

Care for the Mind Amid Chronic Diseases: An Interpretable AI Approach Using IoT

Jiaheng Xie^{1,***}, Xiaohang Zhao^{2,*,**}, Xiang Liu¹, Xiao Fang¹

¹ Lerner College of Business and Economics,
University of Delaware, Newark, DE, USA

² School of Information Management & Engineering,
Shanghai University of Finance and Economics, Shanghai, China

* Corresponding Author: Xiaohang Zhao, xiaohangzhao@mail.shufe.edu.cn

** Equal Contribution

Abstract: Health sensing for chronic disease management creates immense benefits for social welfare. Existing health sensing studies primarily focus on the prediction of physical chronic diseases. Depression, a widespread complication of chronic diseases, is however understudied. We draw on the medical literature to support depression detection using motion sensor data. To connect humans in this decision-making, safeguard trust, and ensure algorithm transparency, we develop an interpretable deep learning model: Temporal Prototype Network (TempPNet). TempPNet is built upon the emergent prototype learning models. To accommodate the temporal characteristic of sensor data and the progressive property of depression, TempPNet differs from existing prototype learning models in its capability of capturing temporal progressions of prototypes. Extensive empirical analyses using real-world motion sensor data show that TempPNet outperforms state-of-the-art benchmarks in depression detection. Moreover, TempPNet interprets its decision by visualizing the temporal progression of depression and its corresponding symptoms detected from sensor data. We further employ a user study and a medical expert panel to demonstrate its superiority over the benchmarks in interpretability. This study offers an algorithmic solution for impactful social good — collaborative care of chronic diseases and depression in health sensing. Methodologically, it contributes to extant literature with a novel interpretable deep learning model for depression detection from sensor data. Patients, doctors, and caregivers can deploy our model on mobile devices to monitor patients' depression risks in real-time. Our model's interpretability also allows human experts to participate in the decision-making by reviewing the interpretation and making informed interventions.

Keywords: TempPNet, interpretable AI, prototype learning, depression detection, motion sensor

1. Introduction

The Fourth Industrial Revolution is characterized by technological advancements in Artificial Intelligence (AI) and big data analytics (Shim et al. 2022). Along with its emergence and progress, numerous IT artifacts are implemented by organizations worldwide to modernize their services, scale up their businesses, or improve the efficiency of data exchange. Meanwhile, these IT artifacts record massive amounts of data, which describe the context and outcomes of users' actions and form unique digital traces for each user (Hedman et al. 2013). As part of this trend, digital traces of wearable sensor signals represent an immense and novel source of ecological data that reflect human behaviors and psychological characteristics (Chau et al. 2020). The collection of such data opens up significant research opportunities for Information Systems (IS) researchers to develop innovative artifacts aimed at addressing critical societal and healthcare challenges (Zhang and Ram 2020).

Chronic disease management is a significant societal challenge that can be addressed with novel IT artifacts. The attention to this area in the IS literature is growing rapidly due to its profound social and economic impacts. In the US alone, 133 million Americans suffer from one or more chronic diseases (Forbes 2022), of which the most common ones are heart diseases, cancer, Alzheimer's, and diabetes (CDC 2022). The population of chronic disease patients in the US is expected to reach over 143 million by 2050 (Ansah and Chiu 2023). Moreover, treatments of chronic diseases account for 90% of the US's \$3.8 trillion annual healthcare cost (CDC 2021).

While treatments of chronic diseases attract the primary research efforts in health sensing, care for depression associated with chronic diseases has not received due attention, although medical experts have repeatedly stressed the urgency of collaborative care for depression and chronic diseases (Katon et al. 2010). Depression is epidemic among chronic disease patients, caused by frustrating long-haul symptoms and complex home regimens. Evidence shows that depression occurs in 17% of cardiovascular cases, 23% of cerebrovascular cases, 27% of diabetes patients, and 40% of cancer patients (CDC 2012). According to the National Institutes of Health (NIH), chronic disease patients are twice as likely to suffer from depression as the general population, and depression could drastically elevate chronic disease severity (NIH 2022). Therefore, mental care amid chronic diseases is essential. Accordingly, our first objective is to *detect the occurrence of depression for chronic disease patients using sensing technologies*. When signs of depression are detected, our model could alert caregivers and doctors to actively intervene and treat chronic disease patients' mental disorders.

A key challenge of using sensing technologies to detect depression is that motion sensors only capture patients' physical movements, while depression is a disorder of mental activities. Nevertheless, medical studies have shown that depressed patients experience various physical symptoms,

such as slow movement or speech, unexplained aches and pains, and lack of energy (NHS 2022, Slooman et al. 1982, Lemke et al. 2000). Such physical symptoms of depression are embodied in walking patterns that can be captured by motion sensors. The association between physical movements and depression supports our detection of depression from motion sensor data.

Due to the high stake and impact of health analytics outcomes, a black-box model is inadequate. Interpretability is essential to increase trust in a machine learning (ML) model, prevent failures, and justify its usage (Moss et al. 2022). Moreover, an interpretable model allows medical experts to understand the reasoning behind the decision and make professional diagnoses and accurate interventions. Therefore, our second objective is to *build an interpretable model for depression detection*. As mentioned above, the medical literature finds that depressed patients present typical walking symptoms (NHS 2022). Therefore, these walking symptoms discovered from sensor data serve as a natural interpretation mechanism for the detection of depression. Discovering prototypical patterns from the input to interpret ML predictions forms an active research forefront: prototype learning (Chen et al. 2019, Ming et al. 2019). However, existing prototype learning methods are inadequate for our study because of the following challenges. First, existing methods predict and interpret from a single object (e.g., an image). In contrast, we need to design a method that can predict and interpret from a time series of multivariate time series sensor data, a data structure beyond the scope of existing time series prototype learning methods (Gee et al. 2019). Second, the severity of depression symptoms evolves over time, such as trending up, peaking, and fluctuating (Dattani et al. 2021). Static prototypes defined by existing prototype learning methods cannot capture such dynamic and temporal patterns of depression symptom evolution. Accordingly, we need to define a novel temporal prototype that can capture such dynamic and temporal progressions of prototypes and design a new prototype learning method that can learn temporal prototypes from motion sensor data.

This study targets the interpretable detection of depression associated with chronic diseases and makes the following contributions. From a managerial perspective, our study is positioned in the health information technology (HIT) research area in IS. We extend sensing IS research with a novel AI-based solution and validate the potential of sensing technologies for depression management. Motivated by the unique challenges at the intersection of ML and its societal environments, our study develops a novel computational method to solve a critical societal problem (i.e. depression detection). From a methodological perspective, we propose a novel interpretable deep learning method that interprets the detection of depression by learning temporal progressions of prototypes from sensor data. Different from state-of-the-art prototype learning methods, our method takes a time series of multivariate time series sensor data as inputs and innovatively learns two types of prototypes from the data: prototypes of depression symptoms and prototypes of temporal symptom

progression. It then detects the depression status of a chronic disease patient and interprets the decision based on these two types of prototypes. Extensive empirical analyses using real-world motion sensor data demonstrate the superior detection performance of our method over state-of-the-art benchmarks. Through a user study and a medical expert panel, we also show that our method outperforms the benchmark in interpretability. From a practical perspective, our proposed method can be implemented as a mobile app or be embedded as a function in existing mobile health apps. It is able to detect and monitor chronic disease patients' depression risk in real-time based on motion sensor data collected by their mobile devices. Once our method detects early signs of depression, it can send alerts to caregivers and doctors to provide timely treatments, such as antidepressants and social support groups.

2. Literature Review

2.1. Health IS and Chronic Disease Management

To mitigate the consequences of chronic diseases, IS scholars, health organizations, and IT companies have been cultivating interdisciplinary intelligent chronic disease management (Dixon-Woods et al. 2013), or HIT. In the past decade, HIT for chronic disease management has evolved from hospital-based electronic health records (EHR) to patient-centric mobile devices and assistive technologies, while the analytics methods have expanded from statistical and econometric models to advanced computational models (Bardhan et al. 2020). These computational HIT studies in IS can be broadly categorized into three areas. First, the majority of IS studies use hospital-based information systems such as EHR, with application areas spanning adverse events prediction (Lin and Fang 2021), hospital readmission (Xie et al. 2021a), among others. The second area touches on analytics on online health communities and health social media (Zhang et al. 2024). These HIT studies aim to utilize online information technologies to connect patients, caregivers, and health professionals, with topics ranging from patient education (Liu et al. 2020) and drug safety surveillance (Xie et al. 2021b), to medication nonadherence (Xie et al. 2022).

The last and emerging health IS research for chronic disease management is related to the increasingly popular mobile apps and wearable sensors (Yu et al. 2022). Despite the active research in the first two categories, mobile health systems design and analytics remain under-explored. Yet, it is an increasingly critical area. “*Wearable sensors and home devices can play a pivotal role not only in monitoring the status of chronic disease patients and predicting adverse events before they occur but also in preventing the onset of chronic diseases*” (Bardhan et al. 2020). Such resulting predictive models can generate fruitful managerial implications such as motivating patients to adhere to treatment protocols or individuals to lead a healthy lifestyle (Liu et al. 2024).

In recent years, a few wearable sensor-based HIT studies have begun to emerge in premier IS journals. For instance, Yu et al. (2022) design an attention network for chronic disease management

using sensor signals collected via mobile apps. Zhu et al. (2020) and Zhu et al. (2021) deploy sensing technologies to the senior care setting where deep learning models can recognize activities of daily living. An MIS Quarterly editorial refers to this type of research as the Type I ML research in IS, where the “*designed IT artifacts are models and algorithms*” and “*the methodological contributions of Type I ML research tend to be novel ML models and algorithms developed to solve important business and societal problems*” (Padmanabhan et al. 2022). The Information Systems department within Management Science also welcomes this type of research and states it “*combine(s) a methodological advance with an important and novel managerial application*” (Simchi-Levi 2020).

2.2. Wearable Sensor Technology and Depression

This study belongs to health IS research that collects high-fidelity, real-time sensing data for intelligent chronic disease management. Studies in this area typically recruit 12 to 683 patients to conduct walking tests (Coelln et al. 2019), during which a participant is instructed to walk for a certain distance while his or her walking signals are recorded by motion sensors. Appendix A summarizes recent health sensing studies related to our study. While the detection of chronic diseases, such as Parkinson’s disease (PD) and diabetes, deserves research attention in the health sensing discipline, mental care is equally critical to chronic care. This is because irritating physical symptoms and tedious long-term disease management cause serious mental complications, leading to depression among 40-50% of chronic disease patients (Reijnders et al. 2008). If left untreated, depression’s impact could lead to greater functional disability, faster physical and cognitive deterioration, increased mortality, poorer quality of life, and increased caregiver distress (Marsh 2013). In this study, *we aim to detect the occurrence of depression for chronic disease patients using motion sensor data.*

Medical studies have provided abundant evidence that motion symptoms are an essential manifestation of depression. Compared to healthy people, depressed patients exhibit the following patterns: slower walk, slower lifting motion of the leg (Sloman et al. 1982), shorter strides, slower gait velocity (Lemke et al. 2000), reduced vertical head movements, more slumped posture, and lower gait velocity (Michalak et al. 2009). Clinical studies have also proven that these physical depression symptoms can be reliably captured via self-conducted walking tests (Vancampfort et al. 2020). This is because, rather than temporary mood swings, major depressive disorder is persistent and causes significant impairment in daily life (Mayo Clinic 2022). For such a persistent disorder, the above-mentioned walking symptoms will be present in walking tests.

