UNIVERSITY OF CALIFORNIA SAN DIEGO

Human-Centered Machine Learning for Healthcare: Examples in Neurology and Pulmonology

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioengineering with a Specialization in Multiscale Biology

by

Vishwajith Ramesh

Committee in charge:

Gert Cauwenberghs, Chair
Nadir Weibel, Co-Chair
Erhan Bilal
Todd P. Coleman
Garrison W. Cottrell
Terrence J. Sejnowski

2020

The dissertation of Vishwajith Ramesh is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2020

DEDICATION

To my family and advisors for their unconditional support

EPIGRAPH

You can't connect the dots looking forward; you can only connect them looking backward. So you have to trust that the dots will somehow connect in your future.

—*Steve Jobs*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Early in my PhD, I learnt that often the most effective way forward was to simply trust that whatever choices I make will somehow connect in my future, that my decisions will all culminate in my success and good well-being. This was my driving philosophy for much of my PhD career. It helped me avoid overthinking the consequences of my choices before they were even made and gave me the courage to take risks. More so than the technical knowledge I gained, learning to trust in my own ability to make good decisions – to trust my gut – was perhaps the most useful takeaway from my time pursuing graduate studies. I confidently pursued new avenues of research without being paralyzed by the fear of failure due to my lack of prior experience. (Before my PhD, I had no experience with artificial intelligence or human-centered design, both core focus areas of this dissertation.) As a result, I have been fortunate to be able to learn a broad range of research skills, from machine learning to design thinking, and to experience a diverse set of professional activities, both academic and entrepreneurial. It later occurred to me that while it took some time to learn to trust that my choices will lead to my eventual success (or at least, the successful completion of my PhD), the people to whom I dedicate this dissertation, my family and advisors, have always had this faith in me from the start.

Perhaps my greatest thanks goes to Dr. Nadir Weibel. I joined Nadir's Human-Centered and Ubiquitous Computing Lab in 2015 just as he was starting it. I was two years into my PhD when I realized the great risk Nadir took when he decided to bring me on as one of his first PhD students; at the time, I did not have a strong computer science background nor the funding to support myself. Nadir had faith in my work ethic from the beginning, and trusted my ability to independently drive my projects forward. Nadir has a knack for identifying students with strong potential and for bringing that potential out, a leadership trait that I learned from him and to which I of course owe my PhD career.

Moreover, the talented individuals Nadir identified and brought into the lab were some of the smartest people I met in graduate school, people who number among my close friends. The passion for research and technology from fellow graduate students like Steven Rick, Danilo Gasques, Janet Johnson, and Tommy Sharkey helped me deal with the stresses of life in academia. I am grateful to Nadir and the rest of the Ubiquitous Computing Lab for their support.

To have Dr. Gert Cauwenberghs as an advisor was a genuine stroke of good luck. His strong experience in academia and technical intelligence helped me solve several day-to-day issues. His enthusiasm for my research and science in general was infectious. Our meetings often felt as though we were colleagues passionately discussing the intricacies of scientific research rather than a mentor simply meeting with his mentee for updates. Like Nadir, Gert always trusted my opinion and know-how. Both Nadir and Gert came from different backgrounds with different expertise and with different years of experience in academia. Together they greatly enhanced my PhD experience and I dedicate this dissertation to them for that reason.

My work truly benefited from a fruitful and years long collaboration with the UCSD Stroke Center. I would be remiss if I did not thank my neurologist collaborators Dr. Brett C. Meyer and Dr. Kunal Agrawal for opening up their clinic to me and my data collection team. I learned a great deal about stroke medicine by shadowing them. In my opinion, it is rare to be able to set up a close collaboration between engineers and doctors. The UCSD ecosystem and the willingness of Dr. Meyer and Dr. Agrawal to explore new ideas for stroke diagnosis and care contributed to a synergistic relationship. To be able to interact face-to-face with stroke patients, the end benefactor of much of my PhD work and the people I aimed to help, was a source of inspiration. I am thankful to Dr. Meyer and Dr. Agrawal for giving me the opportunity to conduct high-impact, translational research in a

real clinical environment.

Lastly, I would like to thank my family – my mother Usha Ramesh, my father Ramesh Santhanam, and my brother Yeshwanth Ramesh (not to mention our pet dog Lilo and birds Tutu and Tikki) – for their support. While they did not always understand my research, their patience and trust in me was invaluable. I am thankful to my father in particular for helping me navigate the complexities of early adulthood and to deal with the setbacks that are natural to any PhD career. Once again, even when I was still learning to trust myself, my parents were unconditionally supportive and always knew I would make the right decisions.

This dissertation is divided into 6 chapters. Chapter 1 outlines the technical and human challenges of applying machine learning in the healthcare domain. Chapters 2 through 5 cover the solutions I developed to address these unique challenges, focusing specifically on problems in neurology and pulmonology. Chapter 6 summarizes the contributions of my work and provides an outlook on the field of artificial intelligence in healthcare.

Chapter 2 is largely a reprint of "Stroke-Associated Hemiparesis Detection Using Body Joints and Support Vector Machines" published in the *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare* by authors Vishwajith Ramesh, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. Chapter 2 also covers as of yet unpublished work (in review) "Robust pose-based identification of weakness in sitting stroke patients with an interpretable machine learning classifier" by Vishwajith Ramesh, Lisa M. Grega, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. This work was supported by the National Science Foundation Graduate Research Fellowships Program (DGE-1650112) and the UCSD Chancellor's Research Excellence Scholarship (formerly the Frontiers of Innovation Scholars Program).

Chapter 3 is largely a reprint of "CoughGAN: Generating Synthetic Coughs that Improve Respiratory Disease Classification" published in the *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* by authors Vishwajith Ramesh, Korosh Vatanparvar, Ebrahim Nemati, Viswam Nathan, Md Mahbubur Rahman, and Jilong Kuang. This work was conducted as part of an internship at Samsung Research America. I would like to acknowledge the entire Digital Health Team for providing me with a valuable and fun internship experience, particularly my mentors Korosh and Jilong.

Chapter 4 covers as of yet unpublished work (in review) "Detecting Motor Symptom Fluctuations in Parkinson's Disease with Generative Adversarial Networks" by Vishwajith Ramesh and Erhan Bilal. This work was conducted as part of two internships at the IBM T.J. Watson Research Center. I would like to acknowledge Dr. Erhan Bilal and the late Dr. Jeremy Rice for the opportunity to work in a prestigious industrial research lab and for providing me with the flexibility to work independently and to learn and implement novel deep learning techniques.

Chapter 5 is largely a reprint of "Assessing Clinicians' Reliance on Computational Aids for Acute Stroke Diagnosis" published in the *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare* by authors Vishwajith Ramesh, Andrew Nguyen, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. It also addresses "Developing Aids to Assist Acute Stroke Diagnosis" published in the *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems – Late Breaking Work* by Vishwajith Ramesh, Stephanie Kim, Hong-An Nguyen, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. Chapter 5 touches on follow-up work being prepared for submission at the time of writing this dissertation, "Designing a Clinical Decision Support System for Acute Stroke Diagnosis" by Vishwajith Ramesh, Stephanie Kim, Hong-An Nguyen, Andrew Nguyen, Gauri Iyer, Lisa M. Grega,

Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. The work covered in this chapter was supported by the NSF GRFP (DGE-1650112) and the UCSD Chancellor's Research Excellence Scholarship. I would like to specifically acknowledge all the students who I have had the pleasure to mentor during the course of my graduate studies, although I always thought of them as equal peers. The human-centered design aspects of my thesis certainly would not have been possible without the diligent efforts of driven and intelligent students (and co-authors) like Lisa M. Grega, Andrew Nguyen, Stephanie Kim, Hong-An "Ann" Nguyen, and Gauri Iyer.

My gratitude to everyone involved in my journey to complete my PhD. I will do my best to pass on my learnings and to aid others embarking on the same journey, much as I have been supported. Thanks!

VITA

2015	B.S. in Bioengineering, University of California Los Angeles
2017	M.S. in Bioengineering, University of California San Diego
2020	Ph.D. in Bioengineering, University of California San Diego

PUBLICATIONS

Vishwajith Ramesh, Andrew Nguyen, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. "Assessing Clinicians' Reliance on Computational Aids for Acute Stroke Diagnosis." *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2020).*

Vishwajith Ramesh, Korosh Vatanparvar, Ebrahim Nemati, Viswam Nathan, Md Mahbubur Rahman, and Jilong Kuang. "CoughGAN: Generating Synthetic Coughs that Improve Respiratory Disease Classification." *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2020).*

Vishwajith Ramesh, Stephanie Kim, Hong-An Nguyen, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. "Developing Aids to Assist Acute Stroke Diagnosis." *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems – Late Breaking Work.* May 2020.

Hesham Mostafa, **Vishwajith Ramesh**, and Gert Cauwenberghs. "Deep supervised learning using local errors." *Frontiers in Neuroscience.* Aug. 2018.

Vishwajith Ramesh, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. "Stroke-Associated Hemiparesis Detection Using Body Joints and Support Vector Machines." *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2018).*

Mustafa Ugur Daloglu, Wei Luo, Faizan Shabbir, Francis Lin, Kevin Kim, Inje Lee, Jiaqi Jiang, Wenjun Cai, **Vishwajith Ramesh**, Mengyuan Yu, and Aydogan Ozcan. "Label-free 3D computational imaging of spermatozoon locomotion, head spin and flagellum beating over a large volume." *Light: Science and Applications.* Aug. 2017.

Steven Rick, **Vishwajith Ramesh**, Danilo Gasques Rodriguez, and Nadir Weibel. "Pervasive Sensing in Healthcare: From Observing and Collecting to Seeing and Understanding." *Workshop on Interactive Systems in Healthcare, ACM Conference on Human Factors in Computing Systems.* May 2017.

Vishwajith Ramesh, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. "Exploring Stroke-Associated Hemiparesis Assessment with Support Vector Machines." *Extended Abstracts of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare.* May 2017.

Vishwajith Ramesh, Steven Rick, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. "A neurobehavioral evaluation system using 3d depth tracking and computer vision: the case of stroke-kinect." *Extended Abstracts of the Society for Neuroscience Annual Conference.* Nov. 2016.

ABSTRACT OF THE DISSERTATION

Human-Centered Machine Learning for Healthcare: Examples in Neurology and Pulmonology

by

Vishwajith Ramesh

Doctor of Philosophy in Bioengineering with a Specialization in Multiscale Biology

University of California San Diego, 2020

Gert Cauwenberghs, Chair
Nadir Weibel, Co-Chair

Machine learning (ML) in healthcare has enabled the automatic detection of diseases from medical images or sensors with high accuracy, often outperforming domain experts. Unfortunately, there is a large variance between how such diagnostic aids perform in research settings and in the real-world. This is due to challenges unique to healthcare that, if unaddressed, limit the usefulness of ML-based software when deployed in hospitals. For example, in human subjects research and clinical trials, subject recruitment and data acquisition are involved processes for both patients and healthcare providers; there are several regulatory and cybersecurity requirements to satisfy to ensure that patient care is not compromised in the pursuit of big data. Without abundant data to train ML models,

it can be difficult to elicit good performance that also generalizes well on unseen data in clinical practice. Moreover, ML tools in hospitals cannot function independently but must integrate with existing workflows. There are ethical considerations with respect to how these tools influence the decision making of clinicians and whether they encourage an over-reliance on predictions.

In this dissertation, we discuss these and other concerns in the context of three focus areas: stroke, respiratory disease, and Parkinson's disease. We present machine and deep learning pipelines for weakness detection in stroke patients from video, respiratory disease classification from audio of coughs, and gait assessment in Parkinson's disease with body sensors. In our efforts, we were cognizant of the technical and human challenges of healthcare. We developed models that not only performed well but also could be trained and rigorously evaluated in a data-conscious way. Our ML solutions ranged from simple leave-one-out approaches to data augmentation with generative adversarial nets. Lastly, we show how ML can more effectively aid medical diagnosis when paired with human-centered design. We describe a clinical decision support system for acute stroke, focusing on the development of an intuitive user interface that balances neurologist assessments with the symptom predictions of our models. This dissertation details novel, human-centered ML techniques for disease diagnosis in neurology and pulmonology, highlighting several lessons learned to benefit the field of machine learning in healthcare at large.

Chapter 1

Applied Machine Learning for Healthcare

1.1 Problems Unique to Healthcare

Machine learning (ML) and deep learning have made significant strides in the past decade, being used to drive cars autonomously [1–3], improve computational photography on smartphones [4, 5], and even create synthetic media [6, 7]. All three examples have fundamentally shifted entire industries and have driven heated discussions as to the implications of artificial intelligence (AI). A consumer vehicle able to drive autonomously based on video sensor feeds motivates a discussion of how such cars handle the safety of pedestrians relative to that of the driver [8–11]. A photograph no longer represents a singular moment in time but is comprised of several slices merged together by deep learning to create a better looking image for social media [4, 5]. The creation of synthetic media, so called "DeepFakes," have drawn considerable push-back from a public fearful of the spread of misinformation from agenda-driven institutions [7, 12–14].

Relative to its impact in other fields, the impact of AI in healthcare and medicine,

while growing, has been restricted. It can be challenging to show that AI tools designed to support medical decision making do not in practice negatively impact patient health or care. Unlike typical drugs or medical devices, there is a greater variance between performance of an AI or ML-based software in artificial testing environments and in actual practice [15]. This discrepancy is due to several technical and human problems presented by AI in medicine that are unique to this space. For example, to be effective in hospitals, AI systems in healthcare must be designed to work alongside a range of medical professions, from medical residents to nurses to experienced physicians. They need to work in conjunction with existing clinical tools and workflows in order to avoid obstructing the ability of providers to care for patients. Recently, the Food and Drug Administration (FDA) developed guidelines for classifying and regulating "software as a medical device" [15, 16]. The FDA recognizes that such challenges do not preclude the use of AI in healthcare nor do they reduce the large impact it has had and can yet have in the field. But it is important to understand them and to build AI-based systems that are cognizant of them. We outline a few of these challenges below.

1.1.1 Technical Challenges

The three examples given earlier in the automotive, photography, and media industries have benefited from a rich and large corpus of labeled images that can be used to train machine learning algorithms [17–21]. ML algorithms are very sensitive to the datasets being used to train them and often reflect the biases and imbalances in the training data. Image datasets are large enough to mitigate the effects of imbalances between classes, and even if such imbalances exist, they can be readily fixed. Consider for example Google's facial recognition technology for unlocking smartphones. There is a well documented imbalance in face datasets between dark-skinned people and light-skinned people [22–25]. To counter this

inequality, Google recorded videos of those with darker skin tones to minimize any racial bias in their face detection algorithms – a fairly straightforward (but ethically questionable) process in which participants were compensated $5 [26–28]. Even if a racial bias was not corrected for, ultimately, inability to unlock a phone via face is an innocuous concern to the consumer.

In healthcare, however, such problems can have grave consequences. Healthcare datasets tend to be magnitudes of order smaller than popular image datasets, which are exceedingly big; there are 60,000 labeled examples in CIFAR-100 over 14 million examples in ImageNet [17, 20]. In human subjects research and clinical trials, it is challenging to recruit and record a large number of participants on the order of millions or even tens of thousands [29–32]. Patient recruitment and the logistics of data acquisition in hospitals and clinics is an involved process with numerous regulatory and cybersecurity hurdles [33, 34]. After all, extensive requirements are necessary to avoid obstructing patient health and care. Acquiring labels, the gold standard target of interest used in supervised learning tasks (e.g. true states of clinical outcomes or true disease phenotypes), requires the domain knowledge of highly trained experts [35]. Data collection must therefore involve the time of doctors, nurses, clinicians, and healthcare providers, in addition to that of patients recorded. These constraints tend to limit the size of healthcare datasets comprised of information from human subjects.

Machine learning algorithms often tend to overfit when trained with little data, performing well on the data used to train the algorithms but not on unseen out-of-sample data; they have poor generalizability [36–39]. The effects of imbalances and biases are especially strong with small datasets [38, 40]. Imbalances in data may lead to underfitting with poor prediction accuracy, especially due to samples not well represented in the training set [37]. Feature rich input (time series from body sensors with high sampling frequency,

for example) in an otherwise small dataset and overparameterized models trained on a finite number of samples suffer from overfitting and low statistical precision – the "curse of dimensionality" [41, 42].

Fixing imbalances is challenging because of the difficulty of consenting and recruiting (sick) patients, especially when considering the ease with which individuals otherwise underrepresented in a facial recognition database can be recruited and recorded. The labels required to train supervised learning algorithms can also be difficult to come by in medicine. An example that highlights all these healthcare-specific data concerns is the electronic health record (EHR). Longitudinal EHR data describes the trajectory of a patient's health over time. Since different patients visit hospitals at different frequencies, events in EHR data are irregularly sampled [35]. Moreover, patients who visit a hospital often comprise more of a EHR dataset; healthy individuals are unlikely to make regular clinical visits and are thus less represented. EHR datasets are inconsistent between different sites due to a lack of standards and semantic interoperability of health data. The inconsistency is worse when considering free-text notes written by doctors, nurses, and other providers [35, 43, 44]. Gold standard labels are also not consistently captured in EHR data, in part due to different standards between healthcare systems. The unavailability of labels further restricts the amount of EHR data usable for training [35].

With biometric facial recognition for unlocking smartphones or computational photography, mistakes made by AI are relatively harmless. Mistakes made by an AI-based clinical decision support system, however, can be pernicious to patient health. Medical errors by humans are considered to be a leading cause of death and AI may only serve to exacerbate this issue [45–48]. Prediction mistakes due to improper training on small and imbalanced datasets may also lead to malpractice lawsuits and large settlements for hospitals and health care providers [49, 50]. It is for these reasons that the issue of small

and imbalanced datasets is of special concern in healthcare.

1.1.2 Human Challenges

While there is an abundance of retrospective studies conducted in controlled settings, AI-based clinical decision support systems deployed in the real world are relatively few and far between. There is a large variance between performance in the two environments due in part to human factors [15]. AI tools in hospitals cannot work independently but in conjunction with existing clinical care workflows, resources, and data management protocols, all of which involve humans [15, 51]. Consider a computer-aided detection system for mammography that was shown to improve breast cancer detection and therefore deployed in practice [52]. Real world diagnostic performance with the system was in fact no better (and, in some cases, worse) than without because of how humans interacted with the system [15, 52].

Poor interpretability of AI or ML-based algorithms can cause confusion and make it difficult to understand the underlying reasons behind predictions. AI often tends to be a black box but its inner workings cannot be nebulous when it comes to healthcare, as it impacts a doctor's trust in the system. Physicians prefer to understand how systems produce recommendations and desire explainability as a feature of clinical decision support systems [53–55]. Systems can be considered explainable if they use non-black box approaches (a rule-based one, for example) or black box algorithms that make interpretable and justifiable predictions [56–59]. Interpretability is particularly important in situations when the AI system disagrees with the clinician; a result that deviates from expected behavior will prompt a need for an explanation [60–62].

Human judgment also introduces biases and errors. Humans can over-extrapolate from small samples, identify false patterns from noise, and be unusually risk averse [15, 63].

There are ethical concerns as well, with respect to how AI tools influence the decision making of clinicians and whether they encourage an over-reliance on predictions, particularly in those with little experience [60, 64]. Robust human factors testings and a human-centered design approach are therefore required for AI-based computational aids to be effective in practice [15].

1.2 Exploring Solutions by Example

We outlined a few technical and human challenges unique to the healthcare space above. In the following chapters of this dissertation, we touch again on these challenges and introduce other concerns in the context of **three specific focus areas: stroke, Parkinson's disease, and respiratory disease.** In **Chapter 2**, we describe our early efforts to develop a clinical decision support system for acute stroke diagnosis. In this work, we developed a machine learning classification pipeline for identifying weakness in sitting stroke patients from video. We discuss the challenges of building a stroke dataset in an active clinical environment and our approach for training and rigorously evaluating the ML pipeline in a data-conscious way. We outline the pros and cons of leave-one-out cross validation, a training strategy that enables maximal use of limited data for training but has high variance. In **Chapter 3**, we discuss an unsupervised data augmentation strategy to address the small, imbalanced dataset concern we brought up in Section 1.1.1. We significantly improved the performance of off-the-shelf classifiers for cough-based respiratory disease diagnosis by adding synthetic cough examples to balance and augment a training set with otherwise underrepresented classes. In **Chapter 4**, we detail a deep learning pipeline to assess the gait of Parkinson's disease patients and to predict their ON/OFF state. We showed that training the network adversarially alongside another network that generated fake samples ultimately reduced overfitting, a concern with applying deep learning on

small datasets. In **Chapter 5**, we discuss again the clinical decision support system for stroke but this time focusing on our efforts to make it human centered. Through a user interface experiment, we showed that domain experts do not seem to over-rely on the system's predictions and can catch the mistakes that it makes, properly weighing their own assessments with the symptom predictions of the algorithms. Human factors testing such as the one we outline in this chapter are necessary to maximize the efficacy of such systems in practice, as mentioned in Section 1.1.2. Lastly, in **Chapter 6**, we discuss the contributions of our work, highlighting how the solutions we developed in three focus areas can benefit the field of machine learning in healthcare at large.

Chapter 2

Identifying Stroke-Associated Weakness from Video

2.1 Motivation and Background

Stroke is a leading cause of disability and death and costs the United States $34 billion dollars a year in missed days of work, treatment, and rehabilitation [65, 66]. Early treatment of stroke improves long-term outcomes and lowers recurrent stroke risk by up to 80% [67]. The standard of care is the administration of a clot dissolving agent, tissue plasminogen activator (tPA), within 4.5 hours of the onset of symptoms. To facilitate early treatment, hospitals see most strokes first in the emergency department (ED); approximately 70% of strokes are admitted as inpatient from the ED [68].

Unfortunately, strokes are often missed or incorrectly diagnosed in the ED. A comprehensive study examining the ED visit records of hospitals across 9 states in the United States found that nearly 13% of stroke cases were diagnostic errors – a diagnosis that was missed or incorrect, as determined by a definitive test or finding done at a later time [69]. This corresponded to 24,000 patients who received an incorrect diagnosis and

were seen on a treat-and-release basis in the ED, only to be later admitted as inpatient with a confirmed stroke diagnosis [69]. Because of the narrow treatment window of tPA and its reduced effects if administered after this window, such missed or delayed diagnoses can prove fatal or hurt a patient's chances of rehabilitation. Another study found that 22% of strokes were missed in the ED [70]. Of these a third presented within the time window for tPA eligibility, but because they were misdiagnosed, they did not receive tPA [70]. A study assessing stroke care in Texas found that 35% of stroke and transient ischemic attack cases were missed in the ED [71]. This last study is of particular interest because most of the patients who presented with acute neurological symptoms in the ED were not examined by neurologists but by ED physicians alone [71]. The high rate of missed strokes by emergency medicine clinicians can be ascribed to the difficulty of distinguishing strokes from stroke mimics [71, 72]. There is a wide range of diseases with symptoms similar to those of stroke – seizure, encephalopathy, and brain tumors, for example – and an experienced vascular neurologist is required to be able to differentiate strokes from mimics [73].

The cost of stroke misdiagnosis is high. 65% of malpractice claims made against clinicians involved a delayed or missed ED diagnoses that harmed patients, and compensation for plaintiffs ranged from $100,000 to $30 million [74, 75]. There is a large cost for hospitals as well, as treating a stroke mimic increases costs by a median of over $5,000 [76]. Missed strokes have been shown to increase hospital length of stay by 2 days, corresponding to approximately $1,600 per day in costs [77]. Most importantly, there is a large human cost – 80,000 to 160,000 preventable deaths or permanent disabilities each year in the U.S. [78].

To reduce the number of incorrect stroke diagnoses made in the ED, we propose a computational system that uses ubiquitous technology and machine learning to capture the expertise of experienced neurologists and translate it to the emergency room. A system that automatically identifies acute stroke symptoms when present can help ED physicians to

not only accurately distinguish stroke from stroke mimics but also reduce time to diagnose or rule out stroke. Such a system especially benefits hospitals in underserved areas without ready access to neurologists nor a comprehensive stroke center.

As a first step towards a clinical decision support system for acute stroke, we focus on automatically identifying body weakness, or hemiparesis.

2.1.1 Hemiparesis – A Key Indicator of Acute Stroke

Hemiparesis is the partial paralysis of the left or right side of the body and is a common, early warning sign of stroke [79–81]. It is characterized by difficulty moving the limbs of the affected side. Hemiparesis is identified and its severity assessed through motor arm and leg tests conducted by a trained stroke specialist as part of the National Institute of Health Stroke Scale (NIHSS) [82, 83]. While the NIHSS motor tests have been found to have strong inter-rater reliability, they require an experienced neurologist to conduct them [84].

Efforts have been made to quantitatively analyze hemiparesis without the need for a stroke specialist. Computational methods to track walks using wearable three-axis accelerometers, reflective markers, and ankle-mounted step watch activity monitors have been shown to reliably identify hemiparetic gait [85, 86]. These techniques were designed to monitor hemiparetic stroke patients outdoors or at their residence and do not require the presence of a neurologist. Unfortunately, the approaches focus on rehabilitation in the long-term rather than diagnosis in the short term, with the goal of tracking improvements in movement over time. Moreover, they require made-to-fit, obtrusive devices to be worn by patients.

In this chapter, we explore means of determining hemiparesis in an unobtrusive way without requiring active motor movements like walking. Our goal is to develop approaches

that can be used in emergency medical settings without ready access to neurologists for diagnosis soon after the incidence of stroke. We first describe a support vector machine (SVM) pipeline trained on body skeletons of stroke patients recorded with the Microsoft Kinect. We then discuss a more sophisticated video-based approach that improved on the initial model. This work was based on the hypothesis that body posture, particularly when sitting, is a meaningful metric for diagnosing hemiparesis in stroke patients [87–89]. In the context of the overall dissertation, we will highlight key problems with clinical data collection as well as the pros and cons of using the leave-one-out training paradigm to measure classifier performance.

2.2 Hemiparesis Detection with the Microsoft Kinect and Support Vector Machines

2.2.1 Building a Clinical Dataset of Hemiparetic Stroke Patients

In collaboration with a team of neurologists from the University of California, San Diego (UCSD) Stroke Center, we recorded 39 stroke subjects at outpatient clinics in San Diego with the Microsoft Kinect v2. The Kinect enabled us to record audio, high-definition video at 60 frames per second, and depth footage of subjects, as well as provided the x, y, z spatial coordinates of 25 body joints over time. The Kinect was mounted on the wall opposite to a patient table. We placed the camera fully facing the subject to avoid occluding parts of the body. Data was collected using ChronoSense software, at a body tracking capture rate of 30 Hz [90]. Figure 2.1 shows our data collection set-up. All subjects agreed to be recorded with sensors by signing a consent form approved by the local Human Research Protections Program office.

We recorded subjects during a set of motor and cognitive evaluation exams – the

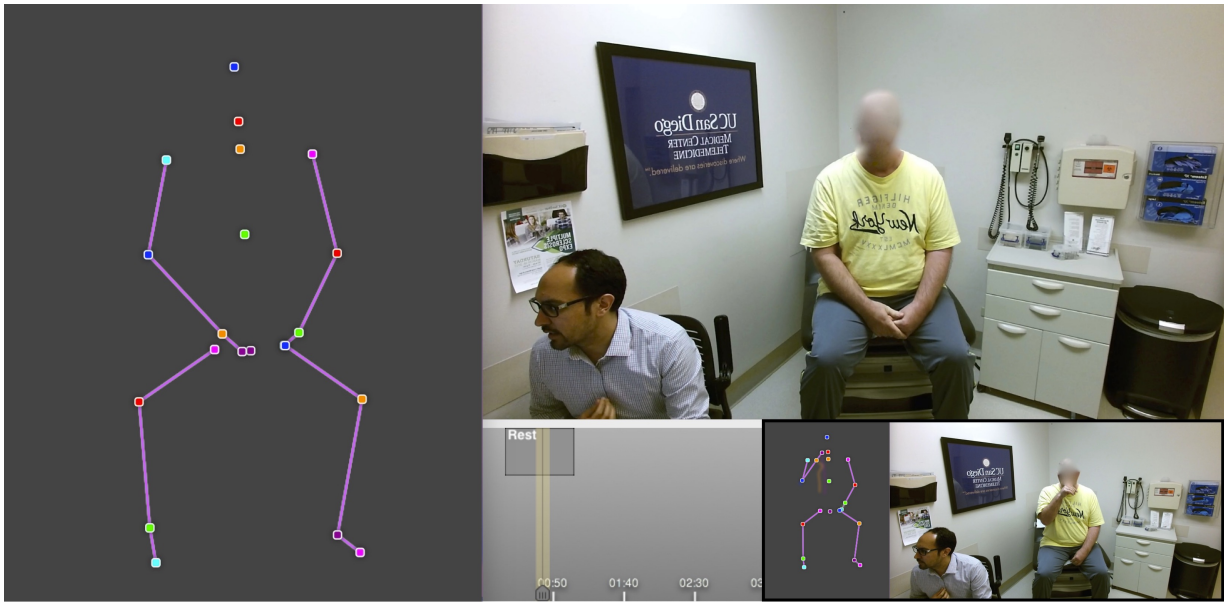


Figure 2.1: Body tracking data collected from a subject in an outpatient clinic with the Microsoft Kinect v2. The subject shown was sitting at rest prior to the start of the NIHSS examination. The representation of the body skeleton obtained with the Kinect is on the left. Note that because the image and skeleton is mirrored, the viewer's right side corresponds to the subject's right side.

