Wearable Sensor-based Chronic Condition Severity Assessment: An Adversarial Attention-based Deep Multisource Multitask Learning Approach

Shuo Yu[1], Yidong Chai[2], Hsinchun Chen[3], Scott J. Sherman[4], Randall A. Brown[5]

[1]Area of Information Systems and Quantitative Sciences, Rawls College of Business,

Texas Tech University, Lubbock, TX 79409

[2]Department of Electronic Commerce, School of Management,

Hefei University of Technology, Hefei, Anhui 230011, China

[3]Department of Management Information Systems, Eller College of Management,

University of Arizona, Tucson, AZ 85721

[4]Department of Neurology, University of Arizona, Tucson, AZ 85721

[5]Hermes Medical Intelligence, LLC, Tucson, AZ 85721

Shuo.Yu@ttu.edu

chaiyd@hfut.edu.cn

hchen@eller.arizona.edu

ssherman@neurology.arizona.edu

randall.brown247@gmail.com

ABSTRACT

Advancing the quality of healthcare for senior citizens with chronic conditions is of great social relevance. To better manage chronic conditions, objective, convenient, and inexpensive wearable sensor-based information systems (IS) have been increasingly used by researchers and practitioners. However, existing models often focus on a single aspect of chronic conditions and are often "black boxes" with limited interpretability. In this research, we adopt the computational design science paradigm and propose a novel Adversarial Attention-based Deep Multisource Multitask Learning (AADMML) framework. Drawing upon deep learning, multitask learning, multisource learning, attention mechanism, and adversarial learning, AADMML addresses limitations with existing wearable sensor-based chronic condition severity assessment methods. Choosing Parkinson's Disease (PD) as our test case due to its prevalence and societal significance, we conduct benchmark experiments to evaluate AADMML against state-of-the-art models on a large-scale dataset containing thousands of instances. We present three case studies to demonstrate the practical utility and economic benefits of AADMML and by applying it to detect early-stage PD. We discuss how our work is related to the IS knowledge base and its practical implications. This work can contribute to improved life quality for senior citizens and advance IS research in mobile health analytics.

Keywords: design science, deep learning, multitask learning, multisource learning, attention mechanism, adversarial learning, mobile health analytics

Wearable Sensor-based Chronic Condition Severity Assessment: An Adversarial Attention-based Deep Multisource Multitask Learning Approach

INTRODUCTION

Recent years have witnessed a significant increase in life expectancy in the United States (79.3 years in 2016, World Health Organization 2016). Given the increase in the senior population, the prevalence of chronic conditions has also become a growing societal concern. In 2014, 60 percent of Americans had at least one chronic condition. Meanwhile, the medical costs related to chronic conditions have also been steadily increasing, currently accounting for 90 percent of the nation's $3.5 trillion in annual healthcare expenditures (Centers for Disease Control and Prevention 2020). Among the various chronic conditions, Parkinson's disease (PD) is the second most common neurodegenerative disorder in the U.S. with symptoms including tremors, rigidity, bradykinesia (slowed movement), and postural instability, and is estimated to cost $15.5 billion per year (Gooch et al. 2017). Similar to many other chronic conditions, PD cannot be cured. However, its early detection can significantly alleviate its progression as medicine and therapies can be effective in treating early-stage PD (Mayo Clinic 2020). With proper management, a 20 percent reduction in PD progression would result in net monetary benefits of $60,657 per patient (Johnson et al. 2013).

Practitioners are eager for new tools that could be deployed with ease in both the clinic and remotely via telemedicine to assist in the management of chronic conditions. For instance, by assessing the severity of PD, such tools could potentially optimize pharmacological treatment, guide the application interventions such as physical and occupational therapies, and determine the need for environmental modifications. Ultimately, they may lead to a measurable decrease in

hospitalizations and early deaths among the PD population. With the development of mobile sensing technologies, wearable sensor-based information systems (IS) could provide researchers and practitioners an objective, convenient, and inexpensive means to assess the severity of chronic conditions (Howcroft et al. 2013). Wearable sensors (e.g., accelerometers, gyroscopes, and microphones) can capture quantitative data with high sensitivity and high granularity (sampling frequency up to 100 Hz, equal to 100 data points per second). They can be attached to the human body in a clinical or home setting to collect an immense amount of detailed data (millions of data points per day). With a machine learning model, the wearable sensor data can be analyzed to reveal early signals of chronic condition progression, thus empowering seniors and their families to seek earlier treatment.

Assessing the severity of chronic conditions is often not a standalone task. Comorbidities are very common in chronic condition patients (70 percent of them have more than one condition) (Buttorff et al. 2017). Even within a single chronic condition, there can be multiple related aspects that need to be considered separately (e.g., motor and non-motor factors of PD). A multi-faceted model that can jointly tackle multiple chronic conditions, or multiple aspects of a chronic condition, would be extremely beneficial for comprehensive and accurate chronic condition assessment. In addition, with the abundance of wearable sensors both in volume (e.g., 100 data points per second) and in variety (e.g., accelerometers, gyroscopes, and microphones), an integrated model that can leverage a broad range of data input types is crucially needed.

Interpretability is another critical issue that haunts many machine learning models, and deep learning models in particular. Partly due to their complex architectures, most deep learning models often simply provide a class label for classification problems or a numeric value for regression problems as the result without suggesting contributing factors that are likely to lead to

the result. Such “black box” models are especially concerning for health applications such as disease diagnostics, because health practitioners are eager to understand how a diagnosis is made in addition to the diagnosis result. From an IS perspective, deep learning models with little interpretability can hardly contribute to people’s understanding of the problem domain. A model that can provide insight into the data and improve the design of domain-specific information systems would be of great interest for health professionals and IS researchers.

While valuable, extant studies on wearable sensor-based chronic condition severity assessment face the following four challenges. First, prevailing approaches focus on assessing a single aspect of a chronic condition (e.g., freezing of gait (FoG) in PD). As chronic conditions may have multiple related aspects, an integrated model that can assess multiple aspects of a condition has the potential to investigate the interplays between them. This can be resolved by a model that supports multitask learning. Second, the majority of extant studies conducted only one type of experimental trial (e.g., walking tests) for assessing chronic condition severities. A model that can integrate multiple types of experimental trials (e.g., walking and standing tests) and multiple types of data sources (e.g., accelerometers, gyroscopes, and microphones) has the potential to allow a more comprehensive assessment. This can be resolved by a model that supports multisource learning. Third, although various deep learning models have been proposed for wearable sensor data in the literature, they face the interpretability issue. Knowing which part of data contributes more to the predictive outcomes could be critical in health contexts, as it could help health professionals interpret model results and prioritize therapies and medications. This can be tackled by employing the attention mechanism in the deep learning model. Fourth, deep learning models are known to require high computational power and involve high training overheads. Novel training algorithms are needed for reduced training time and improved model

performance; one such promising direction is adversarial learning. The above challenges motivate an innovative IT artifact for advanced wearable sensor-based chronic condition severity assessment.

Recently, chronic condition management has become an increasingly significant focus of the IS community (Zhang and Ram 2020; X. Liu et al. 2020). Although mobile technologies have long been an interest of IS researchers (Ghose et al. 2012; Sun et al. 2017), to the best of our knowledge, no existing IS study has proposed a wearable sensor-based solution for chronic condition severity assessment, where the design science paradigm can make a unique contribution. Design science creates and evaluates IT artifacts intended to solve identified business problems (Hevner et al. 2004). Specifically, the computational design science paradigm provides insights on how to design novel computational models and systems to resolve problems with significant business and societal impact (Rai 2017). The successful demonstrations of design novelty and validity are the core of computational design science research (Rai 2017). Following the computational design science paradigm and prior IS research on health analytics (Lin et al. 2017; Zhang and Ram 2020; X. Liu et al. 2020), we propose and rigorously evaluate a novel deep learning framework, Adversarial Attention-based Deep Multisource Multitask Learning (AADMML). Drawing upon deep learning, multitask learning, multisource learning, and attention mechanism, AADMML addresses limitations with existing wearable sensor-based chronic condition severity assessment methods by automatically extracting features from wearable sensor data, integrating multitask and multisource learning into a unified framework, and adopting the attention mechanism for model interpretability. In addition, we propose an innovative adversarial attention competition mechanism in AADMML that reduces model training time and improves model performance. We chose PD as our test case due to its high

prevalence and societal significance, and conducted benchmark experiments to evaluate AADMML against state-of-the-art feature-based and deep learning models on a large-scale dataset. We also present three case studies to demonstrate the practical utility and economic benefit of the proposed model.

The scholarly contributions of this study are three-folds. First, AADMML is one of the first deep learning models that enables a multi-faceted evaluation of chronic conditions, both as a diagnostic tool and in identifying contributing factors. By involving multisource and multitask learning, the model is able to learn from multiple types of inputs to assess multiple aspects of chronic conditions for a more accurate diagnosis. Meanwhile, the attention mechanism improves model interpretability. From an individual perspective, the attention weights identify the factors that contribute to the patient's chronic conditions, which provides evidence for revising the patient's medications and therapies. From a collective perspective, the attention weights reveal the types of clinical experiments that are more significant to certain demographic groups, which helps future clinical experiment and therapy development. Second, we contribute to the deep learning community by proposing a novel adversarial attention competition mechanism that speeds up model training and improves predictive outcomes in attention-based deep multisource learning models. This adversarial attention competition mechanism is generalizable to other application domains that involve deep multisource learning. Third, we contribute to the design science theories, the IS community, and business disciplines by creating an interpretable IT artifact that can deal with novel challenges in the types of inputs as well as in modes of learning tasks, which is critically needed in the rise of AI for business applications. With AADMML, traditional business problems such as customer relationship management have a potential to leverage the unprecedented abundance of data (e.g., telephone voice, live chat, email

communications, social media, etc.) to analyze, interpret, and forecast consumer patterns and behaviors. The same can be applied to other business domains, such as determining premiums based on consumer data for the insurance industry.

The rest of the paper is organized as follows. First, we review the existing literature on chronic conditions and wearable sensor technologies in IS and other related fields, attention-based deep multisource multitask learning, and adversarial learning in deep learning. Then we identify research gaps and questions. Subsequently, we present our testbed and research design. Next, we summarize the results from experiments and case studies. Finally, we conclude this research by discussing our contributions to the IS knowledge base, practical implications, and promising future directions.

RESEARCH BACKGROUND

Our research is guided by the following three streams of literature: (1) research on chronic conditions and wearable sensor technologies in IS and other related fields, (2) deep learning, multitask learning, multisource learning with the attention mechanism, and their applications in health contexts, and (3) adversarial learning in deep learning. We also discuss the research gaps and questions.

Chronic Conditions and Wearable Sensor Technologies

Health Information Technology (HIT) is defined as "a broad concept that encompasses an array of technologies to store, share, and analyze health information" (Baird et al. 2018). Benefiting from the rapid development of HIT such as Health Information Systems (HIS), social media, patient portals, online health communities (OHC), and electronic health records (EHR), information systems and data analytics have been playing an increasingly significant role in the

management of chronic conditions. Table 1 summarizes selected recent IS research based on the HIT examined and the focus of the research.

Table 1. Recent Selected IS Research on Chronic Conditions

Year	Author	HIT Examined	Focus
2020	Bao et al.	Patient portal	Impact of patient-provider engagement on patients' health outcomes
2020	Zhang and Ram	Social media	Identifying and understanding triggers and risk factors that cause asthma exacerbations
2020	X. Liu et al.	Social media	Evidence-backed digital therapeutics with technology-enabled interventions
2020	Savoli et al.	HIS	Effective use of self-management IS for chronic conditions
2020	Son et al.	HIS	Designing a smart asthma management system with Bluetooth-enabled inhalers
2020	Brohman et al.	HIS	How a telemonitoring feedback ecosystem is related to patient behavioral outcomes
2020	Q. Liu et al.	OHC	Mutual impact between patients' and physicians' participation in physician-driven OHC
2019	Chen et al.	OHC	How participants are affected by relationships within an OHC and content exchanged between OHC participants
2017	Lin et al.	EHR	Predicting adverse health events for diabetes patients
2016	Kohli and Tan	EHR	Discussion on integration and analytics of EHR

In addition to behavioral and empirical studies, design science has emerged as a significant branch in designing more effective IT artifacts (analytical models, systems, etc.) for chronic condition management (e.g., Zhang and Ram 2020; X. Liu et al. 2020). The increasing prevalence of mobile devices and sensors (e.g., Apple Watch, Fitbit, etc.) presents a unique opportunity for collecting large-scale, always-on health information for advanced health analytics (Chen et al. 2012). Although mobile and sensor analytics exist in the broader IS field, they focus on mobile Internet and Web browsing activities (Adipat et al. 2011; Ghose et al. 2012), mobile app design and behavior (Hoehle and Venkatesh 2015; Kwon et al. 2016; Sun et al. 2017), and service innovation (Venkatesh et al. 2017; Ye and Kankanhalli 2018). Due to the increasing societal significance of chronic condition management (e.g., PD), designing novel mobile health analytics for chronic condition severity assessment remains an underexplored, yet critically needed, perspective of IS literature.

The constantly evolving characteristics of mobile technology and health analytics require the continuous design of novel algorithms, models, and analytical frameworks. The design science paradigm in IS creates and evaluates IT artifacts intended to solve identified business problems. The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods (Hevner et al. 2004). The computational design science paradigm further guides IS researchers on how to design computational models to resolve emerging practical business problems (Rai 2017). This paradigm suggests that the design of the IT artifact can be informed by domain knowledge and data characteristics, where a relevant mature theory is often absent due to the nascence of the problem domain. The novelty and utility of the designed artifact have to be demonstrated by comprehensive evaluations vis-à-vis selected state-of-the-art benchmark approaches. In addition, the artifact's design should contribute to the IS knowledge base (Rai 2017) through situated implementations of the artifact (e.g., model or system), nascent design theory (e.g., design principles), and other forms (Gregor and Hevner 2013). For example, Lin et al. (2017) incorporated the computational design science paradigm by developing Bayesian multitask learning on EHR data to predict critical hospital adverse events. However, designing a high-utility and high-impact computational IT artifact requires a comprehensive understanding of the problem domain and state-of-the-art methods. We therefore review wearable sensor-based chronic condition severity assessment in Table 2, based on the data source (experimental trials from which sensor data are collected), number of subjects, assessment task, and models.

Table 2. Selected Research on Wearable Sensor-based Chronic Condition Severity Assessment

Year	Author	Data Source	#Subjects	Task	Models
2020	Nemati et al.	1-second coughs, 10-second speech	21	Severity assessment of obstructive lung disease	Logistic Regression, SVM, RF, MLP

2020	Moon et al.	30-second stand, 7-meter walk	524	PD prediction	NN, SVM, KNN, Decision tree, RF
2019	Rastegari et al.	10-meter walk, heel-toe tapping, circling	43	Early diagnosis of PD	SVM, RF, NB, AdaBoost
2019	Polat	Daphnet Freezing of Gait Dataset	16	Detecting freezing of gait (FoG) for PD	Logistic regression
2019	Piau et al.	9-meter walk	125	Fall risk assessment	Regression
2019	von Coelln et al.	10-meter walk, standing posture, timed up and go	683	PD prediction	Cox proportional hazards model
2018	Anand et al.	10-meter walk	25	Detecting PD on/off states	Regression, NB, RF
2017	Millor et al.	30-second chair stand, 3-meter walk	431	Frailty severity assessment	Decision tree
2017	Watanabe et al.	15-meter walk	12	Detecting forefoot load for diabetes	Statistical tests

Notes: SVM: Support Vector Machine; NN: Naïve Bayes; RF: Random Forest; NB: Naïve Bayes; MLP: Multi-Layer Perceptron

In the literature, researchers collect wearable sensor data from experimental trials (e.g., 10-meter walking test, standing posture, 10-second speech, etc.) for chronic condition severity assessment in various contexts, including PD, diabetes, frailty, and falls. Wearable sensor data have two unique characteristics: each data point (e.g., a tri-axial acceleration value) contains very little information about an individual's mobility status, while such data points are generated in an extremely high-velocity manner (e.g., 100 data points per second). These characteristics necessitate features being extracted from wearable sensor data before any statistical or machine learning models can be applied. Given that, most prior studies first manually designed and selected their feature sets, extracted features from wearable sensor data, and then either applied statistical tests to examine the explanatory power of their features, or inputted the features into a traditional machine learning model (e.g., support vector machine (SVM), random forest (RF), naïve Bayes (NB), etc.) for a classification or regression task.

However, current studies on wearable sensor-based chronic condition severity assessment face the following four limitations. First, while valuable, manual design and selection of feature sets is ad hoc and labor-intensive, requires significant domain knowledge, while could lead to

inconclusive results (Hubble et al. 2015). Recently, deep learning has emerged as a significant branch of machine learning and has notably outperformed feature engineering-based approaches in numerous sensory tasks (e.g., image recognition (Krizhevsky et al. 2012), wearable sensor-based activity of daily living (ADL) recognition (Wang et al. 2019), and speech recognition (Amodei et al. 2016), etc.). The outstanding performance of deep learning is largely attributable to its powerful automatic feature learning from complex data (Goodfellow et al. 2016). Given that, wearable sensor-based chronic condition severity assessment has the potential to benefit greatly from deep learning. Second, prevailing approaches focus on a single aspect of a chronic condition (e.g., freezing of gait (FoG) in PD). As chronic conditions may have multiple related aspects (e.g., motor and non-motor factors of PD), we are motivated to assess multiple aspects of a condition in an integrated model, which has the potential to investigate the interplays between them. This falls into the field of multitask learning. Third, although the majority of extant studies conducted only one type of experimental trial (e.g., 10-meter walking) and used it as the only data source, an increasing number of studies conducted multiple types of trials or sensors (e.g., accelerometers, gyroscopes, and microphones) and extracted features from each data source for a more comprehensive assessment of chronic condition severity. These types of experiments fall into the field of multisource learning. Fourth, although various wearable sensor-based deep learning models have been proposed in the literature, they face the interpretability issue. Knowing which part of data contributes more to the predictive outcomes could be critical in health contexts, as it could help health professionals interpret model results and prioritize therapies and medications. An emerging solution to this is the attention mechanism. Therefore, we review deep learning, deep multitask learning, deep multisource learning, and the attention mechanism in the next section.

Deep Learning, Multitask Learning, and Multisource Learning

Deep Learning

As we mentioned, deep learning has become one of the most promising branches of machine learning with numerous successful applications (LeCun et al. 2015). Deep learning methods have also been increasingly embraced by IS researchers in chronic condition management (X. Liu et al. 2020; Zhang and Ram 2020). Benefiting from stacked neural network layers of non-linear transformation (i.e., activation) functions, error correction computations, and backpropagation operations, deep learning models can automatically learn salient feature representations from complex data, especially raw sensory data (LeCun et al. 2015). A brief comparison and discussion between deep learning and feature-based learning methods can be found in Appendix G.

Deep Multitask Learning

Next, we briefly discuss multitask learning and its implementation in deep learning. Multitask learning is a machine learning strategy where multiple related prediction tasks are trained jointly to improve learning performance by leveraging the domain-specific information contained in the training signals of related tasks (Caruana 1998). The key to multitask learning is knowledge sharing among the tasks. Two major types of multitask learning exist in the literature: parameter-based and feature-based (Zhang and Yang 2017). Parameter-based multitask learning uses model parameters (e.g., coefficients in linear models) in one task to help learn model parameters in other tasks via some mechanism (e.g., via regularization). There are four major approaches: low-rank (Ando and Zhang 2005), task clustering (Jacob et al. 2009), task relation learning (Williams et al. 2007), and parameter matrix decomposition (Jalali et al. 2010). In contrast, feature-based multitask learning aims to learn common features across different tasks.

There are two approaches: feature selection (Obozinski et al. 2010) and feature transformation (Caruana 1998). The feature selection approach selects a subset of the original features as a common representation for all tasks. The feature transformation approach applies a linear or nonlinear transformation on the original features and generates a set of features as a common representation for all tasks. Among the above approaches, feature transformation is the most widely adopted approach for deep multitask learning (Goodfellow et al. 2016) because each deep learning layer is essentially performing a nonlinear transformation on its input. Therefore, multitask learning with feature transformation can be easily integrated into existing deep learning models. We summarize selected health research on deep multitask learning in Table 3, based on the data source, tasks, multitask learning strategy, and loss function.