Related to this study, several prior works also recognize the opportunity of detecting depression based on mobility data. Canzian and Musolesi (2015) use GPS mobility traces to detect a patient’s depressive state using SVM. Jacobson and Chung (2020) conduct a study of 31 patients and employ

XGBoost to detect their depressive status based on sensor data collected from these patients. Farhan et al. (2016) recruit 79 participants for sensor data collection and detect their occurrence of depression using regression. Our study differs from these studies in the following aspects. First, these studies employ conventional ML methods, such as SVM, XGBoost, and regression, which have been shown to be less effective in various applications compared to deep learning models (Xie et al. 2022, Yu et al. 2022, Zhu et al. 2021). Due to the high impact of this area and the rapid advances in deep learning, the computational methods for sensor-based depression detection are due a much-needed, timely revisit and upgrade. Second, because of their use of conventional ML methods, these studies have to craft features from sensor data. Such manual design and selection of features are labor-intensive, require significant domain knowledge, and could lead to inconclusive results (Hubble et al. 2015, Yu et al. 2022). Third, the intention of these models is to make predictions, not interpretations. Therefore, they are either black-box models or models that explain predictions based on simple, hand-crafted features from sensor data, such as mean and variance. These features are primitive and not relevant to depression, thus falling short of offering a meaningful and practical interpretation. In order for end users (e.g., doctors, patients, and caregivers) to trust and adopt a prediction model, meaningful interpretation is the key. As our society is progressing toward integration with ML, new regulations have been imposed to require verifiability, accountability, and more importantly, full transparency of algorithm decisions. A key example is the European General Data Protection Regulation (GDPR), which requires companies to provide data subjects the right to an explanation of algorithm decisions. Consequently, *we aim to propose a novel interpretable deep learning model for depression detection.*

2.3. Interpretable Machine Learning and Prototype Learning

The majority of interpretable ML methods offer feature-based interpretation, where hand-crafted features are required, and their interpretation reveals the attribution of each feature to the prediction. These methods can be post-hoc, such as SHAP and LIME (Lundberg and Lee 2017, Kim et al. 2023), or model-based, such as generative additive model (GAM) (Caruana et al. 2015) and wide and deep learning (W&D) (Xie et al. 2023). A related IS study is Kim et al. (2023), who develops an interpretable method based on LIME for the predictions of heart disease, cancer, and fracture. While their study and ours fall into the same field and application domain, our contributions are significantly different from theirs. The method proposed by Kim et al. (2023) is a post-hoc interpretable method that predicts diseases with a predictive model and then interprets its prediction with a separate interpretation model. Our method, on the other hand, is not a post-hoc method. It builds the interpretable module into the predictive model and makes predictions and interpretations at the same time. In addition, Kim et al. (2023) use structured clinical data and

offer feature-based interpretation. Yet, feature-based interpretations are not desirable in sensor-based depression detection studies. This is because, first, although a few studies crafted simple sensor-based features — such as mean, variance, and standard deviation of a segment of signals (Oung et al. 2015) — they do not characterize the symptoms of depression. Second, manually engineering meaningful features from sensor signals requires intensive labor and could even result in inconclusive results (Hubble et al. 2015, Yu et al. 2022). Such a practice also demands rich domain knowledge, which is not easily accessible (Yu et al. 2022). Fortunately, medical literature reveals that depressed patients present typical walking symptoms (Sloman et al. 1982, Lemke et al. 2000, NHS 2022). Without feature engineering, these walking symptoms can be fruitfully leveraged to interpret the detection of depression. For that purpose, an emergent line of interpretable ML research, prototype learning, has been proposed to utilize the prototypical part of a class (prototype) to interpret the prediction. For instance, a prototypical depression walking symptom learned from sensor signals can interpret why a patient is classified as depressed.

Prototype learning was originally proposed to interpret image recognition (Chen et al. 2019). When interpreting how to classify an image, one focuses on parts of the image and compares them with prototypical images from a given class. For instance, radiologists compare suspected tumors in X-rays with prototypical tumor images for the diagnosis of cancer (Chen et al. 2019). The resemblance of an input image to a prototypical class is the mechanism to interpret the classification of the image. Another example is that when humans describe why a bird picture is classified as a clay-colored sparrow, we might reason that the bird’s head and wing bars look like those of a prototypical clay-colored sparrow. Leveraging such a reasoning mechanism, Chen et al. (2019) devise ProtoPNet. To classify a bird picture into a particular bird species with interpretation, this model introduces a prototype layer where K prototypes are assigned to each known species. Each of these prototypes is intended to capture the prototypical parts, or the most salient and typical representation, of the corresponding bird species, such as the head of a clay-colored sparrow or the wing of a cardinal. ProtoPNet embeds each prototype j as a vector p_j in the latent space, and defines *prototype similarity* s_j to measure how strongly prototype j exists in the input bird picture by comparing p_j and the feature maps (extracted via convolutional layers) of the picture in the latent space. The model subsequently classifies the input bird picture based on the weighted sum of the prototype similarities computed between this picture and each prototype. ProtoPNet visualizes each prototype as the most relevant region of the bird picture where the prototype most strongly exists, and interprets why it thinks the input bird picture should be classified as a particular species by identifying several parts from the picture that look like the prototypical parts (prototypes) of the species.

Other prototype learning models have also emerged, such as text-based prototype learning (Ming et al. 2019), tree prototype learning (Nauta et al. 2021), and more. Appendix B contrasts major prototype learning studies with our method. Existing prototype learning methods interpret classifications of static objects, such as images and texts. Accordingly, these methods learn static prototypes (e.g., segments of images) and employ the learned static prototypes to interpret classifications. Therefore, they fail to model temporal progressions of events, which are critical for designing interpretable depression detection methods with wearable sensor data (Bockting et al. 2015, Dattani et al. 2021).

2.4. MTSC, Prototype Learning for MTSC, and Time Series Deep Learning

To account for time series data, the most closely related problem to ours is multivariate time series classification (MTSC), which aims to leverage the measurement of multiple variables over a period of time to assign data points to classes. MTSC has been used to solve classification problems in human activity recognition, diagnosis using electrocardiogram (ECG), electroencephalogram (EEG), magnetoencephalography (MEG), and systems monitoring (Ruiz et al. 2021). The existing MTSC models fall into five categories: distance measures, shapelets, histograms over a dictionary, interval summarising, and deep learning (Ruiz et al. 2021), which are summarized in Appendix C.

Our method differs from existing MTSC models in terms of input data structure, model design, and model interpretability. From the perspective of input data structure, MTSC models classify multivariate time series data. Each input instance is a multivariate time series with t observations: $X_k = \langle X_{1_k}, X_{2_k}, \dots, X_{t_k} \rangle$, where X_{i_k} is a vector of d dimensions for $i = 1, 2, \dots, t$ (Ruiz et al. 2021). Our study, on the other hand, aims to classify a time series of multivariate time series. That is, input data in our study is a time series of walking segments, each of which is a multivariate time series. Moreover, these walking segments are irregularly spaced in time. Figure 2 depicts a visualization of such differences. Our input data structure, yet challenging and beyond the capability of existing MTSC methods, is fairly common in wearable sensor-based physical activity monitoring for two reasons. First, most studies conduct minutes-long walking tests for participants, and these walking tests are performed at arbitrary time points. Second, for home monitoring settings, participants do not wear the mobile device from time to time, e.g., not wearing it during showers and sleep. Also, they would not walk constantly. For instance, the time segments of sitting in front of the computer during working hours are not candidate data points. Time segments when driving are not useful either. Arbitrarily connecting the walking segments to a single multivariate time series is not feasible either, because this will distort the actual progression of health status over time. For example, depression progresses gradually even though some time segments are not observed. Directly connecting the observed segments will create an illusion of the depression status jump or shift instantly.

To accommodate our data structure, we propose the following methodological novelties, in comparison to existing MTSC methods. Our method first detects a series of discrete symptom severities from an irregularly spaced time series of multivariate time series walking segments. We then learn the underlying continuous temporal progressions of symptom severities by treating the detected symptom severities as discrete observations sampled from a carefully designed continuous temporal distribution. The deep learning-based MTSC studies also suffer from setbacks in our context — i.e. from the interpretability perspective, they are black-box models, which cannot offer interpretable insights for health providers. To promote interpretability in MTSC, recent studies resort to prototype learning. For instance, Gee et al. (2019) use ECG data to detect clinical bradycardia. ECG data are multivariate time series, which are treated as 2-D images, where the horizontal dimension represents the time. Then an autoencoder layer and a prototype layer are added for the classification and interpretation. The learned prototypes are snippets of the constructed image that shows which segment of the ECG data indicates typical bradycardia. There are multiple similar studies in recent years. Appendix D compares our method with these MTSC studies in prototype learning.

From the perspective of input data structure, prototype learning methods for MTSC follow the studies in Appendix C that can only process a multivariate time series, such as ECG (Gee et al. 2019, Zhang et al. 2020b), vital signs (Ma et al. 2020), and videos (Trinh et al. 2021). However, our data is a time series of multivariate time series. These multivariate time series are irregularly spaced in time. As discussed in the MTSC section above, to accommodate this data structure, we propose novel method designs.

From the perspective of prototype method design, prototype learning methods for MTSC share a common pitfall of employing static prototypes to interpret classifications (Gee et al. 2019, Ming et al. 2019, Ma et al. 2020, Zhang et al. 2020b, Ghosal and Abbasi-Asl 2021, Trinh et al. 2021), while our method utilizes our proposed dynamic prototypes to interpret depression detection. More concretely, in Section 3.3, we propose dynamic prototypes, namely trend prototypes, to capture temporal progressions of symptoms. For example, the method proposed by Gee et al. (2019) is a representative prototype learning method for MTSC. Specifically, Gee et al. (2019) adopt the prototype learning method for image classification (Chen et al. 2019) by treating ECG data as static images. As discussed at the end of Section 2.3, the prototype learning method proposed by Chen et al. (2019) employs static prototypes. In particular, the prototype in Gee et al. (2019) is a segment of ECG signals. In depression detection, it is necessary to model depression symptoms and temporal progressions of depression symptoms (Dattani et al. 2021). While depression symptoms can be modeled as static prototypes, temporal progressions of these symptoms need to be captured with dynamic prototypes. This temporal dimension of our method’s learned prototypes has practical implications because medical literature suggests that depression symptoms are progressive and

have multiple phases (Bockting et al. 2015, Dattani et al. 2021). As depicted in Figure 1, from the onset phase, a patient’s symptom severity may trend up and progress to acute depression. After the acute phase, some patients’ symptoms may peak, while others could exacerbate. A portion of the patients may recover from depression, and their symptom severity trends down. At any timepoint, a recovered patient is likely to relapse, and symptoms will subsequently recur and their severities may fluctuate. When applying existing prototype learning methods to our study, we are able to determine to what extent a single walking segment is “normal.” However, no depression symptom in a single walking segment does not imply that the patient has no depression because this segment might be during the “normalcy phase,” which could just be transitory and would quickly progress to severe depression. As a result, it is essential to detect and interpret the mental state of a patient based on a time series of walking segments that characterize the full course of symptom progression. Accordingly, we design trend prototypes that depict temporal progressions of symptom prototypes, which necessitate sophisticated function design, distributional inference, and generative process, as we articulate in Section 3.3.

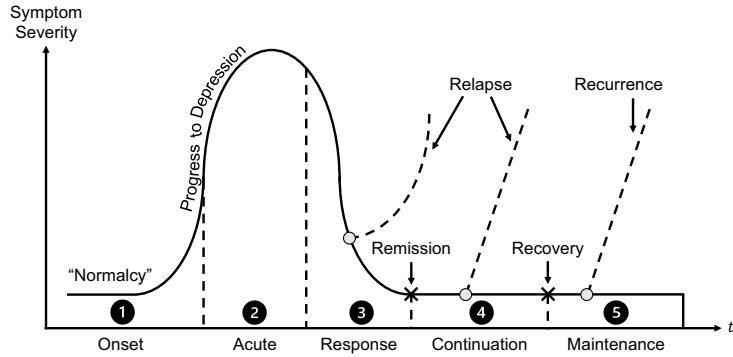


Figure 1 Temporal Progression of Depression (Bockting et al. 2015)

Beyond the common limitation of static prototypes, each prototype learning for MTSC method has its own downsides. When adapted to our study, Ming et al. (2019)’s method only recognizes the entire walking segment as a prototype. This granularity is not desired, as this prototype cannot distinguish multiple symptoms from a single walking segment. Zhang et al. (2020b) use multiple data sources and include blackbox encoders in the feature extractor. Consequently, the prototype is only hidden embedding and cannot be traced back to a segment of the input. Because of this, Zhang et al. (2020b) only apply the prototype idea to prediction instead of interpretation. The prototypes learned in Ghosal and Abbasi-Asl (2021) are defined for each individual variable. It is more appropriate for MTSC problems where each dimension is a hand-crafted feature. In our context, hand-crafted features are not as effective as representation learning (Yu et al. 2022).

The interpretation in Trinh et al. (2021) is done in a post-hoc manner. However, many studies in interpretable ML have shown model-based interpretation, such as ours, is more faithful than post-hoc interpretation (Rudin 2019, Xie et al. 2023). Figure 2 visualizes the differences between our method and existing prototype learning for MTSC methods.

