A penetration-aspiration scale. *Dysphagia*, 11(2):93–98, 1996.

[224] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 556–564. Springer, 2015.

[225] Snehashis Roy, Aaron Carass, Jerry L Prince, and Dzung L Pham. Subject specific sparse dictionary learning for atlas based brain MRI segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 248–255, 2014.

[226] MG Rugiu. Role of videofluoroscopy in evaluation of neurologic dysphagia. *Acta Otorhinolaryngologica Italica*, 27(6):306, 2007.

[227] Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. In *12th International Conference on Artificial Intelligence and Statics*, number 3, pages 448–455, 2009.

[228] C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M. C. Gilardi, and A. Quattrone. Machine learning on brain MRI data

for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. *Journal of Neuroscience Methods*, 222:230–237, 2014.

[229] Rachid Sammouda, Rami Mohammad Jomaa, and Hassan Mathkour. Heart region extraction and segmentation from chest CT images using Hopfield Artificial Neural Networks. In *International Conference on Information Technology and e-Services*, pages 3–8, 2012.

[230] Saman Sarraf, John Anderson, Ghassem Tofighi, et al. DeepAD: Alzheimer' s disease classification via deep convolutional neural networks using MRI and fMRI. *BioRxiv*, page 070441, 2016.

[231] E Schulte, LM Ross, and ED Lamperti. Thieme atlas of anatomy: Neck and internal organs. *Chpt*, 4:304, 2006.

[232] Suman Sedai, Pallab Roy, and Rahil Garnavi. Segmentation of right ventricle in cardiac MR images using shape regression. In *International Workshop on Machine Learning in Medical Imaging*, pages 1–8, 2015.

[233] Han Gil Seo, Byung-Mo Oh, and Tai Ryoon Han. Swallowing kinematics and factors associated with laryngeal penetration and aspiration in stroke survivors with dysphagia. *Dysphagia*, 31(2):160–168, 2016.

[234] Gaurav Sethi and B. S. Saini. Abdomen disease diagnosis in CT images using flexiscale curvelet transform and improved genetic algorithm. *Australasian Physical & Engineering Sciences in Medicine*, 38(4):671–688, 2015.

[235] Reza Shaker. Oropharyngeal dysphagia. *Gastroenterology & Hepatology*, 2(9):633, 2006.

[236] Therese K Shanahan, Jeri A Logemann, Alfred W Rademaker, and B Roa Pauloski. Chin-down posture effect on aspiration in dysphagic patients. *Archives of Physical Medicine and Rehabilitation*, 74(7):736–739, 1993.

[237] Neeraj Sharma, Lalit M Aggarwal, et al. Automated medical image segmentation techniques. *Journal of Medical Physics*, 35(1):3, 2010.

[238] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural

networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, PP(99):1, 2016.

[239] Junji Shiraishi, Lorenzo L Pesce, Charles E Metz, and Kunio Doi. Experimental design and data analysis in receiver operating characteristic studies : lessons learned from reports in radiology from 1997 to 2006. *Radiology*, 253(3), 2009.

[240] Tapas Si, Arunava De, and Anup Kumar. Artificial neural network based lesion segmentation of brain MRI. *Communications on Applied Electronics*, 4(5):1–5, 2016.

[241] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[242] Gurpreet Singh and Lakshminarayanan Samavedham. Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of Parkinson disease. *Journal of Neuroscience Methods*, 256:30–40, 2015.

[243] Satya P. Singh and Shabana Urooj. An improved CAD system for breast cancer diagnosis based on generalized pseudo-zernike moment and Ada-DEWNN classifier. *Journal of Medical Systems*, 40(4):1–13, 2016.

[244] Douglas Sloane and S Philip Morgan. An introduction to categorical data analysis. *Annual Review of Sociology*, 22(1):351–375, 1996.

[245] Christopher D. Smyser, Nico U.F. Dosenbach, Tara A. Smyser, Abraham Z. Snyder, Cynthia E. Rogers, Terrie E. Inder, Bradley L. Schlaggar, and Jeffrey J. Neil. Prediction of brain maturity in infants using machine-learning algorithms. *NeuroImage*, 136:1–9, 2016.