NIHSS [83]. Hemiparesis was identified and its severity assessed through motor arm and leg tests conducted by a trained stroke specialist [82]. We also recorded subjects as they were sitting and waiting for the neurologist to begin these in-person assessments. Subjects sat at rest waiting for the neurologist to begin the NIHSS exam anywhere from 10 seconds to 2 minutes.

Of the 39 subjects, 4 were omitted from this study due to technological errors in data capture and storage. 13 out of 35 subjects suffered from hemiparesis – 8 with right-side weakness and 5 with left-side weakness. 22 subjects were used as non-hemiparetic controls. Controls were healthy subjects who had never had a stroke or were fully recovered (no deficits identified by the neurologist's NIHSS exam). To incorporate as much of the collected data into the training set, we also included subjects who had stroke symptoms unrelated to weakness such as facial droop or speech difficulties. A subject with conversion

disorder but who was otherwise healthy was considered a control too.

Subjects seen at the outpatient clinic are usually several months from the incidence of stroke and are well into their rehabilitation, so their symptoms tend to be less severe. Weakness in each arm or leg is scored on a scale of 0 to 4 [83]. The sum of the NIHSS scores for the left arm and left leg or right arm and right leg ranged from 1 to 4 (out of a maximum of 8) for the 13 hemiparetic outpatient subjects.

2.2.2 Capturing Body Posture at Rest with the Kinect

For each of the 35 outpatient subjects, we obtained a time series of 25 body joints as they were sitting at rest. We focused our analysis on this rest period because we wanted to capture subjects in their natural state; we did not ask them to perform any action or activity. This would allow our system to be deployed in EDs where stroke patients could be monitored while waiting to be evaluated by a specialist. Note that it is often challenging to ask patients admitted to the ED to perform motor tasks because they are likely severely debilitated by their recent stroke.

To account for differences in heights, physicality, and body skeletons mapped by the Kinect, we used body angles, a metric that is relative to each subject and that can be compared between subjects. To extract the value of the core body angles (shown in blue and green in Figure 2.2), we first calculated 3D vectors between 4 core body angles (highlighted in dark red in Figure 2.2). We then applied simple trigonometry. The body angle θ between 3D unit vectors a and b, each connecting 2 body joints, was calculated (in degrees) as: $\theta = \arccos(a \cdot b)$. We monitored core body angles instead of limb angles (about the elbow for example) because even at rest, subjects tended to move their arms and legs. We did not ask subjects to remain perfectly still to avoid being obtrusive.

To account for temporal dynamics of motion, i.e. changes in the 4 core body angles,

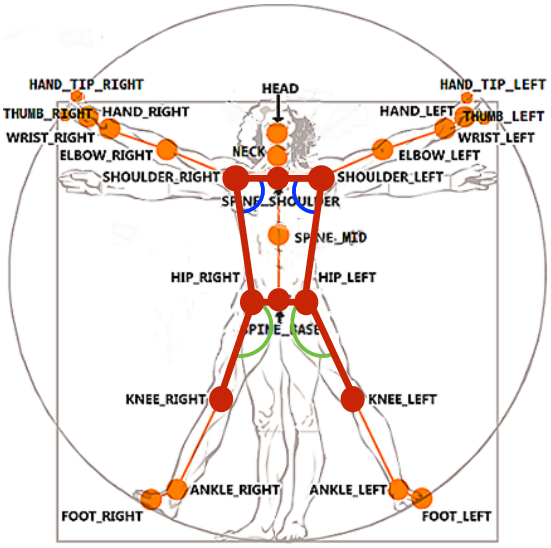


Figure 2.2: Body geometry recorded by the Microsoft Kinect v2. Orange dots represent 3D positions of body joints. Highlighted in blue and green are the 4 core body angles analyzed. The vectors from which these angles were calculated are in red.

we computed the first order derivative or slope of each of the 4 time series. Slope was the difference between 2 consecutive data points divided by $\Delta t = \frac{1}{\text{Frequency of Data Capture (30 Hz)}}$. Slight changes in a body angle over time – fidgets – were captured in the time series of its slope; if no fidgets occurred, slope was 0. Since fidgets happened randomly, if at all, we could not predict when during the rest period they occurred or their amplitude (the amount of change in body angles). We therefore averaged the root mean square (RMS) value of the slope of the 4 core body angles over the entire rest duration.

Our data processing yielded a set of 4 features for each of the 35 subjects. Each feature was obtained by averaging (over the rest period) the time series of the RMS slope of 1 of the 4 core body angles. Our approach was greatly simplified because we were no longer dealing with time series of dynamically changing body angles.

2.2.3 Developing an Initial Hemiparesis Classification Pipeline

In order to predict whether each subject had hemiparesis or no hemiparesis based on 4 average RMS slopes described above, we used the C-support vector classification implementation from the Scikit-learn library (`svc.svm.SVC`) [91]. The benefit of using SVMs over other machine learning algorithms is that they perform well when the number of feature dimensions is much larger than the number of samples in the dataset.

Because of the low sample size of 35 subjects and to avoid overfitting, we used leave-one-out cross validation instead of dividing the dataset into separate training and test sets. Cross validation is an estimate of model generalizability, i.e. the expected fit of a model to unseen data independent of the data used to train it. We used the classification accuracy as a means of summarizing this fit. In the leave-one-out training paradigm, 34 outpatient subjects were used to train a linear SVM, which was subsequently tested with the 1 outpatient subject left out of the training set. This process was repeated 35 times, with each subject left out and tested exactly once. The leave-one-out training paradigm maximized the amount of data for training. We obtained a predicted label per subject that we then compared to the ground truth label. Classification accuracy was calculated as the number of correct predictions divided by the total number of samples.

To account for imbalance between sample sizes of the classes, we weighted cost parameter C of each class i by the inverse of its sample size. Weighting cost in this way greater penalized mistakes in classes with lower sample size: $\frac{\text{Total Samples in Training Set (34 with leave-one-out)}}{\text{Number of Samples of Class } i \text{ in Training Set}}$.

We performed a grid search over all possible combinations of the following values of SVM hyper-parameters. The performance metric optimized was classification accuracy. We tested radial basis function, linear, sigmoid, and polynomial kernels (with 2, 3, or 4 possible degrees for the polynomial kernel). Based on suggestions in literature, we varied penalty parameter C and gamma by orders of magnitude [92]. To minimize overfitting, we

capped C at 1.0 to favor a soft-margin classifier. Possible values for C were therefore 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, and 1.0. Possible values for gamma were $2e{-}9$, $2e{-}6$, $2e-3$, 0.25, 1.0, $2e3$, $2e6$, $2e9$. We report classification accuracy of predictions from the best, most optimal classifier identified by the grid search.

Performance of SVM Classifier

The grid search revealed that a radial basis function kernel (C = 10^{-5}, gamma = $2e{-}9$) gave the best performance in terms of classification accuracy. By applying an SVM algorithm with a radial basis function kernel under the leave-one-out training/testing methodology, we obtained a prediction per subject. We compared each prediction with the corresponding ground truth label from the neurologist to calculate an overall accuracy.

With 35 subjects, the SVM was able to classify hemiparesis and no hemiparesis with 100% accuracy. There were no instances of misclassification. We were able to achieve this accuracy when we averaged the RMS slopes of body angles over the rest period for each subject. Subjects only need to sit at rest without being asked to move, waiting for a neurologist to start the NIHSS examination, for our algorithm to detect hemiparesis with high accuracy. The benefit of such an approach is that it can be implemented in emergency settings where subjects are not likely to move. Gait analysis in ambulances or in the ED, for example, is difficult. Our method was less obtrusive than wearable accelerometers, and did not need to be tailored to each subject because body angles were relative and could be compared between subjects.

Moreover, the results support medical literature that show that hemiparetic patients have difficulty maintaining their posture and balance due to reduced muscle strength on the paretic side [87]. The ability to maintain balance requires postural control, which involves limiting the body's sway and keeping its center of gravity within its base of support.

Studies used a force platform to measure center-of-pressure displacements in sitting stroke patients and found increased postural disturbance in hemiparetic patients and an inability to recover from leaning either forward or to the paretic side while sitting [88, 89]. The Kinect and RMS slope features computed from body angles at rest were seemingly accurate and precise enough to capture subtle fidgets in sitting posture.

Despite promising results, there were significant drawbacks with this work that limited its real-world use. The SVMs were not tested on subjects with severe symptoms, only on recovering outpatient subjects with slight symptoms. The optimal model identified by the grid search likely would not generalize well to patients with severe strokes seen in the ED. Furthermore, accurate Kinect body skeletons (without occlusion of the core body) were required for the classifier to work well. The angle-based machine learning pipeline also involved complicated feature processing and grid search, making the features and results difficult to interpret.

2.3 A Video-Based Machine Learning Pipeline for Identifying Weakness in Sitting Stroke Patients

Despite drawbacks, our early hemiparesis detection efforts showed that a SVM classifier could be used to identify hemiparesis from body posture at rest [93–95]. That is, the way patients orient themselves in 3D space at rest is indicative of body weakness, a conclusion bourne out by medical literature that showed that hemiparetic patients have difficulty maintaining their posture and balance [96, 97]. Inspired by these results, we leveraged the latest in video-based pose estimation to automatically diagnose hemiparesis from 10 seconds of footage of stroke subjects simply sitting at rest, without requiring any movement (Figure 2.3). We obtained high performance with a SVM classifier trained

and rigorously tested with two real-world clinical datasets: the dataset of outpatient clinic subjects described in Section 2.2.1 and a dataset of more severe inpatient subjects. Moreover, our new video-based classifier exceeded the performance of 8 experienced stroke specialists conducting a video-based assessment of hemiparesis.

2.3.1 Obtaining Body Skeletons from Video

In our initial approach described in Section 2.2.2, we used body skeletons (particularly body joint angles) generated by the Microsoft Kinect v2 to classify hemiparesis [93–95]. Unfortunately, the hospital gowns worn by some subjects (especially inpatient subjects, as described later) interfered with the Kinect's depth-based pose estimation approach.

As an alternative, we used VideoPose3D, a deep learning approach for 3D human pose estimation from video published by Pavllo et al. in 2019 [98]. Using the VideoPose3D technique and the high-definition videos recorded, we generated 17-joint body skeletons for

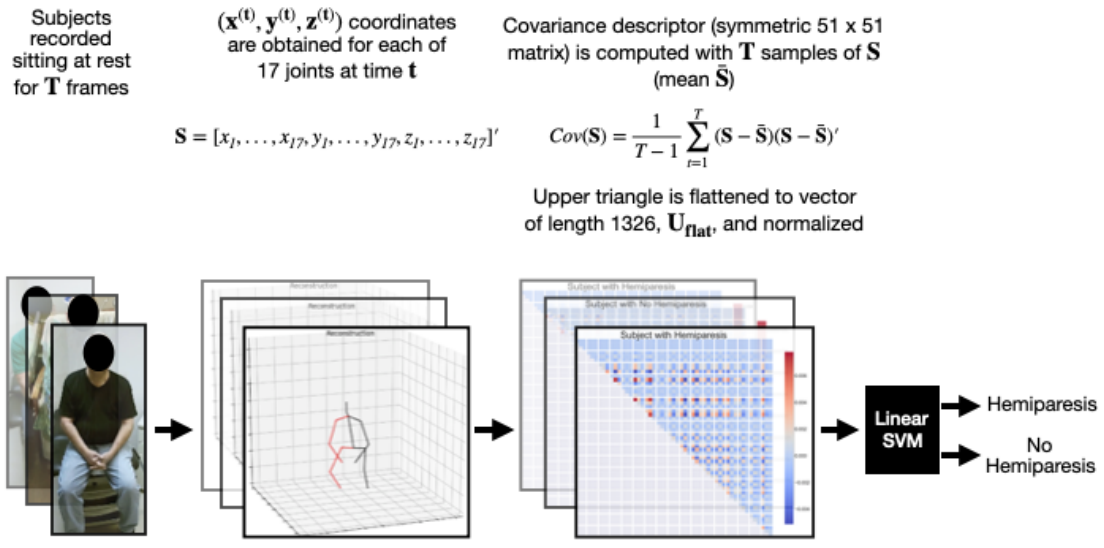


Figure 2.3: Overview of video-based hemiparesis classification approach. Body skeletons are obtained from videos of subjects sitting at rest. The covariances of body joints are used as input into a linear support vector machine classifier for hemiparesis detection.

each subject. All subjects had a rest duration of at least 10 seconds. So we considered only the first 10 seconds of recorded video to avoid omitting subjects from the dataset.

At 60 frames per second for X seconds of video, $T = 60 * X$ total frames. For each frame t out of T, we obtained x (left/right), y (up/down), z (toward/away from camera) spatial coordinates of 17 body joints per subject. We defined $\mathbf{S}$ as the vector of all joint locations per subject:

$$\mathbf{S} = [x_1, ..., x_{17}, y_1, ..., y_{17}, z_1, ..., z_{17}]' \qquad (|\mathbf{S}| = 51)$$

2.3.2 Computing a Feature Descriptor for Body Posture

Body posture at rest has shown to be indicative of hemiparesis [94, 95]. Because we only considered subjects sitting at rest and subjects did not significantly move around during the 10 seconds, we treated each frame of video as independent from the next. Each subject had T samples from which we calculated sample mean $\bar{\mathbf{S}}$.

To obtain a feature descriptor for body posture, we computed the covariance matrix from 3D joint locations for each subject, as described by Hussein et al. in 2013 [99]. A benefit of the covariance matrix is that it can be visualized and interpreted easily with a heat map. The covariance matrix describes whether 3D body joint positions correlate positively with each other, negatively with each, or do not correlate, and to what extent. We calculated the sample covariance as follows:

$$Cov(\mathbf{S}) = \frac{1}{T-1} \sum_{t=1}^{T} (\mathbf{S} - \bar{\mathbf{S}})(\mathbf{S} - \bar{\mathbf{S}})'$$

where $'$ is the transpose operator.

Each subject had T vectors $\mathbf{S}$ with which the covariance matrix for that subject was calculated. T was defined to be the frame rate of video capture, 60 frames per second, multiplied with the duration of video in seconds, X. We considered X values between 1

second and 10 seconds (T between 60 and 600) to gauge how few samples per subject were required for good classification performance. That is, we wanted to identify the least number of frames of video necessary for SVM performance to be better than that of clinician raters.

By definition, the sample covariance matrix is a symmetric $51X51$ matrix so we only consider the upper triangle of $Cov(\mathbf{S})$, $\mathbf{U}$. After flattening, the descriptor has length $|\mathbf{S}|(|\mathbf{S}|+1)/2 = 1326$: $\mathbf{U_{flat}} = [u_1, u_2, ..., u_{1326}]$. In this way, for outpatient subjects, we obtained a set of 35 $\mathbf{U_{flat}}$ descriptors. The feature processing protocol is outlined in Figure 2.3.

2.3.3 A Human Benchmark for Hemiparesis Classification

In order to obtain a baseline against which to compare the performance of our machine learning model, we created and emailed a survey to clinicians at the UCSD Stroke Center. Telestroke evaluations are as accurate as in-person evaluations in diagnosing acute stroke, and video-based motor arm and leg tests in particular having high NIHSS score correlation with bedside evaluations [100, 101]. Motivated by these studies, we asked stroke specialists to view the recorded videos of 35 subjects sitting at rest (for 10 seconds to 2 minutes depending on the subject) as well as undergoing NIHSS motor tests; we obtained a separate set of responses from each type of video shown. The former is also the input into the machine learning pipeline (see Figure 2.3) while the latter is representative of a telestroke NIHSS motor exam. Average accuracy was calculated for each of the two cases by comparing survey participants' responses for whether or not an outpatient subject was hemiparetic with the results from the in-person neurologist's motor exam. (Note that survey participants had never examined any of the 35 subjects previously, in person or over video.)

8 clinicians completed the survey. We found that when examining at footage of subjects at rest, average accuracy across the 8 participants was 64%. Diagnostic performance was better when participants looked at footage of the NIHSS motor exam – 68% average accuracy.

The benefit of a computational diagnostic tool is that it should perform consistently across multiple runs, whereas human evaluation can give inconsistent results based on the individual conducting the NIHSS exam. To gauge this inconsistency, we calculated the inter-rater reliability using the free-marginal multi-rater kappa coefficient, κ_{free}. This type of kappa coefficient is suitable for cases involving more than two raters and for which raters are not informed beforehand of the number of subjects that fall into each category [102]. Both conditions held true in our survey. The equation for the free-marginal multi-rater kappa coefficient is:

$$\kappa_{free} = \frac{\left[\frac{1}{Nn\,(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right)\right] - \left[\frac{1}{k}\right]}{1 - \left[\frac{1}{k}\right]}$$

N is the number of subjects evaluated via video, n is the number of raters, and k is the total number of categories into which the subjects can be placed. We calculated κ_{free} for each of the two sets of participant responses. A κ_{free} value of 0 indicates no consistency in the responses of raters while a value of 1 indicates perfect consistency. Inspired by Fleiss' Rule of Thumb, we used a cutoff of 0.75 for gauging good or bad inter-rater reliability; $\kappa_{free} > 0.75$ was considered good [103].

We calculated the free-marginal multi-rater kappa coefficient, κ_{free}, with $N = 35$ outpatient subjects, $n = 8$ survey participants, and $K = 2$ categories (hemiparetic or not). Kappa for raters examining videos of outpatient subjects at rest, κ_{free_rest}, was 0.69. For raters examining videos of the NIHSS motor exam, $\kappa_{free\_movement}$ was 0.61. Both κ_{free} values were below Fleiss' 0.75 cutoff for good performance.

2.3.4 Video-Based Hemiparesis Classification with SVMs

To classify hemiparesis from covariance descriptors, we once again used the Scikit-learn implementation of a C-support vector machine (`svc.svm.SVC`) [91]. The SVM classifier is known to perform well for small datasets. Moreover, while we explored the use of other kernels, using a linear kernel enabled us to not only maximize classification performance but also let us determine the top 10 features that contributed the most towards classification (that is, the features with the greatest weights `svm.coef_`) [104]. Identifying the support vectors that influence the decision boundary is intuitive only for a linear kernel [105].

Normalization of data has been shown to maximize the performance of SVMs. We explored two strategies:

$$\text{L2 Norm: } \frac{\mathbf{U_{flat}}}{\|\mathbf{U_{flat}}\|_2} = \frac{\mathbf{U_{flat}}}{\sqrt{{u_1}^2 + {u_2}^2 + \cdots + {u_{1326}}^2}}$$

$$\text{Max Norm: } \frac{\mathbf{U_{flat}}}{\max(\mathbf{U_{flat}})} = \frac{\mathbf{U_{flat}}}{\max(u_1, u_2, ..., u_{1326})}$$

To facilitate comparison with the results from human raters, we compared the predictions of the SVM with the hemiparesis diagnoses from the in-person neurologist to compute raw accuracy. We also calculated weighted F1, weighted precision, and weighted recall. For rigor, we report results from 6 different training and testing paradigms: leave-one-out and 90/10, 80/20, 75/25, 66/33, and 50/50 train/test splits. The leave-one-out training paradigm was the same as the one described in Section 2.2.3. Leave-one-out not only maximized the amount of data used for training but also allowed us to identify problematic subjects and possible reasons why the SVM failed to make a correct prediction.

To test the robustness of the classifier and its sensitivity to reducing the number of training samples, we randomly selected 90% of the dataset (31 outpatient subjects) for training and 10% (4 outpatient subjects) for testing. We calculated the aforementioned performance metrics from predictions made on the test set. We repeated this process 20 times for 20 different classifiers, each trained on a random selection of 90% of the dataset

and tested on 10%. We averaged the evaluation metrics across all 20 runs to check how consistently the SVM performed. We repeated this same analysis for smaller training set sizes – 80/20, 75/25, 66/33, and 50/50 train/test splits.

Performance of SVM on Outpatient Subjects Only

Table 2.1 summarizes the results of experiments using outpatient subjects only. The leave-one-out training paradigm had highest accuracy, weighted F1, weighted precision, and weighted recall, all >=80%. As the size of the training set decreased, the performance of the SVM generally decreased. Regardless, all accuracies, including for the 50/50 train/test split, were >=70%. L2 normalization yielded slightly better SVM performance on outpatient subjects than max normalization (∼3% bump in accuracy for leave-one-out).

SVM Performance Compared to Clinicians

80% accuracy obtained with leave-one-out exceeded 64% average accuracy of 8 clinician raters using video footage of subjects sitting at rest to diagnose hemiparesis. The SVM also outperformed clinician raters using video footage of subjects moving their arms and legs as part of the NIHSS motor exams.

80% was obtained with $T = 60 * 10 = 600$ samples per subject. By varying the number of samples used to compute a covariance matrix per subject, we found that as the number of samples increased, the performance of the SVM generally increased as well. Figure 2.4 shows the accuracy and weighted F1 metrics for a linear SVM trained and tested with leave-one-out with respect to the number of seconds of video used as input. We found that we needed at least 8 seconds of video footage to match the best performance of clinician raters (68% accuracy).

Table 2.1: SVM Performance on Outpatient Subjects – 10 Seconds of Video

				Linear SVM, Balanced Class Weight							
				L2 Norm				Max Norm			
Experiment	**Num Train**	**% Train**	**Num Test**	**Test Accuracy**	**Weighted F1**	**Weighted Precision**	**Weighted Recall**	**Test Accuracy**	**Weighted F1**	**Weighted Precision**	**Weighted Recall**
Leave-One-Out	34	97%	1	80.00%	80.30%	81.78%	80.00%	77.14%	77.52%	79.91%	77.14%
90/10 Split	31	89%	4	77.50%	76.81%	80.31%	77.50%	76.25%	75.60%	85.83%	76.25%
80/20 Split	28	80%	7	70.71%	68.96%	74.00%	70.71%	75.00%	74.04%	77.67%	75.00%
75/25 Split	26	74%	9	73.89%	72.90%	76.32%	73.89%	68.89%	68.69%	76.71%	68.89%
66/33 Split	23	66%	12	75.83%	74.75%	77.32%	75.83%	72.08%	72.10%	76.89%	72.08%
50/50 Split	17	49%	18	71.94%	70.49%	72.40%	71.94%	63.89%	63.57%	65.26%	63.89%

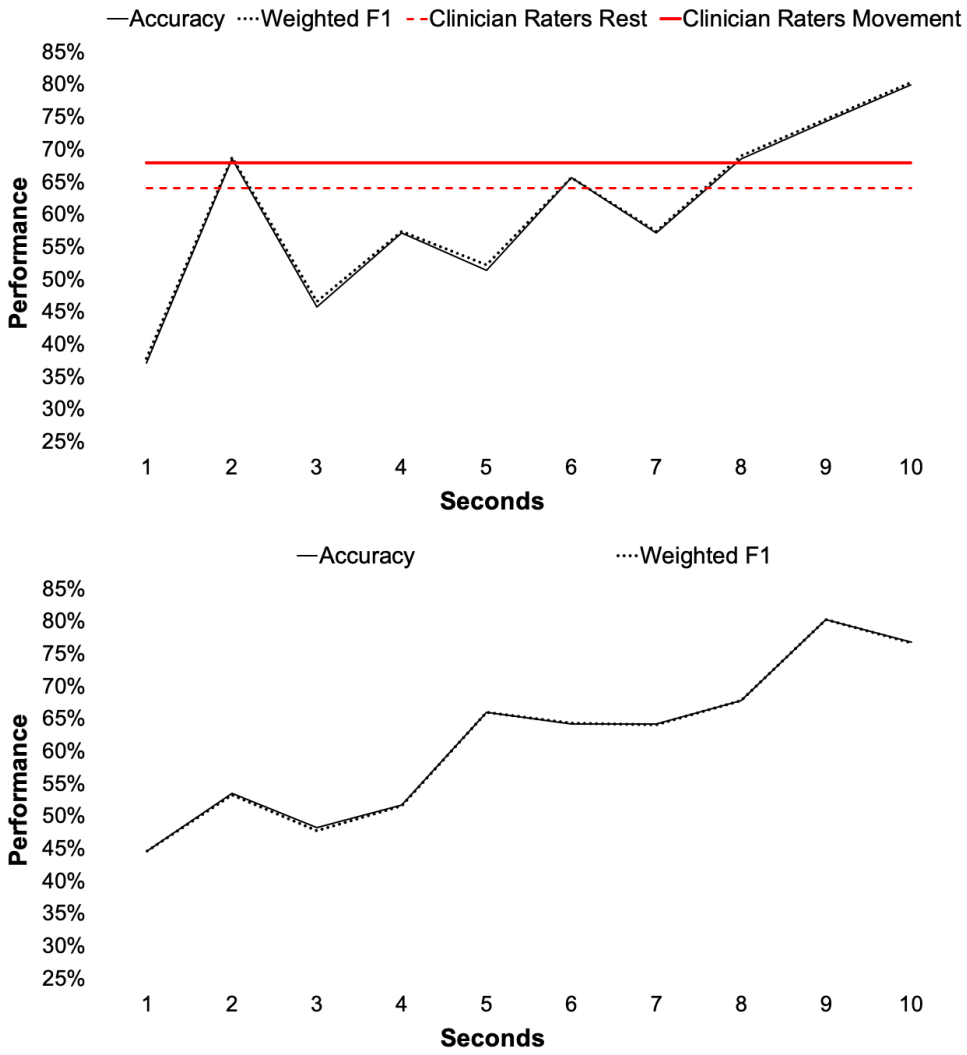


Figure 2.4: Accuracy and weighted F1 of SVM trained using leave-one-out with respect to duration of video used as input. Classifier performance generally increased with the number of frames of video used to calculate covariance descriptors. *Top:* SVM trained on outpatient subjects only. In red are the performance of clinician raters using video to detect hemiparesis, either footage of subjects at rest ("Clinician Raters Rest") or of subjects moving their arms and legs as part of NIHSS motor exams ("Clinician Raters Movement"). Accuracy meets and exceeds human benchmark, for durations greater than 8 s. *Bottom:* SVM trained on outpatient and inpatient subjects.

2.3.5 Incorporating Inpatient Subjects with Severe Weakness

To augment the dataset of outpatient subjects, we incorporated 21 subjects – 12 with hemiparesis and 9 without – seen at inpatient clinics at the UCSD Jacobs and Hillcrest Medical Centers. These subjects were diagnosed with stroke 24 to 48 hours prior to being recorded, and were kept in hospital for observation. Due to the recency of stroke, inpatient subjects had more severe symptoms; the sum of the left limb motor tests or right limb motor tests ranged from 1 to 8 rather than from 1 to 4 as was the case with hemiparetic outpatient subjects. Data collection protocol was identical to the one described for outpatient subjects. With inpatient subjects included, the dataset was comprised of 56 subjects total – 25 hemiparetic subjects and 31 non-hemiparetic controls.

We re-ran the 6 classification experiments detailed in Section 2.3.4 with the larger dataset of outpatient and inpatient subjects. Note that inpatient data collection is ongoing. We do not yet have a complete set of subjects for human raters to evaluate via video as in the outpatient case, so we do not have a human performance benchmark for inpatient subjects.

Performance of SVM with Inpatient Subjects Included

Table 2.2 summarizes the results of the linear SVM when we incorporated inpatient subjects with more severe hemiparesis. With inpatient subjects included, the max normalization strategy yielded marginally better performance than the L2 normalization strategy – 76.79% accuracy for the former as compared to 73.21% for the latter. These metrics were for the leave-one-out experiment, when the training set size was maximized. The best performance was obtained when we split the dataset into 80% training and 20% testing – 79.58% accuracy (averaged across 20 classifiers). Once again, SVM performance generally increased with the duration of video used as input (the number of samples used

Table 2.2: SVM Performance on Outpatient and Inpatient Subjects – 10 Seconds of Video

				Linear SVM, Balanced Class Weight							
				L2 Norm				Max Norm			
Experiment	**Num Train**	**% Train**	**Num Test**	**Test Accuracy**	**Weighted F1**	**Weighted Precision**	**Weighted Recall**	**Test Accuracy**	**Weighted F1**	**Weighted Precision**	**Weighted Recall**
Leave-One-Out	55	98%	1	73.21%	73.15%	75.43%	73.21%	76.79%	76.73%	76.73%	76.79%
90/10 Split	50	89%	6	64.17%	61.93%	66.79%	64.17%	73.33%	71.61%	73.13%	73.33%
80/20 Split	44	79%	12	70.83%	70.33%	74.50%	70.83%	79.58%	78.97%	82.20%	79.58%
75/25 Split	42	75%	14	66.07%	65.61%	70.47%	66.07%	72.14%	71.74%	73.26%	72.14%
66/33 Split	37	66%	19	65.00%	64.60%	68.41%	65.00%	72.37%	71.90%	74.01%	72.37%
50/50 Split	28	50%	28	63.75%	62.88%	66.08%	63.75%	68.39%	67.98%	69.46%	68.39%

to compute a covariance matrix per subject). Performance with 9 seconds of video was slightly better than with 10 seconds, ∼80% versus ∼77% (Figure 2.4).