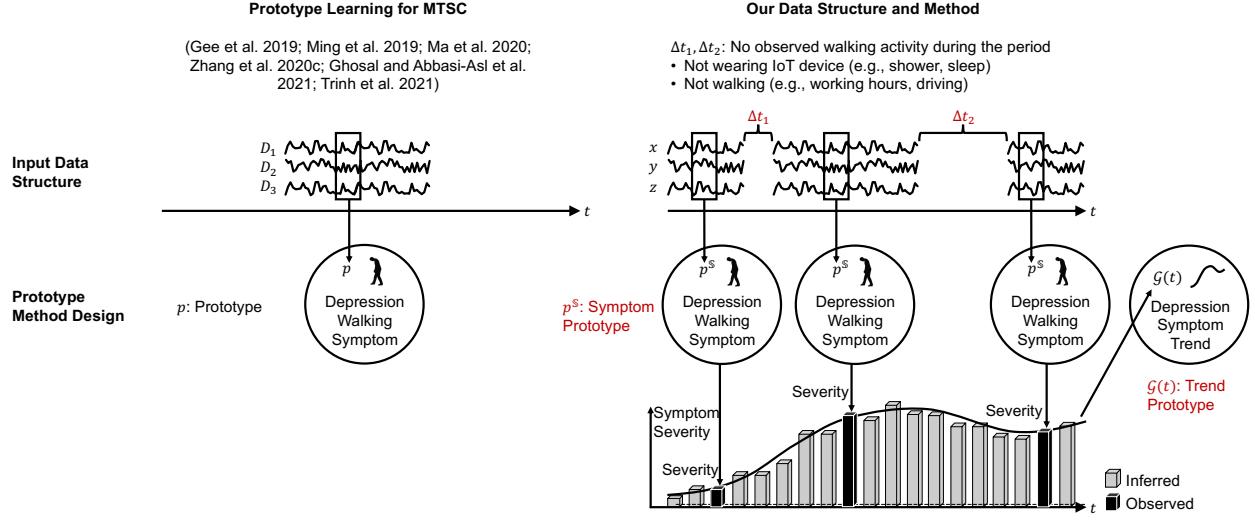


Figure 2 Our Method v.s. Prototype Learning for MTSC

To solve the problem in the right part of Figure 2, it is critical to learn two types of prototypes: symptom prototypes (e.g., short strides and slow gait velocity) and trend prototypes (e.g., symptom severity trending up, trending down, and fluctuating). To learn these prototypes, we need to tackle three methodological challenges. First, while rich sensor data can be collected during a walking segment, it is non-trivial to define prototypes representing depression symptoms, such that their existing strength (how strong a symptom presents in a walking segment) in a walking segment can be interpreted as the snapshot of symptom severity at the time of the walking segment. Second, the elapsed time between two consecutive walking segments is irregular. On the one hand, it is essential to explicitly consider these irregular inter-segment time intervals when inspecting the time series of symptom severities, because they reveal the progression of symptom severities and enable the definition of prototypes representing temporal depression symptom progression. On the other hand, existing prototype learning for MTSC cannot deal with a time series of multivariate time series that are irregularly spaced in time, as indicated by the TP column in Table A.2 and the last column in Table A.4. This gap motivates us to develop a novel and effective method to analyze this new type of input for prototype learning. Lastly, different patients often have varying numbers of walking segments within an observation time window. It is also unknown which phase of depression progression each walking segment is performed at. Consequently, given a prototypical temporal

symptom progression, it is difficult to measure its existing strength in a walking segment sequence because the two objects are not aligned in time. That is, for each walking segment, we do not know against which part of the prototypical progression the walking segment should be compared.

Another uniqueness of the wearable sensor data is that the walking segments are irregularly spaced in time. To design a novel approach to address this challenge, we review deep learning studies that incorporate the temporal dimension in prediction models. They have been applied to a variety of time-series predictions, such as Parkinson’s prediction, cancer prediction, and temperature prediction (Li et al. 2020, Gao et al. 2019, Liu et al. 2018). Appendix E summarizes these studies. Although these “time-aware” models provide guidance to encode temporal information, their purposes are to address the data challenges to improve the *prediction* performance of black-box models, such as LSTM. Distinct from these studies, our study aims to model a time series of irregularly spaced multivariate time series sensor data to improve the *interpretation* and define a new dynamic prototype based on temporal progressions of symptoms, as an extension to the prototype learning models.

3. The TempPNet Approach

3.1. Problem Setup

For a patient u , let $y^{(u)}$ be the patient’s depression status, where $y^{(u)} = 1$ denotes depression, and $y^{(u)} = 0$ represents non-depression. We also observe this patient’s N_u walking segments denoted by $X^{(u)} = \langle X_1^{(u)}, X_2^{(u)}, \dots, X_{N_u}^{(u)} \rangle$ that are recorded at timepoints $\langle t_1^{(u)}, t_2^{(u)}, \dots, t_{N_u}^{(u)} \rangle$. These walking segments can be measured by wearable sensors, such as the accelerometers in smartphones and smartwatches. The i th walking segment $X_i^{(u)}$ observed at timepoint $t_i^{(u)}$ is comprised of a sequence of regularly sampled sensor data: $X_i^{(u)} = \langle a_1, a_2, \dots, a_L \rangle$, where vector a_l is the sensor feature sampled at timepoint l . Each sensor feature a_l is derived from the accelerometer readings recorded in the walking segments. Please see Appendix G for the details of a_l .

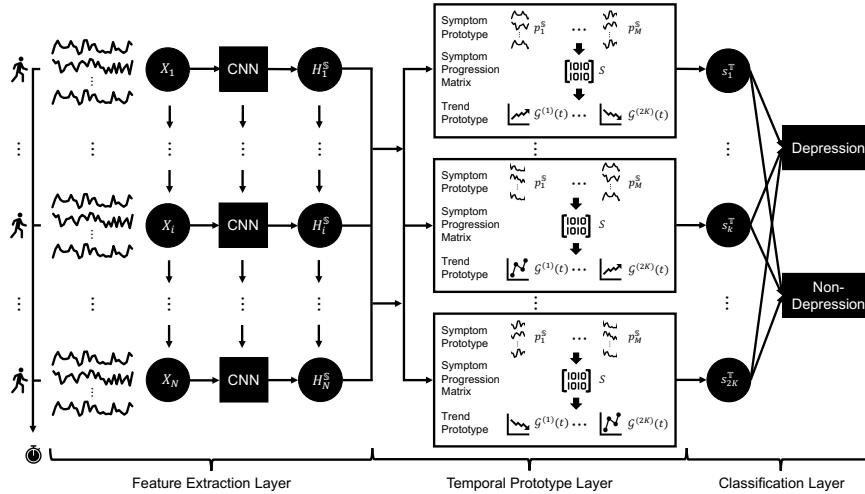
Let \mathcal{U} denote the set of patients. We observe the dataset $\mathcal{D} = \{(X^{(u)}, y^{(u)})|u = 1, 2, \dots, |\mathcal{U}|\}$, where $X^{(u)}$ and $y^{(u)}$ represent the sequence of walking segments and the depression status of patient u , respectively, and $|\mathcal{U}|$ denotes the number of patients. Our objective is to learn a model from \mathcal{D} that can detect the depression status of a new patient, based on sensor data collected from the patient’s walking activities, and interpret the decision. Table 1 summarizes the important notations.

We present the Temporal Prototype Network (TempPNet), an interpretable classifier equipped with a novel temporal prototype layer that detects depression status based on the prototypical temporal progression of walking symptoms. Figure 3 shows the architecture of our model, which features three building blocks: a feature extraction layer that represents a sequence of walking segments as a sequence of features, a temporal prototype layer that defines symptom prototypes

Table 1 Notation

Notation	Description
X_i	The sensor signal sequence of the i th walking segment of a focal patient.
t_i	The timepoint when the walking segment X_i is performed.
H_i^S	The feature matrix extracted from walking segment X_i .
p_m^S	The embedding vector for symptom prototype m .
$s_{m,i}^S$	The existence strength of symptom prototype m in walking segment X_i , defined by Equation 2.
S	The symptom progression matrix detected for the focal patient, defined by Equation 3.
$\mathcal{G}^{(k)}(t)$	The definition of trend prototype k , defined by Equation 6.
s_k^T	The existence strength of trend prototype k in symptom progression matrix S , defined by Equation 9.

and trend prototypes for depression detection and then computes the existing strength of these prototypes in a walking segment, and lastly a classification layer that classifies a patient into depression or non-depression based on the existing strength of trend prototypes.

**Figure 3 TempPNet Architecture**

3.2. The Symptom Prototype and Symptom Progression Matrix

We consider a focal patient u , and thus drop the superscripts and subscripts related to the patient to simplify the notation. Recall that we can observe a sequence of N walking segments for this patient, whose i th walking segment is represented by the sensor signal sequence X_i . To learn an effective representation of X_i , we employ a deep Convolutional Neural Network (CNN) layer (LeCun et al. 1998, Zhang et al. 2020a). Let $H_i^S \in R^{n_o \times n_e}$ denote the learned embedding matrix for the sensor signal sequence X_i , where n_o is the number of patches generated by the deep CNN layer and n_e is the embedding dimension of each patch (Chen et al. 2019). We define H_i^S as

$$H_i^S = \text{CNN}(X_i) \quad (1)$$

where the specifications of the deep CNN layer are articulated in Table A.8. Let $H_{o|i}^S$ denote the o th column of H_i^S . The key benefit of using CNN for learning the embedding of X_i is that each patch $H_{o|i}^S$ has its own receptive field, which can be identified and visualized as a local segment in X_i . Following Chen et al. (2019), we identify the symptoms at the receptive field level. We articulate how to leverage this property to inspect the learned prototypes in Section 3.5.

Similar to prototype learning in image recognition, we aim to detect what prototypical walking symptoms that H_i^S (the representation of X_i) resembles. For this purpose, we need to learn M symptom prototypes that represent typical walking symptoms, such as short strides and slow gait velocity (Lemke et al. 2000, Michalak et al. 2009, Czech and Patel 2019). These symptom prototypes are defined as the latent representation of prototypical waking patterns in the input signal. We embed each symptom prototype m as a vector $p_m^S \in R^{n_e}$ for $m = 1, 2, \dots, M$, which is in the same latent space as H_i^S . Following the common strategy of prototype learning (Chen et al. 2019), symptom prototype vector p_m^S will be learned as model parameters, and then identified as and visualized by the walking segment where it presents most strongly. The interpretation mechanism is to find the prototypical patterns within the input sensor signal that are indicative of depression walking symptoms and thereby informative for depression detection. The middle part of Figure 3 provides an exemplar visualization of the symptom prototypes. Let $s_{m,i}^S$ denote the existing strength of symptom prototype m in walking segment X_i , which can also be understood as the severity of symptom m detected in the walking segment. We define $s_{m,i}^S$ as

$$s_{o|m,i}^S = \exp(\gamma - \|H_{o|i}^S - p_m^S\|_2^2) \quad (2a)$$

$$s_{m,i}^S = \max_{o=1,2,\dots,n_o} s_{o|m,i}^S \quad (2b)$$

where $\gamma < 0$ is an infinitesimal constant to ensure $0 < s_{o|m,i}^S < 1$ for $o = 1, 2, \dots, n_o$. A high value of $s_{o|m,i}^S$ suggests that strong existence of symptom prototype m , or high severity of walking symptom m , is detected in the o th region in walking segment X_i . As a result, $s_{m,i}^S$ measures the overall severity of symptom prototype m in walking segment X_i . Unlike existing prototype learning studies that predict and interpret based on static prototypes solely (Chen et al. 2019, Ming et al. 2019), the severity of symptom prototype $s_{m,i}^S$ is dynamic and forms a temporal progression pattern, as we observe patients' sensor data over time. By collecting the severity scores of all symptom prototypes across the sequence of walking segments, we can construct the symptom progression matrix $S \in R^{M \times N}$ as

$$S = \begin{bmatrix} t_1 & t_2 & \dots & t_N \\ s_{1,1}^S & s_{1,2}^S & \dots & s_{1,N}^S \\ s_{2,1}^S & s_{2,2}^S & \dots & s_{2,N}^S \\ \vdots & \vdots & \ddots & \vdots \\ s_{M,1}^S & s_{M,2}^S & \dots & s_{M,N}^S \end{bmatrix} \begin{array}{l} \text{sym}_1 \\ \text{sym}_2 \\ \vdots \\ \text{sym}_M \end{array} \quad (3)$$

where we label each column by the timepoint when the corresponding walking segment is performed, and each row by the corresponding symptom prototype (sym_m denotes symptom m). In matrix S , the m th row corresponds to the temporal progression of the severities of symptom m , while the i th column corresponds to the contemporaneous distribution of symptom severities observed at timepoint t_i . In other words, if we regard each symptom severity as a random variable that randomly evolves over time, then the column vector S_i is a realization of these M random variables drawn from their joint distribution specific to timepoint t_i . From this perspective, the symptom progression matrix describes how the severities of the M symptoms co-evolve over time. To make an interpretable detection, we need to detect whether such a matrix resembles typical depression or non-depression symptom progression trends such as in Figure 1, which we define as trend prototypes.