Table 3. Selected Health Research on Deep Multitask Learning

Year	Author	Data Source	Tasks	Multitask Learning Strategy	Loss Function
2020	Wang et al.	OCT images	Visual field measurement; Glaucoma diagnosis	Feature transformation	Direct sum of the tasks
2020	Davoodnia et al.	Pressure data	Subject identification; BMI estimation	Feature transformation	Direct sum of the tasks
2019	Lang et al.	Impulse signals	Motion classification; Person identification	Feature transformation	Direct sum of the tasks
2019	Shi et al.	EHR	50 disease diagnosis	Feature transformation	Direct sum of the tasks
2018	Chen	EEG signals	Motion intention recognition (5 tasks)	Feature transformation	Not mentioned
2018	Liao et al.	Gene expression data	12 types of cancer diagnosis	Feature transformation	Not mentioned
2016	Moeskops et al.	Radiology images	Medical image segmentation	Feature transformation	Not mentioned

Notes: EEG: Electroencephalography; BMI: Body Mass Index; OCT: Optical Coherence Tomography

Deep multitask learning has been used in disease diagnosis, motion classification, and medical image segmentation, among other topics. The dominant multitask learning strategy is feature transformation, which creates shared and task-specific layers in a deep learning

architecture. In such a strategy, shared layers learn general features applicable to all tasks, and task-specific layers learn features applicable to a single task. In contrast, deep multitask learning often does not have special designs on the loss function. Most studies use the direct sum of the loss functions of the tasks as the total loss function.

Deep Multisource Learning and the Attention Mechanism

Unlike multitask learning, multisource learning is motivated by the need of data and sensor fusion, as the severity of a chronic condition often cannot be comprehensively assessed with only one data source. For instance, physicians often examine a PD patient's motor status with multiple motor activities, including hand movements, leg agility, arising from a chair, gait, postural stability, speech, among other tests and observations (MDS-UPDRS, Goetz et al. 2007). In the literature, multisource learning is also known as multimodal learning (Baltrusaitis et al. 2019), which refers to learning different information from multiple measurement modalities (e.g., simultaneously recorded audio and video (Chaudhuri et al. 2009), or images and text (Hodosh et al. 2013)). We review recent health research on deep multisource learning in Table 4, based on the data sources, multisource learning strategy, and task.

Table 4. Selected Health Research on Deep Multisource Learning

Year	Author	Data Sources	Multisource Learning Strategy	Task
2020	Zhou et al.	3D MRI image	Early fusion with attention	Brain Tumor Segmentation
2020	Zhang and Shi	MRI and PET image	Early fusion with attention	Alzheimer's disease diagnosis
2020	L. Sun et al.	Video, audio, and text	Early fusion with attention, late fusion	Emotion recognition
2020	Ghaleb et al.	Video and audio	Late fusion	Emotion recognition
2019	Ma et al.	Wearable sensor	Early fusion with attention	Human activity recognition
2019	Xue et al.	Wearable sensor	Early fusion with attention	Human activity recognition
2019	Zhang et al.	Wearable, ambient sensors; audio; WiFi, app log and screen	Early fusion with attention	Mood instability inference

2019	Luo et al.	Disease and gene information	Early fusion	Disease-gene association prediction
2019	de Jong	Radar and video	Early fusion, raw data fusion, late fusion	Human activity recognition
2019	Qiao et al.	EHR	Early fusion	Disease diagnosis prediction
2018	Hu et al.	EMG	Early fusion with attention	Gesture recognition
2018	Yuan et al.	EEG	Early fusion with attention	Seizure detection
2018	Xue et al.	Chest X-ray	Raw data fusion	Radiology report generation
2018	Vlachostergiou et al.	Radiology (brain images)	Early fusion	PD prediction
2017	Radu et al.	Wearable, ambient, physiological sensor	Early fusion, raw data fusion	Activity and context recognition
2017	Zou et al.	fMRI and sMRI	Raw data fusion	Diagnosis of hyperactivity disorder

Notes: EMG: Electromyography; EEG: Electroencephalography; PSG: Polysomnography; fMRI: functional Magnetic Resonance Imaging; sMRI: structural Magnetic Resonance Imaging; PET: Positron Emission Computed Tomography

Deep multisource learning has been applied to human activity recognition, emotion recognition, and disease detection or prediction, among many other medical applications. The data sources include wearable sensor data, video and audio, radiological images, EHR, etc. Three general multisource learning strategies have been studied: raw data fusion, early fusion, and late fusion. Raw data fusion directly concatenates data from multiple sources and uses the merged data as if there is only one source. Early fusion learns source-specific features from each source, while there is a shared model structure that links the source-specific features for the prediction task. Late fusion trains a separate model for each data source. The final prediction result is an integration (e.g., an average) of the results of the models. An abstract illustration of the structures of the above three strategies can be found in Figure 1, assuming we have three sources and one prediction task.

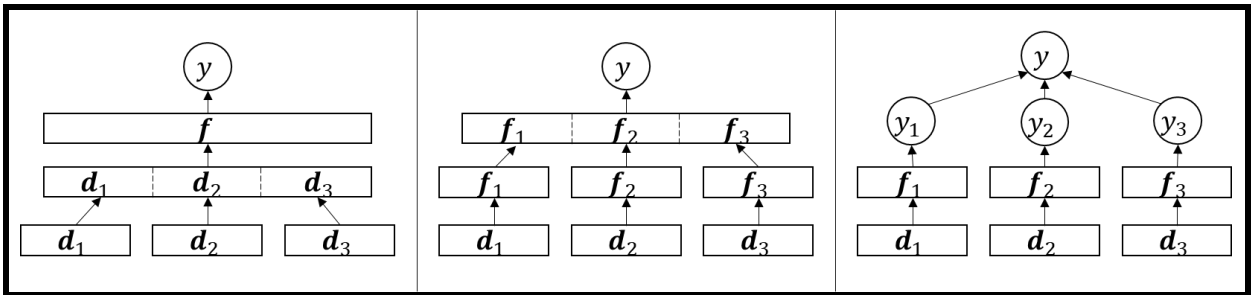


Figure 1. Illustration of the structures of deep multisource learning strategies. Left: Raw data fusion. Middle: Early fusion. Right: Late fusion.

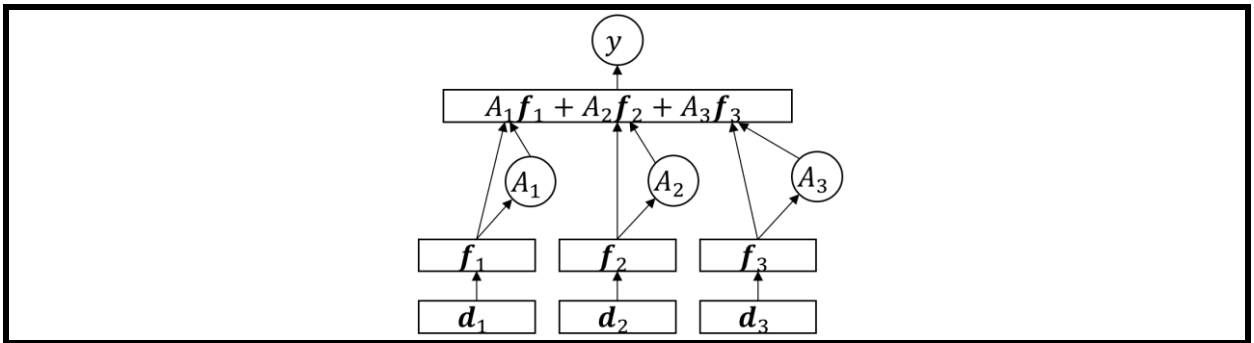


Figure 2. Illustration of attention-based early fusion

Notes: $\boldsymbol{d}_1, \boldsymbol{d}_2, \boldsymbol{d}_3$: raw data from sources 1, 2, 3, respectively; $\boldsymbol{f}$: shared features across the three sources; $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3$: learned features for sources 1, 2, 3, respectively; y: prediction task result; y_1, y_2, y_3: prediction results of the models corresponding to sources 1, 2, 3, respectively; A_1, A_2, A_3: attention weights corresponding to $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3$, respectively. $A_1 + A_2 + A_3 = 1$. Arrows denote the flows of data (tensors) in a deep learning model.

Early fusion is the most widely adopted approach in the literature due to its improved performance over the other two options. We also observed that the attention mechanism is being increasingly adopted as a technique to facilitate deep multisource learning. The attention mechanism was first introduced for machine translation (Bahdanau et al. 2014). This mechanism can be intuitively explained using human biological systems. For example, people tend to focus selectively on parts of an image while ignoring irrelevant information, which can assist in perception (Xu et al. 2015). This mechanism is implemented as an allocation of attention weights among different portions of the input data (Figure 2).

While all attention weights add up to 1, a larger attention weight is allocated if that portion of data is more relevant to the model output (e.g., the words or sentences in EHR that are

more relevant to the diagnosis of a disease, or the type of experimental trial that is more relevant to the assessment of a chronic condition). As such, the attention mechanism significantly improves the interpretability, accountability, and transparency of deep multisource learning models, which is critical for applications that influence human lives such as chronic condition severity assessment (Chaudhari et al. 2019).

Extant studies with multiple data sources typically allocate attention weights among their data sources, i.e., each data source is assigned a scalar attention weight (e.g., Wu et al. 2018; Gao et al. 2020). In this manner, the attention allocation can also be viewed as a "criterion" assessing the quality of the features learned from each data source (e.g., assessing the quality of $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3$ in Figure 1). A larger attention weight is allocated to the data source with features that contribute more towards the model output (e.g., the assessment score of chronic condition severity). Given such attention weights, multiple data sources could "compete" with each other for a larger attention weight by each source forcing its model branch to learn features that are more relevant to the model output, thus improving overall model performance. However, to the best of our knowledge, none of the extant studies have proposed an explicit attention competition mechanism in a deep multisource learning model. One of the potential solutions to this issue is adversarial learning.

Adversarial Learning in Deep Learning

Adversarial learning in deep learning refers to a learning process that multiple components in a deep learning model compete with each other in a zero-sum game to improve their performance (Goodfellow et al. 2014). Each component in the competition is designed to update its parameters towards the direction that "wins more" in the zero-sum game. Over multiple iterations, a Nash equilibrium is achieved with overall model performance improved.

The most well-known example of adversarial learning is generative adversarial networks (GAN) (Goodfellow et al. 2014) where the competition involves a generator and a discriminator. The generator aims to learn the data distribution in the dataset and generate synthetic data samples to confuse the discriminator. The discriminator aims to discriminate synthetic data samples from real data samples and to not be confused by the generator. In the adversarial learning process, the generator and the discriminator are improved alternately in multiple iterations until the point at which the discriminator can no longer discriminate synthetic data samples. At this point, the performances of both the generator and the discriminator are improved, and the two networks can be used for downstream tasks (e.g., the generator can be used to generate synthetic data similar to real data, and the discriminator can be used for classification tasks). Adversarial learning has been a hot research topic in the deep learning community. A summary of recent health studies applying adversarial learning is listed in Table 5, based on the data source, task, and the model components competing with each other.

Table 5. Selected Health Research using Adversarial Learning

Year	Author	Data Source	Task	Model Components Competing with Each Other
2020	Ghassemi et al.	Brain MRI	Brain tumor classification	Generator vs. discriminator
2020	de Bois et al.	Structural variables	Glucose Prediction	Generator vs. discriminator
2020	Y. Sun et al.	Breast MRI	Generating synthetic MRI	Generator vs. discriminator
2019	Balabka	Wearable sensor	Human activity recognition	Generator vs. discriminator
2019	Jin et al.	Brain CT	Transforming CT to MRI images	Generator vs. discriminator
2019	Tang et al.	Chest X-ray	Generating synthetic X-ray images	Generator vs. discriminator
2018	Sun et al.	EHR	Identifying susceptible portions of EHR	Generator vs. generator
2018	Wang et al.	Wearable sensor	Generating synthetic wearable sensor data	Generator vs. discriminator
2017	Choi	EHR	Generating synthetic EHR	Generator vs. discriminator
2017	Hwang et al.	EHR	Disease prediction	Generator vs. discriminator

Note: CT: Computed tomography

Prior adversarial learning studies on health applications used radiological images, EHR, and wearable sensors as data sources. Most of the model component competitions occurred between a generator and a discriminator, which is a typical setting for GAN. To the best of our knowledge, no prior studies applied adversarial learning on an attention-based deep multisource learning model in the form of adversarial attention competition mechanism to improve the features being learned from each data source as well as overall model performance.

Research Gaps and Questions

Based on the above literature review, we have identified the following research gaps. First, past studies on wearable sensor-based chronic condition severity assessment manually designed and selected the feature sets for their context. While valuable, this process is laborious and ad hoc, and the extracted features could lead to inconclusive results. Second, although past studies have proposed models that employ deep multitask learning and attention-based deep multisource learning, such models do not comprise an attention competition mechanism for more relevant features being learned from each data source. Based on the above research gaps, we ask the following research questions:

- How can we design a deep multitask multisource learning framework for wearable sensor-based chronic condition severity assessment?
- How can we employ the attention mechanism in the framework to provide insights into the data and improve model interpretability?
- How can we adopt adversarial learning to improve the performance of attention-based deep multisource multitask learning?

RESEARCH DESIGN

Guided by prior studies and the identified research gaps, we propose a novel Adversarial Attention-based Deep Multisource Multitask Learning (AADMML) framework for wearable sensor-based chronic condition severity assessment. Our research design comprises three major components: (1) data collection and preprocessing, (2) the proposed AADMML framework, and (3) model evaluation using benchmark experiments and case studies. Each is discussed in detail in the following subsections.

Data Collection and Preprocessing

A large publicly available dataset, the mPower dataset (Bot et al. 2016), is used as the research testbed. mPower is an observational smartphone-based study developed using Apple's ResearchKit library to evaluate the feasibility of remotely collecting sensor data to reflect people's PD severities. We obtained the following categories of data in the mPower dataset: (1) demographics, (2) PD severity assessment survey, (3) accelerometer data of walking and standing experiments (denoted as acc-walk and acc-stand, respectively), (4) gyroscope data of walking and standing experiments (denoted as gyro-walk and gyro-stand, respectively), and (5) microphone data of a speech experiment. The PD severity assessment survey includes a subset of questions from the Movement Disorder Society-Sponsored Universal Parkinson Disease Rating Scale (MDS-UPDRS, Goetz et al. 2008). More specifically, the survey consists of 6 out of the 13 questions in MDS-UPDRS Part 1: Non-Motor Aspects of Experiences of Daily Living (nM-EDL), and 10 out of the 13 questions in MDS-UPDRS Part 2: Motor Aspects of Experiences of Daily Living (M-EDL). Each question has a numerical response score between 0 (normal) and 4 (severe). As the literature has suggested that the sum of response scores of nM-EDL questions and of M-EDL questions can be used separately to assess two related aspects of PD (Martinez-

Martin et al. 2015), we choose "nM-EDL" and "M-EDL" as the two PD severity assessment tasks in this study.

Regarding the walking and standing experiments, each test instance corresponds to two experimental trials: (1) 20-step outbound walking in a straight line, and (2) 30-second standing still. Each experimental trial generates an accelerometer sample and a gyroscope sample, each of which is a series of data points. A data point is a single reading from a tri-axial accelerometer or gyroscope, which is a vector with three components corresponding to the three orthogonal axes (x, y, z) in the three-dimensional physical space. A data point collected at time i, $\boldsymbol{a}_i$, is represented as $\boldsymbol{a}_i = [a_{\mathrm{x},i} \quad a_{\mathrm{y},i} \quad a_{\mathrm{z},i}]^T$. 100 data points are collected in one second. As a data sample is a series of data points, it can be represented as $\boldsymbol{x} = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_l]$, where l is the length (the number of data points) in data sample $\boldsymbol{x}$, which depends on the amount of time the subject needed to complete the experimental trial. We use the accelerometer and gyroscope samples collected from the walking and standing experiments as four data sources in this study (acc-walk, acc-stand, gyro-walk, gyro-stand), as walking patterns and standing stability are critical factors that can reflect PD severity (Hubble et al. 2015), and accelerometers and gyroscopes reflect different aspects of motion patterns (translational and rotational, respectively).

Regarding the speech experiment, each subject is instructed to say "five, four, three, two, one" loudly and the smartphone microphone records the waveform audio data. Following state-of-the-art practices in audio processing (Zeng et al. 2019) and deep learning (Deng and Yu 2014), we transform the waveform audio data into mel-scaled spectrograms as a preprocessing step. Spectrograms are matrices that describe how sound frequencies change temporally. Each matrix entry is an amplitude value for a certain sound frequency at a certain time. The spectrograms are generated using the librosa package in Python (McFee et al. 2015). In

alignment with the notation for accelerometer and gyroscope samples, we use $a_{f,i}$ to denote the element at the f-th row and the i-th column in a spectrogram matrix, which represents the amplitude of the f-th sound frequency at time i. Hence, $\boldsymbol{a}_i$ is represented as $\boldsymbol{a}_i = [a_{1,i} \quad \cdots \quad a_{F,i}]^T$ where F is the total number of sound frequencies. Resembling accelerometer and gyroscope data samples, audio data samples can also be represented as $\boldsymbol{x} = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_l]$, where l is the time length of the spectrogram.

To summarize, each instance contains five data sources (acc-walk, acc-stand, gyro-walk, gyro-stand, and mic) and two assessment tasks (eM-EDL and M-EDL). The dataset consists of 2,471 instances. A summary of descriptive statistics can be found in Table 6.

Table 6. Dataset Descriptive Statistics

	N = 2,471	Mean	Std.	Max.	Min.
Demo-graphics	Age	51.44	17.66	81	18
	Gender	Male: 69.4%; Female: 30.6%			
MDS-UPDRS Survey	Task 1: nM-EDL	4.88	3.63	18	0
	Task 2: M-EDL	4.55	5.17	24	0
Wearable Sensor Data Duration	Acc-walk (length)	1,985	700	3,195	568
	Acc-stand (length)	3,041	42	3,239	2,719
	Gyro-walk (length)	1,890	676	3,062	550
	Gyro-stand (length)	2,923	111	3,140	2,566
	Mic (seconds)	9.53	0.28	11.36	7.04

Note: Std.: Standard Deviation; Max.: Maximum; Min.: Minimum

We split the dataset into a 90% experiment subset and a 10% hold-out subset. The 90% experiment subset is used for model training, validation, and testing, while the 10% hold-out subset is reserved for case studies using trained models. This split can help avoid biased results. To preserve the information in wearable sensor data while speeding up the training process, we pad zeroes in all data samples such that all accelerometer and gyroscope samples have a length of 3,840, which exceeds their maximum length. For spectrograms, we pad zeroes such that all the samples have a duration of 12 seconds. With the default setting in the librosa (n_mels = 128),

we obtained a spectrogram with size 128×1034 for each sample. The zero-padding is a widely adopted technique in deep learning for processing data samples with varying lengths (Yenter and Verma 2017; Bin et al. 2018). In the next subsection, we discuss the four stages of our AADMML framework.