2.4 Contributions to the Field

A computational system designed to aid acute stroke diagnosis needs to capture the expertise of neurologists and translate it outside of the stroke center and into (underserved) emergency departments. As the first part of such a system, we developed a machine learning pipeline that used 3D body posture at rest (described by the covariance matrix of body joints) to detect hemiparesis or weakness with high accuracy, exceeding that of stroke specialists conducting a video-based hemiparesis assessment. A neurologist is required to conduct the NIHSS exam reliably. Given that we obtained high performance with sitting subjects without requiring them to perform any particular exercise, we believe that out approach is primed to be taken outside of stroke centers and into places where neurologists are not readily available. Our work here paves the way for a more comprehensive clinical decision support system for stroke diagnosis that considers other acute symptoms as well, like facial droop and speech difficulty. Ultimately, we hope that this work reduces misdiagnosis of stroke in the emergency department and overall improves stroke patient outcomes.

2.4.1 Making the Most of Small Datasets with Leave-One-Out

In this work, we collected a novel dataset of 35 outpatient stroke subjects across the stroke severity spectrum. In Section 2.2.3, we used a leave-one-out cross validation approach to train and test an SVM model, along with a grid search of parameters to find the optimal classifier. The benefit of leave-one-out was that it enabled us to use the most amount of data possible for training – 34 subjects out of 35. Leave-one-out cross validation

is a low bias technique, and yielded perfect accuracy in differentiating between left-side hemiparesis, right-side hemiparesis, and no hemiparesis from the slope of body angles (as reported in Section 2.2.3) [36]. The leave-one-out technique has certain drawbacks however. Leave-one-out results are often over optimistic especially on small datasets [38,40]. Typically, overfitting refers to a hypothesis or model (with several parameters) that is too complex and perfectly fits the training data generalizing poorly on test data. With leave-one-out coupled with an extensive grid search, overfitting comes from having tested too many hypotheses [36, 37, 106]. For these reasons, when we developed a video-based approach using covariance features as input, we more rigorously evaluated the classifier by splitting the dataset into training and test. The high performance obtained with a linear SVM trained on just 50% of the outpatient data – 72% accuracy, greater than the 68% benchmark from 8 domain experts – was a strong indicator that the model was not overfitting and could generalize well (see Section 2.3.4).

Note that we also reported results from leave-one-out as it let us identify problematic subjects by analyzing SVM performance on a per-subject basis. To better represent subjects with severe stroke in our dataset, we incorporated 21 inpatient subjects. The accuracy dropped from 80% to 77% when we trained and tested the SVM classifier on the augmented dataset with leave-one-out. This in part was due to errors introduced by the addition of inpatient subjects. Some inpatient subjects had such severe weakness (in conjunction with other severe stroke symptoms) that they were unable to sit upright and were recorded lying down, as can be seen in Figure 2.5. Other subjects were recorded sitting abnormally. We believe that the SVM was not able to classify these subjects correctly due to their aberrant postures at rest; the SVM was trained on a dataset that skewed in favor of outpatient subjects with less severe symptoms. The leave-one-out strategy revealed the type of patients missing in our dataset, allowing us to focus our future data collection efforts on those lying

down or sitting abnormally due to the high severity of their symptoms. This is required to translate our approach to the emergency department, where potential stroke patients are often brought in by ambulances on gurneys and so are lying down rather than sitting up right. We hope to improve the generalizability of the SVM model by accommodating these types of body postures.

The example of hemiparesis detection in stroke subjects with machine learning reveals the pros and cons of the leave-one-out training strategy. It enables maximal use of data for training and reveals problematic subjects and gaps in a dataset, but is only an estimate of generalizability and can still overfit. It is not replacement for a more rigorous evaluation strategy involving train/test splits, but did allow us to focus on recording subjects with greater weakness going forward.

2.4.2 Importance of Interpretability in Medicine

In Chapter 1, we discussed the fact that physicians prefer clinical decision support systems that enable interpretation of the reasons underlying predictions [53–55]. So we were conscious of explainability when developing our hemiparesis classification pipelines. For example, we used a SVM with linear kernel in order to be able to identify the features most useful for classification; identifying support vectors is difficult with nonlinear kernels that implicitly map features to high-dimensional spaces that are difficult to comprehend or visualize [105]. 20 classifiers were each trained on a random selection of 90% of outpatient and inpatient subjects and tested on a random 10%, as described in Section 2.3.4. We identified the 10 features weighted the highest by all 20 trained SVM classifiers.

Moreover, despite its simplicity as a feature descriptor, the covariance matrix of 3D body joints not only had high discriminatory potential yielding robust results but also could be easily visualized. 3D orientations of wrists and ankles relative to each other and

Figure 2.5: Hemiparetic subjects missed by the SVM (false negatives). On the left are severely debilitated subjects who were unable to sit upright or were sitting abnormally, possible reasons for the erroneous predictions. The SVM was trained primarily using videos of subjects sitting upright, like the two subjects on the right. Inferred body joints are displayed in red and black dots.

to other joints were the most discriminatory for hemiparesis classification. To visualize how these covariances were different between hemiparetic and non-hemiparetic subjects, we averaged the covariance matrices for 25 hemiparetic subjects and plotted the average as a heat map. We did the same for 31 non-hemiparetic subjects, and highlighted the top 10 features on the 2 heat maps. See Figure 2.4.2.

Several of the top 10 most discriminatory features for classification were the variances of left ankle, left wrist, right ankle, and right wrist and the covariances of these joints with

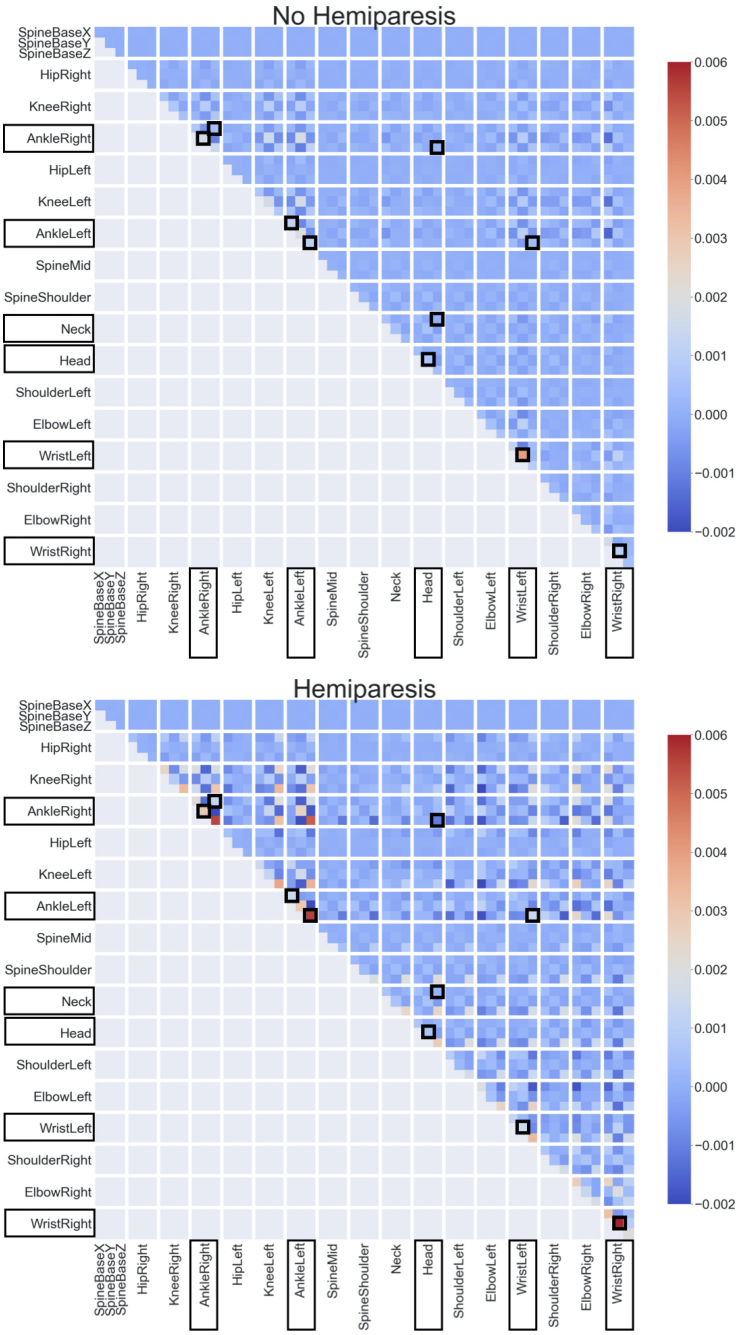


Figure 2.6: Heat map of average covariance matrix for outpatient and inpatient subjects without hemiparesis (*top*) and with hemiparesis (*bottom*). Top 10 most discriminatory features for classification are highlighted in black.

respect to each other. This result emphasized that weakness primarily impacts the arms and legs; the NIHSS motor exams are used to identify weakness for this reason. Moreover, limb positions even at rest are seemingly related to hemiparesis. We therefore believe that as long as we are able to record the 3D positions of the limb joints well, the SVM classifier approach should generalize well.

The variance of y (which represents up/down height) of the left wrist was highly positively in control subjects but 0 in hemiparetic subjects. That is, subjects without hemiparesis had a larger range of values for the heights of the left wrist, whereas hemiparetic subjects generally kept their left wrist positioned the same (vertically). The opposite was true for variance of y of the right wrist. Such asymmetries in the heat maps of the two classes were related to the side of hemiparesis with which a patient was affected. Hemiparesis is the partial paralysis of one side of the body or the other, and we did not consider which side was weak in our classification experiments. We chose to focus on binary classification instead of three-class because in practice, the side of hemiparesis is not relevant to the decision to administer stroke treatment.

Interpreting the linear SVM and the covariance feature descriptor enabled us to identify whether the most discriminatory features made intuitive sense, allowing us to validate whether the machine learning pipeline followed a clear logic – a sanity check to confirm that the SVM is not simply learning a one-off rule unique to the datasets that were used to train it.

2.4.3 Benchmarking Performance Against Domain Experts

A simple, off-the-shelf SVM classifier with linear kernel was able to identify hemiparetic and non-hemiparetic control subjects with 80.0% accuracy when trained with leave-one-out, using outpatient subjects only. Initially, we did not know whether this

accuracy was impressive as we were missing a benchmark against which to compare SVM performance. We therefore recruited 8 stroke specialists who conducted a video-based assessment of hemiparesis. When these video raters assessed videos of subjects moving their arms and legs as part of the NIHSS exam, average accuracy for hemiparesis detection was 68%. When they assessed videos of subjects sitting at rest, average accuracy dropped to 61%. This would make sense as the former is typically used by stroke specialists in practice to identify hemiparesis. Identifying weakness from subjects at rest is not the standard and requires years of experience with paralyzed patients and pattern recognition. The clinician raters also had the benefit of viewing more than 10 seconds of video of sitting subjects, when available. The input into the SVM was truncated at 10 seconds for all subjects to avoid excluding subjects with short wait times from the dataset. Regardless, the SVM clearly exceeded the performance of domain experts.

Note that the performance results reported in Table 2.1 and Table 2.2 were averaged across 20 classifiers, for all experiments except leave-one-out. Each classifier was tested on a different, random selection of subjects. So on average, the SVM performed well regardless of the subjects chosen to test it. This would indicate that the linear SVM was not overfitting on the subjects used for training, and highlights the consistency of results obtained with the covariance feature descriptor. By comparison, human raters were not consistent. 8 raters assessing videos of the NIHSS motor exam and of subjects sitting at rest had inter-rater reliability less than the cutoff for good performance (see Section 2.3.3). Performance therefore varied depending on the stroke specialist conducting the assessment, whereas the SVM classifier consistently did well.

2.5 Acknowledgements

This chapter is largely a reprint of "Stroke-Associated Hemiparesis Detection Using Body Joints and Support Vector Machines" published in the *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare* by authors Vishwajith Ramesh, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. It also covers as of yet unpublished work (in review) "Robust pose-based identification of weakness in sitting stroke patients with an interpretable machine learning classifier" by Vishwajith Ramesh, Lisa M. Grega, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. This work was supported by the National Science Foundation Graduate Research Fellowships Program (DGE-1650112) and the UCSD Chancellor's Research Excellence Scholarship (formerly the Frontiers of Innovation Scholars Program).

Chapter 3

Cough-Based Respiratory Disease Diagnosis

3.1 Motivation and Background

Respiratory diseases, characterized primarily by difficulty breathing, are some of the most prevalent illnesses in the world. In the United States alone, chronic obstructive pulmonary disease (COPD) is the third leading cause of death [107]. Asthma affects more than 25 million people in the United States, of which 6 million are children [108].

The human diagnosis of respiratory diseases is challenging and accurately assessing airway sounds depends on extensive clinical training and experience [109, 110]. Accurate testing not only requires several assessment modalities such as auscultatory examinations, spirometry, fluid analysis, and medical imaging but also trained clinicians to obtain and assess results. The difficulty of accurate testing is highlighted by the poor inter-rater reliability of clinicians' diagnoses, particularly those made through radiographic interpretation and auscultation [111–114].

Errors in asthma or COPD diagnosis (particularly under diagnosis) affect treatment

and even mortality [114]. Moreover, COPD is progressive and presents with slow-to-develop symptoms that are easily dismissed by individuals but gradually get debilitating; not catching COPD early will delay vital treatment and increase morbidity. Automatic diagnosis of respiratory diseases, particularly those that use machine learning techniques and work in the wild, is a growing field that shows much promise in addressing the challenges of human diagnosis. Unfortunately, these efforts are hindered by the small, imbalanced datasets prevalent in healthcare applications.

To address these shortcomings, we drew inspiration from the high correlation of coughs with respiratory diseases, the ubiquity of smartphones for passively capturing coughs, and data augmentation using unsupervised generative models (Figure 3.1). In this chapter, we discuss our efforts to develop random forest (RF) and support vector machine (SVM) classification pipelines that used coughs recorded using a smartphone microphone to distinguish between adults that were healthy and those with major respiratory diseases, namely asthma, COPD, and chronic cough. We developed and trained a generative adversarial network (GAN) per disease class in order to synthesize raw audio samples of coughs – "CoughGAN." We discuss how balancing and augmenting a small, real-world clinical dataset with synthetic samples can boost the performance of the classifiers and reduce overfitting.

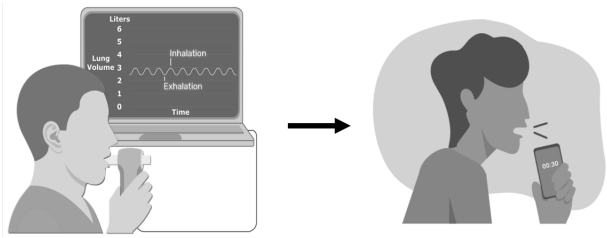


Figure 3.1: Specialized medical equipment like spirometers can be replaced with smartphones for automatic, cough-based respiratory disease diagnosis.

3.1.1 Automatic Diagnosis of Respiratory Diseases

Badnjevic et al. developed an expert diagnostic system to automatically identify asthma and COPD in clinical settings [115]. The system used a symptom-based questionnaire to screen out healthy patients and a classifier for distinguishing between COPD and asthma. The results of the questionnaire and a pulmonary function test (vital capacity, forced vital capacity, forced expiratory volume in the first second, etc.) were used to train a one-layer artificial neural network (ANN). The authors were able to classify COPD and asthma with over 95% accuracy [115]. Despite the decent performance, the primary limitation of this work is that it was designed to be deployed in clinical settings. The features used were the results of manually filled questionnaires and tests conducted by a trained clinician. So this work has some of the same aforementioned limitations as human diagnosis.

Rather than using the results of medical tests, several studies have proven that coughs can reliably discriminate between respiratory diseases such as asthma, COPD, and pertussis [116, 117].

Amrulloh et al. showed that cough sounds could be used to differentiate between asthma and pneumonia in children using ANNs [118]. The benefit of using voluntary coughs is that they do not require any medical testing, which can especially be challenging to perform on children. The authors recorded coughs using a microphone and computed features commonly used in audio signal processing – Mel-frequency cepstral coefficients (MFCCs), entropy, zero crossing rate (ZCR), etc. With a small subject pool of eighteen and the leave-one-out training paradigm, the authors obtained 94% accuracy, 100% specificity, and 89% sensitivity [118].

Another study by Porter et al. expanded on this work and used cough sounds to differentiate between asthma, pneumonia, lower respiratory tract disease, croup, and

bronchiolitis in children [119]. The authors were able to show a high level of agreement between the results of their automated approach and the diagnoses of four pediatric clinicians [119]. These positive results lend credence to the authors' idea that during a cough event, there is an unimpeded column of air between the lungs and atmosphere. The authors hypothesized that this column propagates sounds generated inside the lungs and that pathophysiological changes caused by different respiratory conditions modulate the sound quality. This enabled the use of the signal processing techniques common to speech recognition to analyze the cough sounds [119, 120]. Cough features are overall far simpler to obtain as they do not require any medical tests to be performed, nor do they require any specific equipment.

The ubiquity of smartphones and smart wearables has opened the door for health monitoring through mobile sensing, especially for pulmonary patients [121, 122]. Because of the rich acoustic data that can be captured via smartphone microphones, studies have used signal processing and deep learning techniques to reliably perform smartphone-based spirometry to quantitatively assess lung inhalation and exhalation [123, 124]. The benefit of such tests is that they do not require experienced medical professionals nor expensive equipment to conduct, allowing for an inexpensive alternative for respiratory disease assessments. Such technology makes the continuous identification of cough events and cough data capture far simpler and more feasible in the wild than in clinics [125, 126].

3.1.2 Limitations of Respiratory Disease Datasets

The primary reason for the use of the k-fold or leave-one-out training paradigm in the aforementioned studies was the relatively small size of the healthcare datasets collected by the authors; by comparison, datasets like MNIST or CIFAR that are commonly used in machine and deep learning applications are orders of magnitude larger [115, 118, 119].

Leave-one-out enables maximal use of the little data that is available for training and has low bias. Unfortunately, k-fold/leave-one-out has high variance that may lead to unreliable estimates especially for classifiers with a large number of classes. So the good k-fold/leave-one-out performance obtained in prior work with several respiratory disease classes may not necessarily translate to good performance on unseen data [127, 128].

Furthermore, particularly for imbalanced data, when an instance of a minority label is removed as part of leave-one-out, the model will most likely pick (one of) the majority class(es) [127]. Unfortunately, healthcare datasets tend to be imbalanced, as either there are far more healthy subjects than sick or data collection and subject recruitment in clinical settings is biased towards sick subjects because healthy individuals do not often go to a hospital. The inconsistencies associated with k-fold/leave-one-out motivate the use of typical train-test dataset splits. The standard, more rigorous training paradigm that uses a training set, an optional validation set, and a separate test set that the classifier has never seen is a better indicator of the generalizability of a trained model. But the issues with the small size of healthcare datasets and their class imbalance still need to be addressed.

A data augmentation strategy would solve both concerns, enabling the use of a more rigorous evaluation method that uses a test set unseen by the model. For audio, synthetic examples can be generated by adding noise, changing pitch, or changing speed. However, for classifying respiratory diseases, it is not yet clear what aspects of audio signals would be relevant for diagnosis; modifying audio randomly may lead to the corruption of features that are required to distinguish between healthy and different respiratory conditions.

Therefore, in this work, we used generative adversarial networks (GANs) as a way to cleverly engineer synthetic examples similar to real examples. Because the technique is unsupervised, synthetic examples that preserved class label information could be created without having to specify individual transformations. Their use in healthcare applications

such as those involving medical images and electrocardiograms show that GANs can be used to augment datasets without sacrificing diagnostic ability; in fact, adding synthetic examples created by GANs for training have been shown to boost test performance [129–133]. We demonstrated that the same held true for respiratory disease classification.

3.2 A Respiratory Disease Classification Pipeline

3.2.1 Building a Dataset of Coughs

In collaboration with a pulmonary clinic at the Brigham and Women's Hospital, we recruited 45 subjects – 5 were healthy, 19 were diagnosed with asthma, 17 with COPD, and 4 with chronic cough (a cough that persisted for more than 8 weeks). Each subject was asked to voluntarily cough as many times for two minutes. The audio was recorded with sampling frequency 44100 Hz using a Samsung Galaxy Note 8 smartphone that was held 3 feet in front of the subjects sitting on a table. Study procedures were approved by an Institutional Review Board (IRB) and the subjects gave informed consent for their recordings.

For each subject, three annotators manually annotated cough events by going through the recorded audio and identifying and saving each single cough event – defined as a rapid, sharp expulsion of air from the lungs sometimes directly preceded by inhalation; the duration of a single cough was typically 0.5 seconds. Note that because the coughs were voluntarily produced, subjects had a varying number of cough examples each. In total, there were 2832 annotated cough examples: 237 healthy, 1106 asthma, 1231 COPD, and 258 chronic cough.

The dataset collected varied in the number of subjects with a particular condition – there were only 5 healthy subjects but 19 with asthma, for example – and in the number of

cough examples per subject – from 11 to 168 coughs. We addressed the former imbalance by adjusting the class weights of the costs of the classifier models (described in Section 3.2.4) and the latter imbalance in this dataset preparation step. To discourage models from simply predicting the class with the larger number of coughs, we only used 11 coughs (the least number of coughs produced) from each subject to train the classifiers. So for the healthy vs. asthma experiment, there were a total of $11 * 19 = 209$ (44 healthy, 165 asthmatic) examples in the training set and $5 * 11 = 55$ (11 healthy, 44 asthmatic) in the test.

3.2.2 Creating Synthetic Coughs with GANs to Augment Dataset

We leveraged a generative model for 1 second audio synthesis – WaveGAN – to capture cough audio patterns and to generate synthetic coughs [134]. Because the coughs collected were roughly 0.5 seconds in duration, we modified the WaveGAN's generator to create audio samples that were 8192 data points in length; this corresponded to 0.5 seconds at a sampling frequency of 16384 Hz. We modified the discriminator network to accept 0.5 second coughs as input as well. We trained the GAN with randomly selected batches of *batch_size*. Figure 3.2 summarizes this architecture, detailed below. We implemented CoughGAN in TensorFlow [135].

Generator Architecture

The generator accepted noise sampled from the uniform distribution, with input shape $batch\_size \times 100$. The first layer of the generator was a fully connected dense layer with rectified linear unit (ReLu) activation and 32768 neurons. The output of this layer was reshaped to $batch\_size \times 16 \times 2048$. We then used four one-dimensional transposed convolution layers to upsample the noise. All four layers used ReLU activation and one-dimensional filters of length 25 with stride of 4 (with TensorFlow's "SAME" padding

algorithm). The first of the four layers had 1024 filters, the second 512 filters, the third 256 filters, and the fourth 128 filters. We added a last one-dimensional transposed convolution layer with tanh activation and 1 filter of length 25 and stride 2. This layer reshaped the output of the previous layer, $batch\_size \times 4096 \times 128$, to $batch\_size \times 8192 \times 1$ (for 1 audio channel).

Discriminator Architecture

The discriminator consisted of five one-dimensional convolution layers with leaky ReLU activation and filters of length 25 (with TensorFlow's "SAME" padding algorithm). The first layer had 64 filters with stride 2, reshaping the input $batch\_size \times 8192 \times 1$ to $batch\_size \times 4096 \times 64$. The remaining four layers used 128, 256, 512, and 1024 filters with stride 4. As outlined in the WaveGAN paper, we randomly shuffled the phase of the output of each of these layers (by -2 to 2 radians). This was done to discourage the discriminator from trivially learning to reject synthetic examples by the artifact frequencies caused by transposed convolutions in the generator [134]. The output of the fifth convolution layer was flattened and connected to a single logit using a dense layer. This logit described whether an input cough was real or synthetic.

CoughGAN Training

For training, we used the Wasserstein distance with the gradient penalty strategy proposed by Gulrajani et al. to encourage GAN loss convergence [136]. To minimize the Wasserstein distance, we used the Adam optimizer with a learning rate of 0.0001. We also used a $batch\_size$ of 64 samples. We trained a separate CoughGAN per label (healthy, asthma, COPD, chronic cough) and used the raw audio samples of the corresponding label when training, after resampling (at 16384 Hz) and normalizing the decoded audio waveforms.

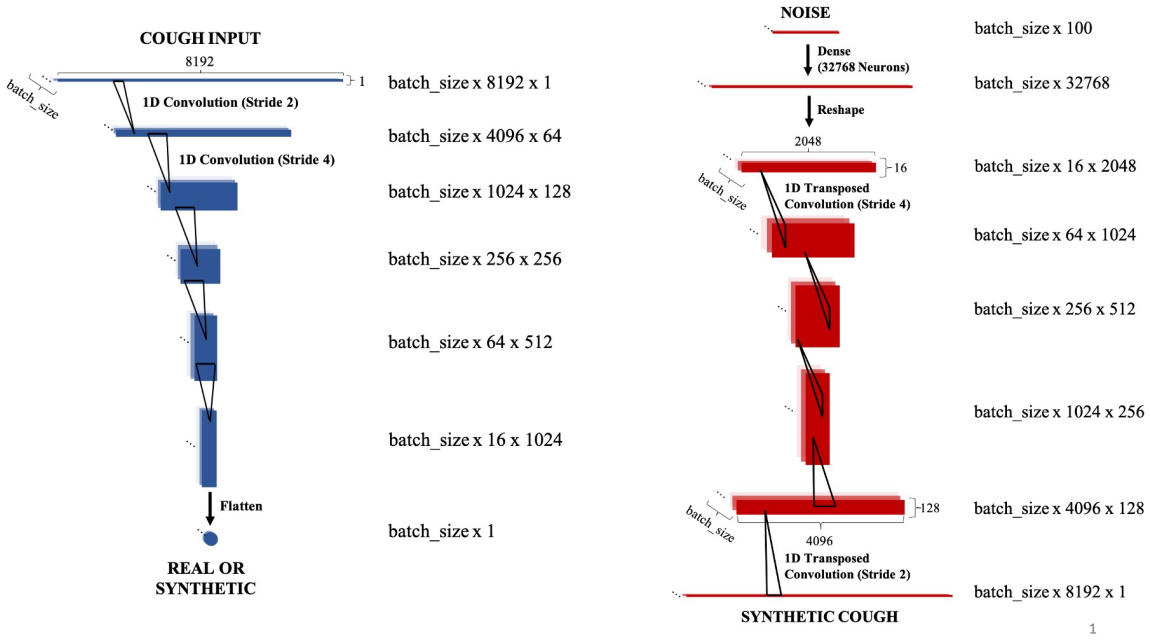


Figure 3.2: CoughGAN discriminator (in blue) and generator (in red) architectures. The discriminator takes cough examples as input and essentially determines whether they are real or synthetic. The generator transforms noise into synthetic cough examples.

During training, we calculated the short-time frequency transform (STFT) of the real batch used as input and of the synthetic batch generated at that training step (with frame size of about 30 milliseconds and frame step of about 15 milliseconds). We plotted the least-squares distance between the two (training step on the horizontal axis). We trained each GAN until this difference between real and synthetic examples converged to a minimum value. This comparison was done per class label.

Evaluating Synthetic Cough Quality

After the CoughGANs were trained, we used the trained generators to create 1000 synthetic samples per label. Each sample was passed through a low-pass filter and encoded. We used a low-pass filter at 5000 Hz to remove high frequency noise in the generated audio,

an occasional issue when using WaveGAN. In order to ensure that the generator was not simply producing the same synthetic cough samples repeatedly, we listened to each of the 1000 synthetic coughs per label as a check of diversity.

For each of the four CoughGANs trained, the plot of the least-squares distance between the STFT of the real and synthetic batches converged to a minimum, steady value (Figure 3.3). Note that the least-squares distance metric is a quantitative measure of the similarity between real and synthetic coughs. We used this metric as an indicator for when to stop CoughGAN training and the fact that the least-squares distance curves converge to a low value around zero at the end of training shows that synthetic coughs are similar to real coughs. (A value of zero would mean that the synthetic coughs are exactly the same as real coughs, akin to oversampling.) In this way, the CoughGANs were successfully trained to mimic real coughs. This similarity is confirmed by Figure 3.4, which shows the average spectrogram generated from randomly selected real and synthetic coughs for the four classes. Figure 3.4 shows that differences between the real spectrograms of the four types of coughs were maintained by the generator when creating synthetic coughs. For example, the spectrogram for real asthma coughs has higher frequencies than the spectrogram for real healthy coughs. This difference can also be seen between the spectrograms for synthetic asthma and synthetic healthy coughs. We thus preserved class label information when creating synthetic coughs.

Furthermore, listening to the synthetic audio samples revealed that the trained CoughGANs successfully generated realistic sounding coughs. Listening to the audio also confirmed that the generator was not trained to simply output a single cough waveform. A diverse range of coughs were generated for the four labels. The more quantitative least-squares distance assessment, coupled with listening to the coughs and looking at the spectrograms, showed that synthetic coughs were similar to real ones.