[246] Youyi Song, Ling Zhang, Siping Chen, Dong Ni, Baiying Lei, and Tianfu Wang. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. *IEEE Transactions on Biomedical Engineering*, 62(10):2421–2433, 2015.

[247] Barbara C Sonies. Instrumental procedures for dysphagia diagnosis. In *Seminars in speech and language*, volume 12, pages 185–198. © 1991 by Thieme Medical Publishers, Inc., 1991.

[248] C Spampinato, S Palazzo, D Giordano, M Aldinucci, and R Leonardi. Deep learning for automated skeletal bone age assessment in X-ray images. *Medical Image Analysis*, 36:41–51, 2017.

[249] N. Speybroeck. Classification and regression trees. *International Journal of Public Health*, 57(1):243–246, 2012.

[250] M Srinivas and C Krishna Mohan. Medical images modality classification using multi-scale dictionary learning. In *International Conference on Digital Signal Processing*, number August, pages 621–625, 2014.

[251] M Srinivas and C Krishna Mohan. Classification of medical images using edge-based features and sparse representation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 912–916, 2016.

[252] M. Srinivas, R. Ramu Naidu, C. S. Sastry, and C. Krishna Mohan. Content based medical image retrieval using dictionary learning. *Neurocomputing*, 168:880–895, 2015.

[253] M Srinivas, Debaditya Roy, and C Krishna Mohan. Discriminative feature extraction of X-ray images using deep convolutional neural networks. *Icassp 2016*, pages 917–921, 2016.

[254] Catriona Steele, Melanie Peladeau-Pigeon, Emily Barrett, and Talia Wolkin. The risk of penetration–aspiration related to residue in the pharynx. *American Journal of Speech-Language Pathology*, pages 1–10, 06 2020.

[255] Catriona M Steele, Woroud Abdulrahman Alsanei, Sona Ayanikalath, Carly EA Barbon, Jianshe Chen, Julie AY Cichero, Kim Coutts, Roberto O Dantas, Janice Duivestein, Lidia Giosa, Ben Hanson, Peter Lam, Caroline Lecko, Chelsea Leigh, Ahmed Nagy, Ashwini M. Namasivayam, Weskania V. Nascimento, Inge Odendaal, Christina H. Smith, and Helen Wang. The influence of food texture and liquid consistency modification on swallowing physiology and function: a systematic review. *Dysphagia*, 30(1):2–26, 2015.

[256] Catriona M Steele, Gemma L Bailey, Tom Chau, Sonja M Molfenter, Mohamed Oshalla, Ashley A Waito, and Dana CBH Zoratto. The relationship between hyoid and laryngeal displacement and swallowing impairment. *Clinical Otolaryngology*, 36(1):30–36, 2011.

[257] Catriona M Steele and Julie AY Cichero. Physiological factors related to aspiration risk: a systematic review. *Dysphagia*, 29(3):295–304, 2014.

[258] Catriona M Steele, Melanie Peladeau-Pigeon, Ahmed Nagy, and Ashley A Waito. Measurement of pharyngeal residue from lateral view videofluoroscopic images. Technical report, ASHA, 2020.

[259] Cara E Stepp. Surface electromyography for speech and swallowing systems: measurement, analysis, and interpretation. *Journal of Speech, Language, and Hearing Research*, 55(4):1232–1246, 2012.

[260] Debra M Suiter, Steven B Leder, and David E Karas. The 3-ounce (90-cc) water swallow challenge: a screening test for children with suspected oropharyngeal dysphagia. *Otolaryngology–Head and Neck Surgery*, 140(2):187–190, 2009.

[261] Heung-II Suk and Dinggang Shen. Deep learning-based feature representation for AD/MCI classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 583–590, 2013.

[262] Wenqing Sun, Tzu-Liang Tseng, Jianying Zhang, and Wei Qian. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*, 2016.