3.3. Detecting Prototypical Depression Symptom Progression Trends

A trend prototype is a prototypical depression or non-depression trend. We aim to learn $2K$ trend prototypes: K for the depression category and the remaining K for the non-depression category. We view each trend prototype as a prototypical continuous co-evolution trajectory of the M symptom severities over time, such as trending up, trending down, and fluctuating. Therefore, a trend prototype can be defined as a continuous vector-valued function of time. Let $\mathcal{G}^{(k)}(t)$ denote the k th trend prototype, which maps a time scalar to a vector of symptom severities of length M . The mathematical details of $\mathcal{G}^{(k)}(t)$ will be given in Section 3.3.1. We offer a particular interpretation of trend prototypes: if the focal patient’s symptom progression matrix S strongly resembles trend prototype k , then the symptom severities S_i detected at timepoint t_i should be close to $\mathcal{G}^{(k)}(t_i)$. Based on this interpretation, the existing strength of trend prototype k in symptom progression matrix S can be measured as the overall closeness between S_i and $\mathcal{G}^{(k)}(t_i)$ over observation timepoints $\langle t_1, t_2, \dots, t_N \rangle$. The implementation of this is articulated in Section 3.3.2.

3.3.1. The Definition of Trend Prototypes: A trend prototype is a continuous function of time that characterizes the evolution of symptom severities. This goal necessitates a continuous function that can trace any freely-valued trajectory over time. Studies have shown that any trajectory over time can be decomposed into a summation of sine and cosine curves in the frequency domain (Xu et al. 2020, Wang et al. 2021a). Inspired by those works, we design the trend prototype as follows. Let $\mathcal{G}_m^{(k)}(t)$ denote the severity of symptom m at timepoint t given by trend prototype k . Leveraging the time encoding method in Xu et al. (2020), we compute $\mathcal{G}_m^{(k)}(t)$ as

$$\begin{aligned}\mathcal{G}_m^{(k)}(t) &= \sigma \left(\sqrt{\frac{1}{n_d}} \sum_{j=1}^{n_d} p_{k,m,j}^T \cos(\omega_j t + \theta_j) + \sqrt{\frac{1}{n_d}} \sum_{j=1}^{n_d} p_{k,m,n_d+j}^T \sin(\omega_j t + \theta_j) \right) \\ &= \sigma(\Phi(t)^T p_{k,m}^T)\end{aligned}\quad (4)$$

where $p_{k,m}^{\mathbb{T}} = [p_{k,m,1}^{\mathbb{T}}, p_{k,m,2}^{\mathbb{T}}, \dots, p_{k,m,n_d}^{\mathbb{T}}]^T \in R^{2n_d}$ is the coefficient vector specific to trend prototype k and symptom prototype m , $\Phi(t)$ is the time encoding function (Xu et al. 2020) that transforms timepoint t from a scalar to a numeric vector, defined as

$$\Phi(t) = \sqrt{\frac{1}{n_d}} \begin{bmatrix} \cos(\omega_1 t + \theta_1), \cos(\omega_2 t + \theta_2), \dots, \cos(\omega_{n_d} t + \theta_{n_d}), \\ \sin(\omega_1 t + \theta_1), \sin(\omega_2 t + \theta_2), \dots, \sin(\omega_{n_d} t + \theta_{n_d}) \end{bmatrix}^T \quad (5)$$

which is parameterized by frequency vector $\omega = [\omega_1, \omega_2, \dots, \omega_{n_d}]$ as well as phase vector $\theta = [\theta_1, \theta_2, \dots, \theta_{n_d}]$, and lastly $\sigma(\cdot)$ is the sigmoid function defined as $\sigma(x) = 1/(1 + \exp(-x))$ to ensure $0 < \mathcal{G}_m^{(k)}(t) < 1$ such that $\mathcal{G}_m^{(k)}(t)$ and $s_{m,i}^{\mathbb{S}}$ are in the same range and therefore comparable to each other. Let $\sigma^{-1}(\cdot)$ denote the inverse of the sigmoid function. Namely, $\sigma^{-1}(x) = \log \frac{x}{1-x}$. Our design of $\Phi(t)$ ensures that $\sigma^{-1}(\mathcal{G}_m^{(k)}(t))$ is able to trace a freely-valued trajectory over time so that the temporal progressions of symptom severities can be seamlessly modeled by a continuous function. By learning the three vectors $p_{k,m}^{\mathbb{T}}$, ω and θ as model parameters, while treating n_d as a hyper-parameter, Equation 4 amounts to learning a prototypical temporal progression of symptom m in the frequency domain. In a compact form, trend prototype k can be written as Equation 6, where $p_k^{\mathbb{T}} \in R^{M \times 2n_d}$ is the coefficient matrix that specifies the function representation of trend prototype k , and is obtained by row stacking the transpose of vectors $p_{k,m}^{\mathbb{T}}$ for $m = 1, 2, \dots, M$. We can view $p_k^{\mathbb{T}}$ as the embedding form of trend prototype k that is analogous to $p_m^{\mathbb{S}}$, the embedding form of symptom prototype m . The depression trend prototypes could be characterized by a rising symptom severity. Such a severity may have deviations from time to time while keeping the upward trend. These trend prototypes correspond to the onset and acute phases in Figure 1. The non-depression trend prototype may be characterized by a downward severity trend, suggesting a previously depressed patient is recovering. The non-depression trend prototype could also be a curve fluctuating around a fixed level, indicating the patient never had depression symptoms or has fully recovered from prior depression. When a non-depressed patient relapses, depression trend prototypes could reappear in his or her walking sensor data. In the subsequent discussion, we will frequently use $\sigma^{-1}(\mathcal{G}^{(k)}(t))$, which means applying the inverse of the sigmoid function element-wisely on $\mathcal{G}^{(k)}(t)$. Therefore, we formally define Equation 7 to simplify the notation.

$$\mathcal{G}^{(k)}(t) = \sigma(p_k^{\mathbb{T}} \Phi(t)) \quad (6)$$

$$\tilde{\mathcal{G}}^{(k)}(t) = \sigma^{-1}(\mathcal{G}^{(k)}(t)) = p_k^{\mathbb{T}} \Phi(t) \quad (7)$$

3.3.2. Detecting the Existing Strength of a Trend Prototype: Let $s_k^{\mathbb{T}}$ denote the existing strength of trend prototype k in symptom progression matrix S . Intuitively, $s_k^{\mathbb{T}}$ should be derived by comparing S_i and $\mathcal{G}^{(k)}(t_i)$ for each timepoint t_i , where the former is the “observed”

symptom severities detected at t_i , the latter is its corresponding “ideal” values in the prototypical progression patterns captured by trend prototype k . The larger the former deviates from the latter across time, the less likely trend prototype k exists in the symptom progression matrix. However, we argue that for interpretability concerns, S_i should be compared to $\mathcal{G}^{(k)}(t_i - t_0^{(k)})$ rather than $\mathcal{G}^{(k)}(t_i)$, where $t_0^{(k)}$ is a latent variable indicating when the trend starts in the patient’s timeline. The computation and rationale of $t_0^{(k)}$ are articulated below.

Recall that Equation 4 defines $\mathcal{G}^{(k)}(t)$ as a function on the domain $t \in (-\infty, +\infty)$. However, to understand what temporal progression that trend prototype k captures, it is necessary to ensure that only a finite segment of $\mathcal{G}^{(k)}(t)$ carries useful information, so that only this segment needs to be visualized to inspect trend prototype k . For this purpose, we need to “bound” the trend by a starting time and an ending time. We treat $t = 0$ as the starting time of trend prototype k , and only work with the right part of $\mathcal{G}^{(k)}(t)$ defined on the non-negative domain $t \in [0, +\infty)$, as shown in Figure 4. This design has two implications. Continuing with the example in Figure 4, on one hand, $\mathcal{G}^{(k)}(0)$ should be treated as the initial state of trend prototype k , and it is the relative timepoint measuring how much time has elapsed since $t = 0$ that truly matters. On the other hand, we can only observe the absolute timepoints when the focal patient performs walking segments in reality (e.g., t_1 , t_2 , and t_3). In general, we cannot assume that t_1 , the absolute timepoint of the first walking segment performed by the patient, is the starting time of trend prototype k in the patient’s timeline, because t_1 is determined by when the first walking segment is performed as well as the range of the observation window. However, the depression progression trend of a patient $\mathcal{G}^{(k)}(t)$ would start regardless of whether a walking segment is performed or not. That unobserved trend starting time is noted as $t_0^{(k)}$. If we were able to observe the zero-th walking segment X_0 at absolute timepoint $t_0^{(k)}$, and detect symptom severities S_0 from it by computing $s_{m,0}^S$ in accordance with Equation 2 for $m = 1, 2, \dots, M$, then S_0 should be compared to $\mathcal{G}^{(k)}(0)$. In this sense, S_i , the symptom severity vector detected at t_i in the patient’s timeline, should be compared to $\mathcal{G}^{(k)}(t_i - t_0^{(k)})$, the corresponding prototypical values in the trend’s timeline. In Figure 4, the detected symptom severity is compared with trend prototype k by aligning the two timelines at a latent personalized trend starting time $t_0^{(k)}$.

With $t_0^{(k)}$ at hand, only relative timepoints given by $t_i - t_0^{(k)}$ for $i = 1, 2, \dots, N$ carry information. As a result, we set $t_1 = 0$ and measure other timepoints including $t_0^{(k)}$ relative to t_1 without loss of generality. To facilitate the learning of interpretable trend prototypes, we additionally impose the constraint $t_0^{(k)} < t_1 = 0$. If $t_0^{(k)} > 0$, the symptom severities detected before $t_0^{(k)}$ need to be compared to the left part of $\mathcal{G}^{(k)}(t)$ defined on domain $t \in (-\infty, 0)$, which we intend to avoid for interpretability concerns. For example, it is undesired to compare S_1 with $\mathcal{G}^{(k)}(t_1 - t_0^{(k)}) = \mathcal{G}^{(k)}(-t_0^{(k)})$. Allowing such a comparison is in conflict with our interpretation of $t = 0$ as the starting time of trend prototype

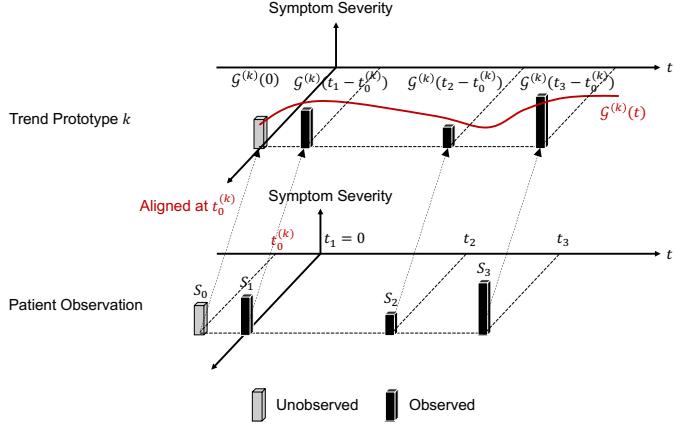


Figure 4 Latent Trend Starting Time

k . Moreover, it also renders the learned function $\mathcal{G}^{(k)}(t)$ hard to inspect, because the progression patterns within the interval $t \in (-t_0^{(k)}, 0)$ also encode some prototypical patterns summarized from data. Given that $t_0^{(k)}$ varies across patients, the meaningful part of $\mathcal{G}^{(k)}(t)$ is different for different patients. In contrast, if we enforce $t_0^{(k)} < 0$, all detected symptom severities will be compared to the right part of $\mathcal{G}^{(k)}(t)$ defined on the domain $t \in (0, +\infty)$, making $t = 0$ a natural patient-independent starting point to inspect $\mathcal{G}^{(k)}(t)$. Armed with the notion of $t_0^{(k)}$, we proceed to the definition of s_k^{\top} by assuming a pre-computed $t_0^{(k)}$, and then discuss how to infer $t_0^{(k)}$ by the end of this section, where we introduce another design to impose an effective trend ending time.

As pointed out by Chen et al. (2019), a prototype can be viewed as a cluster, and its existing strength in an instance is essentially a closeness measurement between the instance and the cluster. To develop a principled definition of s_k^{\top} , we draw inspiration from the Gaussian Mixture Model (Murphy 2022), a classic clustering algorithm, which measures the closeness between an instance and a cluster in terms of how likely the instance can be generated by the Gaussian component characterized by the cluster. By making an analogy in our context, we view the columns of S as being generated from a time-varying distribution characterized by trend prototype k at aligned timepoints $t_i - t_0^{(k)}$ for $i = 1, 2, \dots, N$. Different from the Gaussian Mixture Model which deals with instances described by freely-valued vectors, the entries of S are bounded in the interval $(0, 1)$. To properly specify the generation of the columns of S , we leverage the logistic-normal distribution, whose samples fall between $(0, 1)$. We introduce this distribution first.