The Adversarial Attention-based Deep Multisource Multitask Learning (AADMML) Framework

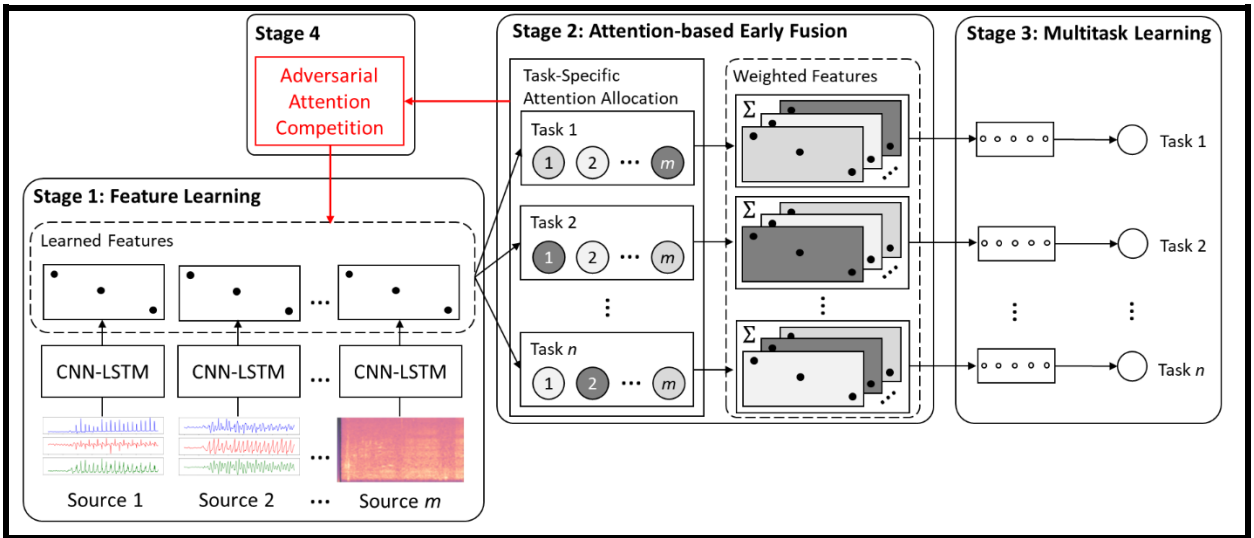


Figure 3. The AADMML Framework

Figure 3 illustrates the proposed AADMML framework. Our main methodological contribution is Stage 4: Adversarial Attention Competition, which will be discussed in detail later. The AADMML framework is flexible and adaptive to the actual number of data sources (m) to learn from and the actual number of tasks (n) to assess. Conceptually, there are four stages in the AADMML framework. Stage 1 is designed for feature learning. A separate CNN-LSTM is applied on each data source to extract source-specific features from the m sources of complex data (e.g., accelerometers, gyroscopes, and microphones) (Trigeorgis et al. 2016; Zeng et al. 2019). Stage 2 is designed to learn the attention allocation across the data sources for each task, which can be interpreted as the relative importance of each data source. At the end of Stage 2, the features learned in Stage 1 and the attention weights in Stage 2 are integrated as a set of task-

specific features. Stage 3 is multitask learning for the n related tasks (e.g., two related aspects of a chronic condition). Stage 4 is our proposed adversarial attention competition mechanism that allows each source's attention to compete with each other, thus improving the quality of the learned source-specific features in Stage 1. We discuss the details in each stage in turn.

We introduce a series of mathematical notations to improve description clarity. We use $\boldsymbol{X} = \{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(d)}, \ldots, \boldsymbol{X}^{(D)}\}$ to denote the set of multisource data that include accelerometer signals, gyroscope signals, and audio spectrograms, where D is the total number of instances and $\boldsymbol{X}^{(d)}$ is the d-th instance. We let $\boldsymbol{X}^{(d)} = [\boldsymbol{x}_1^{(d)}, \ldots, \boldsymbol{x}_i^{(d)}, \cdots, \boldsymbol{x}_m^{(d)}]$, where m is the total number of data sources and $\boldsymbol{x}_i^{(d)}$ is the data sample from the i-th source. As defined before, each $\boldsymbol{x}_i^{(d)}$ is a series of tri-axial acceleration values, tri-axial gyroscope values, or a list of amplitudes of sound frequencies, i.e., $\boldsymbol{x}_i^{(d)} = [\boldsymbol{a}_{i,1}^{(d)}, \boldsymbol{a}_{i,2}^{(d)}, \cdots, \boldsymbol{a}_{i,l}^{(d)}]$. We use $\boldsymbol{Y} = \{\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(d)}, \ldots, \boldsymbol{y}^{(D)}\}$ to denote the set of ground-truth assessment scores, so $\boldsymbol{y}^{(d)} = \left[y_1^{(d)}, \ldots, y_j^{(d)}, \ldots, y_n^{(d)}\right]$, where n is the total number of tasks and $y_j^{(d)}$ is the assessment score of the j-th task.

Stage 1: Feature Learning

Due to the characteristics of wearable sensor data, features need to be learned from raw data before they can be used for assessment tasks. Both accelerometer, gyroscope data and spectrograms can be formalized as matrices, where each row of the matrix is a temporal sequence and each column shows patterns across sensor axes (accelerometers or gyroscopes) or a range of sound frequencies (spectrograms). As CNN excels in automatic feature extraction from local signals and LSTM is effective in the temporal modeling of long sequences (LeCun et al. 2015), we propose a CNN-LSTM that leverages the advantages of both CNN and LSTM for feature learning. For simplicity, we discuss the CNN-LSTM at an abstract level. We consider a

data sample, $\boldsymbol{x}_i^{(d)}$, as an example. We first segment $\boldsymbol{x}_i^{(d)}$ into multiple consecutive snippets $\boldsymbol{s}$, where $\boldsymbol{s}$ is also a series of tri-axial acceleration values, tri-axial gyroscope values, or a list of amplitudes of sound frequencies, i.e., $\boldsymbol{x}_i^{(d)} = [\boldsymbol{s}_{i,1}^{(d)}, \dots, \boldsymbol{s}_{i,q}^{(d)}, \cdots, \boldsymbol{s}_{i,r}^{(d)}]$, where r is the number of snippets that $\boldsymbol{x}_i^{(d)}$ is segmented into. The CNN processes each snippet $\boldsymbol{s}_{i,q}^{(d)}$ and outputs a feature representation $\boldsymbol{o}_{i,q}^{(d)}$. After all snippets are processed by the CNN, we obtain a sequence of feature representations, $[\boldsymbol{o}_{i,1}^{(d)}, \dots, \boldsymbol{o}_{i,q}^{(d)}, \cdots, \boldsymbol{o}_{i,r}^{(d)}]$. The LSTM takes this sequence to analyze its sequential information and outputs $\boldsymbol{R}_i^{(d)}$, which is the learned features of sample $\boldsymbol{x}_i^{(d)}$. The detailed specifications of CNN and LSTM can be found in Appendix A.

Stage 2: Attention-based Early Fusion

In this stage, we aim to learn the task-specific attention allocation for each data source, which reflects the quality of the learned features (Ma et al. 2019; Xue et al. 2019). We show the attention allocation process for task j $(j = 1,2, \dots, n)$ as all tasks follow the same process. As suggested by Ma et al. (2019), with the learned features of all data sources $(\boldsymbol{R}_1^{(d)}, \dots, \boldsymbol{R}_i^{(d)}, \dots, \boldsymbol{R}_m^{(d)})$ as the input, the attention weight for source i, $A_{j,i}^{(d)}$, is given by:

$$u_{j,i}^{(d)} = \tanh\left(\boldsymbol{W}_{j,i}^{\mathrm{a}} \boldsymbol{R}_{j,i}^{(d)} + b_{j,i}^{\mathrm{a}}\right),$$

$$A_{j,i}^{(d)} = \frac{\exp\left(u_{j,i}^{(d)} w_j\right)}{\sum_{i=1}^{m} \exp\left(u_{j,i}^{(d)} w_j\right)} \tag{1}$$

where $\left\{\boldsymbol{W}_{j,i}^{\mathrm{a}}, b_{j,i}^{\mathrm{a}}, w_j\right\}$ are the parameters of the attention allocation process. Then, the m-source features and attention weights $(\boldsymbol{R}_1^{(d)}, \dots, \boldsymbol{R}_i^{(d)}, \dots, \boldsymbol{R}_m^{(d)}; A_{j,1}^{(d)}, \dots, A_{j,i}^{(d)}, \dots, A_{j,m}^{(d)})$ are summarized as a set of integrated features, $\boldsymbol{C}_j^{(d)}$, as follows:

$$\boldsymbol{C}_j^{(d)} = \sum_{i=1}^{m} A_{j,i}^{(d)} \cdot \boldsymbol{R}_i^{(d)}.$$

$\boldsymbol{C}_j^{(d)}$ is then used in the next stage for the learning of task j.

Stage 3: Multitask Learning

We use two stacked fully connected layers to get the predicted assessment score of task j, $\hat{y}_j^{(d)}$:

$$\boldsymbol{F}_j^{(d),1} = \text{sigmoid}\left(\boldsymbol{W}_j^1 \cdot \boldsymbol{C}_j^{(d)} + b_i^1\right),$$

$$\boldsymbol{F}_j^{(d),2} = \text{sigmoid}\left(\boldsymbol{W}_j^2 \cdot \boldsymbol{F}_j^{(d),1} + b_j^2\right),$$

$$\hat{y}_j^{(d)} = \boldsymbol{W}_j^{\text{L}} \cdot \boldsymbol{F}_j^{(d),2} + b_j^{\text{L}},$$

where $\{\boldsymbol{W}_j^1, b_i^1\}$ are the parameters for the first layer, $\boldsymbol{F}_j^{(d),1}$ is the output of the first layer, $\{\boldsymbol{W}_j^2, b_j^2\}$ are the parameters for the second layer, $\boldsymbol{F}_j^{(d),2}$ is the output of the second layer, and $\{\boldsymbol{W}_j^{\text{L}}, b_j^{\text{L}}\}$ are the parameters for the final output layer.

The selection of the loss function depends on the nature of the tasks. As a common practice for regression tasks, we use the mean squared error (MSE) loss functions to evaluate model outputs for each individual task, and use the sum of all tasks' weighted MSE losses as the total loss function (Lang et al. 2019; Shi et al. 2019). As such, the total loss for the d-th instance, $\mathcal{L}^{(d)}$, is:

$$\mathcal{L}^{(d)} = \sum_{j=1}^{n} \omega_j \left(\hat{y}_j^{(d)} - y_j^{(d)}\right)^2 \quad (2)$$

where $\hat{y}_j^{(d)}$ is the model-predicted assessment score for task j, $y_j^{(d)}$ is the ground-truth assessment score for task j, and ω_j is the loss weight for task j. By default, all ω_j are set to 1.

Stage 4: Adversarial Attention Competition

Our design of the adversarial attention competition mechanism is inspired by adversarial learning. In an attention-based multisource learning model, each data source has a model branch that extracts source-specific features from input data. In our study, an example of such a model branch is the CNN-LSTM corresponding to each data source in Stage 1. For each task, attention weights adding up to 1 are allocated among the data sources. We design the competition such that each data source aims to increase its allocated attention weight. As a data source's attention weight reflects the relative relevance of that source, the model branch corresponding to the data source has to adjust its parameters such that more relevant features can be extracted from the data source in order to increase its attention weight. Given that all attention weights have to add up to 1, an increase in one data source's attention weight will result in decreases in other data sources' attention weights. Consequently, to increase their attention weights, the other data sources also update their corresponding model branches to extract more relevant features from the data sources. The relevancy of features extracted from each data source gets increasingly improved over multiple training iterations. Compared with a model without adversarial attention competitions (model parameters are only updated with the guidance (back-propagation) of task loss functions), AADMML's parameters are updated with the guidance of both task loss functions and adversarial attention competitions, which can potentially reduce the number of training iterations and improve model performance.

We implement the adversarial attention competition mechanism as an iterative process. In each iteration, each data source alternately updates the parameters of its model branch in the direction of increasing its attention weight. As we have m sources in total, we randomly select the order in which the m sources update their parameters. To improve training efficiency, we

perform adversarial attention competitions once after the model parameters have been updated by the traditional loss function back-propagation for K rounds. We further refer to K as the "adversarial learning factor," which is a hyperparameter to be set. Although this is a zero-sum game (all attention weights have to add up to 1), each data source learns more relevant features after the adversarial attention competition process, thus improving overall model performance.

As described above, the adversarial attention competition mechanism is part of the parameter learning process of the AADMML framework. Algorithm 1 presents the complete pseudocode for the AADMML parameter learning algorithm. Lines 15 to 21 summarize the adversarial attention competition, which is our core methodological novelty. We use the widely applied Adam optimizer (a gradient descent-based optimizer) to iteratively learn model parameters, and use the batch training technique to improve training efficiency. We applied L2-regularization on all deep learning parameters to avoid overfitting as suggested by the literature (Goodfellow et al. 2016). Full model specifications can be found in Appendix A.

Algorithm 1: AADMML Parameter Learning Algorithm	
Input: ($\boldsymbol{X}$, $\boldsymbol{Y}$); total number of batches, B; number of instances in each batch, E; adversarial learning factor, K; parameter L2-regularization weight, λ	
Output: $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_m, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_n, \boldsymbol{\psi}\}$, where $\boldsymbol{\theta}_i$ refers to the parameters of i-th source's model branch in Stage 1, $\boldsymbol{\phi}_j$ refers to j-th task's parameters in Stages 2, and $\boldsymbol{\psi}$ refers to the parameters in Stage 3.	
Initialize $\boldsymbol{\Theta}$	
1.	**while** not converged **do**
2.	**for** $b = 1, \dots, B$ **do**
3.	**for** $k = 1, \dots, K$ **do**
4.	**for** $e = 1, \dots, E$ **do**
5.	Compute loss $\mathcal{L}^{(e)}$ by Equation (2)
6.	**for** $j = 1, \dots, n$ **do**
7.	**for** $i = 1, \dots, m$ **do**
8.	Compute $A_{j,i}^{(e)}$ by Equation (1)
9.	**end for** (i)
10.	**end for** (j)
11.	**end for** (e)
12.	Update $\boldsymbol{\Theta} \leftarrow \text{Adam}\left(\nabla_{\boldsymbol{\Theta}} \frac{1}{E} \sum_{e=1}^{E} \mathcal{L}^{(e)} + \lambda \|\boldsymbol{\Theta}\|_2^2, \boldsymbol{\Theta}\right)$
13.	**end for** (k)

14.	# Adversarial Attention Competition Mechanism
15.	**for** $i = 1, \dots, m$ **do**
16.	Compute $A_i^{(e)} \leftarrow \sum_{j=1}^{n} A_{j,i}^{(e)}$
17.	**end for** (i)
18.	Determine the source updating order $I \leftarrow \text{Permute}(1,2,\dots,m)$
19.	**for** i in I **do**
20.	Update $\boldsymbol{\theta}_i \leftarrow \text{Adam}\left(\nabla_{\boldsymbol{\theta}_i} \frac{1}{E} \sum_{e=1}^{E} (-A_i^{(e)}), \boldsymbol{\theta}_i\right)$ by Equation (1)
21.	**end for** (i)
22.	**end for** (b)
23.	**end while**
24.	**return** $\boldsymbol{\Theta}$

Experiment and Case Study Design

To adhere to the computational design science paradigm guidelines, we systematically evaluate our proposed AADMML framework for chronic condition severity assessment with three experiments and three case studies. Experiment 1 evaluates AADMML's ability to assess PD severity against state-of-the-art benchmark models, including feature-based machine learning models and alternative deep multisource and/or multitask learning models. For feature-based machine learning models, we compile the feature set by selecting the most common features from the PD literature (Hubble et al. 2015), which can be found in Appendix B. For deep learning models, we evaluate different multisource learning strategies (late fusion, early fusion, attention-based early fusion, and adversarial attention-based early fusion) in deep multitask or single task learning settings. Detailed descriptions of the benchmark models can be found below in Table 7.

Table 7. Summary of Experiment 1 (Comparison against Benchmark Models)

Category	Method	Description
Feature-based, Non-ensemble	Decision tree (DT)	Non-ensemble and ensemble machine learning models. A set of features (Appendix B) is extracted from each data source. Features extracted from different sources are used together. A separate model is trained for each task.
	K-nearest neighbors (KNN)	
	Support vector machines (SVM)	
Feature-based, Ensemble	Extra-trees (ETS)	
	Random forest (RF)	

	AdaBoost (ADA)	
	Gradient boosting model (GBM)	
	XGBoost (XGB)	
	Non-ensemble Voting (NEV)	An ensemble machine learning model by averaging the results from the above non-ensemble models (DT, KNN, SVM).
Deep learning	Single task, late fusion (STLF)	Deep learning models. For single task models, a separate model is trained for each task. For multitask models, a joint model is trained for all tasks. We investigate different multisource strategies (late fusion, early fusion, and attention-based early fusion). In particular, MTAEF is AADMML without the adversarial attention competition mechanism.
	Single task, early fusion (STEF)	
	Single task, attention-based early fusion (STAEF)	
	Multitask, late fusion (MTLF)	
	Multitask, early fusion (MTEF)	
	Multitask, attention-based early fusion (MTAEF)	

Experiment 2 is a sensitivity analysis that evaluates how the selection of hyperparameters affects AADMML's performance. To avoid exponential numbers of hyperparameter combinations, we use our proposed AADMML as a baseline (model specifications in Appendix A) and adjust only one type of hyperparameter at a time. We adjust the size of CNN kernels, number of CNN layers, size of LSTM hidden neurons, structure of LSTM, learning rate, dropout rate, adversarial learning factor, relative weight of Task 1 loss function, and weight of parameter L2-regularization in the AADMML model as suggested by literature (Goodfellow et al. 2016). Detailed description of the sensitivity analysis can be found in Table 8.

Table 8. Summary of Experiment 2 (Sensitivity Analysis)

Hyperparameter	Options	Model Name	Hyperparameter	Options	Model Name
Size of CNN kernels	8	HP-CNN (8)	Number of CNN layers	1	HP-CNN-L1
	16	HP-CNN (16)		2	HP-CNN-L2
	32	HP-CNN (32)		3	HP-CNN-L3
	64	HP-CNN (64)		4	HP-CNN-L4
Size of LSTM hidden neurons	8	HP-LSTM (8)	Structure of LSTM	1-layer LSTM	HP-LSTM-L1
	16	HP-LSTM (16)		2-layer LSTM	HP-LSTM-L2
	32	HP-LSTM (32)		1-layer Bi-LSTM	HP-BLSTM-L1
	64	HP-LSTM (64)		2-layer Bi-LSTM	HP-BLSTM-L2
Learning rate	10^{-1}	HP-LR-1	Dropout rate	0.00	HP-DR-0
	10^{-2}	HP-LR-2		0.25	HP-DR-25
	10^{-3}	HP-LR-3		0.50	HP-DR-50

	10^{-4}	HP-LR-4		0.75	HP-DR-75
Adversarial learning factor	1	HP-AT-1	Relative weight of Task 1 loss function (Weight of Task 2 loss function as 1)	0.5	HP-T1LF-0.5
	3	HP-AT-3		1.0	HP-T1LF-1.0
	5	HP-AT-5		1.5	HP-T1LF-1.5
	7	HP-AT-7		2.0	HP-T1LF-2.0
Weight of parameter L2-regularization	10^{-3}	HP-WL2R-3			
	10^{-4}	HP-WL2R-4			
	10^{-5}	HP-WL2R-5			
	10^{-6}	HP-WL2R-6			

For Experiments 1 and 2, we apply ten-fold cross-validation on the experiment subset of mPower on all models for an objective assessment of model performance. The ten-fold cross-validation is done in the following manner: for each time, eight out of the ten folds of the experiment subset is the training set, one fold is the validation set, and the remaining fold is the test set. The above process is repeated ten times for cross-validation and the results we report are from the test set of each time.

We report the following metrics: mean absolute error (MAE), root mean squared error (RMSE), logarithm mean absolute error (LMAE), and logarithm root mean squared error (LRMSE); all of these are commonly used metrics for regression tasks (Eigen et al. 2014; Kendall et al. 2017). If $\hat{y}_j^{(d)}$ is the model-predicted assessment score for task j in the d-th instance, $y_j^{(d)}$ is the ground-truth assessment score for task j in the d-th instance, and D is the total number of instances, the above metrics are defined as follows:

$$\mathrm{MAE}_j = \frac{1}{D}\sum_{d=1}^{D}\left|\hat{y}_j^{(d)} - y_j^{(d)}\right|, \qquad \mathrm{RMSE}_j = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{y}_j^{(d)} - y_j^{(d)}\right)^2},$$

$$\mathrm{LMAE}_j = \frac{1}{D}\sum_{d=1}^{D}\left|\log\left(\hat{y}_j^{(d)} + 1\right) - \log\left(y_j^{(d)} + 1\right)\right|, \qquad \mathrm{LRMSE}_j = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\log\left(\hat{y}_j^{(d)} + 1\right) - \log\left(y_j^{(d)} + 1\right)\right)^2}.$$

For all four metrics, smaller is better. We also run paired t-tests to identify statistically significant differences between AADMML and benchmark models at p-value threshold of 0.05 (*), 0.01 (**), and 0.001 (***).