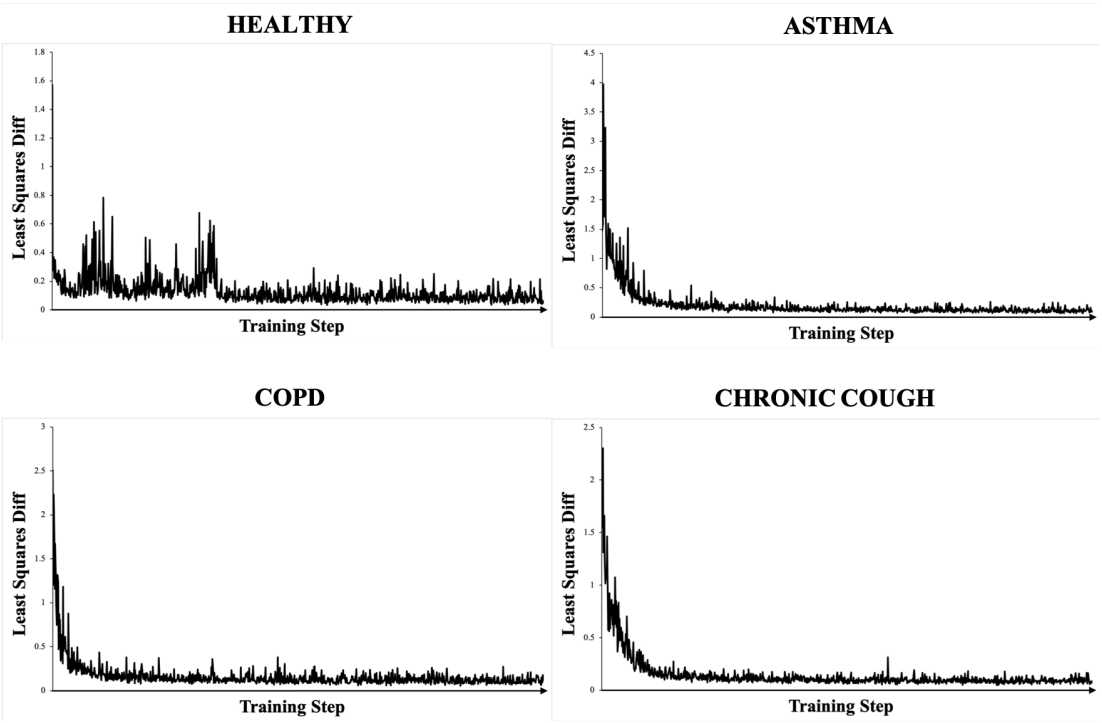


Figure 3.3: Least-squares difference between the short-time frequency transform of a batch (of size 64) of synthetic examples and a batch of real examples, during training. All four curves converged to a minimum, steady value by the end of training.

3.2.3 Extracting Audio Features from Coughs

With the pyAudioAnalysis library, we extracted short-term feature sequences for each cough, be it real or synthetic [137]. We used a frame size of 50 milliseconds and a frame step of 40 milliseconds (20% overlap). We computed, for each time frame, features used in the aforementioned studies and also used commonly in speech processing tasks: ZCR, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, and 13 MFCCs. Since there were slight differences in how the three annotators segmented coughs, a cough event could occur at the start, middle, or end of the audio, depending on the annotator. We therefore averaged each of the 21 features over the time frames corresponding to a cough example.

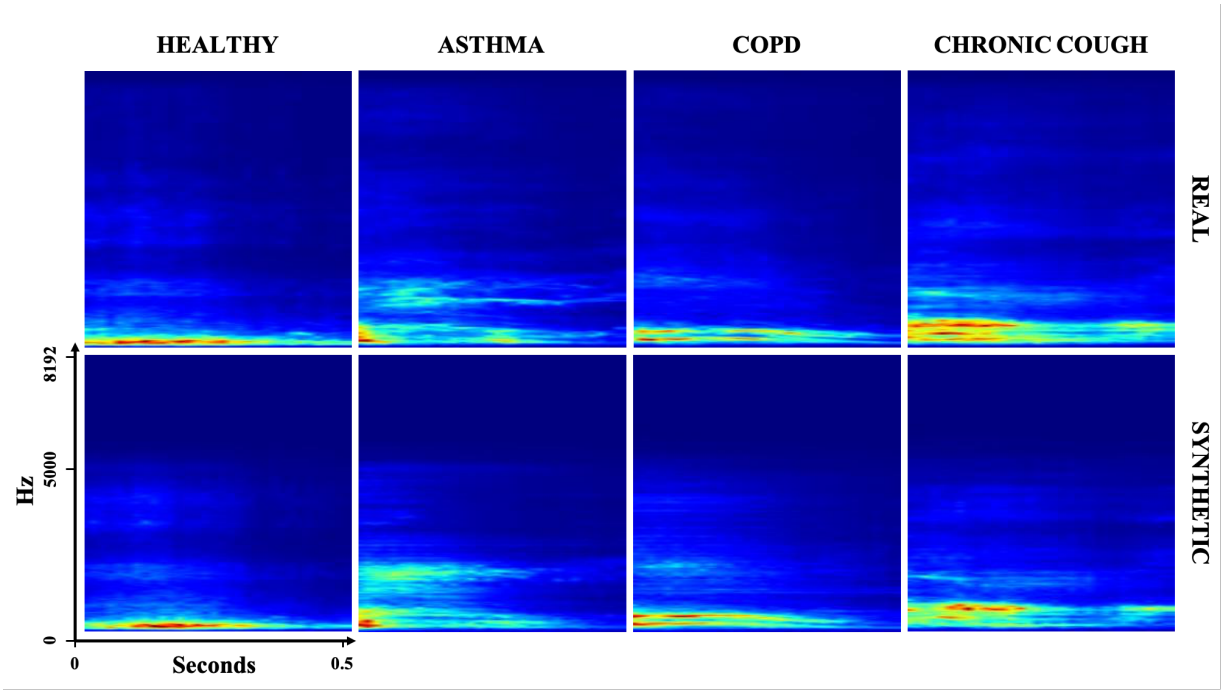


Figure 3.4: Spectrograms of real and synthetic coughs, for healthy, asthma, COPD, and chronic cough conditions. The vertical axis is the frequency ranging from 0 to 8192 Hz. The horizontal axis is time from 0 to 0.5 seconds. Each spectrogram was computed by taking the average of the spectrograms of 64 randomly chosen audio samples with the corresponding label.

3.2.4 Effects of Dataset Augmentation on Classification

We used two common supervised-learning models, a SVM and RF, for cough-based classification and to evaluate the effectiveness of adding synthetic examples. The SVM and RF models were implemented in Scikit-learn [91]. We ran four binary, medically-relevant classification experiments: healthy vs. asthma, healthy vs. COPD, healthy vs. chronic cough, and asthma vs. COPD.

For each experiment, to ensure that there were subjects from each class in both the training and testing of the classifier models, we randomly picked approximately 20% of the subjects for each of the two classes for a test set, with the remaining 80% of each class for training. For example, for the healthy vs. asthma experiment, we selected 1 out of 5

healthy subjects and 4 out of 19 asthmatic subjects, giving a test set of 5 subjects and training set of 19. We repeated this selection process 5 times and ran the analyses for each of the 5 random selections for the training and test sets. Essentially, the model was trained, tuned, and evaluated 5 times independently. For each of these runs, we computed the test accuracy, F1 score, true positive rate (TPR) or sensitivity, false positive rate (FPR), false negative rate (FNR), and true negative rate (TNR) or specificity. The healthy class was the "negative" class for the first three experiments, while the asthma class was the "negative" class for the asthma vs. COPD experiment. We averaged these performance metrics over the 5 runs. This was done to account for any variation in performance that might occur due to the specific subjects randomly chosen for each run.

Note that we intentionally limited the work presented here to binary classification results. Multi-class classification presents unique challenges that, coupled with the greater complexity of deep learning models like convolutional neural networks, would have made it difficult to isolate and analyze the performance improvements obtained via GAN-based data augmentation.

Tuning Models

We were conscious of overfitting given the small size of our dataset. So we chose features and tuned hyperparameters to maximize the performance we could obtain from each baseline model – trained without synthetic examples – for the four binary classification experiments.

Motivated by the feature selection used in prior work, we used average ZCR, spectral centroid, and 12 MFCCs as features for the SVM [119]. To maximize RF performance and to reduce overfitting, we did not use all 21 features computed in the feature extraction step. Instead, we ran a principal component analysis (PCA) dimensionality reduction using 12

components, which together explained >91% of the variance in the feature space (of the training set); each example therefore had 12 transformed features rather than 21 or 14, as was the case for the SVM. The features that account for most of the variance in the data were the first 11 MFCCs and spectral spread or spectral entropy (depending on the particular run and randomly selected training set). We obtained the best performance of the baseline RF by using these 12 features, as opposed to the 14 used for the SVM or all 21 extracted features.

We used balanced class weights to address the imbalance in the number of subjects with a particular condition. The cost parameter for each of the two classes in a classification experiment was weighed by the inverse of the class frequency in the input data; misclassifications of examples belonging to the smaller class were penalized more than misclassifications of examples belonging to the larger class. Balanced class weights discouraged the SVM and RF from simply predicting the dominant class for every example in the test set. Furthermore, a grid search revealed that a linear SVM outperformed SVMs with polynomial, radial basis function, or sigmoid kernels for the four binary classification experiments. In the grid search for the RF, we tested 5 to 100 trees and maximum depths of 4 to 12 for each tree. We found that 5 estimators and a depth of 4 gave the best performance – at least 60% test accuracy, above random chance prediction for binary classification. The baseline RF was prone to overfitting – high train accuracy but low test accuracy – with larger numbers of estimators and greater depths of trees.

Augmentation Strategies

To determine if adding synthetic examples to the training set would yield improvements in classifier performance, we tested two augmentation strategies:

1. *Balanced*: We randomly added synthetic examples until the two classes in the

training set for an experiment were balanced. For instance, for the healthy vs. asthma experiment, there were 165 asthma examples and 44 healthy examples in the training set. We randomly selected 121 synthetic healthy examples out of the 1000 generated. We then added them to the training set to balance the two classes at 165 examples (so 330 examples total in the augmented training set).

2. *Balanced and Doubled*: We randomly added synthetic examples until the two classes were not only balanced but also doubled in the training set. For the healthy vs. asthma experiment, after balancing the two classes by adding 121 randomly chosen synthetic healthy examples, we added 165 synthetic asthma examples and an additional 165 synthetic healthy examples. So the augmented balanced and doubled training set for this experiment consisted of 330 asthma and 330 healthy examples (660 total).

In general, when compared to the baseline SVM and RF models trained without synthetic examples, the "Balanced" augmentation strategy had better test accuracy, F1 score, TPR, and TNR (and smaller FPR and FNR). The "Balanced and Doubled" strategy yielded even larger improvements. While we saw this pattern for all four experiments, we show only the results for the healthy vs. asthma experiment in Table 3.1. (We did not see any significant improvements in test performance past doubling the training set size, so we do not report on those augmentation strategies – "Balanced and Tripled," for example – in this chapter.)

Table 3.1 shows the results for the healthy vs. asthma experiment. A SVM model (with linear kernel and balanced class weight) had 43% test accuracy and 59% F1. The baseline SVM model had a very low number of true negatives (the negative class being the healthy one) and a high number of false positives; the SVM trivially predicted the larger class for many of the test examples, despite our efforts to discourage this behavior by balancing class weights. A baseline RF (5 trees each with depth 4 and balanced class

Table 3.1: SVM and RF Results for HEALTHY VS. ASTHMA Experiment

HEALTHY VS. ASTHMA								
Model	**Test Accuracy**	**F1**	**TPR**	**FPR**	**FNR**	**TNR**	**Train Accuracy**	**Train - Test**
WITHOUT SYNTHETIC								
SVM	43%	59%	52.3%	94.5%	47.7%	5.5%	75%	32%
RF	67%	79%	76.8%	74.5%	23.2%	25.5%	92%	25%
WITH SYNTHETIC – BALANCED TRAINING SET								
SVM	63%	74%	69.1%	63.6%	30.9%	36.4%	87%	24%
RF	71%	82%	80.0%	63.6%	20.0%	36.4%	96%	24%
WITH SYNTHETIC – BALANCED AND DOUBLED TRAINING SET								
SVM	74%	83%	81.8%	58.2%	18.2%	41.8%	90%	16%
RF	73%	83%	81.4%	61.8%	18.6%	38.2%	95%	22%

weight) had better TNR than the SVM, as well as better test accuracy (67%) and F1 (79%).

Adding synthetic examples and balancing the training set increased the SVM test accuracy by 20% and the F1 by 15%. The RF test accuracy improved to 71% and the F1 to 82%. Balancing and doubling the training set using synthetic examples further increased the SVM test accuracy by 30% and F1 by 24%, while the RF test accuracy only increased by 6% and F1 by 4%. Compared to the baseline, both augmentation methods increased TPR and TNR and reduced FPR and FNR. The train minus test accuracy, an indicator for overfitting, decreased for both SVM and RF too.

Table 3.2 shows the performance metrics for "Balanced and Doubled" augmentation for healthy vs. asthma, healthy vs. COPD, healthy vs. chronic cough, and asthma vs. COPD experiments. The "Balanced and Doubled" augmentation strategy generally yielded the largest increases in test accuracy, F1, TPR, and TNR compared to the "Balanced" strategy. Table 3.2 shows whether these metrics were higher or lower than the metrics for the baseline models trained without any synthetic examples (in parentheses). For example, in Table 3.2, the "Train – Test" accuracy of the SVM for healthy vs. asthma is 16% using

Table 3.2: SVM and RF Performance After Balancing and Doubling Training Set

Classification	Model	Test Accuracy	F1	TPR	FPR	FNR	TNR	Train Accuracy	Train - Test
HEALTHY VS. ASTHMA	SVM	74% (+31%)	83% (+24%)	81.8% (+29.5%)	58.2% (-36.3%)	18.2% (-29.5%)	41.8% (+36.3%)	90% (+15%)	16% (-16%)
	RF	73% (+6%)	83% (+4%)	81.4% (+4.6%)	61.8% (-12.7%)	18.6% (-4.6%)	38.2% (+12.7%)	95% (+3%)	22% (-3%)
HEALTHY VS. COPD	SVM	74% (+14%)	83% (+12%)	82.5% (+15.8%)	60.0% (-8.3%)	17.5% (-15.8%)	40.0% (+8.3%)	92% (+13%)	18% (-2%)
	RF	69% (+6%)	80% (+4%)	78.3% (+3.8%)	66.7% (-13.3%)	21.7% (-3.8%)	33.3% (+13.3%)	93% (-2%)	24% (-8%)
HEALTHY VS. CHRONIC COUGH	SVM	76% (+15%)	74% (+19%)	70.0% (+21.7%)	18.3% (-8.3%)	30.0% (-21.7%)	81.7% (+8.3%)	86% (+3%)	10% (-12%)
	RF	63% (+22%)	64% (+21%)	68.3% (+23.3%)	41.7% (-20.0%)	31.7% (-23.3%)	58.3% (+20.0%)	96% (-2%)	32% (-23%)
COPD VS. ASTHMA	SVM	62% (-0%)	60% (+3%)	65.5% (+9.5%)	42.3% (+10.5%)	34.5% (-9.5%)	57.7% (-10.5%)	73% (+5%)	11% (+5%)
	RF	57% (+8%)	55% (+7%)	54.5% (+6.4%)	40.0% (-9.1%)	45.5% (-6.4%)	60.0% (+9.1%)	86% (+1%)	29% (-7%)

the balanced and doubled dataset. It is 32% using the un-augmented dataset (Table 3.1, under "WITHOUT SYNTHETIC"). Balancing and doubling the dataset reduced the "Train – Test" accuracy by 16%.

Across all four experiments, balancing and doubling the training set increased the test accuracy, F1, TPR, and TNR. We saw corresponding decreases in the FPR and FNR except in the case of the SVM in the COPD vs. asthma experiment, when the FPR actually went up by 10%. Moreover, the train minus test metric decreases for all experiments for both models, again except for the SVM in the COPD vs. asthma experiment (up by 5%).

For the healthy vs. asthma, COPD, or chronic cough experiments, the test accuracy is greater than 70%. But for the COPD vs. asthma experiment, the test accuracy is around 60% and adding synthetic examples did not improve the performance by much (not at all for the SVM test accuracy).

Generally, augmenting the dataset best helped the performance of the SVM, decreasing the FPR and increasing the TNR. The baseline RF performed better than the

baseline SVM and therefore had less room to improve with augmentation. But even if the increase in test accuracy or F1 was marginal for the RF, the TNR increase was substantial. In the healthy vs. asthma experiment, the RF test accuracy increased by only 6% (an increase of 9%) but the TNR went up (and the FPR down) by 13%.

3.3 Contributions to the Field

Given the challenges of collecting data in the wild, obtaining good diagnostic performance with a small number of cough examples while using data augmentation to boost performance is important. We developed a classification pipeline to distinguish between healthy and three major respiratory diseases based on cough samples acquired using a smartphone. We demonstrated that augmenting the dataset with synthetic cough samples boosted classifier performance. We were able to tune a SVM classifier to distinguish between healthy and asthma/COPD with 74% test accuracy and 83% F1, between healthy and chronic cough with 76% test accuracy and 75% F1, and between COPD and asthma with 62% test accuracy and 60% F1. While these results were not as high as the >90% metrics reported in previous studies (Section 3.1.1), we were able to obtain decent, generalizable performance with a more rigorous training paradigm than k-fold/leave-one-out cross-validation. Our classification pipeline also used coughs recorded with a smartphone and did not require a clinician or specialized medical equipment, making diagnosis easier. When combined with the latest in automatic cough identification, our work may contribute to the development of a smartphone-based early warning system for catching respiratory disease quickly and accurately.

3.3.1 Augmenting Small and Imbalanced Datasets with GANs

Augmenting the dataset had overall positive effects on the performance. Balancing the dataset by adding synthetic examples to the class with the smaller size increased the number of true negative predictions; the added samples discouraged the models from always predicting the majority class label. Balancing the training set, rather than doubling it, boosted TNR the most (from 6% to 36% for the SVM); the TNR did not go up significantly by doubling the dataset (only from 36% to 42% for the SVM). Doubling the dataset provided the two models with additional examples during training, which generally reduced train minus test accuracy while simultaneously increasing test accuracy, an indication that the models trained on synthetic data do not overfit. Creating synthetic examples in an unsupervised way with GANs are a great way to fill in gaps in a small and imbalanced clinical dataset without sacrificing the discriminatory potential of features. As shown for respiratory disease classification, doing so boosts classifier performance and reduces overfitting.

3.4 Acknowledgements

This chapter is largely a reprint of "CoughGAN: Generating Synthetic Coughs that Improve Respiratory Disease Classification" published in the *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* by authors Vishwajith Ramesh, Korosh Vatanparvar, Ebrahim Nemati, Viswam Nathan, Md Mahbubur Rahman, and Jilong Kuang. This work was conducted as part of an internship at Samsung Research America.

Chapter 4

Gait Assessment in Parkinson's Disease

4.1 Motivation and Background

The standard for assessing motor symptoms associated with Parkinson's disease (PD) – tremor, postural instability, gait difficulty, and bradykinesia – is the motor examination section of the Unified Parkinson's Disease Rating Scale, or UPDRS Part III. This test is conducted in person by clinicians and has good inter-rater reliability [138]. However, Parkinson's disease is characterized by sporadic symptoms that are often not observed in clinic. For example, freezing of gait (FoG), which causes postural imbalance and frequent falling, has been shown to be difficult to elicit during visits to the clinic. More broadly, it is well established that the range and severity of symptoms experienced by a patient in their home environment do not always agree with measurements taken in clinic [139–141]. Discontinuous monitoring through clinical visits and in-person assessments does not capture PD symptoms and their progression completely. This is especially evidenced by the fact that patients' self-assessments of their improvement over time in response to treatment do not agree with their UPDRS scores from in-person appointments [142].

ON/OFF cycles are another defining characteristic of PD. A patient treated with a

dopamine precursor drug like levodopa experiences the ON state when the drug is active and motor symptoms are less severe. As the drug wears off and motor symptoms worsen, the patient transitions to the OFF state [143]. When a patient visits a neurologist at a clinic, they are either ON, OFF, or transitioning into one of the states. The neurologist can identify the patient's state during a visit using the symptom severity measured with the UPDRS exam. But without continuous monitoring, it is not possible to capture the relevant dynamics – time in each state, severity of the OFF state, frequency – of these ON/OFF cycles. Cycles are dependent on the progression of PD and are often unique to individual patients. Understanding the dynamics of these cycles is important because they inform clinicians of the effectiveness of their treatment regimen and the degree of rehabilitation of the patient in a personalized way [144–146].

The standard for tracking ON/OFF cycles more continuously than in-person clinical visits is the Hauser diary, a major endpoint in PD clinical trials [145, 147, 148]. Patients monitor the severity of their symptoms and report their ON/OFF state in a home diary every 30 minutes over the span of several days. PD patients are able to determine their own ON/OFF state well, by correctly perceiving their non-motor and especially their motor function [149]. Despite this, Hauser diaries are often inaccurate and unreliable [145,150,151]. Poor patient compliance, recall bias, and diary fatigue are common, well-established problems with both paper and electronic diaries [145, 150, 152]. Due to poor adherence by patients and the fact that diaries only measure the duration of time spent in a state and not the severity of impairment, Hauser diaries are a limited source of information about patients' physical functions at home.

Moreover, individual UPDRS tasks can benefit from a more objective quantification of symptoms. Gait itself is scored on a scale of 0 to 4 from "Normal" to "Severe". Examiners are asked to consider stride amplitude, stride speed, height of foot lift, heel strike during

walking, turning, and arm swing [153]. Assessing all of these independently and then combining them into a single score is a very subjective process and involves a good deal of intuition derived from the experience of the examiner.

Wearable sensors address all of the above concerns [150]. They have been shown to have high biomechanical resolution for quantitatively assessing gait impairment in PD, so they have a high degree of clinical applicability [139]. They especially benefit from being able to be taken out of clinic and worn continuously at home or "in-the-wild". In other words, as Parkinson's disease patients are simply going about their day, their gait metrics and motor symptoms can be continuously tracked and ON/OFF cycles automatically monitored. It is important to be able to measure Parkinson's disease outcomes objectively without bias, reliably, and in an unobtrusive way. In-the-wild tracking of the progression of PD in patients via wearables increases the relevance of any clinical visits and improves overall patient management by clinicians [139, 150].

4.1.1 Machine Learning for PD Symptom Assessment

There have been several successful attempts at tracking PD symptoms using wearable sensors and machine learning. Sama et al. were able to detect and score bradykinesia, the slowness of movement, using a waist-worn accelerometer and a support vector regression model [154]. By using machine learning to characterize the walks and strides of patients, they were able to detect and score bradykinesia in 12 subjects with high accuracy and correlation. However, the work used a small subject pool of only 12. And while they were able to achieve high accuracies, the authors used a leave-one-subject-out methodology in which each subject was used to not only train the model but also test it once [154]. The model therefore was trained and tested on signals from the same set. Without out-of-sample testing where the entire data from a subject is set aside, models that use few subjects are

prone to overfitting.

Another similar study by Rodriguez-Martin et al. used support vector machines (SVMs) to detect FoG events from a single waist worn sensor that can be worn at home, to address the lack of FoG events that occur during clinical visits [139–141]. The study's subject pool was larger at 21. Their model also used a leave-one-subject-out approach but did not perform as well as current standards for detecting FoG. And while their alternative, patient-specific model outperformed the standard, it was personalized in such a way that it involved training on 50% of a patient's data and testing on the other 50%. While more nuanced, this personalized model was simply a modified leave-one-subject-out approach that used data from 20 subjects and half of the sensor data from 1 test subject to train the algorithm, with the remaining half used to test the model [140]. The model was therefore trained and tested on the same (test) subject's data.

More recent work also used similar approaches. Studies by Rastegeri et al., Rovini et al., and Chomiak et al. tested several common machine learning algorithms for sensor-based gait analysis and diagnosis of PD, including SVMs, random forest, and naive Bayes [155–157]. All three studies used a cross-validation strategy to train their models (five-fold, ten-fold, and Monte Carlo cross-validation, respectively) and reported high performance (accuracy $>$95%). As in the studies by Sama et al. and Rodriguez-Martin et al., this training paradigm enabled the authors to make the most use of the small sample data that they had collected – 10 healthy controls and 10 PD subjects in Restegeri et al., 30 healthy and 30 PD in Rovini et al., and 9 controls and 21 PD in Chomiak et al. [140, 154–157]. Note however that while models trained via cross-validation have lower bias, they also tend to have high variance, giving unreliable estimates particularly for classification involving multiple classes [127, 128]. Furthermore, notably with leave-one-out, removing one or more examples corresponding to the minority label as part of training will encourage the model

to predict the majority class [127]. This is of high concern with healthcare datasets as they tend to be imbalanced; either there are more healthy than sick subjects or subject recruitment in clinics favors sick subjects since healthy individuals do not often go to a hospital. The accuracy values reported in the aforementioned studies are therefore merely upper bounds on real world performance. These inconsistencies motivate the use of a more rigorous training paradigm – one that uses a training set, an optional development set or a validation set for tuning hyperparameters, and a separate, independently collected test set that the classifier has never seen. This more rigorous paradigm is a better indicator of the generalizability of a trained model. However, the small size of healthcare datasets and their class imbalance remain problematic – there is not enough data to split into multiple sets for training, validating, and testing of models.

Despite their shortcomings, the aforementioned studies demonstrated the ease with which signals can be acquired using wearable sensors. And while not ideal, the leave-one-subject-out methodology did seem to demonstrate a connection between the time series accelerometer signals captured during scripted events like walking and PD symptoms like bradykinesia or FoG.

4.1.2 Dataset Augmentation with GANs

The benefit of using support vector machines – the machine learning technique used in the studies by Sama et al. and Rodriguez-Martin et al. – is that they have relatively few parameters to train [140, 154]. Complex deep neural networks have several weight and bias parameters to update and hyperparameters (number of layers, neurons per layer, filters, etc.) to tune. The challenge with using deep nets for small datasets like the ones in the aforementioned studies is that it is easy for models to overfit [36–39]. Such models perform well on the data used to train them but not on unseen out-of-sample test data; they have

poor generalizability. Overfitting is exacerbated by the high sampling rate of wearable sensors, which greatly increases the dimensionality of features – the so-called "curse of dimensionality" [41, 42].

Datasets typically used in deep learning like the MNIST handwritten digit database and the CIFAR image dataset have at least several tens of thousands of examples each for training and testing [20, 21]. By comparison, the Sama et al. study had 12 subjects and Rodriguez-Martin et al. 21 [140, 154]. Unfortunately, in healthcare it is challenging to acquire large samples of subjects on the order of thousands of patients or greater. This is due to difficulties associated with patient recruitment and with logistics of data acquisition, which can be involved for both clinicians and patients [29–32].

Despite small sample sizes, deep learning applications in PD have shown promise. Camps et al. used an 8 layer 1D convolutional neural network (CNN) trained on inertial signals from 21 PD patients to detect FoG events [141]. They were able to obtain 90% geometric mean between specificity and sensitivity, outperforming other state-of-the-art wearable-based methods that used SVMs by 7 to 18% [141, 158, 159]. In order to avoid the problem of overfitting, Camps et al. used a data augmentation strategy to stochastically quadruple their training dataset size [141]. Their strategy involved shifting and rotating the windowed time series inertial signals for a subset of the samples in their dataset, in order to create a transformed subset that they then added to the training dataset. Because of the stochastic nature of this strategy, the modified instances were different for each training epoch, adding noise to the training process and preventing their CNN model from overfitting.

In line with the concept of data augmentation, we propose to use generative adversarial networks (GANs) to create realistically generated artificial or "fake" samples that can be used with real samples during training. GANs involve the use of two neural networks,

a generator and a discriminator, that play an adversarial minimax game. Fundamentally, a discriminator network is trained to output whether a sample is real or fake in order to minimize its loss function (similar to traditional CNNs). Concurrently, a generator is trained to create fake samples that "fool" the discriminator and maximize the discriminator's loss [160, 161]. We incorporate PD symptom assessment (a regression model) into this adversarial game in order to use deep learning on a dataset with a relatively small subject pool with reduced overfitting.

Other studies have shown this to be the case. Odena, A. experimented with a similar type of GAN as the one we propose to use here [162]. An SGAN was designed to learn not only a generative model but also a classifier simultaneously. On MNIST data, Odena, A. was able to show that the classifier component of the SGAN had better classification accuracy on restricted datasets than a regular CNN. In fact, even with as little as 25 training examples, the SGAN outperformed the CNN [162]. The poor performance of the CNN can be attributed to greater overfitting on the small training sets compared to the GAN. Therefore, we believe that GANs will enable us to obtain better performance on smaller datasets with deep neural networks. GAN-based data augmentation strategies have also shown promise in healthcare-focused classification tasks. Synthetic or "fake" examples used in conjunction with real examples from both small and large training datasets have resulted in increased test performance, as shown by Frid-Adar et al. for computed tomography scans of 182 liver lesions, Ratner et al. for a dataset of 1,506 mammograms, and Golany et al. for 109,492 electrocardiograms [129, 130, 163].