Let $x \sim \mathcal{N}(\mu, \Sigma)$ denote a random vector drawn from the M -dimensional normal distribution of mean $\mu \in R^M$ and covariance $\Sigma \in R^{M \times M}$. Let $z = \sigma(x)$, which is obtained by transforming x element-wisely with the sigmoid function such that $0 < z_m = \sigma(x_m) < 1$ for $m = 1, 2, \dots, M$. Then, z follows $\mathcal{LN}(\mu, \Sigma)$, the M -dimensional logistic-normal distribution with mean μ and covariance Σ , and its density can be computed as Equation 8, where $\sigma^{-1}(z)$ means applying the inverse of the

sigmoid function element-wisely on z , and $\det[\Sigma]$ is the determinant of matrix Σ . The derivation of Equation 8 can be found in Appendix H.

$$\mathcal{LN}(z|\mu, \Sigma) = \frac{1}{\prod_{m=1}^M z_m(1-z_m)} \frac{\exp\left(-\frac{1}{2}(\sigma^{-1}(z) - \mu)^T \Sigma^{-1}(\sigma^{-1}(z) - \mu)\right)}{\sqrt{(2\pi)^M \det[\Sigma]}} \quad (8)$$

Now, consider a focal patient whose symptom progression matrix exhibits the presence of trend prototype k to some extent. In this case, we regard S_i as a realization drawn from the contemporaneous distribution of symptom severities given by $\mathcal{LN}(\tilde{\mathcal{G}}^{(k)}(t_i - t_0^{(k)}), I)$, the M -dimensional logistic-normal distribution characterized by mean $\tilde{\mathcal{G}}^{(k)}(t_i - t_0^{(k)})$ and covariance I (the identity matrix), where $\tilde{\mathcal{G}}^{(k)}(t)$ is defined by Equation 7, and $t_0^{(k)}$ is the personalized trend starting time explained previously. In our setting, S_i is compared to $\mathcal{G}^{(k)}(t_i - t_0^{(k)})$ at the aligned timepoint $t_i - t_0^{(k)}$, while $\tilde{\mathcal{G}}^{(k)}(t)$ depicts the mean series of a time-varying logistic-normal distribution which characterizes some prototypical symptom progression patterns subject to a sigmoid transformation. We assume that the covariance of this time-varying distribution is the constant identity matrix for parsimonious concerns.

The generative process of the symptom progression matrix is summarized in Appendix F. Then, we define $s_k^{\mathbb{T}}$ (Equation 9) as a scalar proportional to the log-likelihood of generating the column vectors of S in accordance with Appendix F. $S_{m,i}$ is the entry at row m and column i in matrix S . The second step of Equation 9 is derived by expanding $\mathcal{LN}(S_i|\tilde{\mathcal{G}}^{(k)}(t_i^{(k)}), I)$ using Equation 8 and then simplifying the resulting terms. The intuition is that the more likely the columns of S can be observed from the generative process characterized by $\mathcal{G}^{(k)}(t)$, the stronger the evidence is that trend prototype k exists in the symptom progression matrix of the focal patient. Equation 9 ensures $0 < s_k^{\mathbb{T}} < 1$, which means that the existence strength of trend prototypes has the same value range with symptom prototypes as defined by Equation 2. Indeed, by comparing Equations 9 and 2, an obvious analogy can be established between the definition of $s_k^{\mathbb{T}}$ and $s_{o|m,i}^{\mathbb{S}}$.

$$\begin{aligned} s_k^{\mathbb{T}} &= \sigma\left(\log \prod_{i=1}^N \mathcal{LN}(S_i|\tilde{\mathcal{G}}^{(k)}(t_i^{(k)}), I)\right) \\ &= \sigma\left(\sum_{i=1}^N \sum_{m=1}^M \log \frac{1}{S_{m,i}(1-S_{m,i})} - \frac{NM}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^N \|\sigma^{-1}(S_i) - \tilde{\mathcal{G}}^{(k)}(t_i^{(k)})\|_2^2\right) \end{aligned} \quad (9)$$

Lastly, we complete the definition of $s_k^{\mathbb{T}}$ by specifying the inference procedure of $t_0^{(k)}$. Because $t_0^{(k)}$ is used to generate the symptom progression matrix S , given S , we should be able to inversely infer $t_0^{(k)}$. Based on this intuition, we introduce an inference network, that takes S and the observation timepoints $\langle t_1, t_2, \dots, t_N \rangle$ as input, and outputs $t_0^{(k)}$ as a negative scalar. Specifically, the inference network is defined in Equation 10. \oplus is the concatenation operator, $\Phi(\cdot)$ is the time encoding function defined by Equation 5 but parameterized separately, GRU is a Gated Recurrent Unit

layer (Cho et al. 2014) used to capture the temporal information in the symptom progression matrix augmented by time features, and lastly $h_i^T \in R^{ne}$ is the hidden state of the GRU layer at step i . In Equation 10b, h_N^T is the last hidden state of the GRU layer, which serves as a feature vector summarizing the information contained in S and $\langle t_1, t_2, \dots, t_N \rangle$, $w_k \in R^{ne}$ is a learnable parameter specific to trend prototype k , and $n_w > 0$ is a hyperparameter. Equation 10 enforces that $-n_w < t_0^{(k)} < 0$. We impose this constraint to make the learned trend prototype easier to interpret. Consider the case where the observation window is 1 week, which means that the time difference between the first and the last walking segments is at most 1 week. If we set $n_w = 1$ (time is measured in weeks), then $t_0^{(k)}$ is at most 1 week before the first walking segment, which implies that the symptom severities detected from the last walking segment will be compared to $\mathcal{G}^{(k)}(2)$ in the most extreme scenario. Consequently, we only need to inspect the segment of $\mathcal{G}^k(t)$ defined on the interval $t \in [0, 2]$, since only this segment has been compared to some real data. In this sense, n_w reflects our belief on which segment of $\mathcal{G}^k(t)$ can be learned from data with reasonable qualities.

$$h_i^T = \text{GRU}(h_{i-1}^T, S_i \oplus \Phi(t_i)) \quad (10a)$$

$$t_0^{(k)} = -n_w \sigma(w_k^T h_N^T) \quad (10b)$$

3.3.3. Superior Properties of the Trend Prototype: First, no matter how irregularly the focal patient has performed walking segments over time, the symptom progression matrix detected from these walking segments can always be compared to each trend prototype in a consistent manner. The irregularity of inter-segment time intervals is explicitly considered, because each trend prototype is a continuous function of time, and therefore the differences in inter-segment time intervals are reflected in the evaluation differences of function values. Second, while trend prototype k is learned from irregularly spaced point observations of symptom severities, it can be inspected as a complete temporal symptom progression by evaluating $\mathcal{G}^{(k)}(t)$ at regularly and densely spaced timepoints. Moreover, a trend prototype does not need to be attached to a single patient. Instead, walking segments gathered for different patients could be used to learn different segments of a trend prototype, which together pinpoint a complete temporal symptom progression spanning the observation window.

3.4. Learning Objective

The classification layer of TempPNet computes the probability of depression given the input sensor signal X of a focal patient, which is defined as

$$P(y=1|X) = \sigma\left(\sum_{k \in \mathcal{K}^+} s_k^T - \sum_{k \in \mathcal{K}^-} s_k^T\right) \quad (11)$$

where $\mathcal{K}^+ = \{1, 2, \dots, K\}$, $\mathcal{K}^- = \{K+1, K+2, \dots, 2K\}$, and s_k^T is defined by Equation 9. The above definition imposes the relationship that a high probability of depression is due to the detection of strong overall existing strength of depression trend prototypes, while weak overall existing strength of non-depression trend prototypes. The objective function of TempPNet to be minimized is defined based on the binary cross-entropy loss with two additional regularization terms. Specifically,

$$\mathcal{L}_{\text{TempPNet}} = -\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \log P(y^{(u)}|X^{(u)}) + \lambda_S R_S + \lambda_T R_T \quad (12)$$

where $P(y^{(u)}|X^{(u)})$ is computed in accordance with Equation 11 for patient u , R_S and R_T respectively denote the regularization term for the symptom prototypes and the trend prototypes, and lastly, λ_S and λ_T are the hyperparameters balancing the influence of the regularization terms. Let \mathcal{U}^+ denote the set of depressed patients, and \mathcal{U}^- the set of non-depressed patients. For symptom prototypes, we define R_S as

$$R_S = \frac{1}{M} \sum_{m=1}^M \left(\frac{1}{|\mathcal{U}^-|} \sum_{u \in \mathcal{U}^-} \frac{1}{N_u} \sum_{i=1}^{N_u} s_{m,i|u}^S - \frac{1}{|\mathcal{U}^+|} \sum_{u \in \mathcal{U}^+} \frac{1}{N_u} \sum_{i=1}^{N_u} s_{m,i|u}^S \right) \quad (13)$$

where $s_{m,i|u}^S$ is the symptom severity computed in accordance with Equation 2 for patient u , and $\frac{1}{N_u} \sum_{i=1}^{N_u} s_{m,i|u}^S$ is the average severity level of symptom m detected in the walking test sequence of patient u . Minimizing R_S enforces that the average symptom severity level should be on average higher in the depressed group than it is in the non-depressed group for all symptoms, and thereby imposes our expected interpretation of symptom prototypes. For trend prototypes, we define R_T as

$$R_T = \frac{1}{|\mathcal{U}^+|} \sum_{u \in \mathcal{U}^+} \left(\max_{k \in \mathcal{K}^-} s_{k|u}^T - \min_{k \in \mathcal{K}^+} s_{k|u}^T \right) + \frac{1}{|\mathcal{U}^-|} \sum_{u \in \mathcal{U}^-} \left(\max_{k \in \mathcal{K}^+} s_{k|u}^T - \min_{k \in \mathcal{K}^-} s_{k|u}^T \right) \quad (14)$$

where $s_{k|u}^T$ is the trend existing strength computed in accordance with Equation 9 for patient u . Minimizing the first (second) summation term in Equation 14 encourages the model to detect strong existing evidence for at least one depression (non-depression) trend prototype from the walking segment sequence of each depressed (non-depressed) patient, while weak existing evidence for all non-depression (depression) trend prototypes from the same walking segment sequence.

3.5. Prototype Visualization

TempPNet interprets its decision according to the following mechanism. If the focal patient is classified into the depression class, we must have $P(y = 1|X) > 0.5$ under the decision threshold 0.5, which implies that $\sum_{k \in \mathcal{K}^+} s_k^T > \sum_{k \in \mathcal{K}^-} s_k^T$, i.e., the overall existing strength of depression trend prototypes must be stronger than non-depression trend prototypes. In this case, we can interpret the

decision made by TempPNet by inspecting the depression trend prototypes that strongly present in the walking segment sequence of the focal patient, i.e., trend prototypes with values of s_k^T . Different from existing prototype learning methods, where prototypes are learned in a latent space that is not directly interpretable (Chen et al. 2019), the trend prototypes learned by TempPNet can be directly inspected by visualizing them as trajectories of symptom severities evolving over time. Moreover, as explained in Section 3.3.2, to inspect one trajectory of a trend prototype, it is enough to visualize its segment for the time interval $t \in [0, n_w + n_T]$, where n_w is the hyperparameter defined in Equation 10b, and n_T is the observation window defined as the maximum span from the timepoint of the first walking segment to the timepoint of the depression label.

However, trend prototype $p_{k,m}^T$ does not stand alone. It depicts the typical temporal progression pattern of symptom m . Therefore, visualizing the underlying symptom prototype m is critical for a thorough understanding of model decisions. Given that the learned embedding vector of symptom prototype m is p_m^S , we leverage the sensor data to offer an interpretation for p_m^S in the following approach. First, we search for the particular walking segment where symptom prototype m presents most strongly across all patients. Mathematically, this means solving for the combination $(u^*, i^*) = \arg \max_{u,i} s_{m,i|u}^S$, where $u \in \mathcal{U}$ and $i \in \{1, 2, \dots, N_u\}$. Second, based on the identified sensor signal sequence $X_{i^*}^{(u^*)}$, we use Equation 2 to further search for the particular patch o^* where symptom prototype m presents most strongly. Mathematically, $o^* = \arg \max_o s_{o|m,i^*}^S$, where $o \in \{1, 2, \dots, n_o\}$ and the patient index u^* is omitted for simplicity. Lastly, we inspect symptom prototype m by visualizing the receptive field of this patch, which corresponds to a local segment in $X_{i^*}^{(u^*)}$ that contributes mostly to the computation of $H_{o^*|i^*}^S$, the embedding vector of this patch defined by Equation 1. To this end, we adopt the gradient-based approach (Luo et al. 2016). For each timepoint l in $X_{i^*}^{(u^*)}$, we compute $\partial H_{o^*|i^*}^S / \partial a_l$, which is the Jacobian matrix of the patch embedding vector w.r.t. the sensor feature recorded at timepoint l . By measuring the overall importance score of a_l relative to $H_{o^*|i^*}^S$ as $\|\partial H_{o^*|i^*}^S / \partial a_l\|$, the Frobenius norm of the Jacobian matrix, we can pinpoint the receptive field of patch o^* as a local segment in $X_{i^*}^{(u^*)}$, within which the importance scores of sensor signal inputs are above a pre-specified threshold such as zero.