In Experiment 3, we specifically investigate what the attention mechanism and the adversarial attention competition mechanism bring to a deep multisource multitask learning framework. We repeat the ten-fold cross validation in Experiment 1 for ten times and plot the averaged changes in loss function values with increasing numbers of training iterations for AADMML, the same model without adversarial attention competition (MTAEF), and MTAEF without the attention mechanism (MTEF). In addition, we show the minimum validation losses and the numbers of model training iterations where the minimum validation losses are achieved for each model, run paired t-tests (AADMML vs. MTAEF and AADMML vs. MTEF) to show whether AADMML significantly outperforms the others with regard to validation losses, and discuss their implications. We also demonstrate how the attention weights change for AADMML and MTAEF in Appendix E (MTEF does not involve attention weights).

The three case studies demonstrate the practical utility of the proposed AADMML. Case study 1 presents two examples that the AADMML accurately assessed the subjects' PD severity. We show the wearable sensor data diagrams of the five sources, allocated task-specific attention weight on each source, PD severity assessment results, and ground-truth PD severity scores. We also discuss how the attention weights can help improve model interpretability.

Case study 2 applies AADMML in the context of identifying early-stage PD patients due to its significance in PD treatment and progression alleviation. We formularize the early-stage PD identification problem as follows: given a population of subjects with no PD or mild PD, can a model accurately identify those subjects with mild PD based on wearable sensor data? We then

discuss how much net monetary benefit would be generated by identifying mild PD and subsequent proper management based on the results generated by AADMML and other best-performing models. We control the false alarm rates (i.e., the proportions of false reported mild PD subjects out of all reported mild PD subjects) and then present the averaged recall rate (i.e., the proportion of true reported mild PD subjects out of all mild PD subjects) to estimate the number of mild PD subjects that each model can identify.

Case study 3 reports the statistics of overall attention weights as well as those grouped by age or gender. At an aggregated level, the attention weights demonstrate population patterns that could contribute to the development of new treatments or therapies.

EXPERIMENT AND CASE STUDY RESULTS

Experiment 1: AADMML vs. State-of-the-art Benchmark Models

Experiment 1 benchmarks AADMML's performance in assessing nM-EDL and M-EDL against state-of-the-art benchmark models, including feature-based machine learning models and deep learning models. Table 9 presents the results. Best results are highlighted in boldface. We also demonstrate how AADMML performs with different demographic groups in Appendix H.

Table 9. PD Severity Assessment Results against Benchmark Models

Category	Model	Task 1: nM-EDL				Task 2: M-EDL			
		MAE	RMSE	LMAE	LRMSE	MAE	RMSE	LMAE	LRMSE
Feature-based, Non-ensemble	DT	2.067***	3.367***	0.309***	0.506***	2.662***	4.953***	0.501	0.743
	KNN	2.235***	3.947***	0.348***	0.538***	2.908***	4.848***	**0.488**	0.736
	SVM	2.856***	3.855***	0.358***	0.530**	3.794***	5.549***	0.765***	0.907***
Feature-based, Ensemble	ETS	1.600	2.372**	0.316*	0.512*	2.107*	3.075**	0.713***	0.895***
	RF	1.538	2.328*	0.325**	0.512*	2.144**	3.096*	0.721***	0.900***
	ADA	2.835***	3.565***	0.380***	0.517*	4.143***	4.583***	0.760***	0.923***
	GBM	1.589	2.603***	0.383***	0.547***	2.224***	3.732***	0.528	0.741
	XGB	1.533	2.539***	0.388***	0.578***	2.170**	3.062*	0.523	0.744
	NEV	2.390***	3.117***	0.346***	0.516*	3.353***	4.300***	0.718***	0.889***
Deep Learning	STLF	1.743**	2.429**	0.341***	0.554***	2.326***	3.335***	0.551**	0.782**
	STEF	1.759**	2.450**	0.337**	0.552***	2.403***	3.199**	0.536*	0.753

	STAEF	1.757***	2.446***	0.346***	0.542***	2.154**	2.991	0.532*	0.752
	MTLF	1.673*	2.438**	0.323*	0.531***	2.276***	3.007*	0.514	0.755*
	MTEF	1.570	2.410**	0.311	0.513*	1.998	2.932	0.516	0.742
	MTAEF	1.523	2.370*	0.318	0.515*	2.074*	**2.859**	0.518	0.735
	AADMML	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	2.867	0.507	**0.732**

Notes: * p<0.05; ** p<0.01; *** p<0.001

In Task 1: assessment of nM-EDL Sum, the proposed AADMML outperformed all feature-based and other deep learning models in all metrics. In Task 2: assessment of M-EDL Sum, AADMML outperformed all feature-based and other deep learning models in MAE and LRMSE. The improved performance of AADMML over feature-based models shows the advantage of deep learning's automatic feature extraction without manual design and selection of feature sets. The improved performance of AADMML over other deep learning models (e.g., MTAEF, which is AADMML without adversarial attention competitions) provides evidence that adversarial learning further improves the state-of-the-art attention-based deep multisource multitask learning, leading to a more accurate assessment of PD severity.

Experiment 2: Sensitivity Analysis of AADMML

In Experiment 2, we adjust the hyperparameters in the proposed AADMML and test the model performance after the adjustments. The results are summarized in Table 10. The model names and results in boldface indicate the hyperparameters we chose among the different options. The significance levels are calculated against the chosen hyperparameters. For comparison, the sensitivity analysis results of select feature-based and deep learning benchmarks can be found in Appendix F.

Table 10. Sensitivity Analysis Results

Model	Task 1: nM-EDL				Task 2: M-EDL			
	MAE	RMSE	LMAE	LRMSE	MAE	RMSE	LMAE	LRMSE
HP-CNN (8)	1.764**	2.259	0.314	0.497	1.929	2.879	0.516	0.745
HP-CNN (16)	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-CNN (32)	1.603	2.245	0.306	0.484	1.889	2.901	0.544*	0.725

HP-CNN (64)	1.600	2.114	0.310	0.495	1.974	2.965	0.511	0.739
HP-CNN-L1	1.567	2.208	0.294	0.502	1.978	2.905	0.509	0.739
HP-CNN-L2	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-CNN-L3	1.543	2.178	0.301	0.492	1.934	2.840	0.511	0.735
HP-CNN-L4	1.595	2.212	0.308	0.491	2.114*	2.893	0.515	0.751
HP-LSTM (8)	1.517	2.237	0.308	0.513*	2.058	2.951	0.521	0.751
HP-LSTM (16)	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-LSTM (32)	1.528	2.184	0.301	0.486	2.074	2.932	0.505	0.730
HP-LSTM (64)	1.514	2.170	0.319	0.501	2.133*	2.973	0.524	0.745
HP-LSTM-L1	1.698**	2.516**	0.314	0.515**	2.277***	2.968	0.527	0.773**
HP-LSTM-L2	1.581	2.345*	0.318	0.516**	2.047	3.001	0.514	0.761*
HP-BLSTM-L1	1.652	2.249	0.295	0.503	1.920	2.907	0.507	0.717
HP-BLSTM-L2	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-LR-1	1.805***	2.268	0.333**	0.516**	2.160**	3.045	0.655***	0.864***
HP-LR-2	1.542	2.176	0.315	0.491	2.020	2.915	0.513	0.742
HP-LR-3	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-LR-4	1.511	2.144	0.307	0.498	2.002	2.930	0.516	0.749
HP-DR-0	1.631	2.355*	0.324*	0.512*	2.087	3.085**	0.531	0.753
HP-DR-25	1.576	2.192	0.310	0.506	2.010	2.978	0.524	0.730
HP-DR-50	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-DR-75	1.574	2.246	0.308	0.496	1.972	2.887	0.518	0.737
HP-AT-1	1.613	2.297	0.322*	0.502	2.062	2.922	0.520	0.756
HP-AT-3	1.544	2.131	0.309	0.491	2.007	2.858	0.515	0.741
HP-AT-5	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-AT-7	1.520	2.129	0.305	0.483	1.979	2.861	0.512	0.743
HP-T1LF-0.5	1.667*	2.295*	0.313	0.521**	1.840	2.756	0.504	0.726
HP-T1LF-1.0	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-T1LF-1.5	1.485	2.066	0.294	0.492	2.042	3.128**	0.522	0.772**
HP-T1LF-2.0	1.496	2.091	0.295	0.477	2.122*	3.166**	0.562***	0.810***
HP-WL2R-3	1.585	2.327*	0.329**	0.527***	2.108*	3.158**	0.551***	0.783***
HP-WL2R-4	1.538	2.261	0.301	0.517*	2.121**	3.063*	0.527	0.757*
HP-WL2R-5	**1.515**	**2.140**	**0.302**	**0.491**	**1.974**	**2.867**	**0.507**	**0.732**
HP-WL2R-6	1.528	2.186	0.303	0.493	1.964	3.001*	0.513	0.747

Notes: * p<0.05; ** p<0.01; *** p<0.001

We can observe that hyperparameters do influence AADMML model performance, although none of them is a decisive factor. Meanwhile, some options of hyperparameters demonstrate advantages consistently in most evaluation metrics (e.g., choosing a learning rate of

10^{-3} generally leads to improved performance). We hope this practice can help IS researchers interested in wearable sensor data and deep learning to design deep learning models.

Experiment 3: Changes of Loss Function Values in Model Training

In Experiment 3, we plot the averaged values of loss functions corresponding to both tasks (Task 1: nM-EDL; Task 2: M-EDL) with increasing numbers of iterations in the training of AADMML, MTAEF, and MTEF to investigate the effect of adversarial attention competition and attention weights in general. Results can be found in Figure 4. We also show the minimum validation losses and the numbers of model training iterations where the minimum validation losses are achieved for each model in Table 11. Finally, we demonstrate how attention weights change with increasing training iterations in Appendix E.

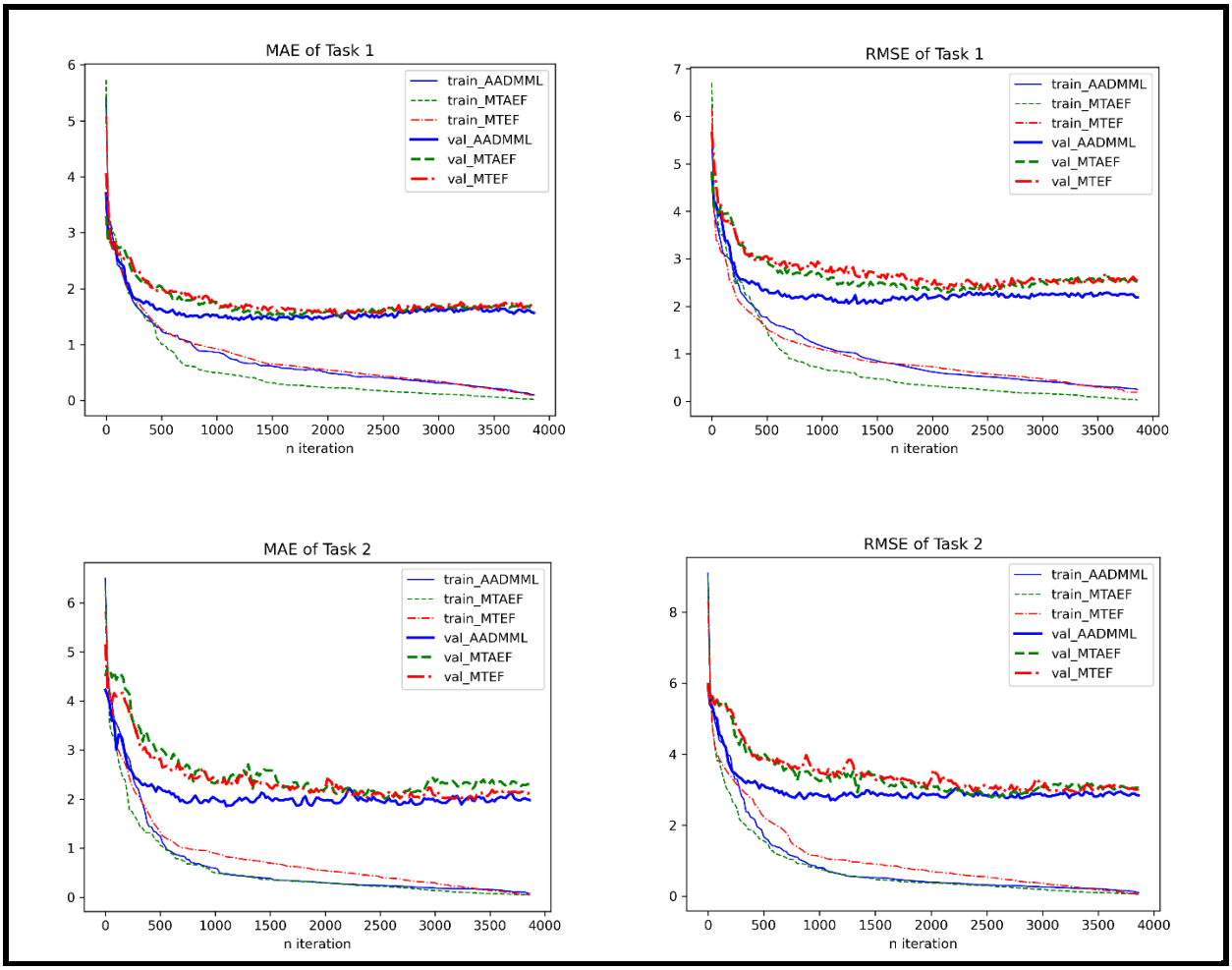


Figure 4. Changes of Loss Function Values in Model Training

Table 11. Minimum Validation Losses for Each Model

Model		Task 1: nM-EDL		Task 2: M-EDL	
		MAE	RMSE	MAE	RMSE
AADMML	Validation Loss	**1.439**	**2.057**	**1.856**	**2.704**
	# Iteration	**1200**	**1320**	**1120**	**1120**
MTAEF	Validation Loss	1.504*	2.296**	2.001*	2.771
	# Iteration	2100	2140	2600	2600
MTEF	Validation Loss	1.527*	2.359**	1.974*	2.909*
	# Iteration	2140	2460	3400	3400

Notes: * p<0.05; ** p<0.01; *** p<0.001

We choose MAE and RMSE as the representatives of loss functions due to space constraints. Model names with a "val" prefix represent the values of loss functions on the validation dataset, while names with a "train" prefix represent those on the training dataset. From Figure 4, we can observe that up to a certain number of iterations, both the training and validation losses are decreasing. However, after the validation losses achieve minima, they are increasing despite training losses still decreasing. In other words, the training losses are typically monotonically decreasing, while the validation losses are U-shaped. The increase of validation losses is a signal of model overfitting with excessive training iterations. A typical practice in deep learning training is to identify the number of iterations where the validation loss minimum is achieved and use it as the indicator that the model is adequately trained and converged.

We can observe that the validation losses (the thicker lines) of AADMML decrease much faster than those of MTAEF and MTEF. The validation losses of AADMML approach minima in between 1,200 and 1,320 iterations in task 1 and about 1,120 iterations in task 2. By contrast, those of MTAEF approach minima after 2,100 iterations in task 1 and after 2,600 iterations in task 2, and those of MTEF approach minima after 2,140 iterations in task 1 and after 3,400 iterations in task 2. This shows a clear advantage in applying the adversarial attention

competition as it speeds up model training. In addition, the validation loss minima of AADMML are less than those of MTAEF, which means more precise predictions by AADMML. To summarize, the adversarial attention competition mechanism both reduces the number of training iterations and achieves less validation loss minima, collectively leading to improved model performance.

Case Study 1: Examples of AADMML PD Severity Assessment

We present two representative examples of wearable sensor-based PD severity assessment in Figure 5, and illustrate how the attention weights can be used as indicators of specific activities that health professionals and patients should pay more attention to, which improves model interpretability. In both examples, we show the wearable sensor data charts for the five data sources (acc-walk, acc-stand, gyro-walk, gyro-stand, and mic), task-specific attention weights learned by AADMML, PD severity assessment results (nM-EDL and M-EDL) by AADMML, and ground-truth nM-EDL and M-EDL reported by the subjects. In the first example, the subject was a male aged 63 with mild PD symptoms (nM-EDL = 6, M-EDL = 10). Based on the experimental trials, the AADMML accurately assessed the subject's nM-EDL to be 6.5 and M-EDL to be 9.8. In addition, the attention network allocated the highest attention on acc-walk (0.401 and 0.325), which can be a signal that the subject's translational motion in walking contributes the most to the assessment result and should be further investigated. In contrast, the second subject was a male aged 34 with intermediate PD symptoms (nM-EDL = 14, M-EDL = 8). The AADMML assessed the second subject's nM-EDL to be 12.9 and M-EDL to be 7.5, and allocated more attention on the subject's gyroscope sources (gyro-walk: 0.342 and 0.308; gyro-stand: 0.197 and 0.202). This can be interpreted as the subject's rotational motion patterns are more heavily correlated with the PD severity assessment results. Given such results,

the attention weights can help stakeholders interpret the model's decision-making process, and potentially guide effective therapies to be designed for the subject's specific activities (e.g.,

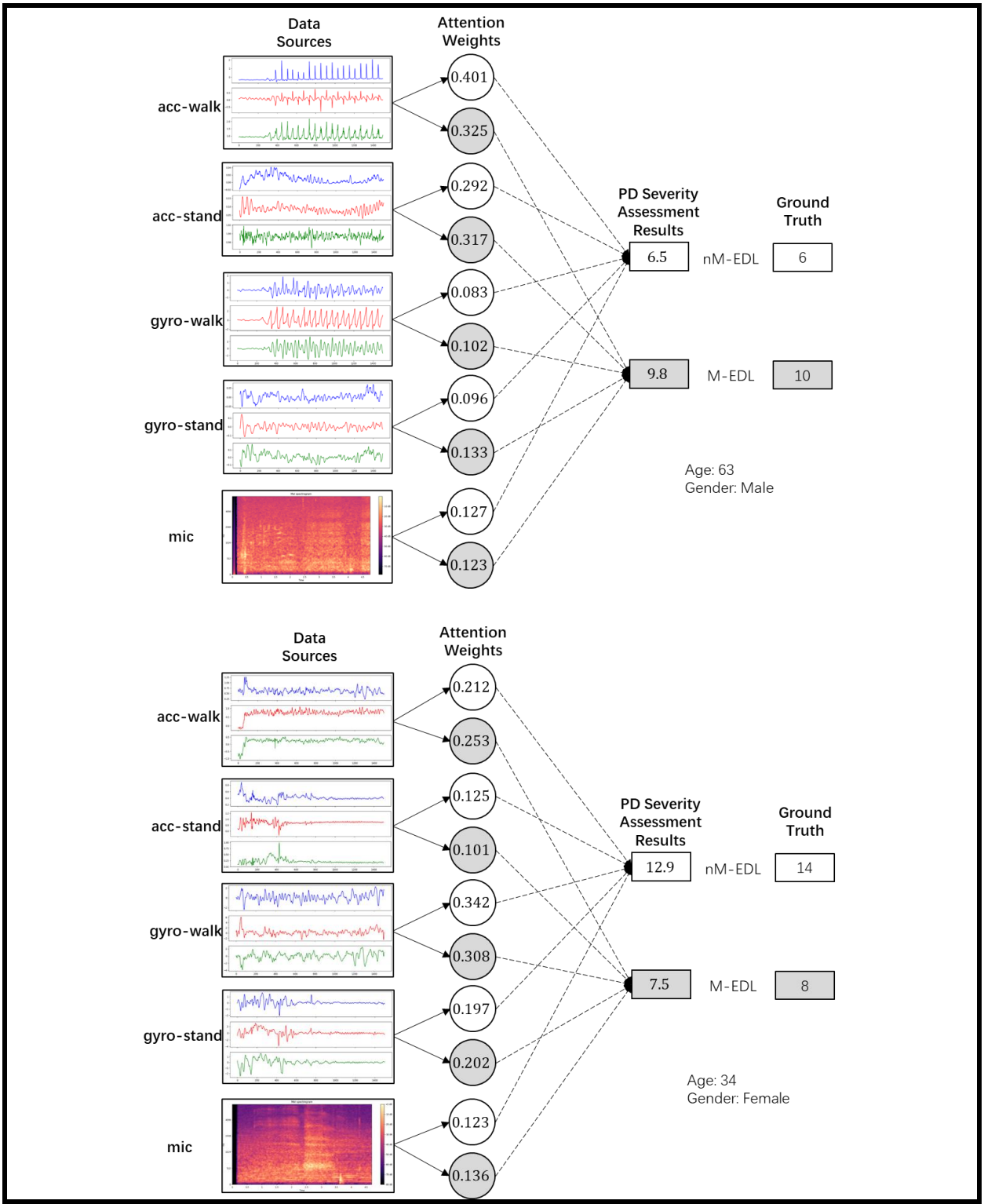


Figure 5. Examples of AADMML PD Severity Assessment

walking or standing). For example, larger attention weights on gyroscope signals may indicate therapies should be designed over the patient's joints, which account for rotational motion.