[263] Wenqing Sun, Tzu-Liang Bill Tseng, Wei Qian, Jianying Zhang, Edward C Saltzstein, Bin Zheng, Fleming Lure, Hui Yu, and Shi Zhou. Using multiscale texture and density features for near-term breast cancer risk analysis. *Medical Physics*, 42(6):2853–2862, 2015.

[264] Wenqing Sun, Bin Zheng, Fleming Lure, Teresa Wu, Jianying Zhang, Benjamin Y. Wang, Edward C. Saltzstein, and Wei Qian. Prediction of near-term risk of developing breast cancer using computerized features from bilateral mammograms. *Computerized Medical Imaging and Graphics*, 38(5):348–357, 2014.

[265] Livia Sura, Aarthi Madhavan, Giselle Carnaby, and Michael A Crary. Dysphagia in the elderly: management and nutritional considerations. *Clinical Interventions in Aging*, 7:287, 2012.

[266] L. P. Suresh, Subhransu Sekhar Dash, and Bijaya Ketan Panigrahi. Artificial intelligence and evolutionary algorithms in engineering systems. *Advances in Intelligent Systems and Computing*, 324:109–117, 2015.

[267] Inga Suttrup and Tobias Warnecke. Dysphagia in parkinson's disease. *Dysphagia*, 31(1):24–32, 2016.

[268] J A K Suykens and J Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[269] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[270] Claire Takizawa, Elizabeth Gemmell, James Kenworthy, and Renée Speyer. A systematic review of the prevalence of oropharyngeal dysphagia in stroke, parkinson's disease, alzheimer's disease, head injury, and pneumonia. *Dysphagia*, 31(3):434–441, 2016.

[271] A Thomas, K Erlandsson, Anthonin Reilhac, Alexandre Bousse, Daniil Kazantsev, Stefano Pedemonte, Kathleen Vunckx, Simon Arridge, Sebastien Ourselin, and Brian F Hutton. A comparison of the options for brain partial volume correction using PET/MRI. In *IEEE Nuclear Science Symposiwn and Medical Imaging Conference*, pages 2902–2906, 2012.

[272] Kim-Han Thung, Chong-Yaw Wee, Pew-Thian Yap, Dinggang Shen, and the Alzheimer's Disease Neuroimaging Initiative. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. *NeuroImage*, 91:386–400, 2014.

[273] Dong Ping Tian. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 8(4):385–395, 2013.

[274] Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V. Hajnal, and Daniel Rueckert. Multiple instance learning for classification of dementia in brain MRI. *Medical Image Analysis*, 18(5):808–818, 2014.

[275] Tong Tong, Robin Wolz, Zehan Wang, Qinquan Gao, Kazunari Misawa, Michitaka Fujiwara, Kensaku Mori, Joseph V. Hajnal, and Daniel Rueckert. Discriminative dictionary learning for abdominal multi-organ segmentation. *Medical Image Analysis*, 23(1):92–104, 2015.

[276] T Torheim, E Malinen, K Kvaal, H Lyng, U G Indahl, E K F Andersen, and C M Futsaether. Classification of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines. *IEEE Transactions on Medical Imaging*, 33(8):1648–1656, 2014.

[277] Huynh Tri, Gao Yaozong, Kang Jiayin, Wang Li, Zhang Pei, Shen Dinggang, and Alzheimer's Disease Neuroimaging Initiative. Multi-source information gain for random forest: an application to CT image prediction from MRI data. In *International Workshop on Machine Learning in Medical Imaging*, pages 321–329, 2015.

[278] Chih Fong Tsai. Image mining by spectral features: A case study of scenery image classification. *Expert Systems with Applications*, 32(1):135–142, 2007.

[279] Jolien GJ van der Kruis, Laura WJ Baijens, Renée Speyer, and Iris Zwijnenberg. Biomechanical analysis of hyoid bone displacement in videofluoroscopy: a systematic review of intervention effects. *Dysphagia*, 26(2):171–182, 2011.