Inspired by the prior reasoning, we employed adversarial training to develop a neural network-based regression model that could predict the postural instability and gait disorder score (PIGD) of a Parkinson's disease subject wearing a lumbar inertial sensor. The proposed models were evaluated on their ability to predict PIGD score as well as to

accurately classify ON/OFF states from inferred PIGD scores (i.e. predicted PIGD scores should be greater for the OFF state than for the ON state). We used a modified loss function that takes into account ON/OFF states of subjects to encourage the network to learn meaningful features that were relevant to Parkinson's disease state, instead of simply learning the difference between various subjects. The models were tested rigorously with a dataset collected independently and at a different clinic from the one used for training. Adversarial training of a GAN led to better performance compared to typical training of a CNN, and GAN model outperformed clinicians when determining ON/OFF state from PIGD scores.

4.2 Assessing Gait and Predicting ON/OFF State with Deep Learning

4.2.1 Building a Clinical Dataset of Parkinson's Disease Patients

Data was collected from PD patients at two different sites: Tufts University and Spaulding Rehabilitation Hospital. In both sites, subjects were recorded while they performed the different tasks required for the UPDRS test under the supervision of a trained clinician. Each task was scored on a scale from 0 to 4, and then added to create the total UPDRS score. The PIGD sub-score was calculated by adding only the scores relevant to posture and gait: arising from chair, gait, freezing of gait, postural stability, and posture.

Subjects were outfitted with APDM Opal inertial sensors strapped around the limbs and torso with stretchable bands. The APDM sensors recorded accelerometer, gyroscope, and magnetometer data at 128 Hz over time, as subjects underwent the UPDRS test.

We briefly describe the two studies – Study 1 and Study 2 – below. For more details

about data acquisition, consult Erb et al. [150, 164].

1. **Study 1** – The subjects in this study were recruited at Tufts University. 35 subjects were recorded over 2 visits, one visit when they were in the ON state and one when they were in the OFF state. In each state/visit, they performed the full battery of UPDRS Part III tests and their sensor data and UPDRS scores were recorded. As per protocol, the UPDRS Part III was timed to take place either immediately prior to a subject's next dose of levodopa or immediately after the next dose, with self-reported confirmation that the subject was feeling OFF or ON, respectively. When reporting a state, subjects assessed the severity of their non-motor and motor functions to determine whether they felt the re-emergence of symptoms associated with the wearing off of levodopa.

   Motivated by studies that showed the feasibility of remotely assessing PD symptoms with video conferencing software, we recruited 2 clinicians to score the symptoms of Study 1 subjects from video [165]. We compared the PIGD scores from each of these video raters to those of the live rater (the clinician who conducted the UPDRS Part III exam in person) to calculate two coefficients of determination or R^2 (reported in Section 4.3.2).

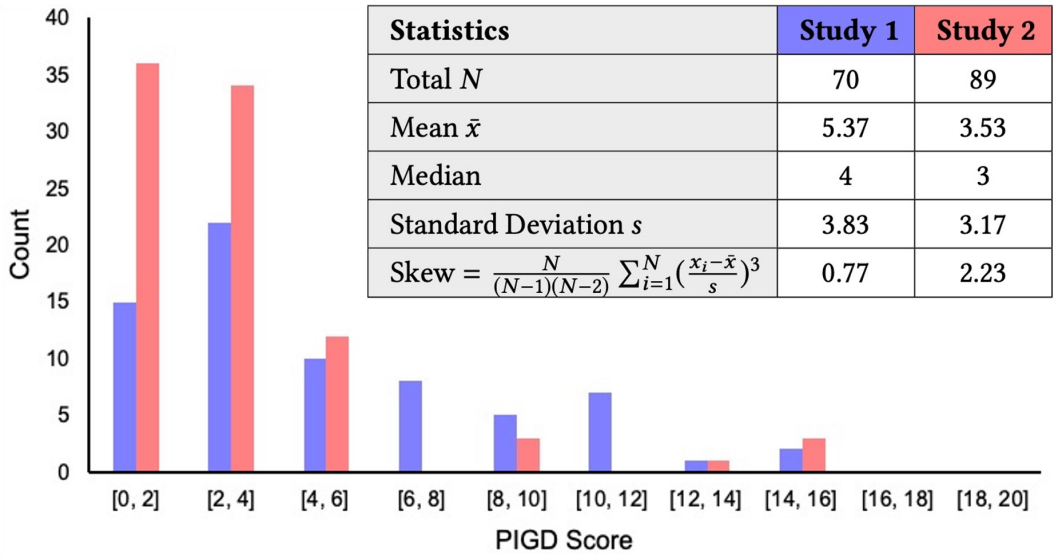
2. **Study 2** – The subjects in this study were recruited at Spaulding Rehabilitation Hospital. A total of 26 subjects were recorded, but 3 were omitted due to missing UPDRS clinician scores. Each of the 23 subjects remaining was recorded up to 5 times over 6 hours, the approximate duration of a full ON/OFF cycle. As a consequence, subjects could have been ON, OFF, or somewhere in between ("TRANSITIONING TO ON" or "TRANSITIONING TO OFF") in each recording/visit. Furthermore, not all subjects completed a full cycle. For example, there were subjects who were recorded only when they were ON or transitioning into the ON state. Other subjects had at least

1 visit out of (up to) 5 when they were ON and at least 1 recording when they were OFF. In each recording, the subjects underwent the UPDRS Part III test and their sensor data and UPDRS scores were collected. The sensor setup used was the same as in Study 1.

There were 70 visits (35 subjects, 2 visits each) for Study 1, each with a self-reported "ON" or "OFF" state. For Study 2, there were 89 recordings with an "ON," "OFF", "TRANSITIONING TO ON", or "TRANSITIONING TO OFF" self-reported state, corresponding to 23 subjects with up to 5 visits each. Figure 4.1 shows the distribution of PIGD scores for each study, along with a table of statistics including mean and skew.

Feature Processing

As part of the UPDRS Part III, subjects walked for 2 minutes along a straight line, turned around, returned to the examiner, and repeated this process. They were outfitted



Statistics	Study 1	Study 2
Total N	70	89
Mean $\bar{x}$	5.37	3.53
Median	4	3
Standard Deviation s	3.83	3.17
Skew = $\frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} (\frac{x_i - \bar{x}}{s})^3$	0.77	2.23

Figure 4.1: Distributions of PIGD scores for study visits. Both distributions favor lower PIGD scores; distribution for Study 1 is skewed less than for Study 2.

Figure 4.2: Position of an APDM Opal inertial sensor attached to the lumbar region using a stretchable belt. The sensor was placed on the lower back of the subject. Accelerometer, gyroscope, and magnetometer data was collected as subjects walked back and forth for 2 minute (as part of the UPDRS Part III test) while wearing this sensor.

with several sensors, including a lumbar inertial measurement unit (IMU) worn about the torso (Figure 4.2). In this work, we considered only this sensor because:

1. It is close to the center of mass of the human body. The trunk is therefore the best sensor location for assessing standing balance, walking stability, and posture identification [166–170].

2. Wearing a lumbar sensor around the torso towards the front (above the anterior superior iliac spine) has been shown to be more comfortable for subjects [171].

The lumbar IMU included an accelerometer to measure acceleration in the X, Y, Z directions and a gyroscope to measure rotation speed around the X, Y, Z axes. We therefore obtained 6 time series as the subjects walked and turned. By using the gyroscope

values, we detected and cut out turn events; turns were defined as time periods when the rotation about the X axis was larger than 120 degrees. We were therefore left with only snippets of straight walking events. Since subjects walked back and forth, each walk was an example and each subject had several walks or examples per clinic visit. To simplify our approach, we considered only the 3 acceleration signals.

Some subjects walked faster than others, so we truncated the 3 IMU acceleration signals at 3 seconds. Doing so ensured that we maintained the same data dimensions across all subjects. We found that a greater cutoff reduced the number of examples overall, since subjects generally completed a single straight walk in around 3 seconds. On the other hand, a smaller cutoff resulted in shorter examples with less information, ultimately giving poorer model performance. A 3 second cutoff gave us a good number of examples without encouraging overfitting or sacrificing performance.

The timeseries IMU signals were filtered using a high-pass Butterworth filter (0.25 cutoff) to remove drift and gravity effects. Lastly, we converted the time series to log spectra using a Fourier transform. It was natural to look at the Fourier decomposition of the time series signals because walking is a periodic activity; this periodicity was reflected in the log spectra [172, 173].

At the end of feature processing, we obtained 4178 examples from the Study 1 dataset and 5405 examples from the Study 2 dataset. Each subject had multiple walks or examples per visit. Each example consisted of 3 spectra (acceleration along the X, Y, Z directions) and was 384 data points long (3 seconds multiplied by 128 Hz, the frequency of data capture of the IMU). The dimensions of the data were therefore 4178 x 3 x 384 for training and 5405 x 3 x 384 for testing. Each example had an associated PIGD sub-score and UPDRS score, which were unique to a subject as well as a visit.

4.2.2 A Neural Network Trained With and Without an Adversary

We used the larger subject pool recorded at Tufts University to train our deep learning pipelines, and the Study 2 dataset to test them. While both dataset distributions were concentrated around lower PIGD scores rather than higher scores, Study 1 dataset had more evenly distributed scores. Study 2 dataset had few subject recordings with PIGD scores in the middle; a disproportionate majority of visits had low score labels (below 4) with only a few visits with higher scores (greater than 10). Moreover, skewness was smaller for Study 1 than for Study 2, indicating that the distribution of Study 1 was less asymmetric (Figure 4.1). This in particular was important for generalization. A pipeline trained with the more skewed distribution of Study 2 scores would not have performed well when tested with examples not well represented by that dataset, namely walks for subjects with high scores.

The networks were trained to take 3 second walks as input and to output a PIGD score. Parameters were updated using the Adam optimization algorithm, an extension to stochastic gradient descent; Adam is a standard for minimizing a parametrized (non-convex) objective function or "loss" in a computationally effective way [174]. The technique was also used by Salimans et al. in "Improved Techniques for Training GANs" [175]. In this paper, Salimans et al. outlined several methods to encourage loss convergence in the minimax game played by the GAN, an otherwise challenging model to train [175]. One such technique was the historical averaging learning rule, which involved keeping a running average of the parameters of the last few models during training. Any updates that yielded parameters significantly different from this historical average were discouraged (with an L2 cost added to the objective function) to improve convergence. Our implementation of this learning rule was the same as the one by Salimans et al. [175]. While not required, we used historical averaging during the training of the CNN primarily because we used it when

training the GAN discriminator. This was done to more directly compare the performance of the two techniques and to better understand the effects of adversarial training.

Our deep learning pipelines were developed using Lasagne and Theano, Python libraries to build and train neural networks and to work with mathematical expressions involving large multi-dimensional arrays [176, 177].

CNN Architecture

CNN performance was the baseline against which we compared the GAN's performance. In order to minimize overfitting, the CNN architecture was not deep – 2 1D convolutional layers and 2 fully connected layers, with the last fully connected layer also serving as the output of the pipeline. Hyperparameters like batch size and learning rate were empirically determined based on what values yielded the fastest convergence of training losses and the best accuracy metrics.

We trained with mini-batches containing 400 examples from two visits in Study 1. In the 1D convolutional layers, 32 filters of size 3 operated along the last dimension of the input with stride 1 and padding 1. The 400 X 3 X 384 output of the second convolutional layer was passed into a fully connected layer with 512 units. In order to minimize overfitting we applied dropout regularization with probability of 0.5 for all three hidden layers. The last fully-connected layer had 2 units for output. Figure 4.3 summarizes the CNN architecture.

The convolutional layer weights were initialized as an orthogonal matrix [178]. For the two fully connected layers we used a He initializer with weights sampled from the uniform distribution [179]. All layers except the last output layer used the rectified linear unit (ReLU) activation function. We applied weight normalization to the hidden layers (instead of batch normalization in order to avoid adding noise to an already noise-sensitive regression model) [180]. Biases were initialized as 0's and the norms of the parametrized

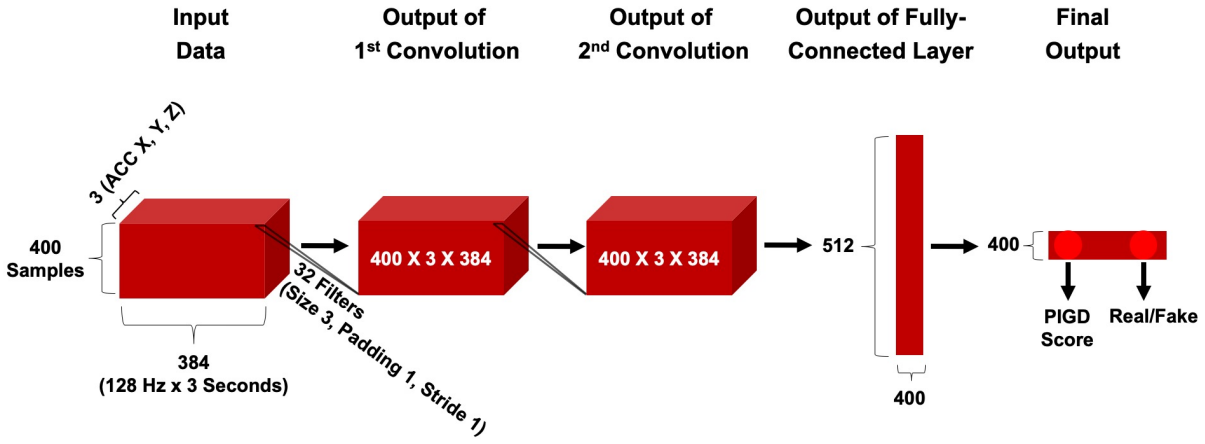


Figure 4.3: CNN and GAN discriminator architecture. When training the CNN, we disregarded the "Real/Fake" output. Dropout with probability 0.5 was applied to the output of 2 convolutional and 1 fully-connected layers.

weights, g, as 1's.

CNN Training

We used the Study 1 dataset to train the pipeline. Each subject in this dataset had examples from two visits, one visit when they were ON and one when they were OFF. We trained the CNN (and GAN) with data from 25 out of 35 subjects recorded at Tufts University; 10 subjects were randomly selected for a development set and their walks left out to be used as a check for convergence.

We trained in mini-batches of size $N = 400$. We first randomly selected 200 examples out of 4178. A walk from one of the 25 subjects corresponded to one of the two visits for that subject. If the example came from the first visit, we randomly selected a walk from the second visit for the same subject, and vice versa. Subjects may have appeared multiple times in a mini-batch, but every example from one visit was paired with an example from the opposite visit. We therefore obtained mini-batches of size $2 * 200 = 400 = N$.

The goal was to learn to differentiate between not only subjects but also the two visits of a subject. Note that the difference in score between subjects was often larger than the difference in scores between the two visits of the same subject. We therefore included the mean squared error (MSE) of the difference between two visits in the standard mean squared loss objective function:

$$loss\_train = \alpha * \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \beta * \frac{1}{N}\sum_{i=1}^{N}((y_{i_1} - y_{i_2}) - (\hat{y}_{i_1} - \hat{y}_{i_2}))^2$$

$$loss\_train = \alpha * MSE(y_i, \hat{y}_i) + \beta * MSE(y_{i_1} - y_{i_2}, \hat{y}_{i_1} - \hat{y}_{i_2})$$

y_i is the predicted PIGD score of a subject's walk, provided by one of the two units in the output layer. $\hat{y}$ is the ground truth PIGD score of the example. The first summation term is the MSE between predicted values and ground truth values, weighted by hyperparameter α.

In the second summation term, we first calculated the difference between the predicted PIGD score for a walk corresponding to a subject's first visit, y_{i1}, and the predicted PIGD score for a walk from the subject's second visit, y_{i2}. We then calculated this same difference but between the ground truth PIGD values, $\hat{y}_{i_1} - -\hat{y}_{i_2}$. Lastly, we calculated the mean squared error between these two terms, weighted by hyperparameter β. We calculated the second MSE in this way because we wanted to encourage the CNN (and GAN) to learn to make predictions such that the regressed PIGD score of one visit had the same inequality relationship with respect to that of the other visit. That is, if a subject was OFF during their first visit and ON during their second, their PIGD score for the first visit should be higher than that of their second. The predictions made for this example should therefore reflect the same inequality relationship.

In addition to α and β, other hyperparameters relevant for CNN training included

the learning rate and the number of epochs over which we trained. We set α and β to 1.0 to prioritize differentiating subjects and differentiating visits for a given subject equally during training. The learning rate was fixed at 0.01.

The 10 development set subjects recorded at Tufts University that were not used to train the pipeline were used to compute an error at the end of every epoch. We picked these 10 subjects out of the 35 randomly once before training; the 10 subjects remained fixed during training so that we could calculate a loss per epoch and compare it between different epochs. At the end of each epoch, we inputted all walk examples from these 10 subjects into the pipeline. We then computed a MSE loss between the predicted and ground truth labels of these 10 subjects. We used this development MSE loss as a way of determining the number of epochs over which to train. We trained until this loss converged to a steady-state value (that remained the same for at least 1000 epochs). Randomly selecting a different set of 10 subjects at the end of every epoch would have made it difficult to reliably compare loss between epochs; it would have been challenging to gauge whether the model was being properly trained if changes in loss could have been due to the subjects randomly chosen for a given epoch.

GAN Architecture

The architecture of the GAN consisted of two neural networks: the generator and the discriminator. The discriminator had the same architecture as the CNN described earlier. This allowed us to directly compare the performance of the CNN and GAN discriminator and to isolate and understand the effects of training with and without data augmentation via a generator network. Henceforth, "GAN discriminator" refers to a shallow neural network trained with an adversarial generator to predict PIGD score and identify real walks from fake ones. "CNN" is the same shallow neural network but trained without a

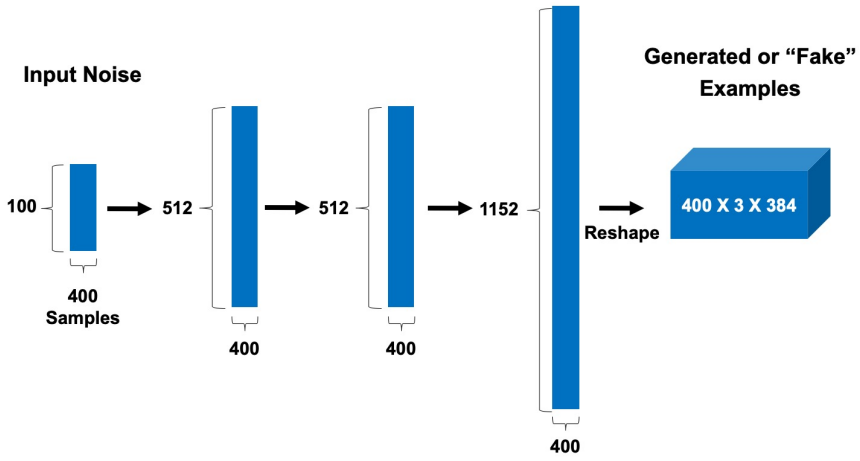


Figure 4.4: GAN generator architecture. Input noise sampled from the uniform distribution was fed into 3 fully-connected layers, and the output of the last was reshaped to the dimensions of the input accepted by the discriminator.

generator that predicts PIGD score alone.

The generator neural network accepted input noise with dimensions 400 X 100, sampled from the uniform distribution between 0 and 1 as in the study by Salimans et al. [175]. The generator architecture consisted of 3 fully connected layers. The first 2 layers had 512 units, used the ReLU activation function and weight normalization [180]. The last layer served as the output layer and consisted of 1152 units with no nonlinearity and L2 regularization. Weights were initialized using the He initializer with the uniform distribution, biases were initialized as 0's, and the weight norms (g) as 1's [179]. Figure 4.4 summarizes the GAN generator architecture.

GAN Training

In the CNN (trained without an adversarial network), we used only one of two output units in the network; this unit provided the predicted PIGD score. However, when training the GAN, we made use of the second output unit to distinguish real walk examples

from fake ones created by the generator. The generator itself was trained to fool the discriminator. The training paradigm is detailed below and summarized in Figure 4.5.

The discriminator was trained to minimize the following loss, designed to accommodate both PIGD score regression and real/fake example identification:

$$loss\_disc = \gamma * loss\_train + \delta * (\frac{1}{N}\sum_{i=1}^{N}(D(x_i) - 1)^2 + \frac{1}{N}\sum_{i=1}^{N}(D(G(z_i)))^2)$$

The function $loss\_train$ was previously defined for training the original CNN, except weighted by the hyperparameter γ. $D(\bullet)$ is the sigmoid output of the discriminator, and identifies whether a sample is real, 1, or fake, 0. Thus, $D(G(z_i))$ is the discriminator output when a fake walk example $G(z_i)$ is passed in as input. Fake examples were created from noise z_i by the generator $G(\bullet)$. $D(x_i)$ is the output when a real walk example, x_i, is passed in as input. (i ranges from 0 to the batch size.) The discriminator was essentially trained to push the value of $D(x_i)$ to 1 and $D(G(z_i))$ to 0.

The generator was trained to minimize the following loss, designed to fool the discriminator by pushing $D(G(z_i))$ to 1:

$$loss\_gen = \frac{1}{N}\sum_{i=1}^{N}(D(G(z_i)) - 1)^2$$

The interplay between training the discriminator and generator is outlined in Figure 4.5. The training protocol was the same as that of the CNN, with a learning rate fixed at 0.01 and data from 10 subjects used as a test for convergence.

Parameters γ and δ allowed us to control the game between the discriminator and generator. For example, a GAN discriminator trained with $\gamma > \delta$ would behave similarly to the original CNN; the loss term weighted by γ and using unlabeled, fake examples from the generator would not be weighted as highly as $loss\_train$, which used real examples. In

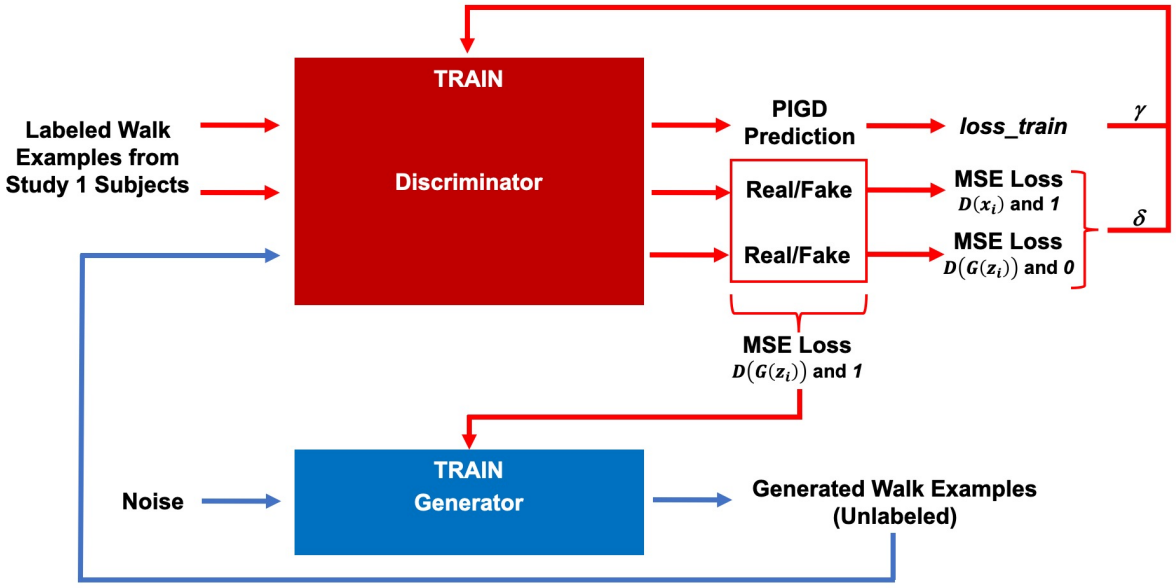


Figure 4.5: GAN training paradigm. The discriminator predicted a PIGD score as well as a real/fake label for every walk example input. The generator provided fake walk examples. These outputs were used to compute loss terms for training the discriminator and generator.

the opposite scenario, with $\delta > \gamma$, the generator would have a greater than equal influence on the training of the discriminator. In our testing, we used the following combinations of (γ, δ): $(1.0, 1.0)$, $(0.1, 0.5)$, $(1.0, 0.5)$, $(0.5, 1.0)$, and $(0.1, 1.0)$.

4.2.3 Testing Models and Analyzing their Output

We evaluated the CNN and GAN discriminator on their ability to determine ON or OFF state from predicted scores – "ON/OFF accuracy." For a given subject, the OFF state should have a higher PIGD score than the ON state. The same should hold true for the scores provided by clinicians. After training, we tested the models using the Study 1 development set described in Section 4.2.2 as well as the Study 2 dataset.

Note that each subject had several walks in a recording/visit. We averaged the

predicted PIGD scores of all walk examples for a given visit. We therefore obtained 2 predicted scores per Study 1 subject, one for each of their visits. For Study 2, we obtained up to 5 predicted scores per subject.

Performance on Study 1 Development Set

Study 1 subjects reported themselves as either ON or OFF in each visit. For the 10 development set subjects, we compared the predicted PIGD score for the OFF state with the predicted PIGD score for the ON state; we counted the number of times out of 10 that the former was greater than the latter. We repeated this analysis with scores provided by the clinician rater. We gauged clinician performance via both the PIGD sub-score and the overall UPDRS score.

In addition to ON/OFF accuracy, we report the coefficient of determination, R^2. R^2 was calculated using 20 predicted PIGD scores from the CNN or GAN discriminator and 20 ground truth PIGD scores from the clinician.

Performance on Study 2 Dataset

Study 2 subjects could have been ON, OFF, transitioning into the ON state, or transitioning into the OFF state in each of their visits. These labels were self-reported, and it was not a requirement that each subject had visits satisfying all 4 criteria. Therefore, to simplify our approach, we averaged predicted PIGD scores for visits when the subject was ON or TRANSITIONING TO ON, μ_{ON}. We similarly averaged the predicted scores for visits when the subject was OFF or TRANSITIONING TO OFF, μ_{OFF}. Out of 23 subjects, 17 had at least 1 visit when they were ON or TRANSITIONING TO ON and at least 1 visit when they were OFF or TRANSITIONING TO OFF. Of these 17, we counted the number of times $\mu_{OFF} > \mu_{ON}$. That is, to calculate ON/OFF accuracy, we counted the number of correct ON/OFF determinations made by the CNN or GAN discriminator

Table 4.1: Performance of CNN, GAN Discriminator, and Clinician Rater

		Study 1 Development Set		**Study 2 Dataset**	
		ON/OFF Accuracy	R^2	**ON/OFF Accuracy**	R^2
GAN Discriminator	(γ, δ)				
	(1.0, 1.0)	100%	0.51	65%	-0.62
	(0.1, 0.5)	100%	0.49	76%	-0.42
	(1.0, 0.5)	90%	0.57	71%	-0.33
	(0.5, 1.0)	90%	0.51	71%	-0.29
	(0.1, 1.0)	100%	0.51	76%	-0.43
CNN	–	70%	0.52	53%	-0.60
Clinician Rater – PIGD	–	100%	–	65%	–
Clinician Rater – UPDRS	–	100%	–	88%	–

and divided by total number of examples. We also considered the clinician scores $\hat{\mu}$, and assessed whether $\hat{\mu}_{OFF} > \hat{\mu}_{ON}$. $\hat{\mu}$ could be either PIGD or UPDRS. Lastly, as for the Study 1 development set, we computed R^2 with 89 predicted and 89 ground truth PIGD scores (corresponding to all 23 Study 2 subjects).

4.2.4 Baseline Performance Without Adversarial Training

We trained the CNN model for 5000 epochs on the Study 1 dataset. Figure 4.6 shows $loss\_train$ (with $\alpha = \beta = 1.0$) plotted over epochs. We similarly plotted the MSE loss curve computed using 10 Study 1 subjects at the end of every epoch (as described in Section 4.2.2). The fact that both curves converged to some steady-state minimum value confirmed that 5000 epochs was enough to complete training.

As detailed in Section 4.2.3, we used 10 subjects from Study 1 and 17 subjects from Study 2 to test whether predicted PIGD scores could be used to correctly determine ON/OFF states. Table 4.1 shows that in 7 occurrences out of 10, the CNN regressed a PIGD score that was greater for the OFF state than for the ON state, yielding an

ON/OFF accuracy of 70%. For comparison, the clinician had 100% accuracy for the Study 1 development set.

The CNN's ON/OFF accuracy for Study 2 subjects was 53%. In this case, the clinician evaluated on the PIGD sub-score had an accuracy of 65%. In other words, the tests the clinician conducted to assess postural instability and gait difficulty yielded scores that were sometimes erroneously larger for the ON state than for the OFF state. These mistakes happened for 6 out of 17 subjects. The clinician rater made fewer mistakes when the total UPDRS was taken into account – 2 mistakes out of 17 or 88% accuracy. Note that the UPDRS score included tasks unrelated to postural instability and gait difficulty, like speech and cognition.

4.2.5 Performance when Trained Alongside Adversarial Network

We trained the GAN model for 5000 epochs on the Study 1 dataset. In Figure 4.6, $loss\_disc$ and $loss\_gen$ (with $\alpha = \beta = \gamma = \delta = 1.0$) were plotted over epochs. Figure 4.6 also shows the MSE loss curve from 10 Study 1 subjects. All curves converged to some steady-state minimum value in 5000 epochs. Convergence occurred for all combinations of γ and δ as well.