Case Study 2: Economic Benefit Analysis

Among the numerous uses of wearable sensor-based chronic condition severity assessment, one promising direction is identifying early-stage PD patients. In this section, we discuss the potential economic benefit of identifying early-stage PD patients. The purpose of this case study is to investigate the models' ability to identify early-stage PD patients from a population of potential early-stage PD patients (subjects either without PD or with early-stage PD). To achieve this, we first reorganized the 10% hold-out subset of mPower to include only subjects with nM-EDL < 5 and M-EDL < 10. The thresholds are a proportional rescale of the "mild PD" criteria in the literature (Martinez-Martin et al. 2015) as mPower only contains a subset of the MDS-UPDRS questionnaire. This generates a subset of 131 instances out of the total 248 instances in the hold-out subset. Among the 131 instances, we further label those who have been professionally diagnosed as PD patients (indicated in the mPower demographics table) as "early-stage PD." The remainder is labeled as "normal." As such, the dataset is further grouped as 40 "early-stage PD" and 91 "normal" instances. After this reorganization, we run the already trained models on the 131 instances to investigate how they generate nM-EDL and M-EDL assessment scores, which will be further translated to "early-stage PD" or "normal." The precision-recall tradeoffs of the models are conducted by adjusting the minimum sum of nM-EDL and M-EDL required for being classified as "early-stage PD."

In addition to AADMML, we chose four best-performing benchmark models from Experiment 1 (RF, ETS, MTAEF, MTEF) for comparison. To fully investigate the model's performance at different precision levels, we control their precision rates at all possible levels

from 0.4 to 1.0[1] and calculate their recall rates as the proportions at which the models can successfully identify early-stage PD patients. For instance, a recall rate of 80 percent means the model can identify 80 out of 100 early-stage PD patients. For a comprehensive comparison, we report the averaged recall rate over all possible precision rates for each model. The Precision-Recall Curve (PR Curve) and the area under the curve (AUC) for the models can be found in Appendix D.

For the purpose of this economic benefit analysis, we assume that 60,000 U.S. citizens become new early-stage PD patients every year (Heusinkveld et al. 2018) and assume that each correctly identified early-stage PD patient with proper management would result in net monetary benefits of $60,657 (Johnson et al. 2013). Meanwhile, if a model misclassified a normal person as "early-stage PD" (i.e., a false alarm), we roughly assume there is a $500 false alarm cost[2] (Wilkins et al. 2012) due to unnecessary clinic visits and other redundant screenings. For each model, we calculate the averaged recall rate, number of identified early-stage PD patients, economic benefit from correctly identified early-stage PD patients, cost for false alarms, net economic benefit, and AADMML's advantage in net economic benefit. Results are summarized in Table 12.

Table 12. Economic Benefit Analysis Results

Model	Recall	#Identified Early-Stage PD Patients	Economic Benefit from PD Patients	Cost for False Alarms	Net Economic Benefit	AADMML's Advantage in Net Economic Benefit
AADMML	**0.8000**	**48,000**	**$2,911,536,000**	**$14,869,000**	**$2,896,667,000**	**N/A**
RF	0.7096	42,577	$2,582,593,089	$14,473,000	$2,568,120,089	$328,546,911
ETS	0.7231	43,385	$2,631,603,945	$14,663,000	$2,616,940,945	$279,726,055
MTAEF	0.7596	45,577	$2,764,564,089	$14,873,500	$2,749,690,589	$146,976,411
MTEF	0.7365	44,192	$2,680,554,144	$14,858,500	$2,665,695,644	$230,971,356

Notes: **# Identified Early-Stage PD Patients** = Recall × # New Early-Stage PD Patients (60,000); **Economic Benefit from PD Patients** = # Identified Early-Stage PD Patients × Cost for Each Case

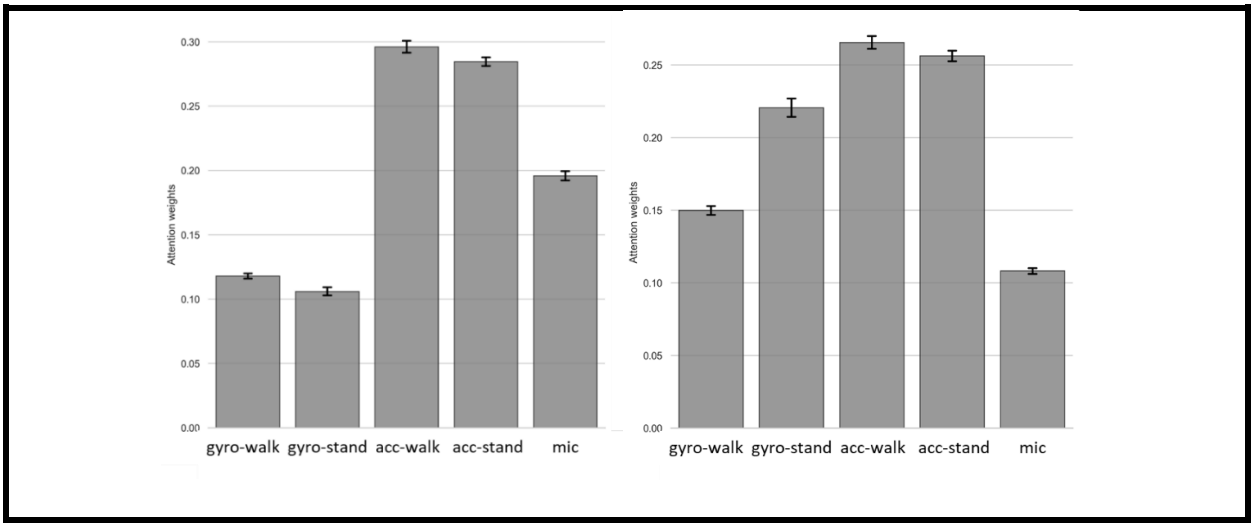
[1] Precisions less than 0.4 are not considered as at such precision levels all models achieve identical recalls.
[2] https://www.eurekalert.org/pub_releases/2007-02/ra-lpd022207.php. Retrieved on October 10, 2021.

($60,657); **Cost for False Alarms** = (# Identified Early-Stage PD Patients / Precision – # Identified Early-Stage PD Patients) × Cost for Each Case ($500); **Net Economic Benefit** = Economic Benefit from PD Patients – Cost for False Alarms. All values are averaged across precision levels from 0.4 to 1.0.

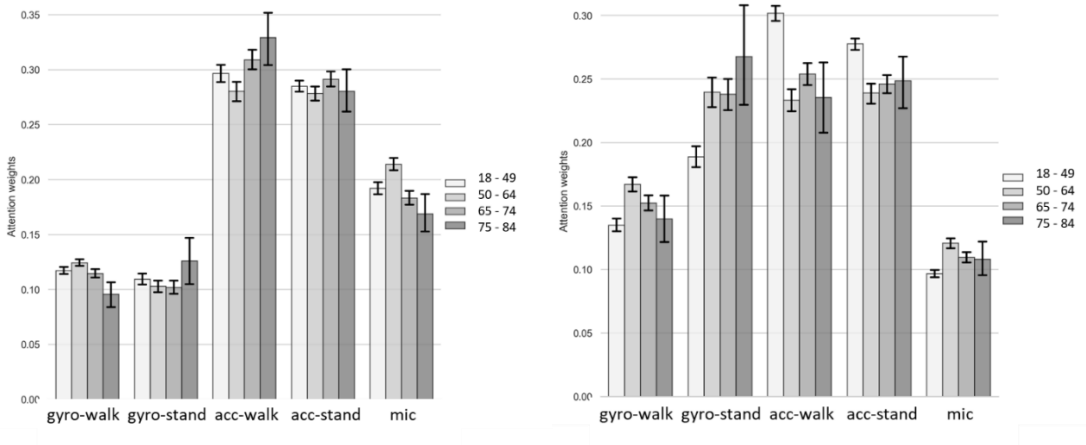
By assessing PD severities based on walking and standing tests, AADMML could identify about 48,000 out of the 60,000 new early-stage PD patients every year, leading to an estimated net economic benefit exceeding $2.8 billion. Compared to the best-performing benchmarks, AADMML can recognize at least 2,000 more early-stage PD patients and generate at least an additional $146 million in extra benefits. In summary, this case study demonstrates the significance of PD severity assessment as well as the advantage of AADMML over competing benchmark models.

Case Study 3: Attention Weights Statistics

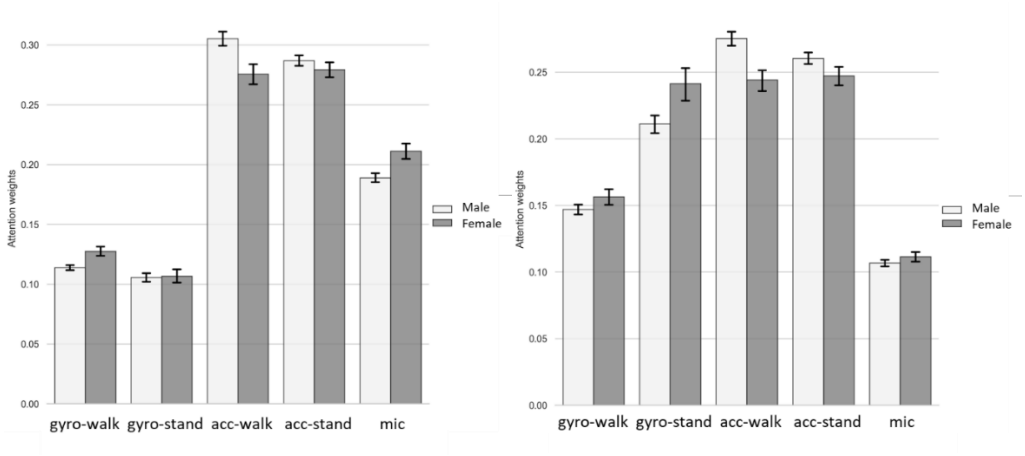
One of the major advantages of AADMML is that it not only assesses the severity of PD, but also reports attention weights that reflect the relative significance of different data sources. In this case study, we report the statistics of the attention weights estimated in the 10% hold-out subset of mPower. We report the overall results on assessing nM-EDL and M-EDL as well as the results further grouped by age or gender (Figure 6). The age groups are chosen following a professional PD report (Lewin Group 2019). The I-shaped marks indicate 95% confidence intervals.



Overall Attention Weights (Left: nM-EDL; Right: M-EDL)



Attention Weights by Age Group (Left: nM-EDL; Right: M-EDL)



Attention Weights by Gender (Left: nM-EDL; Right: M-EDL)

Figure 6. Aggregated Attention Weights Statistics

We can observe that accelerometers (acc-walk, acc-stand) are highly weighted in assessing both nM-EDL and M-EDL. Mic contributes more in assessing nM-EDL, while gyroscopes (gyro-walk, gyro-stand) contribute more in assessing M-EDL. Across different age groups, we notice that the attention weights on gyro-stand for those aged 75 or over are larger compared to other ages in assessing M-EDL, which may indicate that posture stability is a critical factor for older PD patients. Meanwhile, attention weights on acc-walk for those aged 49

or below are larger compared to other ages in assessing M-EDL, which may indicate that walking ability more closely relates to younger PD patients. Between males and females, accelerometers are more significant for males in assessing both tasks, while gyroscopes are more significant for females in general. We hope the above findings show how attention weights can be interpreted and are helpful in future PD diagnosis and therapy development.

DISCUSSION AND CONCLUSIONS

Advanced information technologies such as mobile sensing devices are transforming healthcare for both providers and receivers (Chen et al. 2012). Consumer electronics, such as iPhone, Apple Watch, and Fitbit, and specialized mobile sensor-based caregiving platforms proposed by industry and academia, have opened up a new and promising multi-disciplinary area of research: mobile health analytics. With the vast amount of personal wearable data being collected and stored, mobile health analytics has been receiving increasing focus in the IS community. To a large extent, however, we lack specialized IT artifacts for mobile health analytics to better harness the abundant data for advanced business intelligence.

Benefiting from the recent advance and success of deep learning, we designed a novel Adversarial Attention-based Deep Multisource Multitask Learning (AADMML) framework for more comprehensive chronic condition severity assessment. We proposed an innovative adversarial attention competition mechanism that contributes to the methodology of attention-based multisource multitask learning. We selected PD as our test case due to its significance and prevalence and collected a large-scale dataset, mPower, as our testbed. We conducted rigorous evaluations against state-of-the-art feature-based and deep learning models to validate AADMML's performance. In addition, we presented three case studies that further illustrate the

utility and impact of our proposed AADMML. Next, we discuss the contributions to the IS knowledge base, practical implications, and future directions of this study.

Contributions to the IS Knowledge Base

IS scholars have suggested that novel IT artifacts should contribute prescriptive knowledge to the IS knowledge base (Nunamaker et al. 1990; Hevner et al. 2004). Our major contributions are four-fold: a situated implementation for chronic condition management, a novel mechanism for interpretable deep learning, model design guidelines for future IS researchers, and theoretical and design science contributions. We discuss each below.

Chronic condition management. As indicated in *MIS Quarterly*'s recent special issue on the role of information systems and analysis in chronic disease prevention and management, chronic condition management is clearly a significant domain of interest for IS scholars. Our proposed AADMML is a situated implementation for wearable sensor-based chronic condition severity assessment, which is a critical step towards chronic condition management. By leveraging the novel adversarial attention competition mechanism, the AADMML outperformed traditional feature-based machine learning models as well as other deep learning models, and therefore significantly enhanced wearable sensor-based chronic condition severity assessment. Although we selected PD as a test case, the AADMML can be easily applied on many other chronic conditions (e.g., dementia or frailty) by choosing the proper data sources and assessment tasks that best reflect the respective chronic condition.

Interpretable deep learning. Compared to traditional feature-based machine learning models, deep learning has been a game-changer in terms of its predictive power (e.g., prediction accuracy) and transferability (e.g., a trained deep learning model can be applied for related problems with minimal hyperparameter tuning), especially for many complex sensory tasks. The

improved predictive power is even more important for health-related applications. However, deep learning models (and many complex feature-based models, such as ensemble models) face the interpretability issue: it is very difficult for the model to provide a clear rationale for its predictions, or for humans to inspect what independent variables cause the results. This issue prevents deep learning models from being widely used in many health or medical applications, as such models are relatively weak in informing health professionals which possible factors could potentially lead to a predicted result. The introduction of the attention mechanism is a significant step towards interpretable deep learning, which helps us understand which part of the input data is more relevant to the model output. As an improvement of the current attention-based deep multisource learning models, our proposed adversarial attention competition mechanism is not constrained in the AADMML. Instead, it is a generalizable mechanism that is applicable to other topics that involve multiple data sources and can apply deep learning models. We believe this core methodological contribution pushes the state-of-the-art of interpretable deep learning and multisource learning.

Model design guidelines. Deep learning models, including our proposed AADMML, offer a wide range of options in their structures and hyperparameters. In this study, we showed a good example of building a deep learning model with CNN and LSTM, where CNN is designed to learn the local features of grid-like data while LSTM is designed to model the long temporal sequences. In addition, we conducted extensive sensitivity analyses to investigate how the changes in hyperparameters (size of CNN kernels, number of CNN layers, size of LSTM hidden neurons, structure of LSTM, learning rate, dropout rate, adversarial learning factor, relative weight of Task 1 loss function, and weight of parameter L2-regularization) would impact the

model performance. We hope this practice can facilitate model selection and building in related problem domains for future IS research.

Theoretical and design science contributions. Understanding intertwined interdependencies among different contributing factors and leveraging different types of data sources is an issue in many business problems. Meanwhile, interpretability is key to AI-enabled business decision making as predictive outcomes without traceable reasons are not as informative for human decision makers. The core contribution of this work is in creating an interpretable IT artifact that can deal with novel challenges in the types of inputs as well as in modes of learning tasks. Unlike traditional machine learning models that learn only one task at a time, AADMML is able to simultaneously learn from seemingly distinct task components for more comprehensive decision making. Numerous business problems that employ deep learning-enabled speech recognition and natural language processing, such as customer relationship management and the insurance industry, can benefit largely by applying AADMML to process consumer data that are inherently high-volume and high-variety.

Practical Implications

Since Former President Obama launched the Precision Medicine Initiative in 2015, healthcare research and practice have become increasingly more precise, prompt, and personalized. With the development of sensing technologies, researchers and practitioners are able to collect, store, and analyze various types of sensor data from senior citizens to assess their chronic condition severities and provide proper medical interventions. Senior citizens, their families, and health professionals can all benefit from AADMML, a generalizable and interpretable framework for mobile health analytics. We discuss major practical implications for those stakeholders below.

Senior citizens. Senior citizens face difficulties in their independent living, partly due to the lack of a convenient and reliable approach to better understand their chronic condition severities. In addition, they are now placed at serious risk of contracting COVID-19 due to the necessity of an in-person visit to the clinic. Assistive tools such as instantiations of the AADMML framework can empower senior citizens by enabling them to conduct mobility tests at home, thus complementing a phone or video telemedicine visit to improve the quality of the interaction to be comparable to a traditional evaluation with physical examinations by the healthcare provider. In addition, even without a clinical visit, senior citizens can obtain precise assessments of the progression of existing chronic conditions or the risks of potential chronic conditions, improving their confidence in living independently.

Families. Traditionally, home-dwelling senior citizens require one or more caregivers to ensure that their health conditions are properly monitored, and unforeseen accidents are resolved in a timely manner (e.g., falling down, acute onset of symptoms, etc.). With a remote wearable sensor system equipped with analytical engines such as the AADMML framework, senior citizens' families can attain real-time reports of their loved ones' daily routines and chronic condition severity assessment results without disrupting them, leading to peace of mind.

Health professionals. The AADMML framework is a flexible tool for advanced clinical decision support at the point-of-care for health professionals. It can be integrated into health professionals' clinical routines by conducting mobility tests (e.g., walking and standing tests) to support their diagnoses. Even for highly specialized movement disorder specialists that perform walking and postural tests and utilize the MDS-UPDRS for routine management and clinical trials, the application of the AADMML framework may divulge non-intuitive information that assist in prognosticating falls and hospitalizations. For instance, the attention weights of the five

sources (acc-walk, acc-stand, gyro-walk, gyro-stand, and mic) correspond to the relative significance of five aspects of the patient (translational motion in walking, translational motion in standing, rotational motion in walking, rotational motion in standing, and speech), respectively, thus can provide guidance in designing physical therapies, applying assistive devices such as walkers or wheelchairs, adjustment of medication, among others.