4. Empirical Analyses

4.1. Research Context and Data Collection

We have obtained the National Health and Nutrition Examination Survey (NHANES) dataset. The NHANES dataset includes depression labels, measured by PHQ-9, for a variety of chronic disease patients. It also includes time series wearable sensor data collected by smartwatches. The background, data types, data collection process and preprocessing of the NHANES dataset is articulated in Appendix I. Our final dataset for analysis includes participants with at least one chronic disease and participated in the wearable sensor measurement, encompassing 1,069 patients

(42% are depressed). We split this dataset into 60% for training, 20% for validation, and 20% for test. To show our method’s generalizability, we obtained a second dataset: mPower, a smartphone-based study that collects daily motion sensor signals for 916 chronic disease patients (Bot et al. 2016). We replicate all the empirical analyses for the mPower dataset (Appendix L). The results are consistent with the NHANES dataset.

4.2. Depression Detection Evaluation

The benchmark selection and hyperparameter settings are shown in Appendix J. We adopt F1-score, precision, and recall as the evaluation metrics. The best model should have the highest F1-score. The detection performance, input, and interpretability of our model and the benchmarks are reported in Tables 2 and 3. These performances are the mean of 10 experimental runs. We also report the standard deviations of the performances. We first compare with the commonly used machine and deep learning models in sensing studies. Compared to the best deep learning model (RNN), our model increases F1-score by 0.084. This increase is attributed to our model’s capability of capturing temporal symptom progression and depression symptoms. Compared to the leading feature-based ML model (SVM), TempPNet boosts F1-score by 0.128. This performance enhancement is due to our model’s ability to learn effective features from the raw sensor signal.

Table 2 Detection Performance Comparison with Machine and Deep Learning Methods (NHANES)

Model	Input	Interpretable	F1-score	Precision	Recall
TempPNet (Ours)	Raw sensor	Yes	0.778 ± 0.009	0.764 ± 0.011	0.792 ± 0.016
CNN	Raw sensor	No	0.686 ± 0.014	0.664 ± 0.146	0.710 ± 0.163
RNN	Raw sensor	No	0.694 ± 0.013	0.684 ± 0.174	0.704 ± 0.176
KNN	Features	No	0.492 ± 0.000	0.978 ± 0.000	0.328 ± 0.000
SVM	Features	No	0.650 ± 0.000	0.985 ± 0.000	0.485 ± 0.000
Random forest	Features	No	0.562 ± 0.056	0.611 ± 0.228	0.602 ± 0.157
AdaBoost	Features	No	0.508 ± 0.000	0.979 ± 0.000	0.343 ± 0.000
XGBoost	Features	No	0.629 ± 0.000	0.984 ± 0.000	0.463 ± 0.000

As described in the literature review, our study is related to MTSC and prototype learning for MTSC. Although our model significantly differs from these models in data structure and model design, we still apply those studies to our problem in order to show the successful design choice of our model and validate our methodological novelty. Compared to regular MTSC models without temporal progressions of prototypes (Fawaz et al. 2020), our model increases F1-score by 0.102. This result proves that capturing the prototypes and their temporal progressions assists in detection performance. Compared to the best-performing prototype learning for MTSC model (Ming et al. 2019), TempPNet improves F1-score by 0.073. Such a significant performance gain indicates that capturing the temporal progressions of prototypes greatly contributes to depression detection.

Table 3 Detection Performance Comparison with MTSC and Prototype Learning for MTSC (NHANES)

Model	Interpretable	Progression of Prototype	F1-score	Precision	Recall
TempPNet (Ours)	Yes	Yes	0.778 ± 0.009	0.764 ± 0.011	0.792 ± 0.016
Chen et al. (2019)	Yes	No	0.698 ± 0.013	0.686 ± 0.144	0.710 ± 0.165
Ming et al. (2019)	Yes	No	0.705 ± 0.012	0.707 ± 0.179	0.704 ± 0.176
Gee et al. (2019)	Yes	No	0.672 ± 0.017	0.646 ± 0.135	0.701 ± 0.160
Ma et al. (2020)	Yes	No	0.695 ± 0.014	0.681 ± 0.177	0.710 ± 0.180
Fawaz et al. (2020)	No	No	0.676 ± 0.016	0.646 ± 0.127	0.755 ± 0.142

Since our model consists of multiple critical design components, we further perform ablation studies to show their effectiveness, as reported in Table 4. We first remove the latent trend starting time design ($t_0^{(k)}$). We also remove the trend prototype design. After removing the trend prototype, the model loses the capability of detecting temporal symptom progression. Consequently, we test two options: using the last symptom severity to detect depression and using the average symptom severity over time to detect depression. Table 4 suggests that removing any design component will significantly hamper the detection accuracy, proving that our design choice is optimal.

Table 4 Ablation Studies (NHANES)

Model	F1-score	Precision	Recall
TempPNet (Ours)	0.778 ± 0.009	0.764 ± 0.011	0.792 ± 0.016
TempPNet removing offset $t_0^{(k)}$	0.694 ± 0.024	0.731 ± 0.020	0.661 ± 0.036
Remove trend prototype using last symptom severity	0.711 ± 0.022	0.719 ± 0.021	0.703 ± 0.025
Remove trend prototype using average symptom severity	0.724 ± 0.012	0.708 ± 0.013	0.740 ± 0.017

To reduce noise in the sensor data and avoid overfitting, sensor-based prediction studies usually downsample the sensor signals (Sigcha et al. 2020). We test the effect of different sample rates in Table 5: 10 Hz, 20 Hz, and 30 Hz. The results suggest that 10 Hz signal frequency achieves the best performance. Therefore, we use the 10 Hz signal frequency for all the other analyses.

Table 5 Analysis of Signal Frequency (NHANES)

Signal Frequency	F1-score	Precision	Recall
10 Hz	0.778 ± 0.009	0.764 ± 0.011	0.792 ± 0.016
20 Hz	0.754 ± 0.011	0.742 ± 0.015	0.766 ± 0.018
30 Hz	0.747 ± 0.013	0.735 ± 0.014	0.761 ± 0.017

Certain pre-existing chronic diseases and depression may have a correlation because they share similar walking symptoms. To make sure that our model is actually detecting depression instead of other pre-existing chronic diseases, we perform evaluations that are conditioned on whether a patient has a chronic disease or not. Since our dataset contains numerous chronic diseases, we report two conditions (have or not have) for each disease to be concise in the main manuscript.

In Appendix K, we further select two diseases (diabetes and kidney disease) to showcase our model’s performances when conditioned on more nuanced disease severity levels. If our model indeed detects depression, we expect that, conditioned on each disease, our model’s performance on depression detection should remain consistently high. If our model only detects pre-existing chronic diseases, conditioned on having (or not having) the disease, the performance should be very low because, in this group, the model has not seen different values of the outcome, thus unable to update parameters. Table 6’s results prove that given having (or not having) any chronic disease, our model is able to accurately detect depression consistently. The results of evaluations on more nuanced disease severity levels in Appendix K show the same conclusion, suggesting our depression detection is robust in any pre-existing disease severity levels. Therefore, our model indeed detects depression rather than pre-existing chronic diseases.

Table 6 Evaluations Conditioned on Chronic Diseases (NHANES)

Disease	Status	F1-score	Precision	Recall	Disease	Status	F1-score	Precision	Recall
HBP	Yes	0.806 ± 0.021	0.806 ± 0.032	0.807 ± 0.021	Gout	Yes	0.741 ± 0.036	0.716 ± 0.059	0.769 ± 0.036
HBP	No	0.770 ± 0.024	0.752 ± 0.039	0.789 ± 0.025	Gout	No	0.779 ± 0.021	0.767 ± 0.038	0.793 ± 0.022
CVD	Yes	0.807 ± 0.023	0.767 ± 0.038	0.853 ± 0.026	Heart	Yes	0.759 ± 0.045	0.743 ± 0.042	0.778 ± 0.065
CVD	No	0.769 ± 0.016	0.768 ± 0.029	0.771 ± 0.013	Heart	No	0.779 ± 0.016	0.763 ± 0.030	0.797 ± 0.013
Diabetes	Yes	0.777 ± 0.028	0.769 ± 0.056	0.787 ± 0.014	Stroke	Yes	0.839 ± 0.055	0.810 ± 0.080	0.873 ± 0.043
Diabetes	No	0.775 ± 0.035	0.758 ± 0.064	0.796 ± 0.024	Stroke	No	0.772 ± 0.022	0.764 ± 0.038	0.782 ± 0.022
Kidney	Yes	0.809 ± 0.037	0.807 ± 0.061	0.813 ± 0.018	Emphysema	Yes	0.825 ± 0.053	0.778 ± 0.087	0.882 ± 0.027
Kidney	No	0.763 ± 0.040	0.757 ± 0.081	0.773 ± 0.023	Emphysema	No	0.769 ± 0.020	0.747 ± 0.032	0.794 ± 0.019
Asthma	Yes	0.795 ± 0.015	0.735 ± 0.046	0.869 ± 0.039	Bronchitis	Yes	0.805 ± 0.035	0.742 ± 0.055	0.881 ± 0.026
Asthma	No	0.767 ± 0.017	0.782 ± 0.033	0.754 ± 0.016	Bronchitis	No	0.752 ± 0.022	0.746 ± 0.038	0.759 ± 0.021
Celiac	Yes	0.755 ± 0.093	0.892 ± 0.102	0.689 ± 0.175	COPD	Yes	0.811 ± 0.025	0.756 ± 0.043	0.876 ± 0.018
Celiac	No	0.783 ± 0.018	0.772 ± 0.033	0.795 ± 0.016	COPD	No	0.767 ± 0.020	0.762 ± 0.036	0.773 ± 0.021
Arthritis	Yes	0.808 ± 0.022	0.738 ± 0.031	0.894 ± 0.022	Cancer	Yes	0.766 ± 0.024	0.777 ± 0.059	0.764 ± 0.080
Arthritis	No	0.758 ± 0.031	0.793 ± 0.029	0.728 ± 0.049	Cancer	No	0.782 ± 0.023	0.765 ± 0.047	0.801 ± 0.017

4.3. Interpretation of Depression Detection

Beyond depression detection, TempPNet is capable of interpreting why a patient is classified as depressed by presenting the contributing temporal symptom progression (trend prototype) and the corresponding walking symptom (symptom prototype). Figure 5 shows the trend prototypes that our model learned from the NHANES dataset. These trend prototypes are the prototypical depression or non-depression trends. They are learned in a data-driven manner and show representative trends. For each picture, the x-axis is time, and the y-axis is symptom severity.

Trend prototypes (1)-(5) are depression trend prototypes. They represent the severities of depression symptoms trending up. Some of them have deviations from time to time in the upward trend, representing temporary symptom relief and deterioration of depression. This conforms with the typical depression trend (Bockting et al. 2015). Trend prototypes (6)-(10) are non-depression trend prototypes, where (7)-(10) represent trending down and (6) represents fluctuating at a certain level.

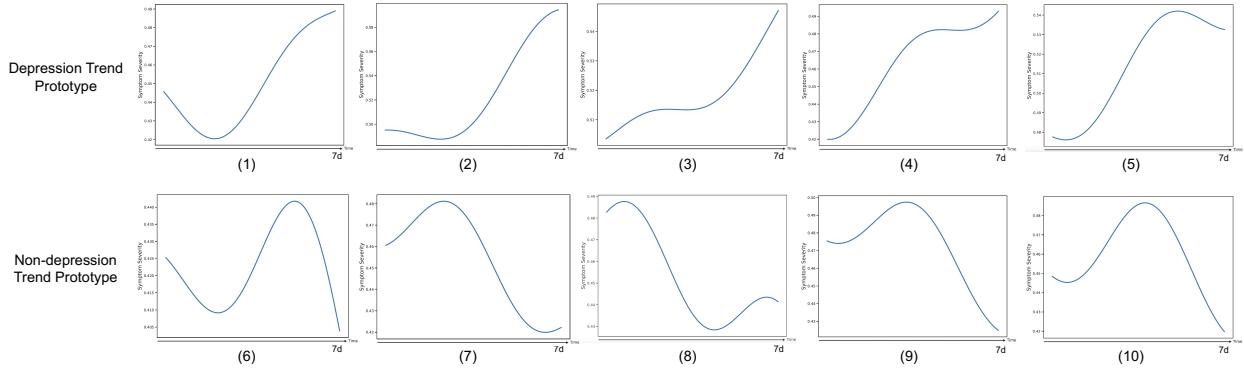
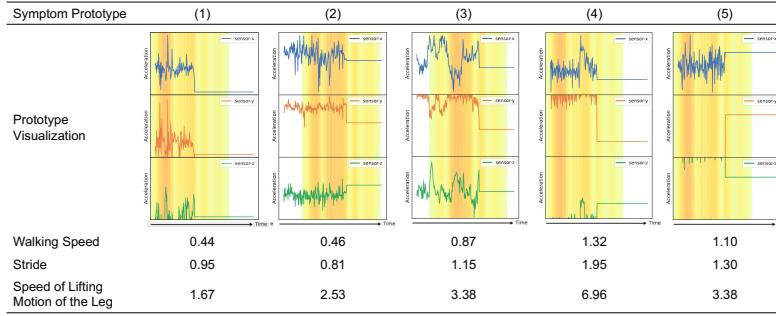


Figure 5 Trend Prototypes

These are not typical depression trends. The trend prototypes are learned using all the patients' data rather than relying on a single patient's data. Each patient's observed walking data could be at different stages of a trend — some at the rising stage, some at the stable stage, among others. Together they depict a complete trend. Multiple patients could also share the same trend prototype if their symptom severity levels are at the same stage (e.g., all on the rise). Each trend prototype is coupled with underlying symptoms. Figure 6 shows the symptom prototypes that our model learned.