Table 4.1 shows ON/OFF accuracy results of the GAN discriminator for all 5 combinations of (γ, δ) tested: $(1.0, 1.0)$, $(0.1, 0.5)$, $(1.0, 0.5)$, $(0.5, 1.0)$, and $(0.1, 1.0)$. The GAN discriminator's performance varied based on the values of (γ, δ) tested, more so for Study 2 than for Study 1. At best, the GAN discriminator outperformed the CNN by 30% and matched the clinician's perfect performance on the Study 1 development set, for $(\gamma = 1.0, 0.1, \delta = 0.5, 1.0)$. For Study 2, the discriminator always outperformed the CNN, doing 12% to 23% better in determining ON/OFF state. When $(\gamma = 0.1, \delta = 0.5, 1.0)$, the discriminator also outperformed the clinician evaluated on PIGD, by 11%. For Study 2,

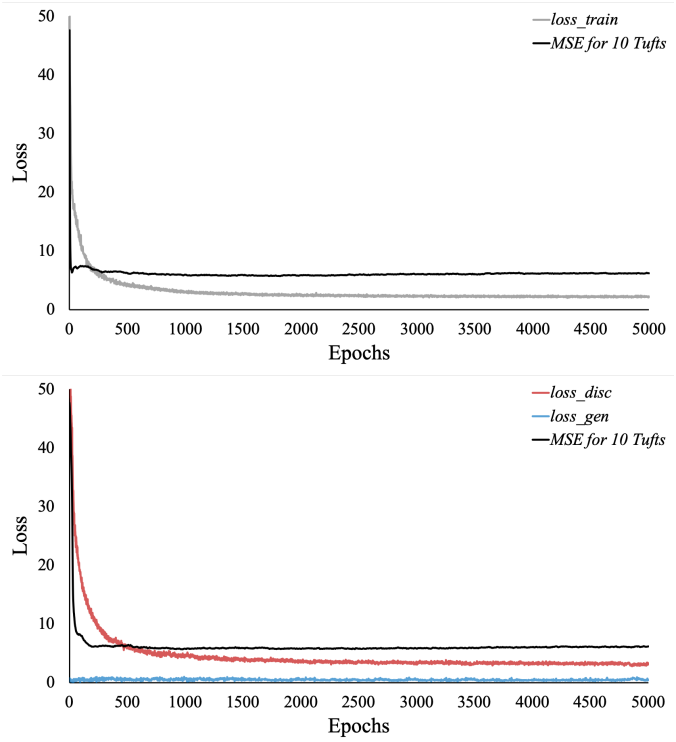


Figure 4.6: CNN and GAN loss curves plotted over training epochs. (Top) CNN loss curves. *loss_train* is plotted over epochs in gray, and the MSE loss curve for 10 Study 1 subjects in black. (Bottom) GAN loss curves. *loss_disc* is plotted in red and *loss_gen* in blue over epochs. MSE loss curve for 10 Study 1 subjects is shown in black. All 3 curves are for the instance when for $(\gamma = 1.0, \delta = 1.0)$. Note that convergence occurred for the 4 other (γ, δ) combinations tested as well.

using the full UPDRS score to make ON/OFF predictions resulted in the best clinician performance – 88%, 12% better than the GAN discriminator's best ON/OFF accuracy. The full UPDRS score contains additional information about the subject's symptoms that go beyond gait (e.g. speech, tremor, bradykinesia of the upper limbs, etc.).

4.3 Contributions to the Field

In this work, we set out to show that we could teach neural networks to predict PIGD score from a single lumbar accelerometer. By using separate training and test sets,

we were better able to evaluate the generalizability of the resulting models as compared to other work that used less rigorous training paradigms on a single dataset. There was novelty in this work that distinguished it from other machine or deep learning sensor-based approaches for tracking PD symptoms. First, all models were trained and tested with two independently collected datasets. The results reported in this paper were obtained without using less rigorous cross-validation training paradigms that potentially inflate performance (detailed in Section 4.1.1). Second, adversarial training maximized the performance of the discriminator, a shallow neural network that was otherwise prone to overfitting due to the small size of the training set (a common problem with deep learning in healthcare applications). Lastly, PIGD scores alone were used to accurately predict ON/OFF states from just a few walking events. Going forward, the challenge is to be able to translate this work from the clinic to the home. We envision a system that could be used to reliably assess symptoms and track ON/OFF cycles continuously. Doing so would better inform clinicians as to the state of their patients and aid patients in their rehabilitation, ultimately improving their quality-of-life.

4.3.1 Reducing Overfitting with Adversarial Training

We tested the ability of adversarial training to improve out-of-sample performance of a shallow neural network on the premise that generating synthetic samples from noise can successfully augment real clinical datasets. Our models were evaluated on their ability to predict PIGD scores that were greater for visits that were OFF than for visits that were ON. The ability to predict ON/OFF – "ON/OFF accuracy" – was used as a performance metric for both the GAN discriminator and CNN to facilitate a comparison with clinicians. Clinicians in practice can use the PIGD or UPDRS scores to determine ON/OFF state of a patient. Self-reported ON/OFF labels were used as ground truth because: (1) it has been

shown that patients' perception of their motor functions successfully model PD severity on par with the clinically objective UPDRS exam and (2) self-reported states are often used as a clinical endpoint in PD clinical trials [149].

GAN discriminator's ability to predict ON/OFF state was greater than that of the CNN due to adversarial training. A fully trained discriminator outperformed a fully trained CNN on Study 1 development set subjects, matching a clinician rater's 100% ON/OFF accuracy. Similarly, GAN discriminator always outperformed the CNN on Study 2 subjects, at best doing 23% better. Since both pipelines were trained on the same data, the results suggest that the performance difference was due to the influence of an adversarial generator during discriminator training. In fact, the GAN discriminator did the best – making only 4 mistakes total (76% accuracy) on Study 2 – when $\delta > \gamma$. That is, the GAN discriminator did better when the generator had greater than equal influence on discriminator training, when the second term in loss_disc (computed using generated examples) was weighted more than the first term. Because augmenting the dataset led to better performance, we argue that the CNN must have performed poorer because of overfitting, a conclusion supported by prior work [162]. GANs help counter the curse of dimensionality problem of applying deep learning techniques on the small datasets prevalent in healthcare applications. After all, a shallow neural network trained alongside an adversary yielded better performance than a network with the same shallow architecture trained normally. Overall, GANs are a promising technique to properly analyze walk data and assess postural instability and gait disorder in a way that enables ON/OFF prediction.

4.3.2 Pros and Cons of Deep Learning Compared to Clinicians

When tested with Study 2 subjects, the GAN discriminator generally outperformed a clinician rater evaluated using their PIGD scores. Besides the instance when performance

equaled the rater ($\gamma = 1.0, \delta = 1.0$), the GAN discriminator made fewer mistakes. The human rater was ultimately still better when considering the full UPDRS score (88% accuracy). However, the clinician had the advantage of assessing symptoms that were not captured by the PIGD sub-score, more information than was available to the neural networks. Note that we were interested in developing models that can be taken outside the clinic to automatically monitor patients in-the-wild. So even though clinicians outperformed the GAN discriminator when they used the full UPDRS score to predict ON/OFF state, clinical visits are discontinuous whereas the trained models can be deployed as part of a larger, wearable-based system to track PD symptoms continuously, our future goal.

We developed networks that could regress PIGD scores useful for accurately determining ON/OFF state. Doing so came at the cost of reduced R^2; the models' PIGD predictions for test data were not close to scores from the live clinician rater. R^2 values as reported in Table 4.1 were generally very low. For Study 1, CNN R^2 was 0.52 and GAN discriminator R^2 was between 0.49 and 0.57 based on (γ, δ). For Study 2, CNN R^2 was only -0.60 and GAN discriminator R^2 ranged from -0.62 to -0.29. (The GAN discriminator generally had greater (more positive) R^2 than the CNN. For Study 2, CNN R^2 was 0.31 or approximately 50% less than the best GAN discriminator R^2, likely due to overfitting by the CNN.) In other words, even if a predicted score for the OFF state was greater than that of the ON state, the score values themselves were not similar to ground truth values.

Note however that the PIGD sub-score is very noisy. As mentioned in Section 4.2.1, we obtained PIGD scores from two video raters for Study 1 subjects. We calculated 2 R^2 values by comparing PIGD scores from each video rater to those of the live clinician. R^2 values were 0.24 and -0.04 for an average of only -0.10. Clinicians themselves score PIGD inconsistently. Moreover, our endpoint was the ability to determine ON/OFF state reflective of subjects' self-reported health. We were not trying to replicate a clinician's

UPDRS exam and did not expect high correlation with the live rater's scores. (To improve R^2, the second term in loss_train would have to be removed and the models retrained. For example, for the Study 1 dataset, simply decreasing the impact of the second term by setting $\alpha = 1.0$ and $\beta = 0.5$ increased CNN R^2 to 0.78 and GAN discriminator R^2 to 0.72.)

4.3.3 Limitations to Consider Before Real-World Deployment

Another drawback of this study is that the models described were trained on walk sensor data collected in a clinic under a data collection protocol (subjects walked back and forth for 2 minutes). Walks that occur at home are shorter in both duration and distance, and in-the-wild walks are likely diverse and erratic. The current models may not generalize well to walks collected outside of a clinic. We are therefore collecting in-the-wild sensor data to incorporate into our training dataset; subject recruitment and data collection is ongoing. Mode collapse, a growing concern with GANs, was unaddressed in this work but needs to be quantified before deployment [181]. Note however that the empirical improvements provided by adversarial training would not have been as large if the generator were producing only a limited range of diverse samples; GAN discriminator performance was still better than CNN performance.

4.4 Acknowledgements

This chapter covers as of yet unpublished work (in review) "Detecting Motor Symptom Fluctuations in Parkinson's Disease with Generative Adversarial Networks" by Vishwajith Ramesh and Erhan Bilal. This work was conducted as part of two internships at the IBM T.J. Watson Research Center.

Chapter 5

Design Considerations for a Clinical Decision Support System for Stroke

5.1 Motivation and Background

Acute stroke diagnosis presents several challenges that motivate the development and use of computational aids. Stroke diagnosis typically requires a neurological examination done by a clinician in a hospital – the National Institute of Health Stroke Scale (NIHSS) [182, 183]. The NIHSS is a set of motor and cognitive tests designed to quickly and quantitatively assess impairments. It is a widely adopted exam that has been shown to be useful for rehabilitation planning. However, it is not a perfect test for diagnosis. Follow-up imaging revealed that 21% of patients treated for stroke based on the results of the NIHSS did not in fact have stroke-associated brain infarcts [184]. Moreover, subtleties between certain symptoms like neglect and weakness make them difficult to differentiate using the NIHSS, especially when conducted by inexperienced clinicians or emergency services personnel. In fact, the diagnosis of stroke by emergency medical services (EMS) is only 50% accurate with a similarly low sensitivity of 57.5% – EMS personnel tend to miss strokes [76, 185].

In addition, while useful for identifying large vessel occlusions, the NIHSS cannot properly identify stroke mimics, false positive stroke cases that represent up to 30% of acute stroke hospital admissions [186, 187]. Stroke mimics are non-vascular conditions (e.g. brain tumors, migraine, and conversion, a psychiatric disorder) that present with similar symptoms as acute ischemic stroke. With a specificity of only 52%, the NIHSS cannot properly filter out stroke mimics, often resulting in unnecessary diagnostic procedures that waste neurologists' and patients' time [186].

The efficient triage of stroke patients in the emergency department is important because ideally, acute ischemic stroke is diagnosed within 3 hours from the presentation of symptoms. Tissue plasminogen activator (tPA), the only FDA-approved therapy for ischemic stroke, must be administered in that time for it to be effective [188]. However, due to the narrow therapeutic time window, tPA administration happens in $<5\%$ of acute stroke cases [189–192]. While the NIHSS assessment is designed to be completed in <8 minutes, the increasing demand for immediate and accurate diagnoses has led to the emergence of computational approaches to aid stroke evaluation [193].

Computational diagnostic aids quicken the recognition of stroke and help reduce errors by providing decision support in a limited amount of time [194]. Several researchers have developed machine and deep learning-based systems for stroke diagnosis to support the NIHSS and improve clinical outcomes [195–197]. Recent work capitalizes on advances in machine learning and ubiquitous technology (namely wearables and depth cameras that provide body pose) to computationally identify and reliably quantify the degree of different stroke-related deficits [93–95]. Our goals were to: (1) shed more light on how to best integrate such artificial intelligence (AI) diagnostic tools into existing hospital workflows such that they support clinicians, (2) understand how AI-based tools might change current practices for stroke diagnosis, and (3) inform proper design of the next generation of AI

tools in clinical settings.

5.1.1 Over-Dependence on Technological Aids

When deciding to adopt digital healthcare technology, hospitals and clinics often prioritize cost and ability to improve workplace efficiency. However, the risk that that technology may build blind trust and negatively impact the decision making process of clinicians is often ignored [198]. In fact, over-dependence on technology was discovered to be an unintended adverse consequence of clinical information systems; Campbell et al. and Ash et al. both found that clinicians could not work efficiently without computerized systems and that they had false expectations about data accuracy and processing [199, 200]. It was this latter issue that concerned us about AI diagnostic tools. Various diagnostic AI techniques have been proven to be capable of solving numerous clinical problems yet clinicians are hesitant to include them in their decision making processes [201]. It is important to balance computational approaches for diagnosing stroke with clinicians' assessments, to augment them rather than replace them. This balance is necessary to promote the adoption of AI tools in healthcare environments.

Our neurologist collaborators were concerned that practitioners outside the field of neurology like EMS and less experienced neurologists like residents may become over-dependent on technological aids for stroke diagnosis, ignoring their own assessments. Motivated by this concern, we decided to better understand clinician reliance on software designed to improve decision making in acute stroke diagnosis. To do so, we took a similar approach as a study done in the area of diabetes care. Sims et al. developed and tested a new type of interface for a diabetes-specific clinical decision support system that displayed laboratory results in a color-coded way for the quick understanding of diabetics' metabolic control [202]. In their paper, Sims et al. evaluated the effects of their dashboard on the

identification of patients who required adjustments in their treatment [202]. They measured the amount of time clinicians spent on each patient's case and compared their dashboard to existing laboratory reporting interfaces [202]. They asked doctors to perform timed mock clinical tasks with and without their dashboard and to rate their confidence in interpreting diabetes-related test results. By uncovering the effects of their dashboard on clinicians' decisions for adjusting treatment, the authors learned how to best integrate their support system into existing workflows [202]. No such study existed for computational aids being developed for stroke diagnosis.

In this chapter, we describe the development of an initial user interface (UI) for a stroke diagnostic tool based on prior work and feedback from a focus group of 10 stroke clinicians. We designed and conducted an online experiment in which we evaluated whether computational aids influenced an otherwise standard, video-based stroke diagnosis. We discovered that clinicians considered computational aids only when the aids confirmed or denied their loosely held beliefs.

5.2 Measuring Clinician Reliance on a Computational Aid for Stroke

In order to understand how computational aids affect stroke clinicians' decision making, we conducted a focus group at the UCSD Stroke Center. The results of the focus group informed the design of a prototype UI for an acute stroke diagnosis computational aid that we then evaluated in an experimental setting.

The focus group consisted of 10 clinicians working at the UCSD Stroke Center, including registered nurses, medical residents, and experienced neurologists. Participants were asked to: (1) outline concerns about the introduction of AI-based diagnostic technology

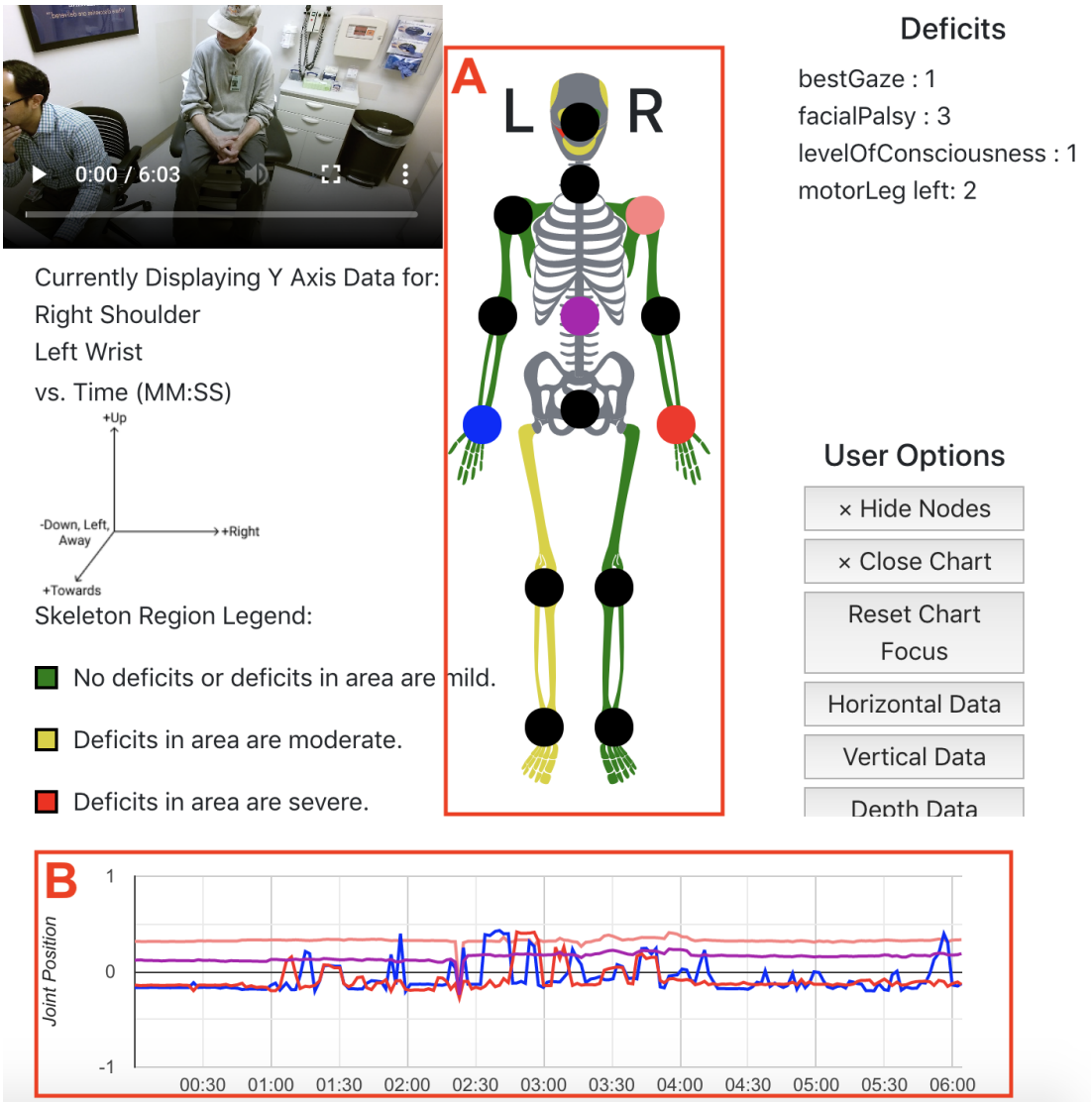


Figure 5.1: The initial prototype of the system's UI that we evaluated in this study. In the top left, we show the footage recorded of a stroke patient undergoing an in-person NIHSS. Deficits identified are listed in the top right. (A) We used a 3-color skeleton to display symptoms and severity (red for most severe, yellow for moderate symptoms, and green for no deficits). Clickable nodes represent the joints for which we could display spatial position data. (B) The position data of the left wrist (blue), right wrist (red), core (purple), and right shoulder (peach) over time are displayed on the bottom graph. The user can choose whether to have horizontal (left/right), vertical (up/down), or depth (away/towards) spatial information displayed on the Y axis.

into their clinic, (2) identify gaps in the information typically available to them, and (3) list additional information that would be valuable for acute diagnosis.

The focus group revealed that a tool that supports clinicians during stroke evaluation needs to quickly convey information. Focus group participants also expressed the desire to see quantitative data about the stroke symptoms identified by a computational system. For example, they wanted to be able to see graphs that could be used to understand how the system detected certain stroke symptoms. Finally, while the prospect of a computational aid was received positively, clinicians worried about the use of such a tool by less experienced clinicians. Specifically, they wondered how less experienced stroke practitioners would balance their own assessments with the predictions of a computational aid, especially in cases of disagreements between the two.

5.2.1 Designing a User Interface

Based on feedback from the focus group and prior work done in the setting of user interfaces for stroke diagnosis, we designed the initial prototype UI shown in Figure 5.1.

We wanted to better understand clinician reliance on technological aids in stroke diagnosis to inform the development of a computational pipeline into a well-integrated, real-time system. An understanding of the user would give us guidelines for the design of a final interface. Since a real-time, fully functional system was still under development at the time of the experiment, we filled the UI prototype with previously recorded data from stroke patients, as described in Section 5.2.2.

We detail two key visual features of the UI relevant to the study to measure clinician reliance:

1. Skeleton of Deficits: In earlier work, Gotoh et al. designed a figure with different body parts highlighted to represent the symptoms of a stroke patient [203]. We used

a similar representation and created a skeleton image to easily visualize different deficits of the body parts, shown in Figure 5.1A. An important benefit of using the skeleton was that our system used 3D coordinates of human body joints (obtained with depth cameras such as the Microsoft Kinect or from video with deep learning methods for human pose estimation) to identify symptoms like hemiparesis [95]. A skeleton visual enabled us to place a clickable node on each body joint to graph the joint's X, Y, or Z position in space over time.

2. Plotting Joint Position: Several stroke studies plot the movement and velocity of body joints in the upper extremities and pupil position (obtained via video-oculography) over time [204, 205]. Inspired by this approach and the fact that the NIHSS is comprised primarily of motor tests that assess symptoms ranging from weakness to tremors, we present quantitative position information of key joints to clinicians using graphs like the one in Figure 5.1B. The X axis represents time and the Y axis represents the movement of a joint in X, Y, or Z coordinate space.

To obtain data to fill our initial UI, we recorded 8 stroke patients while a neurologist was conducting the NIHSS at UCSD Stroke Center outpatient clinics. We followed the same data collection protocol outlined in [95]. We used the Microsoft Kinect v2 depth camera to record audio, video, and the 3D position of 25 body joints over time. For each of the 8 patients, we obtained the neurologist's diagnosis, a list of symptoms identified, and their associated NIHSS scores. All patients agreed to be recorded with sensors by signing a consent form approved by the local Human Research Protections Program office. After obtaining the NIHSS results for the 8 patients, we displayed the recorded video, the identified symptoms and their severity, and the recorded body joint position information in the UI prototype as shown in Figure 5.1.

We created a high-fidelity dynamic web interface using HTML, CSS, and React.js.

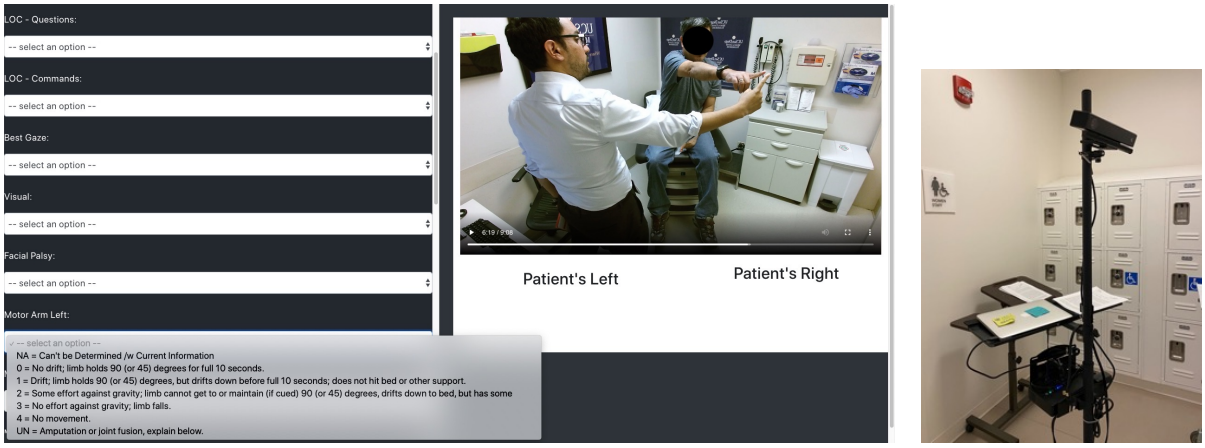


Figure 5.2: Left: First part of our experiment. Participants watched audio and video footage of a neurologist conducting the NIHSS in-person (right side of the UI) and filled out a drop-down menu version of the NIHSS (left side of the UI). Right: Data collection setup in the clinic, with a Microsoft Kinect v2 mounted on the stand of a roll-able computer table.

Users are able to watch and listen to recordings of stroke patients undergoing an NIHSS exam and to view plots of the spatial positions of body joints over the duration of the in-person NIHSS exam. The graphs and the video are synced (as in [206]). So users can jump to any timestamp in the video and be shown the body joint positions at that point in the video, and vice-versa.

We categorically represent a stroke deficit's severity with a 3-color scale to enable quick identification of symptoms. Implementing a 3-color scale using NIHSS scores is challenging because each of the tests in the NIHSS exam has a different scoring system, with some deficits scored out of 3 and others out of 2 or 4 [182]. To address this issue, we designed the skeleton to have 11 components or areas: left/right eyes, left/right face, left/right arm, left/right leg, left/right head, and mouth. The NIHSS test scores associated with each area were added up. For example, the right leg is one area and the related NIHSS tests measure ataxia of the right leg, sensory loss in the right leg, and drift of the right leg. If all 3 of these NIHSS scores were 0, the right leg on the skeleton is colored green. If

at least one of the 3 scores is 1, the area is colored in yellow. Red is used to indicate the highest severity, when the total is greater than or equal to the number of related NIHSS scores. In the case of the motor leg, a total >=3 would be colored red. We do the same analysis for the other 10 areas on the skeleton.

5.2.2 Assessing Clinicians' Reliance

We wished to determine whether a clinical decision support system for stroke diagnosis would influence the decision making of clinicians (across different expertise levels). With this goal in mind, we designed an experiment and survey to evaluate our prototype UI.

Overview of Experiment Structure

In stroke telemedicine, stroke specialists can diagnose a potential stroke case occurring in another location by "calling in" via video conference software. Typically, specialists coordinate with a nurse or member of emergency medicine staff onsite who does an in-person NIHSS. Stroke diagnosis via telemedicine has been shown to be about as effective as diagnosis with an in-person NIHSS, especially when the remote stroke specialist has access to brain imaging [100]. Video-based stroke assessments have similarly been shown to be useful even without brain imaging available [207]. Therefore, in our experiment, we asked participants to watch the recorded footage (video and audio) of stroke patients undergoing an in-person NIHSS and to diagnose stroke to the best of their ability.

Participants were presented with 7 cases, plus 1 more patient who was used as a tutorial to acclimate the user to the experiment's interface. Each case was broken up into 2 parts. In the first, participants filled out a drop-down form of the NIHSS exam while using the recorded footage to identify symptoms and diagnose stroke (see Figure 5.2).

Diagnosing stroke through video is more challenging than interacting with a patient in person, so we gave participants the ability to rate their confidence in their assessments from 1 (least confident) to 7 (most confident) as well as the the option to write general notes and feedback, especially if they felt that they could not properly assess a symptom from video and audio. This approach was inspired by a similar healthcare study related to spirometry interpretation that rated the perceived confidence of participants out of 7 [208]. Studies in non-healthcare fields, namely those related to answering trivia questions and eyewitness identifications, have also used a 7 point scale for confidence [209, 210].

In the second part of each case, participants were shown our prototype UI from Figure 5.1 that listed the deficits identified by the neurologist conducting the in-person NIHSS. Participants' NIHSS scores and confidence ratings from the first part (the left side of Figure 5.2) were saved and carried over. In the second part, participants were given the option to change their answers from the first part in light of the new information provided to them through the UI. This information was namely the body joint position graphs and the identified deficits and associated severity. Participants were told that this data was detected by the computational AI system, but in reality it came were from the original expert neurologist conducting the in-person NIHSS.

To assess reliance on computational aids, we monitored changes made in this second part of the experiment by our participant. Participants were again asked to rate their confidence in their NIHSS diagnoses after using the computational aid. They were also asked if they felt that any of the displayed symptoms were wrong, to which they had the option to respond "Yes" or "No".

After completing the tutorial and all 7 cases, we presented participants with a final, post-experiment survey that included questions regarding: (1) their confidence in the accuracy of the information that was shown by the UI, (2) whether they had difficulty

diagnosing any symptoms with the information provided, (3) what sources of information they generally trusted and regularly used in their diagnoses, and (4) which UI elements they found to be most and least helpful.

The experiment was designed to take 1.5 to 2 hours to complete and could be done in parts to accommodate the busy schedules of clinician participants.