[280] Bram Van Ginneken, Arnaud A. A. Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *12th IEEE International Symposium on Biomedical Imaging*, pages 286–289, 2015.

[281] Gijs van Tulder and Marleen de Bruijne. Combining generative and discriminative representation learning for lung ct analysis with convolutional restricted boltzmann machines. *IEEE Transactions on Medical Imaging*, 35(5):1262–1272, 2016.

[282] Sharon Veis, Jeri A Logemann, and Laura Colangelo. Effects of three techniques on maximum posterior movement of the tongue base. *Dysphagia*, 15(3):142–145, 2000.

[283] K. Velmurugan and S. Santhosh Baboo. Content-based image retrieval using SURF and colour moments. *Global Journal of Computer Science and Technology*, 11(10):1–4, 2011.

[284] Manisha Verma and Balasubramanian Raman. Center symmetric local binary co-occurrence pattern for texture , face and bio-medical image retrieval. *Journal of Visual Communication and Image Representation*, 32:224–236, 2015.

[285] Ge Wang, Mannudeep Kalra, and Colin G Orton. Machine learning will transform radiology significantly within the next 5 years. *Medical Physics*, 2017.

[286] Li Wang, Feng Shi, Yaozong Gao, Gang Li, John H. Gilmore, Weili Lin, and Dinggang Shen. Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation. *NeuroImage*, 89:152–164, 2014.

[287] Ming Wang. Generalized estimating equations in longitudinal data analysis: A review and recent developments. *Advances in Statistics*, 2014, 2014.

[288] Qingzhu Wang, Wenchao Zhu, and Bin Wang. Three-dimensional SVM with latent variable: application for detection of lung lesions in CT images. *Journal of Medical Systems*, 39(1):171, 2015.

[289] Shijun Wang and Ronald M. Summers. Machine learning and radiology. *Medical Image Analysis*, 16(5):933–951, 2013.

[290] Yan Wang, Pei Zhang, Le An, Guangkai Ma, Jiayin Kang, Feng Shi, Xi Wu, Jiliu Zhou, David S Lalush, Weili Lin, and Dinggang Shen. Predicting standard-dose PET image from low-dose PET and multimodal MR images using mapping-based sparse representation. *Physics in Medicine and Biology*, 61(2):791, 2016.

[291] Yaping Wang, Jingxin Nie, Pew Thian Yap, Gang Li, Feng Shi, Xiujuan Geng, Lei Guo, and Dinggang Shen. Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. *PLoS ONE*, 9(1):1–23, 2014.

[292] Chia-Hung Wei, Chang-Tsun Li, and Roland Wilson. A content–based approach to medical image database retrieval. *Database Modeling for Industrial Data Management: Emerging Technologies and Applications*, pages 258–291, 2005.

[293] Andrew R Xavier, Adnan I Qureshi, Jawad F Kirmani, Abutaher M Yahia, and Rohit Bakshi. Neuroimaging of stroke: a review. *Southern Medical Journal*, 96:367–379, 2003.

[294] Min Xian, Yingtao Zhang, and H. D. Cheng. Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains. *Pattern Recognition*, 48(2):485–497, 2015.

[295] Weiying Xie, Yunsong Li, and Yide Ma. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*, 173:930–941, 2016.

[296] Nai Chung Yang, Wei H. Chang, Chung Ming Kuo, and Tsia Hsing Li. A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval. *Journal of Visual Communication and Image Representation*, 19(2):92–105, 2008.

[297] Jianhua Yao, Joseph E Burns, Hector Munoz, and Ronald M Summers. Detection of vertebral body fractures based on cortical shell unwrapping. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 15, pages 509–516, 2012.

[298] Jianhua Yao, Joseph E Burns, and Ronald M Summers. Computer aided detection of bone metastases in the thoracolumbar spine. In *Spinal Imaging and Image Analysis*, pages 97–130. 2015.

[299] Jianhua Yao, Andrew Dwyer, Ronald M. Summers, and Daniel J. Mollura. Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification. *Academic Radiology*, 18(3):306–314, 2011.