Limitations and Future Research

As with other academic studies in emerging applications, there are limitations in this research. First, we tested our AADMML framework only in the context of PD. Future researchers can instantiate the AADMML framework based on their applications and contexts (e.g., dementia, diabetes, etc.) and explore the boundaries of such a framework. Second, the current instantiation of the AADMML framework can utilize only wearable sensor data as its data sources. With the extreme diversity of available health data, an extended model may be preferred in the case that other types of data are more crucial in assessing certain types of chronic conditions (e.g., blood glucose level for diabetes). Third, our current attention mechanism only points out which data source is more relevant to the assessment result. It may be of great significance if a model can pinpoint the exact timestamp in wearable sensor data that leads to the result, which allows more fine-grained inspection for stakeholders. Fourth, although we empirically demonstrated that adversarial attention competitions speed up model training and improves model performance, we have not found a theoretical confirmation. We believe that the above directions of research are of great interest to researchers and practitioners to further enhance senior citizens' quality of life.

ACKNOWLEDGMENT

This study was supported by USA NSF SES-1314631, DUE-1303362, IIP-1622788. Yidong Chai is the corresponding author. Yidong Chai was supported by Program of the National Natural Science Foundation of China (72101079, 91846201).

AUTHOR BIOGRAPHIES

Shuo Yu is an Assistant Professor at Rawls College of Business, Texas Tech University. He received his Ph.D. in Management Information Systems from the University of Arizona. Dr. Yu's research focuses on data mining, deep learning, health analytics, and mobile analytics. His work has been published or accepted in such journals as *Journal of Management Information Systems, IEEE Journal of Biomedical and Health Informatics*, *IEEE Intelligent Systems, ACM Transactions on Management Information Systems*, among others.

Yidong Chai is a Professor at School of Management, Hefei University of Technology. He received his Ph.D. in Management Science and Engineering from Tsinghua University. Dr. Chai's research focuses on machine learning and its applications on healthcare, cybersecurity, e-business, among others. His work has been published or accepted in such journals as *MIS Quarterly, Journal of Management Information Systems, Information Processing and Management, Knowledge Based Systems*, *Applied Soft Computing*, among others.

Hsinchun Chen is Regents Professor and Thomas R. Brown Chair in Management and Technology at the Eller College of Management, University of Arizona. He received his Ph.D. in Information Systems from New York University. He is author or editor of 20 books, 300 journal

papers, and 200 refereed conference articles covering digital library, data/text/web mining, business analytics, security informatics, and health informatics. He served as the lead Program Director of the Smart and Connected (SCH) Program at the National Science Foundation (NSF). Dr. Chen founded the Artificial Intelligence Lab at University of Arizona, which has received $50M+ research funding from the NSF, National Institutes of Health, National Library of Medicine, Department of Defense, Department of Justice, Central Intelligence Agency, Department of Homeland Security, and other agencies. He is a Fellow of ACM, IEEE, and AAAS.

Scott J. Sherman, MD, PhD is Director of the Movement Disorders Clinic at the University of Arizona with 30 years of experience treating patients with Parkinson Disease (PD). He splits his time 50/50 between patient care and research in the field of Movement Disorders including both pharma-sponsored studies and investigator-initiated research. Dr. Sherman is an expert in the clinical evaluation of patient with PD and also in the use of standardized rating scales to quantify clinical phenotypes and disease staging.

Randall A. Brown, MD, MBA is a board-certified Internal Medicine physician who has 30+ years of clinical medical experience providing patient care in academic medical centers in the United States. He collaborated with the University of Arizona Artificial Intelligence Lab from 2010 to 2019 as a medical domain expert advising on healthcare related projects that included data mining and predictive analytics utilizing electronic health records as well as the machine learning processing of healthcare related wireless mobile sensor data. He has retired from clinical medical practice and currently works as a medical advisor with Optum Insight in Walnut Creek,

California as well as president of his independent consulting firm, Hermes Medical Intelligence, LLC.

REFERENCES

Adipat, B., Zhang, D., and Zhou, L. 2011. "The Effects of Tree-View Based Presentation Adaption on Mobile Web Browsing," *MIS Quarterly* (35:1), pp. 99–122.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., and others. 2016. "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," in *Proceedings of the 2016 International Conference on Machine Learning*, pp. 173–182.

Anand, V., Bilal, E., Ramos, V., Naylor, M., Demanuele, C., Zhang, H., Amato, S., Wacnik, P., Hameed, F., Kangarloo, T., Ho, B., Erb, K., and Karlin, D. 2018. "F62. Automatic Detection of ON/OFF States in Parkinson Disease Patients Using Wearable Inertial Sensors," *Clinical Neurophysiology* (129), Elsevier, p. e90.

Ando, R. K., Zhang, T., and Bartlett, P. 2005. "A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data.," *Journal of Machine Learning Research* (6:11), pp. 1817–1853.

Baird, A., Angst, C., and Oborn, E. 2018. "MISQ Research Curation on Health Information Technology," *MISQ Research Curations*, pp. 1–5. (https://www.misqresearchcurations.org/blog/2018/6/20/health-information-technology).

Bahdanau, D., Cho, K., and Bengio, Y. 2014. "Neural Machine Translation by Jointly Learning to Align and Translate," *ArXiv Preprint ArXiv:1409.0473*. (https://arxiv.org/pdf/1409.0473).

Balabka, D. 2019. "Semi-Supervised Learning for Human Activity Recognition Using Adversarial Autoencoders," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pp. 685–688.

Baltrusaitis, T., Ahuja, C., and Morency, L. P. 2019. "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (41:2), pp. 423–443.

Bao C., Bardhan I. R., Singh H., Myer B. A., and Kirksey Kirk. 2020. "Patient-Provider Engagement and its Impact on Health Outcomes: A Longitudinal Study of Patient Portal Use," *MIS Quarterly* (44:2), pp. 699–723.

Bin, Y., Yang, Y., Shen, F., Xie, N., Shen, H. T., and Li, X. 2018. "Describing Video with Attention-Based Bidirectional LSTM,". *IEEE Transactions on Cybernetics* (49:7), pp. 2631-2641.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., …., and Dorsey, E. R. 2016. "The MPower Study, Parkinson Disease Mobile Data Collected Using ResearchKit," *Scientific Data* (3), Nature Publishing Group, p. 160011.

Brohman K., Addas S., Dixon J., and Pinsonneault A. 2020. "Cascading Feedback: A Longitudinal Study of a Feedback Ecosystem for Telemonitoring Patients with Chronic Disease," *MIS Quarterly* (44:1b), pp. 421–450.

Buttorff, C., Ruder, T., and Bauman, M. 2017. *Multiple Chronic Conditions in the United States*, (Vol. 10), Rand Santa Monica, CA.

Caruana, R. 1998. "Multitask Learning," in *Learning to Learn*, pp. 95–133.

Centers for Disease Control and Prevention. 2020. "Health and Economic Costs of Chronic Diseases." (https://www.cdc.gov/chronicdisease/about/costs/index.htm).

Chaudhari, S., Polatkan, G., Ramanath, R., and Mithal, V. 2019. "An Attentive Survey of

Attention Models," *ArXiv Preprint ArXiv:1904.02874*. (https://arxiv.org/pdf/1904.02874).
Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165–1188.
Chen, L., Baird, A., and Straub, D. 2019. "Fostering Participant Health Knowledge and Attitudes: An Econometric Study of a Chronic Disease-Focused Online Health Community," *Journal of Management Information Systems* (36:1), Taylor & Francis, pp. 194–229.
Chen, W., Wang, S., Zhang, X., Yao, L., Yue, L., Qian, B., and Li, X. 2018. "EEG-Based Motion Intention Recognition via Multi-Task RNNs," in *Proceedings of the 2018 SIAM International Conference on Data Mining*, pp. 279–287.
Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. 2017. "Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks," in *Proceedings of the 2017 Machine Learning for Healthcare Conference*, pp. 286–305.
Coffey, R. 2016. "Submucous or Physiological Implantation of Ureter into the Large Intestine," *Urologic and Cutaneous Review* (23:8), pp. 115–123.
Davoodnia, V., Slinowsky, M., and Etemad, A. 2020. "Deep Multitask Learning for Pervasive Bmi Estimation and Identity Recognition in Smart Beds," *Journal of Ambient Intelligence and Humanized Computing*, Springer, pp. 1–15.
de Bois, M., El Yacoubi, M. A., and Ammi, M. 2021. "Adversarial Multi-Source Transfer Learning in Healthcare: Application to Glucose Prediction for Diabetic People," *Computer Methods and Programs in Biomedicine* (199), Elsevier, p. 105874.
de Jong, R. 2019. "Multimodal Deep Learning for the Classification of Human Activity: Radar and Video Data Fusion for the Classification of Human Activity," *Dissertation*.
Deng, L., and Yu, D. 2014. "Deep Learning: Methods and Applications," *Foundations and Trends in Signal Processing* (7:3-4), Now Publishers Inc. Hanover, MA, USA, pp. 197–387.
Eigen, D., Puhrsch, C., and Fergus, R. 2014. "Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network," *Advances in Neural Information Processing Systems* (3), pp. 2366–2374.
Gao, X., Fuli, F., Xiangnan, H., Huang, H., Guan, X., Feng, C., Ming, Z., and Chua, T.-S. 2020. "Hierarchical Attention Network for Visually-Aware Food Recommendation," *IEEE Transactions on Multimedia* (22:6), IEEE, pp. 1647–1659.
Ghaleb, E., Niehues, J., and Asteriadis, S. 2020. "Multimodal Attention-Mechanism For Temporal Emotion Recognition," in *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*, pp. 251–255.
Ghassemi, N., Shoeibi, A., and Rouhani, M. 2020. "Deep Neural Network with Generative Adversarial Networks Pre-Training for Brain Tumor Classification Based on MR Images," *Biomedical Signal Processing and Control* (57), Elsevier, p. 101678.
Ghose A., Goldfarb A., and Han S.P. 2012. "How Is the Mobile Internet Different? Search Costs and Local Activities," *Information Systems Research* (24:3), pp. 613–631.
Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., Olanow, C. W., Rascol, O., Schrag, A., Teresi, J. A., van Hilten, J. J., and LaPelle, N. 2008. "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 (23:15), Wiley Online Library, pp. 2129–2170.

Gonzalez, T. F. 2007. *Handbook of Approximation Algorithms and Metaheuristics*, CRC Press.

Gooch, C. L., Pracht, E., and Borenstein, A. R. 2017. "The Burden of Neurological Disease in the United States: A Summary Report and Call to Action," *Annals of Neurology* (81:4), John Wiley and Sons, pp. 479–484.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*, MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. "Generative Adversarial Nets," *Advances in Neural Information Processing Systems* (27), pp. 1–9.

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact" *MIS Quarterly* (37:2), pp. 337–355.

Heusinkveld, L. E., Hacker, M. L., Turchan, M., Davis, T. L., and Charles, D. 2018. "Impact of Tremor on Patients with Early Stage Parkinson's Disease," *Frontiers in Neurology*, pp. 1–5.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.

Hoehle H., and Venkatesh V. 2015. "Mobile Application Usability: Conceptualization and Instrument Development," *MIS Quarterly* (39:2), pp. 435–472.

Howcroft, J., Kofman, J., and Lemaire, E. D. 2013. "Review of Fall Risk Assessment in Geriatric Populations Using Inertial Sensors," *Journal of NeuroEngineering and Rehabilitation* (10:1), p. 91.

Hu, Y., Wong, Y., Wei, W., Du, Y., Kankanhalli, M., and Geng, W. 2018. "A Novel Attention-Based Hybrid CNN-RNN Architecture for SEMG-Based Gesture Recognition," *PLoS One* (13:10), pp. 1–18.

Hubble, R. P., Naughton, G. A., Silburn, P. A., and Cole, M. H. 2015. "Wearable Sensor Use for Assessing Standing Balance and Walking Stability in People with Parkinson's Disease: A Systematic Review," *PloS One* (10:4), Public Library of Science, p. e0123705.

Hwang, U., Choi, S., Lee, H.-B., and Yoon, S. 2017. "Adversarial Training for Disease Prediction from Electronic Health Records with Missing Data," *ArXiv Preprint ArXiv:1711.04126*. (https://arxiv.org/pdf/1711.04126).

Jacob, L., Vert, J., and Bach, F. 2008. "Clustered Multi-Task Learning: A Convex Formulation," *Advances in Neural Information Processing Systems* (21), pp. 745–752.

Jalali, A., Sanghavi, S., Ruan, C., and Ravikumar, P. 2010. "A Dirty Model for Multi-Task Learning," *Advances in Neural Information Processing Systems* (23), pp. 964–972.

Jin, C. Bin, Kim, H., Liu, M., Jung, W., Joo, S., Park, E., Ahn, Y. S., Han, I. H., Lee, J. Il, and Cui, X. 2019. "Deep CT to MR Synthesis Using Paired and Unpaired Data," *Sensors* (19:10), pp. 1–19.

Johnson, S. J., Diener, M. D., Kaltenboeck, A., Birnbaum, H. G., and Siderowf, A. D. 2013. "An Economic Model of Parkinson's Disease: Implications for Slowing Progression in the United States," *Movement Disorders* (28:3), Wiley Online Library, pp. 319–326.

Kalantarian, H., Alshurafa, N., Pourhomayoun, M., Sarin, S., Le, T., and Sarrafzadeh, M. 2014. "Spectrogram-Based Audio Classification of Nutrition Intake," in *Proceedings of the 2014 IEEE Healthcare Innovation Conference (HIC)*, pp. 161–164.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., and Bry, A. 2017. "End-to-End Learning of Geometry and Context for Deep Stereo Regression," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 66–75.

Kohli, R., and Tan, S. S. L. 2016. "Electronic Health Records: How Can IS Researchers Contribute to Transforming Healthcare?" *MIS Quarterly* (40:3), pp. 553–573.
Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* (25), pp. 1097–1105.
Kwon, H. E., So, H., Han, S. P., and Oh, W. 2016. "Excessive Dependence on Mobile Social Apps: A Rational Addiction Perspective," *Information Systems Research* (27:4), pp. 919–939.
Lang, Y., Wang, Q., Yang, Y., Hou, C., Liu, H., and He, Y. 2019. "Joint Motion Classification and Person Identification via Multitask Learning for Smart Homes," *IEEE Internet of Things Journal* (6:6), IEEE, pp. 9596–9605.
LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *Nature* (521:7553), pp. 436–444.
Lewin Group (2019) *Economic Burden and Future Impact of Parkinson's Disease*. Retrieved from https://www.michaeljfox.org/sites/default/files/media/document/2019 Parkinson%27s Economic Burden Study - FINAL.pdf.
Liao, Q., Ding, Y., Jiang, Z. L., Wang, X., Zhang, C., and Zhang, Q. 2019. "Multi-Task Deep Convolutional Neural Network for Cancer Diagnosis," *Neurocomputing* (348), pp. 66–73.
Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., and Yang, H. J. 2017. "Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach," *MIS Quarterly* (41:2), pp. 473–495.
Liu Q., Liu X., and Guo X. 2020. "The Effects of Participating in Physician-Driven Online Health Community in Managing Chronic Disease: Evidence from Two Natural Experiments," *MIS Quarterly* (44:1b), pp. 391–419.
Liu X., Zhang B., Susarla A., and Padman R. 2020. "Go to YouTube and Call Me in the Morning: Use of Social Media for Chronic Conditions," *MIS Quarterly* (44:1b), pp. 257–283.
Long, M., Cao, Y., Cao, Z., Wang, J., and Jordan, M. I. 2018. "Transferable Representation Learning with Deep Adaptation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (41:12) pp. 3071 - 3085.
Luo, P., Li, Y., Tian, L. P., and Wu, F. X. 2019. "Enhancing the Prediction of Disease-Gene Associations with Multimodal Deep Learning," *Bioinformatics* (35:19), pp. 3735–3742.
Ma, H., Li, W., Zhang, X., Gao, S., and Lu, S. 2019. "Attnsense: Multi-Level Attention Mechanism for Multimodal Human Activity Recognition," in *Proceedings of the 2019 International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3109–3115.
Martinez-Martin, P., Rodriguez-Blazquez, C., Alvarez, M., Arakaki, T., Arillo, V. C., Chana, P., Fernandez, W., Garretto, N., Martinez-Castrillo, J. C., Rodriguez-Violante, M., Serrano-Duenas, M., Ballesteros, D., Rojo-Abuin, J. M., Chaudhuri, K. R., and Merello, M. 2015. "Parkinson's Disease Severity Levels and MDS-Unified Parkinson's Disease Rating Scale," *Parkinsonism & Related Disorders* (21:1), Elsevier, pp. 50–54.
Mayo Clinic. 2020. "Parkinson's Disease." (https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/diagnosis-treatment/drc-20376062).
McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. 2015. "Librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference* (Vol. 8), pp. 18–25.
Millor, N., Lecumberri, P., Gomez, M., Martinez, A., Martinikorena, J., Rodriguez-Manas, L., Garcia-Garcia, F. J., and Izquierdo, M. 2017. "Gait Velocity and Chair Sit-Stand-Sit

Performance Improves Current Frailty-Status Identification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (25:11), pp. 2018–2025.
Moeskops, P., Wolterink, J. M., van der Velden, B. H. M., Gilhuijs, K. G. A., Leiner, T., Viergever, M. A., and Isgum, I. 2016. "Deep Learning for Multi-Task Medical Image Segmentation in Multiple Modalities," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 478–486.
Moon, S., Song, H. J., Sharma, V. D., Lyons, K. E., Pahwa, R., Akinwuntan, A. E., and Devos, H. 2020. "Classification of Parkinson's Disease and Essential Tremor Based on Balance and Gait Characteristics from Wearable Motion Sensors via Machine Learning Techniques: A Data-Driven Approach," *Journal of NeuroEngineering and Rehabilitation* (17:1), pp. 1–8.
Nemati, E., Rahman, M. M., Nathan, V., and Kuang, J. 2020. "Private Audio-Based Cough Sensing for in-Home Pulmonary Assessment Using Mobile Devices," *EAI/Springer Innovations in Communication and Computing* (3), pp. 221–232.
Nunamaker, J. F., Chen, M., and Purdin, T. D. M. 1990. "Systems Development in Information Systems Research," *Journal of Management Information Systems* (7:3), Taylor & Francis, pp. 89–106.
Obozinski, G., Taskar, B., and Jordan, M. I. 2010. "Joint Covariate Selection and Joint Subspace Selection for Multiple Classification Problems," *Statistics and Computing* (20:2), Springer, pp. 231–252.
Piau, A., Mattek, N., Crissey, R., Beattie, Z., Dodge, H., and Kaye, J. 2019. "When Will My Patient Fall? Sensor-Based In-Home Walking Speed Identifies Future Falls in Older Adults," *The Journals of Gerontology: Series A*, pp. 1–6.
Polat, K. 2019. "Freezing of Gait (FoG) Detection Using Logistic Regression in Parkinson's Disease from Acceleration Signals," in *Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1–4.
Qiao, Z., Wu, X., Ge, S., and Fan, W. 2019. "MNN: Multimodal Attentional Neural Networks for Diagnosis Prediction," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 5937–5943.
Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., and Kawsar, F. 2017. "Multimodal Deep Learning for Activity and Context Recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (1:4), pp. 1–27.
Rai, A., 2017. "Editor's Comments: Avoiding Type III Errors: Formulating IS Research Problems That Matter," *MIS Quarterly* (41:2), pp. 3–7.
Rastegari, E., Azizian, S., and Ali, H. 2019. "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* (Vol. 6), pp. 4231–4242.
Savoli A., Barki H., and Pare G. 2020. "Examining How Chronically Ill Patients' Reactions To, and Effective Use of, Information Technology Can Influence How Well They Self-Manage Their Illness," *MIS Quarterly* (44:1b), pp. 351–389.
Shi Z., Zuo W., Chen W., Yue L., Hao Y., and Liang S. 2019. "DPSCAN : Structural Graph Clustering," *Springer Nature Switzerland AG* (2), pp. 53–69.
Son J., Brennan P. F., and Zhou S. 2020. "Data Analytics Framework for Smart Asthma Management Based on Remote Health Information Systems with Bluetooth-Enabled Personal Inhalers," *MIS Quarterly* (44:1b), pp. 285–303.