Recruiting Participants

Our study was conducted in collaboration with the UCSD Stroke Center. To recruit participants, we sent a link to the online experiment and survey to the center, comprised of neurologists, fellows, residents, nurses, and nurse practitioners. Note that none of the experiment's participants were the neurologist conducting the NIHSS in person. The participants of the experiment were not involved in the diagnosis or rehabilitation of any of the patients used in the experiment.

We collected responses from a total of 6 clinicians who used our online prototype. Out of the 6, 4 also finished the post-experiment survey. Table 5.1 shows the roles of the participants: we had responses from 3 stroke fellows, 1 nurse practitioner, and 2 full-time, more experienced, neurologists.

Participant Confidence in Video-Based NIHSS Assessments

Table 5.2 lists how each participant perceived their confidence regarding their ability to diagnose stroke by watching the audio and video footage of the NIHSS exam (in the first part of the experiment). The stroke fellows and experienced neurologists were generally confident in their video-based NIHSS assessments; scores were primarily 5 or higher out of 7. Participant 4, the nurse practitioner, was the least confident with scores of either 4/7 or 5/7.

Table 5.1: Participants and their Clinical Roles

Participant ID	Clinician Position
1	Stroke Fellow
2	Neurologist
3	Neurologist
4	Nurse Practitioner
5	Stroke Fellow
6	Stroke Fellow

Manipulating the Ground Truth

The in-person neurologist's NIHSS assessments were the basis of real-life treatment and rehabilitation planning for the 8 recorded stroke patients. As such, we deemed the in-person scores as "ground truth". We used the in-person NIHSS results in our UI prototype as a proxy to our computational AI-based system. In-person NIHSS diagnoses also tend to be more accurate than video-based diagnoses, so we felt justified in using the former as a gold standard against which to compare the latter [211].

In order to properly measure dependence on the computational aid, we wanted to gauge how clinicians would respond to discrepancies between their own assessments and the results of the system shown on the UI. After all, the final, fully developed system may still make erroneous predictions far from the ground truth. We slightly manipulated the in-person NIHSS results shown to participants in the prototype UI to identify if participants caught the deliberately introduced deviations from the ground truth. It is for this reason that in each case, as well as in the final survey, we asked participants about whether they believed the information showed to them on the prototype UI was accurate.

3 out of the 7 stroke patient cases were randomly chosen and manipulated. In the

Table 5.2: Perceived Confidence in Video-Based NIHSS

Participant ID	**Patient 1**	**Patient 2**	**Patient 3**	**Patient 4**	**Patient 5**	**Patient 6**	**Patient 7**
1	7/7	7/7	7/7	7/7	5/7	7/7	7/7
2	6/7	5/7	5/7	6/7	5/7	4/7	5/7
3	7/7	7/7	7/7	7/7	7/7	7/7	7/7
4	4/7	5/7	5/7	5/7	4/7	4/7	5/7
5	5/7	3/7	3/7	6/7	6/7	4/7	5/7
6	6/7	5/7	6/7	6/7	5/7	6/7	7/7

first case, we slightly changed the NIHSS score of a random symptom identified by the in-person neurologist to another random number. In the second case, for each identified symptom, we changed the score to a random score. In the third case, we listed a deficit that was not identified by the in-person neurologist and so was not actually present in the stroke patient. Participants were given the ability to change their NIHSS assessments from the first part when these manipulated results were shown to them in the second part.

Changes Made After Seeing Prototype UI and Manipulations

For the cases in which we manipulated data, we analyzed how clinicians changed their scores, if they did. We looked for correlations between our manipulations and the changes the clinicians made to their self-reported confidence.

The NIHSS scores and symptoms of Patients 2, 3, and 6 were manipulated as described in Section 5.2.2. For Patient 2, we reduced the motor arm left NIHSS score from 3/4 to 1/4. For Patient 3, a score of 1/4 for the motor arm left NIHSS test and a score of 1/4 for the motor leg left NIHSS test were changed to 2/4 and 3/4, respectively. Lastly, for Patient 6, we added a motor arm right NIHSS score of 2 when the patient was actually healthy (total NIHSS score of 0) and did not have drift in their right arm.

4 out of 6 participants made at least 1 change upon seeing the UI in the second part of the experiment. Participant 1, for example, increased their motor leg score for Patient 3 by 1, from 0. Participant 2 also initially (and erroneously) thought Patient 1 was healthy but changed their answer to the scores presented to them in the UI. Interestingly, Participant 2 initially scored a deficit in Patient 5 that they then removed after seeing the UI, ultimately giving Patient 5 a healthy diagnosis (correct). For Patient 1, Participant 2's confidence decreased from 6/7 to 5/7. For Patient 5, their confidence increased from 5/7 to 6/7. After seeing the UI, Participant 4 changed the motor leg test NIHSS score for

Patient 4 from 1 to 0. But they maintained their total score of 1 and their confidence level of 5/7. Refer to Tables 5.2 and 5.3.

Participants 3 and 5, a neurologist and stroke fellow, respectively, did not make any changes. Participant 3 maintained a confidence rating of 7 for all cases.

Belief in the Accuracy of Displayed Results

We expected clinicians to change some of their decisions in response to being shown manipulated results. We wanted to better understand how unexpected or discordant predictions from the UI would influence participants' belief in the accuracy of results.

Table 5.4 shows the responses of participants when they were asked about the accuracy of the results shown on the UI in the second part of the experiment. Of particular interest are the manipulated patients – Patients 2, 3, and 6. While the results and data shown in the UI for (healthy) Patient 5 were deemed accurate by all of the participants, they unanimously distrusted the manipulated results shown for Patient 6. The belief in the accuracy of the displayed results for Patients 2 and 3 were more mixed.

Final Survey Results

After the experiment was over, clinicians were asked to complete a short survey about the prototype interface and to answer specific UI-related questions.

The survey revealed that despite not being in person, participants did not have difficulty diagnosing most stroke symptoms using audio and video footage and the developed prototype UI. Some participants did have difficulty assessing facial palsy and best gaze symptoms due to lighting conditions in the video. Overall, 3 out of the 4 responding clinicians indicated that the system was "generally accurate" and all 4 considered the interface to be "consistent". For the respondents, the color-coded skeleton proved to be

Table 5.3: Changes Made After Seeing the Prototype UI

	Patient 1		**Patient 2**		**Patient 3**		**Patient 4**		**Patient 5**		**Patient 6**		**Patient 7**	
Participant ID	**Conf.**	**Chg. Made (Y/N)**	**Conf.**	**Chg. Made (Yes/No)**	**Conf.**	**Chg. Made (Y/N)**	**Conf.**	**Chg. Made (Y/N)**	**Conf.**	**Chg. Made (Y/N)**	**Conf.**	**Chg. Made (Y/N)**	**Conf.**	**Chg. Made (Y/N)**
1	7/7	N	7/7	N	7/7	Y	7/7	N	6/7	N	7/7	Y	7/7	N
2	5/7	Y	6/7	N	5/7	N	6/7	N	6/7	Y	3/7	N	7/7	N
3	7/7	N	7/7	N	7/7	N	7/7	N	7/7	N	7/7	N	7/7	N
4	4/7	N	5/7	N	6/7	N	5/7	Y	6/7	N	5/7	N	5/7	N
5	4/7	N	3/7	N	4/7	N	6/7	N	5/7	N	4/7	N	6/7	N
6	6/7	N	6/7	N	6/7	N	6/7	N	6/7	N	6/7	Y	7/7	N

Chg. = Changes, Conf. = Confidence

Table 5.4: Participant Belief in the Accuracy of Displayed Results*

Participant ID	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7
1	No	Yes	No	Yes	Yes	No	No
2	Yes	No	No	Yes	Yes	No	Yes
3	No	No	Yes	No	Yes	No	Yes
4	No	Yes	Yes	Yes	Yes	No	No
5	No	Yes	Yes	Yes	Yes	No	Yes
6	No	Yes	No	Yes	Yes	No	Yes

*Highlighted in grey are the patients with manipulated results.

most helpful for their diagnoses while the joint movement graphs were not considered as important.

The final survey revealed that the joint position graphs were superfluous for diagnosis; the quantitative position information was too granular to be useful to clinicians. Categorical representations of the data would be more appropriate for future UI iterations.

5.3 Contributions to the Field

In this chapter, we discussed our efforts to understand how practitioners would make use of a computational aid created to support stroke diagnosis. Our study uncovered a number of interesting results regarding the effects of a computational support system on decision making and whether clinicians were able to correct for errors made by the system. We showed that generally, clinicians balanced predictions from aids with their own assessments without letting the former dominate their thinking. Clinicians who participated could easily identify when the system was incorrect and did not seem to blindly trust it. Lessons learned from this study provide guidance on properly designing computational aids for healthcare, and specifically how to best integrate them into diagnostic workflows for acute stroke diagnosis.

5.3.1 Effects of Computational Aids on Clinical Decision Making

An important concern raised by the participants of our focus group was the danger of relying too much on computational aids and how doing so would change the diagnostic ability of clinicians. Our results showed that computational aids did affect decision making, but only in specific instances. For example, as discussed in Section 5.2.2, Participant 1's changes for Patient 3 reflected the artificial increase in Patient 3's NIHSS score for the left motor leg. At first glance, it would seem that Participant 1 was completely swayed by our manipulations. However, Participant 1 noted that they suspected Patient 3's NIHSS score in the UI was too high, by looking at footage of the patient lifting her left leg against gravity. Our subtle manipulation of Patient 3 affected Participant 1's diagnosis even though they correctly identified it.

Participant 2's changes for Patient 1 and Patient 5 were also of note. In the case of Patient 1, Participant 2 identified deficits only after they saw what was displayed on the UI, initially misdiagnosing the patient as healthy but later correcting themself. For Patient 5, Participant 2 did the opposite, changing their diagnosis to healthy (correct). Participant 2's decision making process was especially impacted by the data displayed in the UI. Their confidence went down for Patient 1 after seeing the UI, and up for Patient 5 when the UI helped them fix their initial, incorrect assessment. On the other hand, while Participant 4 made a negligible change in the NIHSS motor leg score of Patient 4, their overall assessment did not seem to be affected by the UI.

Participant 1 and Participant 4 were able to incorporate results shown in the UI into their decision making without letting the technological aid dominate their thinking, in contrast with the behavior of Participant 2. **Clinicians seemed inclined to change their diagnoses only when the system definitively confirmed or denied their low confidence or "loosely held" beliefs, a conclusion supported by other work**

in the field of explainable AI [60]. The behavior of Participant 3 further emphasized this idea. Participant 3 was entirely unaffected by the UI, making no changes in the second part of the exam. Participant 3 reported perfect 7/7 confidence in their own assessments in both parts of the experiment, a possible justification for why their decision making was not affected by the UI – Participant 3 had no loosely held beliefs that could have been affected by the UI.

It is unclear why Participant 5 was unaffected by the UI, especially given that they did not identify the manipulations for Patients 2 and 3 and did not have perfect confidence in their own assessments. Further work needs to be done to understand why this individual was unaffected; perhaps issues with the design of the UI itself or the information presented may have discouraged Participant 5 from making use of the computational aid. (Participant 5 did not complete the final post-experiment survey.)

Note also that whether participants changed their responses in the second part of the exam did not relate to their role nor their experience level with the NIHSS. In fact, we did not notice any significant relationship between the roles (fellow, nurse practitioner, or neurologist) and how the UI affected decision making or their belief in the accuracy of results.

Clinicians Correctly Catch Significant Errors in the UI

The absence of an over-dependence on a computational system in our study is an important result to consider when thinking about future computational aids and the difficulty, especially in healthcare, of building an AI-based system that is always correct [64]. We demonstrated on a small scale that clinicians were aware of possible computational mistakes and were able to identify them when they occur. Future work at the intersection of AI and human-computer interaction needs to focus on how to make these instances

obvious; it should be made clear to clinicians that the results presented by computational aids are approximations and should be treated only as additional cues to incorporate in their clinical judgement, rather than a perfectly accurate diagnostic.

5.4 Acknowledgements

This chapter is largely a reprint of "Assessing Clinicians' Reliance on Computational Aids for Acute Stroke Diagnosis" published in the *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare* by authors Vishwajith Ramesh, Andrew Nguyen, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. It also addresses "Developing Aids to Assist Acute Stroke Diagnosis" published in the *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems – Late Breaking Work* by Vishwajith Ramesh, Stephanie Kim, Hong-An Nguyen, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. The chapter touches on follow-up work being prepared for submission at the time of writing this dissertation, "Designing a Clinical Decision Support System for Acute Stroke Diagnosis" by Vishwajith Ramesh, Stephanie Kim, Hong-An Nguyen, Andrew Nguyen, Gauri Iyer, Lisa M. Grega, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. The work covered in this chapter was supported by the NSF GRFP (DGE-1650112) and the UCSD Chancellor's Research Excellence Scholarship.

Chapter 6

Conclusion and Outlook

In **Chapter 1** we outlined the technical and human challenges of machine learning in healthcare. A major technical challenge is the small and imbalanced nature of real-world clinical datasets. It is difficult to rigorously evaluate the generalizability of machine and deep learning models for disease diagnosis and symptom identification trained on poor datasets. A human challenge unique to the healthcare space is the potential for over-reliance on AI-based clinical decision support systems. Through examples in neurology and pulmonology, we highlighted solutions for both these concerns. We summarize these contributions and discuss their broader impacts here.

In **Chapter 2**, we motivated a clinical decision support system for acute stroke diagnosis. As a first step, we developed a machine learning pipeline that used body angles from the Microsoft Kinect depth camera and the leave-one-out training strategy to classify hemiparesis or weakness with perfect accuracy. We addressed limitations with this initial approach – the lack of interpretability of features and the difficulty of obtaining reliable body angles from patients – by developing a video-based classification pipeline. Our video-based approach used the covariance matrix of body joint positions as a feature descriptor, which not only had high discriminatory potential but also could be easily

interpreted. We rigorously evaluated this pipeline by using a test set unseen during training and obtained high accuracy exceeding that of 8 stroke specialists conducting a video-based stroke assessment. This work highlighted the pros and cons of the leave-one-out training paradigm regularly used in machine learning in healthcare. The strategy allows for the maximal use of limited datasets for training and yields high classification accuracy. An unfortunate drawback, however, is that such performance is often overly optimistic and it is possible for models trained with leave-one-out to overfit. Regardless, leave-one-out enabled us to identify gaps in our stroke dataset by evaluating our classifier's performance on a subject-to-subject basis. Doing so enabled us to focus our data collection efforts going forward, namely on those with severe hemiparesis unable to sit upright. More broadly, identifying weakness from video of subjects simply sitting at rest is a highly scalable approach and primed for real-world deployment. We envision deploying this classifier as part of a clinical decision support system for stroke in emergency settings without ready access to neurologists. By automating stroke severity assessment with a real-time system, our goal is to increase speed of diagnosis and reduce missed strokes by emergency medical physicians.

In **Chapter 3**, we leveraged the audio characteristics of coughs to create classifiers that could distinguish common respiratory diseases in adults. Moreover, we built on recent advances in generative adversarial networks (GANs) to augment our dataset with cleverly engineered synthetic cough samples for each class of major respiratory disease, to balance and increase our dataset size. We experimented on cough samples collected with a smartphone from 45 subjects in a clinic. Our CoughGAN-improved support vector machine and random forest models showed up to 76% test accuracy and 83% F1 score in classifying subjects' conditions between healthy and three major respiratory diseases. Adding our synthetic coughs improved the performance we could obtain from a relatively small unbalanced

healthcare dataset by boosting the accuracy over 30%. Our data augmentation reduced overfitting and discouraged the prediction of a single, dominant class. This work was of particular importance because despite the prevalence of respiratory diseases, their diagnosis by clinicians is challenging. Accurately assessing airway sounds requires extensive clinical training and equipment that may not be easily available. Current methods that automate this diagnosis are hindered by their use of features that require pulmonary function tests. Our work paves the way for an automatic, cough-based respiratory disease diagnosis using smartphones or wearables in-the-wild.

In **Chapter 4**, we presented a solution that used a single wearable inertial sensor to automatically assess the gait of a Parkinson's disease patient and predict the postural instability and gait disorder sub-score of the Unified Parkinson's Disease Rating Scale. We showed that for a 2 minute walk test, our deep learning method's predicted PIGD scores could be used to identify a patient's fluctuations in motor state better than a physician evaluated on the same criteria. We used adversarial training via GANs to overcome the curse of dimensionality of the training dataset. Note that the longitudinal progression of Parkinson's disease is monitored using episodic assessments performed by trained physicians in the clinic, whereas our deep learning method has the potential to be deployed in-the-wild for reliable, continuous tracking of patients' symptoms and fluctuations.

In **Chapter 5**, we revisited the clinical decision support system for stroke motivated earlier but from a human-computer interaction perspective. In this work, we developed a high-fidelity user interface for a computational aid designed to support acute ischemic stroke diagnosis. Engaging with practitioners at the UCSD Stroke Center, we conducted an experiment to determine how technology for identifying stroke symptoms affected their diagnostic decision-making processes. Other studies have assessed reliance on clinical decision support systems in fields like diabetes, but no such study exists for stroke diagnosis.

By assessing how clinicians changed their video-based diagnosis when provided with data visualizations and predictions from a machine learning tool, we observed that such computational aids do in fact affect clinicians' decisions but only in cases when the aid directly supported or contradicted their prior beliefs. This experiment is an example of how to measure over-dependence on AI-based technology, an important concern for clinicians. The results emphasized that future computational aids for stroke diagnosis should focus on helping clinicians solidify their decisions rather than only providing them with overly quantitative information that may impede or confuse their judgement.

While the contributions of this dissertation were in the focus areas of stroke, respiratory disease, and Parkinson's disease, the lessons learned and techniques developed can be translated to any healthcare domain in which machine learning can be used. We hope that continued developments in this field are cognizant of the challenges unique to healthcare and medicine because there is a long ways to go for AI to have the same impact in health as it has had in other fields.

Bibliography

[1] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, pp. 120–144, 2009.

[2] F. Falcini, G. Lami, and A. M. Costanza, "Deep learning in automotive software," *IEEE Software*, vol. 34, no. 3, pp. 56–63, 2017.

[3] "Autopilot," 2020.

[4] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[5] S. Byford, "How ai is changing photography," *The Verge*, 2019.

[6] L. Verdoliva, "Media forensics and deepfakes: an overview," *arXiv preprint arXiv:2001.06564*, 2020.

[7] C. Vaccari and A. Chadwick, "'deepfakes' are here. these deceptive videos erode trust in all news media.," *The Washington Post*, 2020.

[8] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[9] J. Leventhal, "In a crash, should self-driving cars save passengers or pedestrians? 2 million people weigh in," *PBS News Hour*, 2018.

[10] M. W. Whalen, D. Cofer, and A. Gacek, "Requirements and architectures for secure vehicles," *IEEE Software*, vol. 33, pp. 22–25, jul 2016.

[11] C. Molina, J. Almeida, L. F. Vismari, R. R. Gonzalez, J. K. Naufal, and J. Camargo, "Assuring fully autonomous vehicles safety by design: The autonomous vehicle control (avc) module strategy," in *2017 47th Annual IEEE/IFIP International Conference on*

Dependable Systems and Networks Workshop (DSN-W), (Los Alamitos, CA, USA), pp. 16–21, IEEE Computer Society, jun 2017.

[12] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media+ Society*, vol. 6, no. 1, p. 2056305120903408, 2020.

[13] A. Ovadya and J. Whittlestone, "Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning," *arXiv preprint arXiv:1907.11274*, 2019.

[14] R. L. Hasen, "Deep fakes, bots, and siloed justices: American election law in a post-truth world," *St. Louis University Law Journal*, 2019.

[15] S. Gerke, B. Babic, T. Evgeniou, and I. G. Cohen, "The need for a system view to regulate artificial intelligence/machine learning-based software as medical device," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–4, 2020.

[16] Software as a Medical Device Working Group, "Software as a Medical Device (SaMD): Clinical Evaluation," 2017.

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

[18] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy, "Openimages: A public dataset for large-scale multi-label and multi-class image classification.," *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[19] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[20] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 and cifar-100 datasets," 2009.

[21] Y. LeCun, "The mnist database of handwritten digits," 1998.

[22] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[23] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019.

[24] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," *arXiv preprint arXiv:2001.00964*, 2020.

[25] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

[26] V. Song, "Report: Google contractors used shady methods to scan dark-skinned people's faces for new pixel 4 feature," *Gizmodo*, 2019.

[27] S. Scrivens, "[update: Google responds to concerns] google is stopping people in the street to buy their face data — possibly for pixel 4's 'face id'," *Android Police*, 2019.

[28] C. Matyszczyk, "Google bought my friend's face for $5," *ZDNet*, 2019.

[29] L. C. Lovato, K. Hill, S. Hertert, D. B. Hunninghake, and J. L. Probstfield, "Recruitment for controlled clinical trials: literature summary and annotated bibliography," *Controlled clinical trials*, vol. 18, no. 4, pp. 328–352, 1997.

[30] G. M. Swanson and A. J. Ward, "Recruiting minorities into clinical trials toward a participant-friendly system," *JNCI Journal of the National Cancer Institute*, vol. 87, no. 23, pp. 1747–1759, 1995.

[31] S. Chowdhury, C. C. Meunier, L. Cappelletti, and T. B. Sherer, "Improving patient participation in parkinson's clinical trials: The experience of the michael j fox foundation," *Clinical Investigation*, vol. 4, no. 2, pp. 185–192, 2014.

[32] S. Berk, B. L. Greco, K. Biglan, C. M. Kopil, R. G. Holloway, C. Meunier, and T. Simuni, "Increasing efficiency of recruitment in early parkinson's disease trials: a case study examination of the steady-pd iii trial," *Journal of Parkinson's disease*, vol. 7, no. 4, pp. 685–693, 2017.

[33] S. Hoffman, "Regulating clinical research: informed consent, privacy, and irbs," *Cap. UL Rev.*, vol. 31, p. 71, 2003.

[34] C. Nebeker, J. Harlow, R. Espinoza Giacinto, R. Orozco-Linares, C. S. Bloss, and N. Weibel, "Ethical and regulatory challenges of research using pervasive sensing and other emerging technologies: Irb perspectives," *AJOB empirical bioethics*, vol. 8, no. 4, pp. 266–276, 2017.

[35] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[36] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[37] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: model selection and overfitting," 2016.

[38] G. C. Smith, S. R. Seaman, A. M. Wood, P. Royston, and I. R. White, "Correcting for optimistic prediction in small data sets," *American journal of epidemiology*, vol. 180, no. 3, pp. 318–324, 2014.

[39] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[40] G. Vandewiele, I. Dehaene, G. Kovács, L. Sterckx, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, J. Decruyenaere, S. Van Hoecke, and T. Demeester, "Overly optimistic prediction results on imbalanced data: Flaws and benefits of applying over-sampling," *arXiv preprint arXiv:2001.06296*, 2020.

[41] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and semiparametric models*. Springer Science & Business Media, 2012.

[42] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007.

[43] H. Sun, K. Depraetere, J. De Roo, G. Mels, B. De Vloed, M. Twagirumukiza, and D. Colaert, "Semantic processing of ehr data for clinical research," *Journal of biomedical informatics*, vol. 58, pp. 247–259, 2015.

[44] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[45] L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, "Errors in health care: a leading cause of death and injury," in *To err is human: Building a safer health system*, National Academies Press (US), 2000.

[46] M. A. Makary and M. Daniel, "Medical error—the third leading cause of death in the us," *Bmj*, vol. 353, 2016.

[47] J. T. James, "A new, evidence-based estimate of patient harms associated with hospital care," *Journal of patient safety*, vol. 9, no. 3, pp. 122–128, 2013.

[48] R. D. Hart, "Who's to blame when a machine botches your surgery?," *Quartz*, 2018.

[49] W. N. Price, S. Gerke, and I. G. Cohen, "Potential liability for physicians using artificial intelligence," *Jama*, vol. 322, no. 18, pp. 1765–1766, 2019.

[50] S. Jha, "Can you sue an algorithm for malpractice? it depends," *Stat News*, 2020.

[51] E. Brynjolfsson and L. M. Hitt, "Beyond computation: Information technology, organizational transformation and business performance," *Journal of Economic perspectives*, vol. 14, no. 4, pp. 23–48, 2000.

[52] C. D. Lehman, R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, and D. L. Miglioretti, "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," *JAMA internal medicine*, vol. 175, no. 11, pp. 1828–1837, 2015.

[53] A. Moxey, J. Robertson, D. Newby, I. Hains, M. Williamson, and S.-A. Pearson, "Computerized clinical decision support for prescribing: provision does not guarantee uptake," *Journal of the American Medical Informatics Association*, vol. 17, no. 1, pp. 25–33, 2010.

[54] R. L. Teach and E. H. Shortliffe, "An analysis of physician attitudes regarding computer-based clinical consultation systems," *Computers and Biomedical Research*, vol. 14, no. 6, pp. 542–558, 1981.

[55] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial intelligence in medicine*, vol. 94, pp. 42–53, 2019.

[56] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI-17 workshop on explainable AI (XAI)*, vol. 8, 2017.

[57] G. O. Barnett, J. J. Cimino, J. A. Hupp, and E. P. Hoffer, "Dxplain: an evolving diagnostic decision-support system," *Jama*, vol. 258, no. 1, pp. 67–74, 1987.

[58] I. Adams, M. Chan, P. Clifford, W. Cooke, V. Dallos, F. De Dombal, M. Edwards, D. Hancock, D. Hewett, and N. McIntyre, "Computer aided diagnosis of acute abdominal pain: a multicentre study.," *Br Med J (Clin Res Ed)*, vol. 293, no. 6550, pp. 800–804, 1986.

[59] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, "Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system," *Computers and biomedical research*, vol. 8, no. 4, pp. 303–320, 1975.

[60] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2019.

[61] D. J. Hilton and B. R. Slugoski, "Knowledge-based causal attribution: The abnormal conditions focus model.," *Psychological review*, vol. 93, no. 1, p. 75, 1986.

[62] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[63] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[64] P. Croskerry, "Clinical cognition and diagnostic error: applications of a dual process model of reasoning," *2*, vol. 14, no. 1, pp. 27–35, 2009.

[65] S. S. Virani, A. Alonso, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, F. N. Delling, L. Djousse, M. S. Elkind, J. F. Ferguson, M. Fornage, S. S. Khan, B. M. Kissela, K. L. Knutson, T. W. Kwan, D. T. Lackland, T. T. Lewis, J. H. Lichtman, C. T. Longenecker, M. Shane Loop, P. L. Lutsey, S. S. Martin, K. Matsushita, A. E. Moran, M. E. Mussolino, A. Marma Perak, W. D. Rosamond, G. A. Roth, U. K. Sampson, G. M. Satou, E. B. Schroeder, S. H. Shah, C. M. Shay, N. L. Spartano, A. Stokes, D. L. Tirschwell, L. B. VanWagner, and C. W. Tsao, "Heart disease and stroke statistics-2020 update: a report from the american heart association.," *Circulation*, vol. 141, no. 9, pp. e139–e596, 2020.

[66] S. Facts, "*Center for Disease Control and Prevention*," 2020.

[67] S.-F. Sung and M.-C. Tseng, "Code stroke: A mismatch between number of activation and number of thrombolysis," *Journal of the Formosan Medical Association*, vol. 113, no. 7, pp. 442–446, 2014.

[68] G. Jackson and K. Chari, "National hospital care survey demonstration projects: Stroke inpatient hospitalizations," *National Health Statistics Reports*, vol. 132, 2020.

[69] D. E. Newman-Toker, E. Moy, E. Valente, R. Coffey, and A. L. Hines, "Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample," *Diagnosis*, vol. 1, no. 2, pp. 155–166, 2014.

[70] A. E. Arch, D. C. Weisman, S. Coca, K. V. Nystrom, C. R. Wira III, and J. L. Schindler, "Missed ischemic stroke diagnosis in the emergency department by emergency medicine and neurology services," *Stroke*, vol. 47, no. 3, pp. 668–673, 2016.

[71] K. A. Kerber, D. L. Brown, L. D. Lisabeth, M. A. Smith, and L. B. Morgenstern, "Stroke among patients with dizziness, vertigo, and imbalance in the emergency department: a population-based study," *Stroke*, vol. 37, no. 10, pp. 2484–2487, 2006.

[72] K. Berekashvili, A. M. Zha, M. Abdel-Al, X. Zhang, J. H. Soomro, S. J. Prater, and J. C. Grotta, "Emergency medicine physicians accurately select acute stroke patients for tissue-type plasminogen activator treatment using a checklist," *Stroke*, pp. STROKEAHA–119, 2019.

[73] N. Goyal, S. Male, A. Al Wafai, S. Bellamkonda, and R. Zand, "Cost burden of stroke mimics and transient ischemic attack after intravenous tissue plasminogen activator treatment," *Journal of Stroke and Cerebrovascular Diseases*, vol. 24, no. 4, pp. 828–833, 2015.

[74] A. Kachalia, T. K. Gandhi, A. L. Puopolo, C. Yoon, E. J. Thomas, R. Griffey, T. A. Brennan, and D. M. Studdert, "Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers," *Annals of emergency medicine*, vol. 49, no. 2, pp. 196–205, 2007.