Note: Each of these symptom prototypes represents a typical depression or non-depression walking pattern. Many gait patterns can be derived from a walking signal. Take three patterns that are frequently referenced in the depression literature as an example: walking speed, stride, and lifting. In our study population, the average walking speed is 0.76. The average stride is 1.08. The average lifting motion is 3.98. For example, symptom prototype 1 shows slower walk, shorter strides, and slower lifting motion than usual. This is indicative of depression. Symptom prototype 4 shows much faster walk than usual, longer stride, and much faster lifting motion. This is indicative of non-depression. The flat line represents zero acceleration. Some walking segments have positive acceleration readings, and some have negative readings. We include flat lines to show which walking segments have positive acceleration (i.e., those above the flat line) and which walking segments have negative acceleration (i.e., those below the flat line).

Figure 6 Symptom Prototypes

The prototype visualization in Figure 6 shows the sensor signal of the symptoms. These symptom prototypes are learned by our method across all patients. Each patient may present none, one, or more of these symptoms in their walking patterns. Prior literature suggests that depression walking symptoms can be reflected in gait patterns, such as walking speed, stride, and speed of the lifting motion of the leg (Sloman et al. 1982, Lemke et al. 2000, NHS 2022). To accompany the interpretation of the detection of a patient, these gait patterns can be computed for the symptom prototypes.

Leveraging the above-learned trend prototypes and symptom prototypes, our model can interpret the detection of depression for every patient. We randomly select two patients (one depressed

and one non-depressed) and showcase TempPNet’s interpretation for them. Figure 7 shows the interpretations of the depressed and non-depressed patients. For simplicity, we only show the trend prototype with the highest existing strength and the corresponding symptom prototype with the highest existing strength in these examples. For the symptom prototype, we also compute a few exemplar gait patterns using GaitPy (Gaitpy 2024) to explain the encoded information from the visualization. The arrows after the gait patterns denote whether a pattern is higher or lower than an average human. They do not imply any trend information (neither go up nor go down).

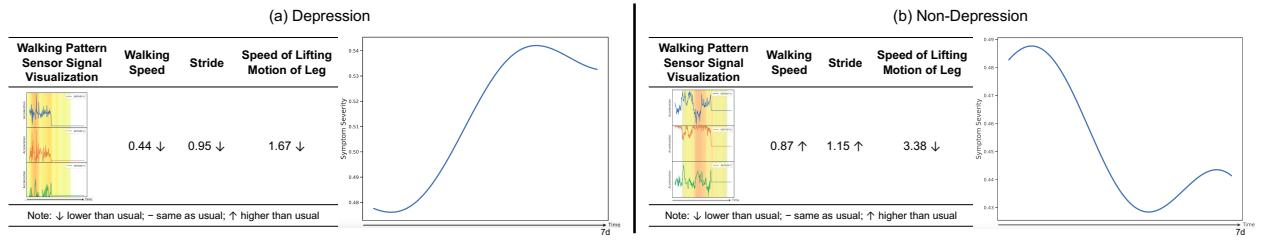


Figure 7 Interpretation of A Depressed and A Non-depressed Patient

TempPNet detects the patient in Figure 7(a) as depressed for two reasons. First, this patient’s walking patterns strongly present a walking symptom like in the left part of Figure 7(a). This walking symptom is manifested as slower-than-usual walking speed, shorter stride, and slower lifting motion of the leg. This symptom conforms with the depression physical symptoms in the literature (Sloman et al. 1982, Lemke et al. 2000, NHS 2022). Second, the severity of this symptom presents a temporal progression pattern like the right part of Figure 7(a). TempPNet believes this temporal symptom progression pattern resembles a typical depression progression pattern. According to the depression progression literature (Bockting et al. 2015, Dattani et al. 2021), this judgment makes sense — this patient’s depression walking symptom first worsens rapidly and then peaks, similar to the onset and acute phases in Figure 1. TempPNet detects the patient in Figure 7(b) as non-depressed for two reasons. First, this patient’s walking patterns strongly present a walking symptom like in the left part of Figure 7(b). This walking symptom is manifested as faster-than-usual walking speed, longer stride, and slower lifting motion of the leg. This symptom does not resemble the typical depression walking symptoms. Second, the severity of this symptom presents a temporal progression pattern like the right part of Figure 7(b). The symptom severity trends down and has fluctuations in the middle. This trend does not resemble a typical depression trend.

As medical experts are aware of what a typical depression walking symptoms and temporal progression look like, our interpretation helps doctors understand why a patient is detected as depressed. When our interpretation matches the medical understanding, as is the case in Figure 7,

doctors would trust our model’s decision more. Based on our interpretation, doctors can also assess what symptoms the patient presents and in what phase of the depression progression the patient is at. This interpretation information coupled with our detection result could inform doctors to take corresponding interventions to treat depression with increased confidence and precision.

4.4. Human Evaluations

4.4.1. Large Scale User Study: The major contribution of TempPNet in interpretation lies in its capability of interpreting temporal symptom progression. To validate that capturing temporal symptom progression improves interpretability of our model, we conduct a user study. We informed the participants that they would be assigned an interpretable ML model to detect depression using sensor data. After filtering out those who failed the attention check and manipulation check (articulated below), there are 66 participants in the user study. We randomly select two patient samples that the model detects as depressed and non-depressed and show the participants how the model interpreted those decisions. We design two randomized groups: one presenting TempPNet’s interpretation, and the other presenting the baseline’s interpretation (without temporal symptom progression).

The first part of the user study collects five control variables: age, education, gender, trust in AI, and health literacy. The health literacy instrument is adopted from Osborne et al. (2013). This user study passed randomization checks. The summary statistics and randomization p -values are reported in Appendix M. Since the interpretation is based on depression walking symptoms, we design a depression knowledge education session for all participants, showing the depression walking symptoms (Appendix M). After reading such information, they are asked to answer four questions (Appendix M) to test their understanding. If they answer these questions incorrectly, an error message will prompt and direct them to read the information and choose again. After they answer the test questions correctly, they have sufficient knowledge to understand the model interpretations.

Next, we inform the participants the context, input, and output. Then, we show the model interpretation to each group respectively, as shown in Appendix M. A manipulation check question is asked to verify whether the participants read the interpretation carefully (“How does the walking speed of the above walking symptom compare to the usual case?”), whose answer can be understood from the arrow sign after the value of the walking speed in the interpretation figure.

Subsequently, we ask the participants to rate the interpretability of the given model. The definition of interpretability varies across domains (Lee et al. 2018). In business analytics and ML (Lee et al. 2018, Miller 2019, Molnar 2020), studies define interpretability as the degree to which a user can trust the cause of a decision (Miller 2019, Molnar 2020, Xie et al. 2023). Trust in ML models also holds paramount importance in healthcare analytics for practitioners’ and patients’ adoption

considerations. Following the literature, we use trust in automated systems as a interpretability metric. We also use other interpretability metrics in the expert evaluation in the next subsection. The measurement scale of trust is adopted from [Jian et al. \(2000\)](#). The Chronbach's Alpha is 0.884, suggesting excellent reliability. An attention check question is added ("Please just select neither agree nor disagree."). The participants' trust in our model (mean = 2.26) is significantly higher than the baseline model (mean = 1.71, $p < 0.001$). This result indicates that interpreting the temporal symptom progression is a critical booster for users' trust in depression prediction models, endorsing our model design.

Afterwards, we show the participants the interpretation of the other model as a comparison. We ask them to choose a model that they would finally adopt. 58 participants (88%) chose TempPNet over the baseline. When asked why they adopt TempPNet rather than the baseline, 43 participants mentioned, in similar wording, that "because this model visualized the walking symptom progression." The above user study results prove that by interpreting the temporal symptom progression, TempPNet improves interpretability, which offers empirical evidence for the contribution of our model innovation.

4.4.2. Medical Expert Panel: The user study employs non-experts who may be unable to judge the clinical meaningfulness of the interpretations. Interpretability also has other objectives. To address these limitations, we employ a medical expert panel. One representative study that employs a medical expert panel is a recent MISQ article by [Kim et al. \(2023\)](#), which also develops an interpretable ML method for healthcare applications. [Kim et al. \(2023\)](#)'s method is a post-hoc method and provides feature-based interpretations, which are undesired in our context, as articulated in Section 2.3. Since both our study and [Kim et al. \(2023\)](#) are IS studies that develop interpretable ML methods for healthcare applications, we follow the evaluation precedence used by [Kim et al. \(2023\)](#) to conduct our expert panel evaluation. Specifically, we employed an expert panel of 8 medical professionals, including 2 clinical psychologists (both having PhDs) and 6 psychiatrists (all having MDs), which is larger than the expert panel utilized by [Kim et al. \(2023\)](#) (4 experts). The size of our expert panel is also larger than or comparable to that of expert panels used in clinical diagnostic research — "*Most panels used two members (29 of 63 papers, 46%), followed by three members (18 of 63 papers, 29%). The maximum reported number of members was nine.*" ([Bertens et al. 2013](#)). Our medical experts have an average of 9 years of clinical experience. We conducted a 20-minute interview with each of the medical experts independently; each interview consisted of a survey and a structured interview that lets the interviewee evaluate the interpretations and prototypes produced by our method. The front part of the survey (introduction of the model and interpretation figures) is the same as the one used in the previous user study, as this part proved

to be easy to understand by both expert and non-expert readers. We randomly split the medical experts into two equally-sized groups. Group 1 read our model’s interpretation (Figure A.7a), and group 2 read the baseline model’s interpretation (Figure A.7b). Following that, we asked 4 questions to evaluate various aspects of interpretability (The question wordings are adopted from Kaur et al. (2020) and Wang et al. (2021b)): 1) How reasonable do you think the model’s interpretation is? 2) How trustworthy do you think the model’s interpretation is? 3) How clinically meaningful do you think the model’s interpretation is? 4) How useful do you think the model’s interpretation is for depression screening practices? All these questions used a scale of 0-4 (0 = “not at all” and 4 = “extremely”).

After the survey, we told each expert there was a model that had learned typical symptom prototypes and trend prototypes from a dataset of wearable sensor signals. We then presented to each expert symptom prototypes produced by our model (Figure 6) and trend prototypes generated by our model (Figure 5). Finally, we asked each expert three questions: 1) Are these prototypes clinically meaningful when interpreting the depression of patients? 2) If using these prototypes to explain the depression case of a patient, are these prototypes useful in assisting your depression screening process? 3) Do you have any further comments on these prototypes and the model?

The survey and interview results reaffirm the better interpretability of our model. In terms of the survey (Figure 8), for reasonability, trustworthiness, clinical meaningfulness, and usefulness for depression screening practices, the answers within the same group are pretty consistent — group 1 experts (presented with interpretations by our method) all rated 3 or 4 (on a scale of 0-4), and group 2 experts (presented with interpretations by the baseline method) all rated 0-2. This proves that our new trend prototypes indeed are meaningful and useful clinically. Experts who read the baseline interpretation in the survey commented unanimously that showing a snapshot of a symptom without knowing its temporal progression is unreliable in understanding the depression status of a patient, which further validates the importance of our introduced trend prototypes. For the structured interview, all 8 experts agreed that prototypes produced our model (Figures 5 and 6) are clinically meaningful and useful in assisting their depression screening process. The experts further commented that our model, if implemented as an app, will be highly useful for their clinical practices. They noted that current depression diagnoses often involve subjectivity, and that patients usually do not describe their feelings well — e.g., whether they are feeling better or worse is difficult to be understood objectively by clinicians. Symptom and trend prototypes learned by our method using objective sensor data, on the contrary, provide an objective assessment of a patient’s depression status and its trend, thereby assisting clinicians’ understanding of the status and progress of a patient’s depression and alleviating subjectivity. The experts also mentioned that using our model’s interpretations in practice would significantly enhance their understanding of patients’ depression conditions.