Sun, L., Lian, Z., Tao, J., Liu, B., and Niu, M. 2020. "Multi-Modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism," in *MuSe 2020 - Proceedings of the 1st International Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*, pp. 27–34.

Sun, M., Tang, F., Yi, J., Wang, F., and Zhou, J. 2018. "Identify Susceptible Locations in Medical Records via Adversarial Attacks on Deep Predictive Models," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 793–801.

Sun, Y., Yuan, P., and Sun, Yuming. 2020. "MM-GAN: 3D MRI Data Augmentation for Medical Image Segmentation via Generative Adversarial Networks," in *Proceedings of the 11th IEEE International Conference on Knowledge Graph, ICKG 2020*, pp. 227–234.

Sun, Z., Dawande, M., Janakiraman, G., and Mookerjee, V. 2017. "Not Just a Fad: Optimal Sequencing in Mobile In-App Advertising," *Information Systems Research* (28:3), pp. 511–528.

Tang, Y.-X., Tang, Y.-B., Han, M., Xiao, J., and Summers, R. M. 2019. "Deep Adversarial One-Class Learning for Normal and Abnormal Chest Radiograph Classification," in *Medical Imaging 2019: Computer-Aided Diagnosis* (Vol. 10950), p. 1095018.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. 2016. "Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204.

Venkatesh, V., Aloysius, J.A., Hoehle, H. and Burton, S., 2017. "Design and Evaluation of Auto-ID Enabled Shopping Assistance Artifacts in Customers' Mobile Phones: Two Retail Store Laboratory Experiments," *MIS Quarterly* (41:1), pp. 83–113.

Vlachostergiou, A., Tagaris, A., Stafylopatis, A., and Kollias, S. 2018. "Investigating the Best Performing Task Conditions of a Multi-Tasking Learning Model in Healthcare Using Convolutional Neural Networks: Evidence from a Parkinson'S Disease Database," in *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2047–2051.

von Coelln, R., Dawe, R. J., Leurgans, S. E., Curran, T. A., Truty, T., Yu, L., Barnes, L. L., Shulman, J. M., Shulman, L. M., Bennett, D. A., and others. 2019. "Quantitative Mobility Metrics from a Wearable Sensor Predict Incident Parkinsonism in Older Adults," *Parkinsonism & Related Disorders* (65), Elsevier, pp. 190–196.

Wang, J., Chen, Y., Gu, Y., Xiao, Y., and Pan, H. 2018. "SensoryGANs: An Effective Generative Adversarial Framework for Sensor-Based Human Activity Recognition," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Wang, J., Chen, Y., Hao, S., Peng, X., and Hu, L. 2019. "Deep Learning for Sensor-Based Activity Recognition: A Survey," *Pattern Recognition Letters* (119), Elsevier B.V., pp. 3–11.

Wang, X., Chen, H., Ran, A. R., Luo, L., Chan, P. P., Tham, C. C., Chang, R. T., Mannil, S. S., Cheung, C. Y., and Heng, P. A. 2020. "Towards Multi-Center Glaucoma OCT Image Screening with Semi-Supervised Joint Structure and Function Multi-Task Learning," *Medical Image Analysis* (63), Elsevier B.V., p. 101695.

Watanabe, A., Noguchi, H., Oe, M., Sanada, H., and Mori, T. 2017. "Development of a Plantar Load Estimation Algorithm for Evaluation of Forefoot Load of Diabetic Patients during

Daily Walks Using a Foot Motion Sensor," *Journal of Diabetes Research* (2017), p. 5350616.
Wilkins, E. J., Rubio, J. P., Kotschet, K. E., Cowie, T. F., Boon, W. C., O'Hely, M., Burfoot, R., Wang, W., Sue, C. M., Speed, T. P., Stankovitch, J., and Horne, M. K. 2012. "A DNA Resequencing Array for Genes Involved in Parkinson's Disease," *Parkinsonism & Related Disorders* (18:4), Elsevier, pp. 386–390.
Williams, C., Bonilla, E. V, and Chai, K. M. 2007. "Multi-Task Gaussian Process Prediction," *Advances in Neural Information Processing Systems* (20), pp. 153–160.
World Health Organization. 2016. "World Health Statistics 2016: Monitoring Health for the SDGs Annex B: Tables of Health Statistics by Country, WHO Region and Globally." (http://www.who.int/gho/publications/world_health_statistics/2016/Annex_B/en/, accessed January 23, 2018).
Wu, C., Gu, Y., and Yu, G. 2019. "DPSCAN: Structural Graph Clustering Based on Density Peaks," in *Proceedings of the 2019 International Conference on Database Systems for Advanced Applications*, pp. 626–641.
Wu, C., Wei, Y., Chu, X., Weichen, S., Su, F., and Wang, L. 2018. "Hierarchical Attention-Based Multimodal Fusion for Video Captioning," *Neurocomputing* (315), Elsevier B.V., pp. 362–370.
Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. 2015. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the 2015 International Conference on Machine Learning*, pp. 2048–2057.
Xue, H., Jiang, W., Miao, C., Yuan, Y., Ma, F., Ma, X., Wang, Y., Yao, S., Xu, W., Zhang, A., and Su, L. 2019. "Deepfusion: A Deep Learning Framework for the Fusion of Heterogeneous Sensory Data," in *Proceedings of the 20th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 151–160.
Xue, Y., Xu, T., Long, L. R., Xue, Z., Antani, S., Thoma, G. R., and Huang, X. 2018. "Multimodal Recurrent Model with Attention for Automated Radiology Report Generation," in *Proceedings of the 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 457–466.
Ye, H. (Jonathan), and Kankanhalli, A. 2018. "User Service Innovation on Mobile Phone Platforms: Investigating Impacts of Lead Userness, Toolkit Support, and Design Autonomy," *MIS Quarterly* (42:1), pp. 165–187.
Yenter, A., and Verma, A. 2017. "Deep CNN-LSTM with Combined Kernels from Multiple Branches For IMDB Review Sentiment Analysis," in *Proceedings of the IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, pp. 540–546.
Yuan, Y., Xun, G., Ma, F., Wang, Y., Du, N., Jia, K., Su, L., and Zhang, A. 2018. "MuVAN: A Multi-View Attention Network for Multivariate Temporal Data," in *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, pp. 717–726.
Zeng, Y., Mao, H., Peng, D., and Yi, Z. 2019. "Spectrogram Based Multi-Task Audio Classification," *Multimedia Tools and Applications* (78:3), Springer, pp. 3705–3722.
Zhang, T., and Shi, M. 2020. "Multi-Modal Neuroimaging Feature Fusion for Diagnosis of Alzheimer's Disease," *Journal of Neuroscience Methods* (341), Elsevier, p. 108795.
Zhang W., and Ram S. 2020. "A Comprehensive Analysis of Triggers and Risk Factors for Asthma Based on Machine Learning and Large Heterogeneous Data Sources," *MIS Quarterly* (44:1b), pp. 305–349.

Zhang, X., Zhuang, F., Li, W., Ying, H., Xiong, H., and Lu, S. 2019. “Inferring Mood Instability via Smartphone Sensing: A Multi-View Learning Approach,” in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1401–1409.

Zhang, Y., and Yang, Q. 2021. “A Survey on Multi-Task Learning,” *IEEE Transactions on Knowledge and Data Engineering* (Early Access), IEEE.

Zhou, T., Ruan, S., Guo, Y., Rouen, I., Apprentissage, L., and Rouen, I. 2020. “A Multi-Modality Fusion Network Based on Attention Mechanism for Brain Tumor Segmentation,” in *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 377–380.

Zou, L., Zheng, J., Miao, C., McKeown, M. J., and Wang, Z. J. 2017. “3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI,” *IEEE Access* (5), pp. 23626–23636.

ONLINE APPENDICES

Appendix A: AADMML Model Specifications

We recognize that model reproducibility is of great significance for a scholarly publication. Therefore, we provide the specifications of the full AADMML model (e.g., selection of layers and hyperparameters). The specifications in Table A1 are applicable for all data sources and assessment tasks.

Table A1. AADMML Model Specifications

Category	Layer Type	Description
Stage 1 (CNN)	Conv2d	Number of kernels: 16; kernel size: 3; stride 1; padding: "same"
	Max pooling	Window size: 2; stride 1; padding: "same"
	Conv2d	Number of kernels: 16; kernel size: 3; stride 1; padding: "same"
	Max pooling	Window size: 2; stride 1; padding: "same"
	Flatten	Flatten to a vector
	Dense	Neurons: 16
Stage 1 (LSTM)	BiLSTM	Num units: 16; return_sequences=True
	BiLSTM	Num units: 16; return_sequences=False
	Dense	Neurons: 32
Stage 2	Multiply	Matrix multiplication; Size: 32
	Tanh	Tanh activation function
	Exponential	Normalize attention weights with a sum of 1
	Multiply	Multiplication of a vector and a scalar
Stage 3	Add	Integrate attention-weighted source-specific features
	Dense	Neurons: 32
	Dropout	Rate: 0.5
	Dense	Neurons: 1

Appendix B: Feature Sets for Feature-based Machine Learning Models

In Tables B1 and B2, we list the features derived from wearable sensor data that are used by feature-based non-ensemble and ensemble machine learning models. These features are among the most widely used features in past literature on wearable sensor-based PD severity assessment (Hubble et al. 2015). Recall that a data sample $\boldsymbol{x}$ from an accelerometer or gyroscope takes the form:

$$\boldsymbol{x} = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_l], \qquad \boldsymbol{a}_i = [a_{\mathrm{x},i} \quad a_{\mathrm{y},i} \quad a_{\mathrm{z},i}]^{\mathrm{T}}, \qquad i \in \{1, \dots, l\},$$

where l is the total number of data points $\boldsymbol{a}_i$ in $\boldsymbol{x}$ and superscript T denotes matrix transpose.

Table B1. Names and Formulas of Features Derived for Feature-based Machine Learning Models from Accelerometers and Gyroscopes

Feature Name	Formula
Mean x-axis values	$u_{\mathrm{x}} = \frac{1}{l}\sum_{i=1}^{l} a_{\mathrm{x},i}$
Mean y-axis values	$u_{\mathrm{y}} = \frac{1}{l}\sum_{i=1}^{l} a_{\mathrm{y},i}$
Mean z-axis values	$u_{\mathrm{z}} = \frac{1}{l}\sum_{i=1}^{l} a_{\mathrm{z},i}$
Standard deviation of x-axis values	$\sigma_{\mathrm{x}} = \sqrt{\frac{1}{l-1}\sum_{i=1}^{l}\left(a_{\mathrm{x},i} - u_{\mathrm{x}}\right)^2}$
Standard deviation of y-axis values	$\sigma_{\mathrm{y}} = \sqrt{\frac{1}{l-1}\sum_{i=1}^{l}\left(a_{\mathrm{y},i} - u_{\mathrm{y}}\right)^2}$
Standard deviation of z-axis values	$\sigma_{\mathrm{z}} = \sqrt{\frac{1}{l-1}\sum_{i=1}^{l}\left(a_{\mathrm{z},i} - u_{\mathrm{z}}\right)^2}$
Mean magnitude	$u_{\lvert\boldsymbol{a}\rvert} = \frac{1}{l}\sum_{i=1}^{l}(\lvert\boldsymbol{a}\rvert_i)$, where $\lvert\boldsymbol{a}\rvert_i = \sqrt{a_{\mathrm{x},i}^2 + a_{\mathrm{y},i}^2 + a_{\mathrm{z},i}^2}$
Standard deviation of magnitude	$\sigma_{\lvert\boldsymbol{a}\rvert} = \sqrt{\frac{1}{l-1}\sum_{i=1}^{l}\left(\lvert\boldsymbol{a}\rvert_i - u_{\lvert\boldsymbol{a}\rvert}\right)^2}$
Mean x-axis jerk	$\alpha_{\mathrm{x}} = \frac{1}{l-1}\sum_{i=1}^{l-1} p_{\mathrm{x},i}$, where $p_{\mathrm{x},i} = a_{\mathrm{x},i+1} - a_{\mathrm{x},i}$
Mean y-axis jerk	$\alpha_{\mathrm{y}} = \frac{1}{l-1}\sum_{i=1}^{l-1} p_{\mathrm{y},i}$, where $p_{\mathrm{y},i} = a_{\mathrm{y},i+1} - a_{\mathrm{y},i}$
Mean y-axis jerk	$\alpha_{\mathrm{z}} = \frac{1}{l-1}\sum_{i=1}^{l-1} p_{\mathrm{z},i}$, where $p_{\mathrm{z},i} = a_{\mathrm{z},i+1} - a_{\mathrm{z},i}$

Standard deviation of x-axis jerk	$\beta_{\mathrm{x}} = \sqrt{\frac{1}{l-2}\sum_{i=1}^{l-1}\left(p_{\mathrm{x},i} - \alpha_{\mathrm{x}}\right)^2}$
Standard deviation of y-axis jerk	$\beta_{\mathrm{y}} = \sqrt{\frac{1}{l-2}\sum_{i=1}^{l-1}\left(p_{\mathrm{y},i} - \alpha_{\mathrm{y}}\right)^2}$
Standard deviation of z-axis jerk	$\beta_{\mathrm{z}} = \sqrt{\frac{1}{l-2}\sum_{i=1}^{l-1}\left(p_{\mathrm{z},i} - \alpha_{\mathrm{z}}\right)^2}$
Mean jerk magnitude	$\alpha_{\lvert\boldsymbol{p}\rvert} = \frac{1}{l-1}\sum_{i=1}^{l-1}(\lvert\boldsymbol{p}\rvert_i)$, where $\lvert\boldsymbol{p}\rvert_i = \sqrt{p_{\mathrm{x},i}^2 + p_{\mathrm{y},i}^2 + p_{\mathrm{z},i}^2}$
Standard deviation of jerk magnitude	$\beta_{\lvert\boldsymbol{a}\rvert} = \sqrt{\frac{1}{l-2}\sum_{i=1}^{l-1}\left(\lvert\boldsymbol{p}\rvert_i - \alpha_{\lvert\boldsymbol{p}\rvert}\right)^2}$
Stride time variability on x-axis	(1) Identify signal peaks in x-axis, $[t_1, t_2, \ldots, t_Q]$; (2) Identify stride times $[v_1, v_2 \ldots, v_{Q-1}]$, where $v_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_{\mathrm{x}} = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(v_i - \bar{v})^2}$
Stride time variability on y-axis	(1) Identify signal peaks in y-axis, $[t_1, t_2, \ldots, t_Q]$; (2) Identify stride times $[v_1, v_2 \ldots, v_{Q-1}]$, where $v_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_{\mathrm{y}} = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(v_i - \bar{v})^2}$
Stride time variability on z-axis	(1) Identify signal peaks in z-axis, $[t_1, t_2, \ldots, t_Q]$; (2) Identify stride times $[v_1, v_2 \ldots, v_{Q-1}]$, where $v_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_{\mathrm{z}} = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(v_i - \bar{v})^2}$
Stride time variability on magnitude	(1) Identify signal peaks in magnitude, $[t_1, t_2, \ldots, t_Q]$; (2) Identify stride times $[v_1, v_2 \ldots, v_{Q-1}]$, where $v_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_{\lvert\boldsymbol{a}\rvert} = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(v_i - \bar{v})^2}$

For the mic source, recall that each spectrogram is a matrix of $F \times l$. We further denote $a_{f,i}$ as the matrix entry at row f and column i. Consistent with a prior audio processing study (Kalantarian et al. 2014), we extracted the following features for non-deep learning models:

Table B2. Names and Formulas of Features Derived for Feature-based Machine Learning Models from Spectrograms

Feature Name	Formula
Mean amplitude across all frequencies	$u = \frac{\sum_{f=1}^{F}\sum_{i=1}^{l} a_{f,i}}{F \cdot l}$
Mean amplitude for frequency f ($f = 1, \ldots, F$)	$u_f = \frac{\sum_{i=1}^{l} a_{f,i}}{l}, \forall\, f = 1, \ldots, F$

Appendix C: Computational Efficiency Experiment

In this section, we empirically test the average amount of time that AADMML and benchmark models (feature-based and deep learning) need to process wearable sensor data and assess PD severity for an instance. The experiment is conducted on a workstation with CPU as "Intel Xeon CPU E5-2640 v4 @ 2.4GHz" and GPU as "Nvidia GeForce GTX 1080 Ti". Table C1 summarizes the results.

Table C1. Average Execution Time Comparison in Milliseconds (ms)

Category	Model	Execution Time (ms)	Category	Model	Execution Time (ms)
Feature-based, Non-ensemble	DT	127.6	Deep Learning	STLF	193.3
	KNN	127.7		STEF	240.5
	SVM	127.6		STAEF	198.2
Feature-based, Ensemble	ETS	127.6		MTLF	141.8
	RF	127.6		MTEF	164.1
	ADA	127.6		MTAEF	144.8
	GBM	127.6		**AADMML**	144.9
	XGB	127.6			
	NEV	127.8			

We notice that all feature-based models spent similar execution times (about 127.6 ms). This is because the major computational cost comes from extracting features (as shown in Appendix B) from wearable sensor and audio data, which accounts for more than 127 ms. Overall, deep multitask models cost significantly less execution time compared to their single task counterparts (e.g., MTLF (141.8 ms) compared to STLF (193.3 ms)), which is additional evidence of the advantages of deep multitask learning. The computational efficiency of AADMML (144.9 ms) is on a par with feature-based models, and is among the best in deep learning models.

Appendix D: Precision-Recall Curves in Case Study 2

In this section, we report the Precision-Recall Curves (PR Curves) in Figure D1 and the Areas Under the Curves (AUC) in Table D1 for AADMML and select benchmark models in Case Study 2. The AADMML outperforms benchmark models in AUC, which indicates its superiority in identifying early-stage PD patients using wearable sensors.