[300] Jianhua Yao, Hector Munoz, Joseph E Burns, and Le Lu. Computer aided detection of spinal degenerative osteophytes on sodium fluoride PET/CT. *Computational Methods and Clinical Applications for Spine Imaging*, pages 51–60, 2014.

[301] Xu Yingying, Lin Lanfen, Hu Hongjie, Yu Huajun, Jin Chongwu, Wang Jian, Han Xianhua, and Chen Yen-Wei. Combined density, texture and shape features of multiphase contrast-enhanced CT images for CBIR of focal liver lesions: a preliminary study. In *Innovation in Medicine and Healthcare 2015*, pages 215–224. Springer, 2016.

[302] Y. Yoo, T. Brosch, A. Traboulsee, D.K.B Li, and R. Tam. Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation. *Mlmi*, pages 117–124, 2014.

[303] Jie Yu, Jaume Amores, Nicu Sebe, Petia Radeva, and Qi Tian. Distance learning for similarity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):451–462, 2008.

[304] Jun Yue, Zhenbo Li, Lu Liu, and Zetian Fu. Content-based image retrieval using color and texture fused features. *Mathematical and Computer Modelling*, 54(3):1121–1127, 2011.

[305] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224, 2015.

[306] Zhenwei Zhang, James L Coyle, and Ervin Sejdić. Automatic hyoid bone detection in fluoroscopic images using deep learning. *Scientific Reports*, 8(1):12310, 2018.

[307] Zhenwei Zhang, Atsuko Kurosu, James Coyle, Subashan Perera, and Ervin Sejdić. A generalized equation approach for hyoid bone displacement and penetration-aspiration scale analysis. 2021.

[308] Zhenwei Zhang, Shitong Mao, James Coyle, and Ervin Sejdić. Automatic annotation of cervical vertebrae in videofluoroscopy images via deep learning. 2021.

[309] Zhenwei Zhang, Subashan Perera, Cara Donohue, Atsuko Kurosu, Amanda S Mahoney, James L Coyle, and Ervin Sejdić. The prediction of risk of penetration–aspiration via hyoid bone displacement features. *Dysphagia*, 35(1):66–72, 2020.

[310] Zhenwei Zhang and Ervin Sejdić. Radiological images and machine learning: trends, perspectives, and prospects. *Computers in biology and medicine*, 108:354–370, 2019.

[311] Dinggang Zhang, Daoqiuang; Shen. Multi modal multi task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59(2):895–907, 2013.

[312] Xiaofeng Zhu, Heung-il Suk, and Dinggang Shen. Sparse discriminative feature selection for multi-class Alzheimer's disease classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 157–164, 2014.

[313] Xiaofeng Zhu, Heung-il Suk, Yonghua Zhu, and Kim-han Thung. Multi-view classification for identification of Alzheimer's Disease. In *International Workshop on Machine Learning in Medical Imaging*, volume 255-262, pages 255–262, 2015.

[314] Xiaojin Zhu. Semi-supervised learning. In *Encyclopedia of Machine Learning*, pages 892–897. 2011.

[315] Xinjian Zhu, Xuan He, Pin Wang, Qinghua He, Dandan Gao, Jiwei Cheng, and Baoming Wu. A method of localization and segmentation of intervertebral discs in spine MRI based on Gabor filter bank. *BioMedical Engineering OnLine*, 15(1):32, 2016.

[316] A Ziegler and M Vens. Generalized estimating equations. *Methods of Information in Medicine*, 49(5):421–425, 2010.

[317] Darko Zikic, Ben Glocker, Ender Konukoglu, Antonio Criminisi, C Demiralp, J Shotton, O M Thomas, T Das, R Jena, and S J Price. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. *Medical Image Computing and Computer-Assisted Intervention*, 15(Pt 3):369–76, 2012.

[318] Kelly H. Zou, Simon K. Warfield, Aditya Bharatha, Clare M C Tempany, Michael R. Kaus, Steven J. Haker, William M. Wells, Ferenc A. Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology*, 11(2):178–189, 2004.