[75] M. J. Moore, J. Stuart, A. Humphreys, and J. A. Pfaff, "To tpa or not to tpa: Two medical-legal misadventures of diagnosing a cerebrovascular accident as a stroke mimic," *Clinical practice and cases in emergency medicine*, vol. 3, no. 3, p. 194, 2019.

[76] J. Jia, R. Band, M. E. Abboud, W. Pajerowski, M. Guo, G. David, C. C. Mechem, S. R. Messé, B. G. Carr, and M. T. Mullen, "Accuracy of emergency medical services dispatcher and crew diagnosis of stroke in clinical practice," *Frontiers in neurology*, vol. 8, p. 466, 2017.

[77] T. E. Madsen, J. Khoury, R. Cadena, O. Adeoye, K. A. Alwell, C. J. Moomaw, E. McDonough, M. L. Flaherty, S. Ferioli, D. Woo, P. Khatri, J. P. Broderick, B. M. Kissela, and D. Kleindorfer, "Potentially missed diagnosis of ischemic stroke in the emergency department in the greater cincinnati/northern kentucky stroke study," *Academic Emergency Medicine*, vol. 23, no. 10, pp. 1128–1135, 2016.

[78] A. Tehrani, H. Lee, S. Mathews, A. Shore, M. Makary, and P. Pronovost, "Diagnostic errors more common, costly and harmful than treatment mistakes," *John Hopkins Medicine*, 2013.

[79] T. C. Weiss, "Hemiparesis - types, treatment, facts and information," 2017.

[80] American Stroke Association, "Hemiparesis," 2019.

[81] S. Association, "Physical effects of stroke," 2013.

[82] T. Brott, H. P. Adams, C. P. Olinger, J. R. Marler, W. G. Barsan, J. Biller, J. Spilker, R. Holleran, R. Eberle, and V. Hertzberg, "Measurements of Acute Cerebral Infarction: A Clinical Examination Scale," *Stroke*, vol. 20, pp. 864–870, July 1989.

[83] National Institute of Neurological Disorders and Stroke, "NIH Stroke Scale," 2016.

[84] B. C. Meyer, T. M. Hemmen, C. M. Jackson, and P. D. Lyden, "Modified National Institutes of Health Stroke Scale for Use in Stroke Clinical Trials," *Stroke*, vol. 33, no. 5, pp. 1261–1266, 2002.

[85] H. Lee, J. Kim, H. Myoung, J. Lee, and K. Lee, "Repeatability of the Accelerometric-Based Method to Detect Step Events for Hemiparetic Stroke Patients," in *EMBC*, 2011.

[86] E. Haeuber, M. Shaughnessy, L. W. Forrester, K. L. Coleman, and R. F. Macko, "Accelerometer Monitoring of Home- and Community-Based Ambulatory Activity After Stroke," *Archives of Physical Medicine and Rehabilitation*, vol. 85, no. 12, pp. 1997–2001, 2004.

[87] D. Popovic and T. Sinkjaer, *Control of Movement for the Physically Disabled: Chapter 4 (Restoring Movement: State of the Art)*, ch. 4. Springer Science and Business Media, 2012.

[88] N. Genthon, N. Vuillerme, J. Monnet, C. Petit, and P. Rougier, "Biomechanical Assessment of the Sitting Posture Maintenance in Patients with Stroke," *Clinical Biomechanics*, vol. 22, July 2007.

[89] D. Nichols, L. Miller, L. Colby, and W. Pease, "Sitting Balance: Its Relation to Function in Individuals with Hemiparesis," *Archives of Physical Medicine and Rehabilitation*, vol. 77, Sept. 1996.

[90] N. Weibel, S. Rick, C. Emmenegger, S. Ashfaq, A. Calvitti, and Z. Agha, "LAB-IN-A-BOX: Semi-Automatic Tracking of Activity in the Medical Office," *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 317–334, 2015.

[91] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[92] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification." *www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf*, 2016.

[93] V. Ramesh, S. Rick, B. Meyer, G. Cauwenberghs, and N. Weibel, "A Neurobehavioral Evaluation System Using 3D Depth Tracking & Computer Vision: The Case of

Stroke-Kinect.," in *Proceedings of Neuroscience 2016, Annual Meeting of the Society for Neuroscience (Poster presentation)*, (San Diego, CA, USA), Nov. 2016.

[94] V. Ramesh, K. Agrawal, B. Meyer, G. Cauwenberghs, and N. Weibel, "Exploring stroke-associated hemiparesis assessment with support vector machines," in *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 464–467, ACM, 2017.

[95] V. Ramesh, K. Agrawal, B. Meyer, G. Cauwenberghs, and N. Weibel, "Stroke-associated hemiparesis detection using body joints and support vector machines," in *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 55–58, ACM, 2018.

[96] D. B. Popovic and T. Sinkjaer, *Control of movement for the physically disabled: control for rehabilitation technology*. Center for Sensory-Motor Interaction (SMI), Department of Health Science and . . . , 2003.

[97] N. Genthon, N. Vuillerme, J. Monnet, C. Petit, and P. Rougier, "Biomechanical assessment of the sitting posture maintenance in patients with stroke," *Clinical biomechanics*, vol. 22, no. 9, pp. 1024–1029, 2007.

[98] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[99] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[100] K. Agrawal, R. Raman, K. Ernstrom, R. J. Claycomb, D. M. Meyer, T. M. Hemmen, R. F. Modir, P. Kachhi, and B. C. Meyer, "Accuracy of stroke diagnosis in telestroke-guided tissue plasminogen activator patients," *Journal of Stroke and Cerebrovascular Diseases*, vol. 25, no. 12, pp. 2942–2946, 2016.

[101] B. M. Demaerschalk, S. Vegunta, B. B. Vargas, Q. Wu, D. D. Channer, and J. G. Hentz, "Reliability of real-time video smartphone for assessing national institutes of health stroke scale scores in acute stroke patients," *Stroke*, vol. 43, no. 12, pp. 3271–3277, 2012.

[102] J. J. Randolph, "Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa.," *Online submission*, 2005.

[103] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[104] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[105] V. Van Belle, B. Van Calster, S. Van Huffel, J. A. Suykens, and P. Lisboa, "Explaining support vector machines: a color based nomogram," *PloS one*, vol. 11, no. 10, p. e0164568, 2016.

[106] A. Y. Ng, "Preventing" overfitting" of cross-validation data," in *ICML*, vol. 97, pp. 245–253, Citeseer, 1997.

[107] American Lung Association, "Chronic obstructive pulmonary disease (copd)," 2019.

[108] Center for Disease Control and Prevention, "Most recent national asthma data," 2019.

[109] C. M. Oliver, S. A. Hunter, T. Ikeda, and D. C. Galletly, "Junior doctor skill in the art of physical examination: a retrospective study of the medical admission note over four decades," *BMJ Open*, vol. 3, no. 4, 2013.

[110] D. Spence, "Bad medicine: chest examination," *Bmj*, vol. 345, p. e4569, 2012.

[111] T. Cherian, E. K. Mulholland, J. B. Carlin, H. Ostensen, R. Amin, M. d. Campo, D. Greenberg, R. Lagos, M. Lucero, S. A. Madhi, K. L. O'Brien, S. Obaro, M. C. Steinhoff, and W. R. W. Group, "Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies," *Bulletin of the World Health Organization*, vol. 83, pp. 353–359, 2005.

[112] C. Bada, N. Y. Carreazo, J. P. Chalco, and L. Huicho, "Inter-observer agreement in interpreting chest x-rays on children with acute lower respiratory tract infections and concurrent wheezing," *Sao Paulo Medical Journal*, vol. 125, no. 3, pp. 150–154, 2007.

[113] A. Machado, A. Oliveira, J. Aparício, and A. Marques, "Respiratory auscultation: (dis)agreement between health professionals," *European Respiratory Journal*, vol. 46, no. suppl 59, 2015.

[114] A. C. Herrera, M. M. de Oca, M. V. L. Varela, C. Aguirre, E. Schiavi, J. R. Jardim, and P. Team, "Copd underdiagnosis and misdiagnosis in a high-risk primary care population in four latin american countries. a key to enhance disease diagnosis: the puma study," *PloS one*, vol. 11, no. 4, p. e0152266, 2016.

[115] A. Badnjevic, L. Gurbeta, and E. Custovic, "An expert diagnostic system to automatically identify asthma and chronic obstructive pulmonary disease in clinical settings," *Scientific reports*, vol. 8, no. 1, p. 11645, 2018.

[116] M. Al-khassaweneh and R. Bani Abdelrahman, "A signal processing approach for the diagnosis of asthma from cough sounds," *Journal of medical engineering & technology*, vol. 37, no. 3, pp. 165–171, 2013.

[117] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PloS one*, vol. 11, no. 9, p. e0162128, 2016.

[118] Y. Amrulloh, U. Abeyratne, V. Swarnkar, and R. Triasih, "Cough sound analysis for pneumonia and asthma classification in pediatric population," in *2015 6th International Conference on Intelligent Systems, Modelling and Simulation*, pp. 127–131, IEEE, 2015.

[119] P. Porter, U. Abeyratne, V. Swarnkar, J. Tan, T.-w. Ng, J. M. Brisbane, D. Speldewinde, J. Choveaux, R. Sharan, K. Kosasih, and P. Della, "A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children," *Respiratory research*, vol. 20, no. 1, p. 81, 2019.

[120] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Annals of biomedical engineering*, vol. 41, no. 11, pp. 2448–2462, 2013.

[121] D. Liaqat, R. Wu, A. Gershon, H. Alshaer, F. Rudzicz, and E. de Lara, "Challenges with real-world smartwatch based audio monitoring," in *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, WearSys '18, (New York, NY, USA), pp. 54–59, ACM, 2018.

[122] M. M. Rahman, V. Nathan, E. Nemati, K. Vatanparvar, M. Ahmed, and J. Kuang, "Towards reliable data collection and annotation to extract pulmonary digital biomarkers using mobile sensors," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 179–188, ACM, 2019.

[123] E. C. Larson, M. Goel, G. Boriello, S. Heltshe, M. Rosenfeld, and S. N. Patel, "Spirosmart: using a microphone to measure lung function on a mobile phone," in *Proceedings of the 2012 ACM Conference on ubiquitous computing*, pp. 280–289, ACM, 2012.

[124] V. Viswanath, J. Garrison, and S. Patel, "Spiroconfidence: determining the validity of smartphone based spirometry using machine learning," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5499–5502, IEEE, 2018.

[125] E. Nemati, M. M. Rahman, V. Nathan, K. Vatanparvar, and J. Kuang, "A comprehensive approach for cough type detection," in *2019 IEEE/ACM International*

Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 15–16, IEEE, 2019.

[126] E. Nemati, M. M. Rahman, V. Nathan, and J. Kuang, "Private audio-based cough sensing for in-home pulmonary assessment using mobile devices," in *Proceedings of the 13th EAI International Conference on Body Area Networks*, 2018.

[127] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[128] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331, 1983.

[129] A. J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," in *Advances in neural information processing systems*, pp. 3236–3246, 2017.

[130] T. Golany and K. Radinsky, "Pgans: Personalized generative adversarial networks for ecg synthesis to improve patient-specific deep ecg classification," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[131] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293, 2018.

[132] G. Chen, Y. Zhu, Z. Hong, and Z. Yang, "Emotionalgan: Generating ecg to enhance emotion state classification," in *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*, pp. 309–313, 2019.

[133] K. Vatanparvar, V. Nathan, E. Nemati, M. Rahman, and J. Kuang, "A Generative Model for Speech Segmentation and Obfuscation for Remote Health Monitoring," in *IEEE International Conference on Body Sensor Networks (BSN)*, 2019.

[134] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *ICLR*, 2019.

[135] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, Andrew-Harp, Geoffrey-Irving, Michael-Isard, Yangqing-Jia, Rafal-Jozefowicz, Lukasz-Kaiser, Manjunath-Kudlur, Josh-Levenberg, Dan-Mane, Rajat-Monga, Sherry-Moore, Derek-Murray, Chris-Olah, Mike-Schuster, Jonathon-Shlens, Benoit-Steiner, Ilya-Sutskever, Kunal-Talwar, Paul-Tucker, Vincent-Vanhoucke, Vijay-Vasudevan, Fernanda-Viegas, Oriol-Vinyals, Pete-Warden, Martin-Wattenberg, Martin-Wicke, Yuan-Yu, and Xiaoqiang-Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[136] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, pp. 5767–5777, 2017.

[137] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[138] J. L. Palmer, M. A. Coats, C. M. Roe, S. M. Hanko, C. Xiong, and J. C. Morris, "Unified parkinson's disease rating scale-motor exam: inter-rater reliability of advanced practice nurse and neurologist assessments," *Journal of advanced nursing*, vol. 66, no. 6, pp. 1382–1387, 2010.

[139] W. Maetzler, J. Domingos, K. Srulijes, J. J. Ferreira, and B. R. Bloem, "Quantitative wearable sensors for objective assessment of parkinson's disease," *Movement Disorders*, vol. 28, no. 12, pp. 1628–1637, 2013.

[140] D. Rodríguez-Martín, A. Samà, C. Pérez-López, A. Català, J. M. M. Arostegui, J. Cabestany, À. Bayés, S. Alcaine, B. Mestre, A. Prats, M. C. Crespo, T. J. Counihan, P. Browne, L. R. Quinlan, G. ÓLaighin, D. Sweeney, H. Lewy, J. Azuri, G. Vainstein, R. Annicchiarico, A. Costa, and A. Rodríguez-Molinero, "Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer," *PloS one*, vol. 12, no. 2, p. e0171764, 2017.

[141] J. Camps, A. Sama, M. Martin, D. Rodriguez-Martin, C. Perez-Lopez, J. M. M. Arostegui, J. Cabestany, A. Catala, S. Alcaine, B. Mestre, A. Prats, M. C. Crespo, T. J. Counihan, P. Browne, L. R. Quinlan, G. ÓLaighin, D. Sweeney, H. Lewy, G. Vainstein, A. Costa, R. Annicchiarico, A. Bayes, and A. Rodríguez-Molinero, "Deep learning for freezing of gait detection in parkinson's disease patients in their homes using a waist-worn inertial measurement unit," *Knowledge-Based Systems*, vol. 139, pp. 119–131, 2018.

[142] M. B. Davidson, D. J. McGhee, and C. E. Counsell, "Comparison of patient rated treatment response with measured improvement in parkinson's disease," *J Neurol Neurosurg Psychiatry*, vol. 83, no. 10, pp. 1001–1005, 2012.

[143] R. Bhidayasiri and D. Tarsy, *Movement disorders: a video atlas*. Springer Science & Business Media, 2012.

[144] A. Lees, "The on-off phenomenon.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 52, no. Suppl, pp. 29–37, 1989.

[145] S. Papapetropoulos, "Patient diaries as a clinical endpoint in parkinson's disease clinical trials," *CNS neuroscience & therapeutics*, vol. 18, no. 5, pp. 380–387, 2012.

[146] T. Rastin, M. Amstrong, A. Gagliardi, A. Grabovsky, and C. Marras, "Communication about off periods in parkinson's disease: A survey of physicians, patients and carepartners.," *Frontiers in neurology*, vol. 10, p. 892, 2019.

[147] R. A. Hauser, J. Friedlander, T. A. Zesiewicz, C. H. Adler, L. C. Seeberger, C. F. O'Brien, E. S. Molho, and S. A. Factor, "A home diary to assess functional status in patients with parkinson's disease with motor fluctuations and dyskinesia," *Clinical neuropharmacology*, vol. 23, no. 2, pp. 75–81, 2000.

[148] R. A. Hauser, F. Deckers, and P. Lehert, "Parkinson's disease home diary: further validation and implications for clinical trials," *Movement Disorders*, vol. 19, no. 12, pp. 1409–1413, 2004.

[149] C. Goetz, S. Luo, and G. Stebbins, "Modeling the effect of patient's perception of non-motor and motor function on parkinson's disease severity: 1173," *Movement Disorders*, vol. 34, 2019.

[150] M. K. Erb, D. R. Karlin, B. K. Ho, K. C. Thomas, F. Parisi, G. P. Vergara-Diaz, J.-F. Daneault, P. W. Wacnik, H. Zhang, T. Kangarloo, C. Demanuele, C. R. Brooks, C. N. Detheridge, N. Shaafi Kabiri, J. S. Bhangu, and P. Bonato, "mhealth and wearable technology should replace motor diaries to track motor fluctuations in parkinson's disease," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[151] J. A. Vizcarra, Á. Sánchez-Ferro, W. Maetzler, L. Marsili, L. Zavala, A. E. Lang, P. Martinez-Martin, T. A. Mestre, R. Reilmann, J. M. Hausdorff, E. R. Dorsey, S. S. Paul, J. W. Dexheimer, B. D. Wissel, R. L. M. Fuller, P. Bonato, H. Tan, B. R. Bloem, C. Kopil, M. Daeschler, L. Bataille, G. Kleiner, J. M. Cedarbaum, J. Klucken, A. Merola, C. G. Goetz, G. T. Stebbins, and A. J. Espay, "The parkinson's disease e-diary: Developing a clinical and research tool for the digital age," *Movement Disorders*, vol. 34, no. 5, pp. 676–681, 2019.

[152] K. E. Lyons and R. Pahwa, "Electronic motor function diary for patients with parkinson's disease: a feasibility study," *Parkinsonism & related disorders*, vol. 13, no. 5, pp. 304–307, 2007.

[153] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. van Hilten, and N. LaPelle, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[154] A. Samà, C. Pérez-López, D. Rodríguez-Martín, A. Català, J. M. Moreno-Aróstegui, J. Cabestany, E. de Mingo, and A. Rodríguez-Molinero, "Estimating bradykinesia severity in parkinson's disease by analysing gait through a waist-worn sensor," *Computers in biology and medicine*, vol. 84, pp. 114–123, 2017.

[155] E. Rastegari, S. Azizian, and H. Ali, "Machine learning and similarity network approaches to support automatic classification of parkinson's diseases using accelerometer-based gait analysis," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[156] E. Rovini, C. Maremmani, A. Moschetti, D. Esposito, and F. Cavallo, "Comparative motor pre-clinical assessment in parkinson's disease using supervised machine learning approaches," *Annals of biomedical engineering*, vol. 46, no. 12, pp. 2057–2068, 2018.

[157] T. Chomiak, W. Xian, Z. Pei, and B. Hu, "A novel single-sensor-based method for the detection of gait-cycle breakdown and freezing of gait in parkinson's disease," *Journal of Neural Transmission*, pp. 1–8, 2019.

[158] A. Samà, D. Rodríguez-Martín, C. Pérez-López, A. Català, S. Alcaine, B. Mestre, A. Prats, M. C. Crespo, and À. Bayés, "Determining the optimal features in freezing of gait detection through a single waist accelerometer in home environments," *Pattern Recognition Letters*, vol. 105, pp. 135–143, 2018.

[159] E. E. Tripoliti, A. T. Tzallas, M. G. Tsipouras, G. Rigas, P. Bougia, M. Leontiou, S. Konitsiotis, M. Chondrogiorgi, S. Tsouli, and D. I. Fotiadis, "Automatic detection of freezing of gait events in patients with parkinson's disease," *Computer methods and programs in biomedicine*, vol. 110, no. 1, pp. 12–26, 2013.

[160] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[161] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[162] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[163] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[164] K. Erb, J. Daneault, S. Amato, P. Bergethon, C. Demanuele, T. Kangarloo, S. Patel, V. Ramos, D. Volfson, P. Wacnik, H. Zhang, D. Karlin, H. Huggins, L. Soll, G. Costante, G. Vergara-Dia, F. Parisi, J. Banghu, C. Brooks, C. Dethridge,

A. Abrami, E. Bilal, V. Caravagio, S. Heisig, R. Norel, E. Pissadaki, J. Rice, B. Ho, K. Thomas, and P. Bonato, "The bluesky project: Monitoring motor and non-motor characteristics of people with parkinson's disease in the laboratory, a simulated apartment, and home and community settings," in *Movement Disorders*, vol. 33, pp. 1990–1990, Wiley 111 River ST, Hoboken 07030-5774, NJ USA, 2018.

[165] T. Stillerova, J. Liddle, L. Gustafsson, R. Lamont, and P. Silburn, "Remotely assessing symptoms of parkinson's disease using videoconferencing: a feasibility study," *Neurology research international*, vol. 2016, 2016.

[166] D. Rodriguez-Martin, A. Sama, C. Perez-Lopez, A. Catala, J. Cabestany, and A. Rodriguez-Molinero, "Svm-based posture identification with a single waist-located triaxial accelerometer," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7203–7211, 2013.

[167] H. Gjoreski, M. Lustrek, and M. Gams, "Accelerometer placement for posture recognition and fall detection," in *2011 Seventh International Conference on Intelligent Environments*, pp. 47–54, IEEE, 2011.

[168] C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," *Sensors*, vol. 10, no. 8, pp. 7772–7788, 2010.

[169] C. V. Bouten, K. T. Koekkoek, M. Verduin, R. Kodde, and J. D. Janssen, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," *IEEE transactions on biomedical engineering*, vol. 44, no. 3, pp. 136–147, 1997.

[170] R. P. Hubble, G. A. Naughton, P. A. Silburn, and M. H. Cole, "Wearable sensor use for assessing standing balance and walking stability in people with parkinson's disease: a systematic review," *PloS one*, vol. 10, no. 4, p. e0123705, 2015.

[171] M. Mathie, J. Basilakis, and B. Celler, "A system for monitoring posture and physical activity using accelerometers," in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, pp. 3654–3657, Ieee, 2001.

[172] M. P. Murray, "Gait as a total pattern of movement: Including a bibliography on gait," *American Journal of Physical Medicine & Rehabilitation*, vol. 46, no. 1, pp. 290–333, 1967.

[173] M. P. Murray, A. B. Drought, and R. C. Kory, "Walking patterns of normal men," *JBJS*, vol. 46, no. 2, pp. 335–360, 1964.

[174] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[175] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, pp. 2234–2242, 2016.

[176] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[177] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016.

[178] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[179] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[180] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in Neural Information Processing Systems*, pp. 901–909, 2016.

[181] S. Arora and Y. Zhang, "Do gans actually learn the distribution? an empirical study," *arXiv preprint arXiv:1706.08224*, 2017.

[182] N. I. of Health, "Nih stroke scale." `https://www.stroke.nih.gov/documents/NIH_Stroke_Scale_508C.pdf`.

[183] B. C. Meyer and P. D. Lyden, "The modified national institutes of health stroke scale: its time has come," *International journal of stroke*, vol. 4, no. 4, pp. 267–273, 2009.

[184] O. Chernyshev, S. Martin-Schild, K. Albright, A. Barreto, V. Misra, I. Acosta, J. Grotta, and S. Savitz, "Safety of tpa in stroke mimics and neuroimaging-negative cerebral ischemia," *Neurology*, vol. 74, no. 17, pp. 1340–1345, 2010.

[185] P. Ramanujam, K. Z. Guluma, E. M. Castillo, M. Chacon, M. B. Jensen, E. Patel, W. Linnick, and J. V. Dunford, "Accuracy of stroke recognition by emergency medical dispatchers andparamedics—san diego experience," *Prehospital Emergency Care*, vol. 12, no. 3, pp. 307–313, 2008.

[186] D. Antipova, L. Eadie, A. Macaden, and P. Wilson, "Diagnostic accuracy of clinical tools for assessment of acute stroke: a systematic review," *BMC emergency medicine*, vol. 19, no. 1, p. 49, 2019.

[187] P. Vilela, "Acute stroke differential diagnosis: stroke mimics," *European journal of radiology*, vol. 96, pp. 133–144, 2017.

[188] ASA, "Why getting quick stroke treatment is important," 2018.

[189] A. Lunagariya, A. Patel, S. Dalal, V. Jani, N. Nagaraja, B. Huisa, T. Hemmen, B. Ovbiagele, and U. Patel, "Abstract tmp84: Substantial rise in tpa utilization for acute ischemic stroke in the united states corresponds with significant improvement in outcomes," *Stroke*, vol. 49, no. Suppl_1, pp. ATMP84–ATMP84, 2018.

[190] M. C. Fang, D. M. Cutler, and A. B. Rosen, "Trends in thrombolytic use for ischemic stroke in the united states," *Journal of hospital medicine*, vol. 5, no. 7, pp. 406–409, 2010.

[191] J. A. Zivin, "Acute stroke therapy with tissue plasminogen activator (tpa) since it was approved by the us food and drug administration (fda)," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 66, no. 1, pp. 6–10, 2009.

[192] J. C. Grotta, "tpa for stroke: important progress in achieving faster treatment," *Jama*, vol. 311, no. 16, pp. 1615–1617, 2014.

[193] K. S. Yew and E. Cheng, "Acute stroke diagnosis," *American family physician*, vol. 80, no. 1, p. 33, 2009.

[194] H. S. Nam, E. Park, and J. H. Heo, "Facilitating stroke management using modern information technology," *Journal of stroke*, vol. 15, no. 3, p. 135, 2013.

[195] V. Rajan, S. Bhattacharya, R. Shetty, A. Sitaram, and G. Vivek, "Clinical decision support for stroke using multi–view learning based models for nihss scores," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 190–199, Springer, 2016.

[196] S. Bacchi, T. Zerner, L. Oakden-Rayner, T. Kleinig, S. Patel, and J. Jannes, "Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: A pilot study," *Academic radiology*, 2019.

[197] J. Yu, D. Kim, H. Park, S.-c. Chon, K. H. Cho, S.-J. Kim, S. Yu, S. Park, and S. Hong, "Semantic analysis of nih stroke scale using machine learning techniques," in *2019 International Conference on Platform Technology and Service (PlatCon)*, pp. 1–5, IEEE, 2019.

[198] K. Robertson, "Mindful use of health information technology," *AMA Journal of Ethics*, vol. 13, no. 3, pp. 193–196, 2011.

[199] E. M. Campbell, D. F. Sittig, K. P. Guappone, R. H. Dykstra, and J. S. Ash, "Overdependence on technology: an unintended adverse consequence of computerized provider order entry," in *AMIA Annual symposium proceedings*, vol. 2007, p. 94, American Medical Informatics Association, 2007.

[200] J. S. Ash, D. F. Sittig, E. G. Poon, K. Guappone, E. Campbell, and R. H. Dykstra, "The extent and importance of unintended consequences related to computerized provider order entry," *Journal of the American Medical Informatics Association*, vol. 14, no. 4, pp. 415–423, 2007.

[201] C. Kambhampati, P. Drew, A. Ramesh, and J. Monson, "Artificial intelligence in medicine," *Annals of the Royal College of Surgeons of England*, vol. 86, no. 5, 2004.

[202] L. L. W. Sim, K. H. K. Ban, T. W. Tan, S. K. Sethi, and T. P. Loh, "Development of a clinical decision support system for diabetes care: A pilot study," *PloS one*, vol. 12, no. 2, p. e0173021, 2017.

[203] F. Gotoh, Y. Terayama, T. Amano, and S. S. C. of the Japan Stroke Society, "Development of a novel, weighted, quantifiable stroke scale: Japan stroke scale," *Stroke*, vol. 32, no. 8, pp. 1800–1807, 2001.

[204] M. A. Murphy, C. Willén, and K. S. Sunnerhagen, "Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass," *Neurorehabilitation and neural repair*, vol. 25, no. 1, pp. 71–80, 2011.

[205] G. Mantokoudis, A. S. S. Tehrani, J. C. Kattah, K. Eibenberger, C. I. Guede, D. S. Zee, and D. E. Newman-Toker, "Quantifying the vestibulo-ocular reflex with video-oculography: nature and frequency of artifacts," *Audiology and Neurotology*, vol. 20, no. 1, pp. 39–50, 2015.

[206] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan, "Chronoviz: a system for supporting navigation of time-coded data," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 299–304, ACM, 2011.

[207] L. H. Schwamm, E. S. Rosenthal, A. Hirshberg, P. W. Schaefer, E. A. Little, J. C. Kvedar, I. Petkovska, W. J. Koroshetz, and S. R. Levine, "Virtual telestroke support for the emergency department evaluation of acute stroke," *Academic Emergency Medicine*, vol. 11, no. 11, pp. 1193–1197, 2004.

[208] R. Parsons, D. Schembri, K. Hancock, A. Lonergan, C. Barton, T. Schermer, A. Crockett, P. Frith, and T. Effing, "Effects of the spirometry learning module on the knowledge, confidence, and experience of spirometry operators," *NPJ primary care respiratory medicine*, vol. 29, no. 1, pp. 1–8, 2019.

[209] C. E. Kimble and S. D. Seidel, "Vocal signs of confidence," *Journal of Nonverbal Behavior*, vol. 15, no. 2, pp. 99–105, 1991.

[210] J. C. Brigham, "Target person distinctiveness and attractiveness as moderator variables in the confidence-accuracy relationship in eyewitness identifications," *Basic and Applied Social Psychology*, vol. 11, no. 1, pp. 101–115, 1990.

[211] A. M. Alasheev, A. Y. Andreev, Y. V. Gonysheva, M. N. Lagutenko, O. Y. Lutskovich, A. V. Mamonova, E. V. Prazdnichkova, and A. A. Belkin, "A comparison of remote and bedside assessment of the national institute of health stroke scale in acute stroke patients," *European neurology*, vol. 77, no. 5-6, pp. 267–271, 2017.