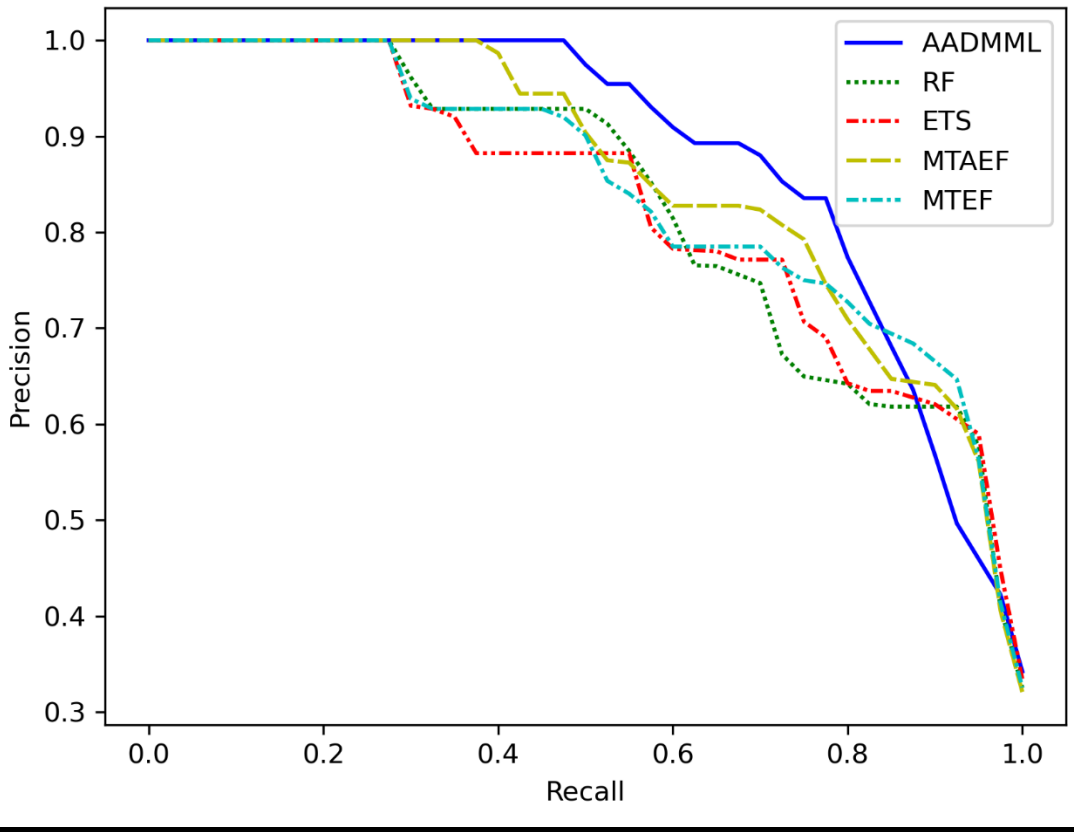


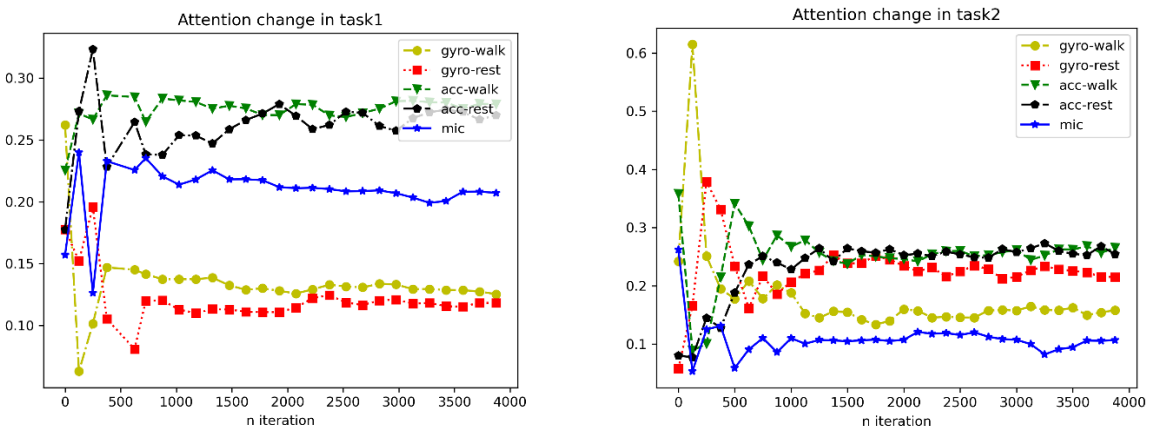
Figure D1. PR Curves in Case Study 2

Table D1. AUC in Case Study 2

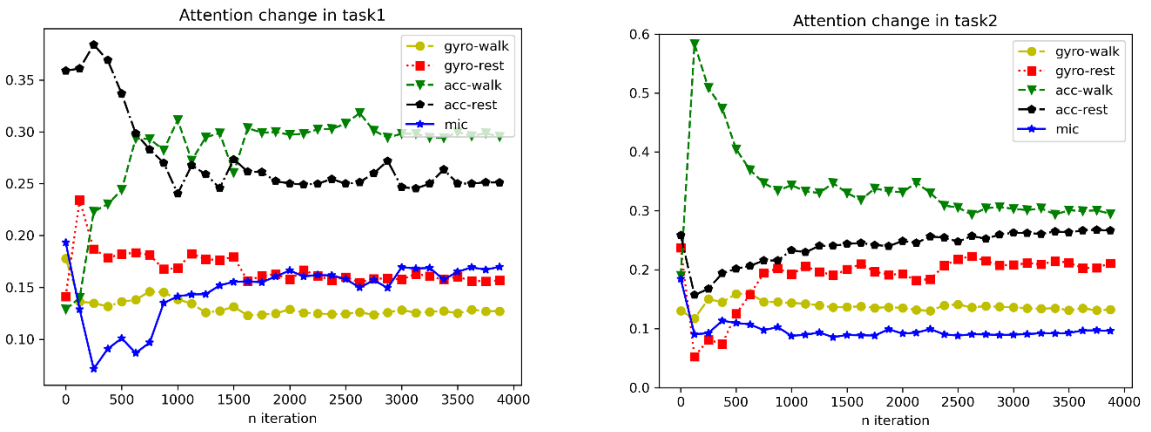
	AADMML	RF	ETS	MTAEF	MTEF
AUC	0.881	0.831	0.829	0.859	0.845

Appendix E: Changes of Attention Weights in Model Training

In this section, we report the how AADMML and MTAEF task-specific attention weights change with increasing numbers of training iterations (Task 1: nM-EDL; Task 2: M-EDL). The top two charts are the results are from AADMML while the bottom two are from MTAEF. We can observe that attention weights in AADMML for both tasks converge with increasing training iterations, which is similar to those of MTAEF.



Results from AADMML



Results from MTAEF

Figure E1. Changes of Task-Specific Attention Weights in Model Training

Appendix F: Sensitivity Analysis for Select Benchmark Models

In this section, we report the sensitivity analysis that reflects the hyperparameter tuning processes for our benchmark models. For feature-based models, we report the sensitivity analysis results in Table F1. The hyperparameters listed are the most influential ones in the Python package scikit-learn. For deep learning models, we only report the results for MTEF and MTAEF (the two models that resemble AADMML the most) in Tables F2 and F3 due to space constraints. The other deep learning benchmark models in Experiment 1 have been tuned with the same search space of hyperparameters. All significance levels are calculated against our AADMML with the best hyperparameters. Hyperparameters and results in boldface are those we finally chose for benchmark models to report in Table 10.

Table F1. Sensitivity Analysis Results for Feature-based Models									
Model & Hyper-parameter	**Option**	Task 1: nM-EDL Sum				Task 2: M-EDL Sum			
		MAE	**RMSE**	LMAE	LRMSE	MAE	RMSE	LMAE	LRMSE
DT criterion	mse	2.275***	3.392***	0.345**	0.513*	3.418***	4.992***	0.728***	0.905***
	friedman_mse	2.356***	3.195***	0.335***	0.523**	3.460***	5.060***	0.720***	0.893***
	mae	**2.067*** **	**3.367*** **	**0.309**	**0.506**	**2.662*** **	**4.953*** **	**0.501**	**0.743**
KNN n_neighbors	**1**	**2.235*** **	**3.947*** **	**0.348*** **	**0.538*** **	**2.908*** **	**4.848*** **	**0.488**	**0.736**
	3	2.478***	3.451***	0.346***	0.505	3.273***	4.806***	0.691***	0.859***
	5	2.672***	3.658***	0.352***	0.527**	3.621***	4.987***	0.748***	0.903***
SVM kernel	linear	3.208***	4.376***	0.397***	0.506	4.471***	6.858***	1.536***	1.365***
	rbf	**2.856*** **	**3.855*** **	**0.358*** **	**0.530** **	**3.794*** **	**5.549*** **	**0.765*** **	**0.907*** **
	sigmoid	2.913***	3.672***	0.375***	0.521**	3.888***	5.455***	0.755***	0.908***
ETS n_estimators	**50**	**1.600**	**2.372** **	**0.316***	**0.512***	**2.107***	**3.075** **	**0.713*** **	**0.895*** **
	100	1.738**	2.303	0.328**	0.515*	2.126*	3.065*	0.713***	0.891***
	200	1.708*	2.304	0.321*	0.527***	2.184**	3.276***	0.725***	0.897***
	300	1.678*	2.392**	0.322	0.516*	2.087	3.082*	0.714***	0.901***
RF n_estimators	**50**	**1.538**	**2.328***	**0.325** **	**0.512***	**2.144** **	**3.096***	**0.721*** **	**0.900*** **
	100	1.599	2.250	0.329**	0.514**	2.266***	3.247***	0.733***	0.911***
	200	1.654	2.245	0.324**	0.513*	2.175**	3.116**	0.728***	0.903***
	300	1.565	2.217	0.312	0.509*	2.164*	3.193***	0.722***	0.893***
ADA n_estimators	**50**	**2.835*** **	**3.565*** **	**0.380*** **	**0.517***	**4.143*** **	**4.583*** **	**0.760*** **	**0.923*** **
	100	2.935***	3.364***	0.369***	0.538***	4.326***	4.671***	0.782***	0.934***
	200	2.751***	3.394***	0.352***	0.540***	4.243***	4.768***	0.768***	0.931***
	300	2.810***	3.369***	0.391***	0.531***	4.259***	4.706***	0.759***	0.939***
GBM n_estimators	50	2.023***	2.728***	0.347***	0.522**	2.854***	4.120***	0.574***	0.799***
	100	1.786**	2.688***	0.324*	0.499	2.624***	4.093***	0.519	0.740
	200	1.680	2.430**	0.364***	0.559***	2.294***	3.844***	0.540*	0.763**
	300	**1.589**	**2.603*** **	**0.383*** **	**0.547*** **	**2.224*** **	**3.732*** **	**0.528**	**0.741**
XGB n_estimators	50	2.017***	3.038***	0.415***	0.594***	2.848***	3.759***	0.652***	0.825***
	100	1.790***	2.753***	0.416***	0.580***	2.508***	3.365***	0.597***	0.799***
	200	1.554	2.616***	0.412***	0.581***	2.129*	3.148**	0.538*	0.761*
	300	**1.533**	**2.539*** **	**0.388*** **	**0.578*** **	**2.170** **	**3.062***	**0.523**	**0.744**

Notes: * p<0.05; ** p<0.01; *** p<0.001

Table F2. Sensitivity Analysis Results for MTEF								
Model	**Task 1: nM-EDL Sum**				**Task 2: M-EDL Sum**			
	MAE	**RMSE**	**LMAE**	**LRMSE**	**MAE**	**RMSE**	**LMAE**	**LRMSE**
HP-CNN (8)	1.868***	2.570***	0.321	0.520**	1.959	2.994	0.525	0.755
HP-CNN (16)	1.559	2.457***	0.322**	0.516**	2.128*	2.903	0.515	0.747
HP-CNN (32)	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**
HP-CNN (64)	1.622	2.416***	0.322*	0.514**	2.115	3.001	0.532	0.753
HP-CNN-L1	1.617	2.556***	0.309	0.542***	2.033	2.997	0.504	0.744
HP-CNN-L2	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**
HP-CNN-L3	1.561	2.412***	0.333***	0.511**	2.074	2.804	0.517	0.748
HP-CNN-L4	1.616	2.543***	0.335***	0.514*	2.207**	2.989	0.523	0.762*
HP-LSTM (8)	1.516	2.601***	0.336***	0.538***	2.088	3.060*	0.532	0.771**
HP-LSTM (16)	1.531	2.454***	0.345***	0.528***	2.067	2.837	0.520	0.754
HP-LSTM (32)	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**
HP-LSTM (64)	1.573	2.462**	0.342**	0.522**	2.310***	3.112*	0.532*	0.764**
HP-LSTM-L1	1.761**	2.904***	0.322	0.537***	2.269***	3.025	0.541**	0.781***
HP-LSTM-L2	1.627	2.665***	0.334**	0.547***	2.134*	3.143***	0.5274	0.777***
HP-BLSTM-L1	1.637	2.546***	0.327*	0.516**	1.963	3.016*	0.509	0.726
HP-BLSTM-L2	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**
HP-LR-1	1.972***	2.546***	0.355***	0.552***	2.285**	3.137**	0.663***	0.889***
HP-LR-2	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**
HP-LR-3	1.555	2.421**	0.320*	0.513**	2.096	2.927	0.504	0.743
HP-LR-4	1.565	2.428***	0.332***	0.532***	2.154**	3.045*	0.533*	0.769**
HP-DR-0	1.633	2.677***	0.344***	0.532***	2.183**	3.161***	0.532	0.770**
HP-DR-25	1.594	2.542***	0.335***	0.539***	2.078	3.128**	0.529	0.731
HP-DR-50	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**
HP-DR-75	1.612	2.497***	0.317*	0.519**	1.997	2.980	0.525	0.747
HP-T1LF-0.5	1.753***	2.596***	0.322*	0.540***	1.954	2.908	0.505	0.743
HP-T1LF-1.0	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**
HP-T1LF-1.5	1.501	2.212	0.306	0.516**	2.165***	3.058	0.529	0.791***
HP-T1LF-2.0	1.526	2.215	0.310	0.505	2.305***	3.437***	0.562***	0.821***
HP-WL2R-3	1.661*	2.640***	0.360***	0.564***	2.310**	3.327***	0.562***	0.805***
HP-WL2R-4	1.566	2.606***	0.317	0.540***	2.324***	3.175***	0.527	0.771**
HP-WL2R-5	1.574	2.416***	0.325**	0.509	2.172**	2.991	0.520	0.750
HP-WL2R-6	**1.570**	**2.410****	**0.311**	**0.513***	**1.998**	**2.932**	**0.516**	**0.742**

Notes: * p<0.05; ** p<0.01; *** p<0.001

Table F3. Sensitivity Analysis Results for MTAEF								
Model	**Task 1: nM-EDL Sum**				**Task 2: M-EDL Sum**			
	MAE	**RMSE**	**LMAE**	**LRMSE**	**MAE**	**RMSE**	**LMAE**	**LRMSE**
HP-CNN (8)	1.753**	2.524***	0.328**	0.516*	2.075	2.905	0.514	0.753
HP-CNN (16)	1.538	2.382*	0.323*	0.506	2.038	2.806	0.521	0.741
HP-CNN (32)	**1.523**	**2.370***	**0.318**	**0.515***	**2.074***	**2.859**	**0.518**	**0.735**
HP-CNN (64)	1.664*	2.330*	0.304	0.526***	2.028	2.896	0.529	0.743
HP-CNN-L1	1.604	2.422**	0.320	0.525***	2.071	3.028*	0.513	0.749
HP-CNN-L2	**1.523**	**2.370***	**0.318**	**0.515***	**2.074***	**2.859**	**0.518**	**0.735**
HP-CNN-L3	1.591	2.318*	0.320	0.513*	1.986	2.898	0.519	0.748
HP-CNN-L4	1.529	2.512***	0.330***	0.512**	2.289**	2.874	0.532	0.749
HP-LSTM (8)	1.526	2.468***	0.320*	0.537***	2.103	2.978	0.537*	0.755*
HP-LSTM (16)	1.577	2.398*	0.315	0.512*	2.076	2.905	0.513	0.742
HP-LSTM (32)	**1.523**	**2.370***	**0.318**	**0.515***	**2.074***	**2.859**	**0.518**	**0.735**
HP-LSTM (64)	1.492	2.434**	0.342***	0.525***	2.198**	3.003	0.531	0.749
HP-LSTM-L1	1.721*	2.754***	0.334**	0.521***	2.328***	2.956	0.535*	0.779**
HP-LSTM-L2	1.609	2.570***	0.324*	0.533***	2.080	2.979	0.523	0.760*
HP-BLSTM-L1	1.598	2.447***	0.310	0.525***	2.081	2.823	0.528	0.717
HP-BLSTM-L2	**1.523**	**2.370***	**0.318**	**0.515***	**2.074***	**2.859**	**0.518**	**0.735**
HP-LR-1	1.574	2.437**	0.325*	0.517**	2.141**	2.969	0.517	0.751
HP-LR-2	1.523	2.370*	0.318	0.515*	2.074*	2.859	0.518	0.735
HP-LR-3	**1.529**	**2.408*****	**0.322****	**0.524*****	**2.095***	**2.888**	**0.527**	**0.751**
HP-LR-4	1.558	2.393**	0.330***	0.537***	2.078*	2.954	0.531	0.762**
HP-DR-0	1.655*	2.593***	0.342***	0.538***	2.183**	3.102**	0.537*	0.764**
HP-DR-25	1.548	2.452***	0.321	0.534***	2.232***	3.044*	0.537*	0.736
HP-DR-50	1.506	2.309	0.321*	0.524**	2.110*	2.881	0.516	0.719
HP-DR-75	**1.523**	**2.370***	**0.318**	**0.515***	**2.074***	**2.859**	**0.518**	**0.735**
HP-T1LF-0.5	1.714**	2.465**	0.339***	0.544***	1.954	2.655**	0.511	0.738
HP-T1LF-1.0	**1.523**	**2.370***	**0.318**	**0.515***	**2.074***	**2.859**	**0.518**	**0.735**
HP-T1LF-1.5	1.511	2.221	0.310	0.507	2.139*	3.039*	0.534	0.776***
HP-T1LF-2.0	1.507	2.205	0.310	0.504	2.184***	3.155**	0.567***	0.804***
HP-WL2R-3	1.626	2.615***	0.337***	0.558***	2.217***	3.182***	0.555**	0.794***
HP-WL2R-4	1.572	2.406**	0.314	0.533***	2.173**	3.124**	0.537*	0.759*
HP-WL2R-5	**1.523**	**2.370***	**0.318**	**0.515***	**2.074***	**2.859**	**0.518**	**0.735**
HP-WL2R-6	1.581	2.351*	0.322*	0.518**	1.987	2.966	0.528	0.745

Notes: * p<0.05; ** p<0.01; *** p<0.001

Appendix G: Comparison between Feature-Based Learning and Deep Learning

The key difference of deep learning compared to feature-based machine learning models is that deep learning models can automatically learn salient feature representations from complex data, especially raw sensory data (LeCun et al. 2015). Table G1 summarizes a comparison among feature-based non-ensemble machine learning, feature-based ensemble learning, and deep learning, in the dimensions of predictive power, transferability, needed dataset size, and computational cost.

Table G1. Comparison between Feature-Based Learning and Deep Learning

	Feature-Based Non-Ensemble Machine Learning	Feature-Based Ensemble Learning	Deep Learning
Predictive Power	Medium	Medium-Strong	Strong
Transferability	Low	Low	Medium
Needed Dataset Size	Small	Medium	Large
Computational Cost	Low	Medium	High

The major advantage of deep learning over feature-based non-ensemble machine learning and feature-based ensemble learning is its predictive power in many classification tasks (e.g., image recognition, speech recognition). For instance, the image recognition error rate dropped from 26 percent to 3.5 percent (Krizhevsky et al. 2012) with the introduction of convolutional neural networks (CNNs) (LeCun et al. 2015). Because no feature engineering is needed, deep learning models often show good transferability for tasks that share identical data characteristics (e.g., image data or wearable sensor data) (Long et al. 2018), whereas feature-based machine learning requires the redesign of feature sets. Although deep learning is often criticized due to its needs for a large amount of data, the exponential amount of data being generated in the big data era (especially sensor data) has largely mitigated that drawback. Regarding deep learning's high computational cost, new hardware and algorithms are constantly being designed to alleviate this concern.

Appendix H: Model Performance with Different Demographic Groups

In this section, we detail how the model performs with different demographic groups, including age and gender. Following Case Study 3, we grouped the subjects into four age groups: 18 to 49, 50 to 64, 65 to 74, and 75 to 84, as well as two gender groups: male and female. Results can be found in Table H1, where the best results within a demographic category are in boldface.

Table H1. Model Performance with Different Demographic Groups

Demographic Groups		Task 1: nM-EDL		Task 2: M-EDL	
		MAE	RMSE	MAE	RMSE
Age	18 – 49	1.450	2.134	**1.842**	**2.747**
	50 – 64	1.641	2.234	2.027	2.902
	65 – 74	1.504	2.129	2.084	2.987
	75 – 84	**1.226**	**1.608**	2.004	2.864
Gender	Male	1.525	2.150	**1.886**	**2.777**
	Female	**1.493**	**2.119**	2.166	3.065

We can see that in different age groups, AADMML performs best in predicting nM-EDL scores for those who are 75 to 84 years old and predicting M-EDL scores for those who are 18 to 49 years old. When measuring nM-EDL for those who are 50 to 64 years old or measuring M-EDL for those who are 65 to 74 years old, the manual MDS-UPDRS questionnaires might be preferred as AADMML is performing slightly worse compared to other age groups. Between male and female, AADMML performs better in predicting nM-EDL scores for female and predicting M-EDL scores for male. We believe these results can help practitioners utilize AADMML as an aiding tool in PD diagnoses.