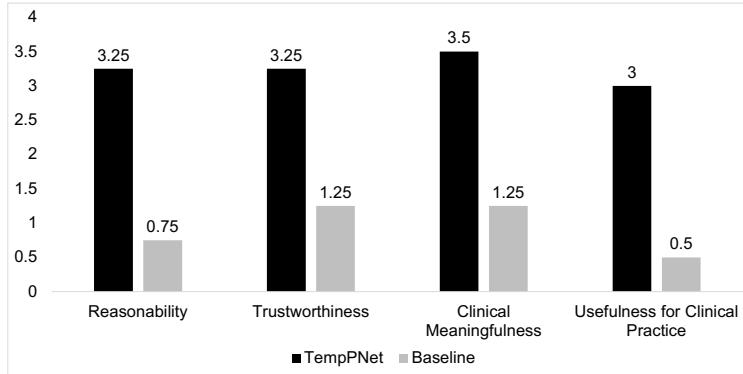


Figure 8 Expert Panel Survey Results (Mean in Each Group)

5. Conclusion

5.1. Summary and Contributions

Depression detection is fundamental to comprehensive chronic care in the health sensing area. Although a few studies tapped this domain, they do not offer a meaningful interpretation of their decisions. Prototype learning methods fall short in modeling the temporal progression of depression. To address these limitations, we propose a novel interpretable deep learning method to detect depression using sensor data while interpreting its decisions based on the temporal progression of depression. We conduct extensive evaluations to demonstrate superior detection performance of our method over state-of-the-art benchmarks and showcase its interpretation of decisions. Our human evaluations show that our method outperforms the benchmark in terms of interpretability.

Our work belongs to the computational genre of design science research in IS, which develops computational methods to solve business and societal problems and aims to make methodological contributions (Rai 2017, Simchi-Levi 2020, Padmanabhan et al. 2022). In this regard, our proposed TempPNet, the designed IT artifact, is a novel prototype learning method that processes a time series of multivariate time series and interprets decisions based on prototypes of temporal symptom progression. To design TempPNet, we innovatively overcome three methodological challenges: learning prototypes of depression symptoms, learning prototypes of temporal symptom progression from walking segments performed at irregular time intervals, and modeling walking segments of varying lengths. This IT artifact design process reveals three general design principles: 1) Capturing the temporal evolution of physical patterns can improve the performance of predictive systems; 2) Identifying typical patterns of an existing class for prediction not only promotes model interpretability but also enhances predictive accuracy; 3) Learning representation from raw physical activities is more useful than feature engineering. These design principles offer computational IS scholars valuable references when searching through a solution space to design novel IT artifacts.

To the HIT literature, we validate the potential of leveraging sensing information technology for depression management amid regular treatments of chronic diseases. Our proposed novel interpretable deep learning method can detect depression associated with chronic diseases using motion sensor data and interpret its decisions. Applying our model, new business models can be developed to assist depression treatment for chronic disease patients in real-time.

5.2. Managerial Implications and Future Work

Our study offers managerial implications for multiple key stakeholders, including doctors, caregivers, and healthcare systems. Doctors can encourage their chronic disease patients to install wellness trackers to actively monitor patients' walking data. The wellness tracker data can be analyzed by our method to detect depression risk on a daily basis. When our method signals a high depression risk, the doctors could review the interpretation given by our model and understand what depression walking symptoms the patient presents and how the patient's depression severity has progressed over time. Timely and personalized interventions can be taken to prevent the deterioration of depression and the negative spill-over effects for physical chronic symptom treatments. Physicians are also concerned about their patient's satisfaction. If the model detects depression correctly, patients' satisfaction will be higher. Our empirical results show that our model significantly outperforms the baselines. Therefore, the satisfaction of our model is likely to be higher than the baseline models.

Caregivers' home care is one of the key determinants of successful chronic disease management. Our method offers an opportunity for caregivers to keep track of the daily mental health updates of their loved ones. Caregivers can use our sensor-based depression detection mobile system for patients as well as themselves. When patients' mental health deteriorates, the caregivers can actively seek medical help, provide emotional support, and look for social support groups for the patients. Taking care of chronic disease patients can be stressful. This depression tracker on their own mobile devices serves as a reminder for self-care. Patients themselves can be aware of their mental health status and be proactive in their health management.

Since depression poses a significant economic burden, our method enables active monitoring of depression and timely intervention, thus allowing for better resource allocation and potentially saving untreated depression costs. There are an estimated 307 million smartphone users in the US ([Statista 2022](#)), among which 62% (190.34 million) use m-health apps that track users' motion data to monitor health conditions ([Vicert 2021](#)). These are consenting users for collecting and using their physical data captured by motion sensors. Nearly 60% of US adults have at least one chronic disease ([Hoffman 2022](#)). Therefore, there are an estimated $190.34 \times 60\% = 114.20$ million consenting m-health app users who suffer from chronic diseases. They represent 34.71% of the

US's 329 million population. The annual direct healthcare cost of untreated mental illnesses in the state of Indiana is \$708.5 million (Taylor et al. 2023). Depression healthcare costs account for around 24.82% of mental illness healthcare costs (APA 2021, AHQR 2022). Therefore, the annual healthcare cost of untreated depression in the state of Indiana is estimated to be \$708.5 million $\times 24.82\% = \$175.85$ million. Based on the population ratio of Indiana to the US (Britannica 2023), the annual healthcare cost of untreated depression in the US is estimated to be \$8.58 billion, among which the chronically-ill consenting m-health app users account for $8.58 \times 34.71\% = \$2.98$ billion. Given such a sizable cost of untreated depression, using our model wisely could have a considerable positive impact on healthcare systems. Specifically, the probability of depression predicted by our model can serve as input for downstream optimization models, which allocate limited medical resources among healthcare systems to maximize the number of depression cases treated.

Our study has limitations and can be extended in various ways. Our proposed method focuses on walking activities, while other activities, such as work productivity and sleep quality, can also serve as indicators of depression. Future research could enhance our method by incorporating data on these other activities, although obtaining such data would incur additional cost and require further user consent. In addition, depression detection is one problem that can be effectively solved by our method. It can be adapted to tackle challenges in many other areas, such as mobile analytics, health information technology, investment portfolio choice, and social media analytics. A detailed discussion of our method's implications for these areas is provided in Appendix N. Future research could collect time series data and adapt our method to address important challenges in these areas.

6. Acknowledgments

Jiaheng Xie and Xiao Fang are supported by the University of Delaware Research Foundation Strategic Initiatives (UDRF-SI) Grant and Alfred Lerner College of Business and Economics Research Grant. Jiaheng Xie and Xiao Fang are not supported by any other funds. Xiaohang Zhao is supported by the Fundamental Research Funds for the Central Universities: “High-Quality Development of Digital Economy: An Investigation of Characteristics and Driving Strategies (Grant Number 2023110139)” and “Intelligent Decision-Making Theories and Methods for Online Digital Platforms (Grant Number 2023110318).”

References

- AHQR (2022) Healthcare expenditures for treatment of mental disorders: Estimates for adults ages 18 and older, u.s. civilian noninstitutionalized population, 2019. https://meps.ahrq.gov/data_files/publications/st539/stat539.pdf?utm_source=chatgpt.com, [Accessed 08-02-2025].
- Ansah JP, Chiu CT (2023) Projecting the chronic disease burden among the adult population in the united states using a multi-state population model. *Frontiers in public health* 10:1082183.
- APA (2021) The economic cost of depression is increasing; direct costs are only a small part. https://www.psychiatry.org/news-room/apa-blogs/the-economic-cost-of-depression-is-increasing?utm_source=chatgpt.com, [Accessed 08-02-2025].
- Bardhan I, Chen H, Karahanna E (2020) Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. *MIS Quarterly* 44(1):185–200.
- Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, Moons KG, Reitsma JB (2013) Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS medicine* 10(10):e1001531.
- Bockting CL, Hollon SD, Jarrett RB, Kuyken W, Dobson K (2015) A lifetime approach to major depressive disorder: The contributions of psychological interventions in preventing relapse and recurrence. *Clinical Psychology Review* 41:16–26.
- Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Doerr M, Pratap A, Wilbanks J, Dorsey ER, Friend SH, Trister AD (2016) The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data* 3(1):1–9.
- Britannica (2023) U.S. States ranked by population: which is largest? — britannica.com. <https://www.britannica.com/topic/largest-U-S-state-by-population>, [Accessed 16-08-2024].
- Canzian L, Musolesi M (2015) Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.
- Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of KDD* Place: New York, NY, USA Publisher: ACM.
- CDC (2012) Mental Health and Chronic Diseases CDC Fact Sheet. Technical report.
- CDC (2021) U.S. healthcare spending attributable to cigarette smoking in 2014. *CDC* 150.
- CDC (2022) Chronic Diseases in America. *CDC* .
- Chau M, Li TM, Wong PW, Xu JJ, Yip PS, Chen H (2020) Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly* 44(2).

- Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C (2019) This Looks Like That: Deep Learning for Interpretable Image Recognition. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.
- Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the Properties of Neural Machine Translation: Encoder-Decoder Approaches.
- Coelln Rv, Dawe RJ, Leurgans SE, Curran TA, Truty T, Yu L, Barnes LL, Shulman JM, Shulman LM, Bennett DA, Hausdorff JM, Buchman AS (2019) Quantitative mobility metrics from a wearable sensor predict incident parkinsonism in older adults. *Parkinsonism & Related Disorders* 65:190–196.
- Czech MD, Patel S (2019) GaitPy: An Open-Source Python Package for Gait Analysis Using an Accelerometer on the Lower Back. *Journal of Open Source Software* 4(43):1778.
- Dattani S, Ritchie H, Roser M (2021) Mental Health. *Our World in Data* .
- Dixon-Woods M, Redwood S, Leslie M, Minion J, Martin GP, Coleman JJ (2013) Improving quality and safety of care using "technovigilance": an ethnographic case study of secondary use of data from an electronic prescribing and decision support system. *The Milbank Quarterly* 91(3):424–454.
- Farhan AA, Yue C, Morillo R, Ware S, Lu J, Bi J, Kamath J, Russell A, Bamis A, Wang B (2016) Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. *2016 IEEE Wireless Health (WH)*.
- Fawaz H, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, Webb GI, Idoumghar L, Muller PA, Petitjean F (2020) Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* 34(6):1936–1962.
- Forbes (2022) Our nation's chronic disease epidemic is getting worse so, who's responsible? *Forbes* URL <https://www.forbes.com/sites/ritanumerof/2022/11/22/our-nations-chronic-disease-epidemic-is-getting-worse-so-whos-responsible/>.
- Gaitpy (2024) gaitpy — pypi.org. <https://pypi.org/project/gaitpy/>, [Accessed 16-08-2024].
- Gao R, Huo Y, Bao S, Tang Y, Antic SL, Epstein ES, Balar AB, Deppen S, Paulson AB, Sandler KL, Massion PP, Landman BA (2019) Distanced LSTM: Time-Distanced Gates in Long Short-Term Memory Models for Lung Cancer Detection. Suk HI, Liu M, Yan P, Lian C, eds., *Machine Learning in Medical Imaging*.
- Gee AH, Garcia-Olano D, Ghosh J, Paydarfar D (2019) Explaining deep classification of time-series data with learned prototypes. *CEUR workshop proceedings* (NIH Public Access).
- Ghosal GR, Abbasi-Asl R (2021) Multi-modal prototype learning for interpretable multivariable time series classification. *arXiv preprint arXiv:2106.09636* .
- Hedman J, Srinivasan N, Lindgren R (2013) Digital traces of information systems: Sociomateriality made researchable. *Proceedings of the 34th International Conference on Information Systems. ICIS 2013* (Association for Information Systems. AIS Electronic Library (AISeL)).

- Hoffman D (2022) chronicdisease.org. https://chronicdisease.org/wp-content/uploads/2022/04/FS_ChronicDiseaseCommentary2022FINAL.pdf, [Accessed 16-08-2024].
- Hubble RP, Naughton GA, Silburn PA, Cole MH (2015) Wearable Sensor Use for Assessing Standing Balance and Walking Stability in People with Parkinson's Disease: A Systematic Review. *PLOS ONE* 10(4):e0123705.
- Jacobson NC, Chung YJ (2020) Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. *Sensors* 20(12):3572.
- Jian JY, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4(1):53–71.
- Katon WJ, Lin EH, Von Korff M, Ciechanowski P, Ludman EJ, Young B, Peterson D, Rutter CM, McGregor M, McCulloch D (2010) Collaborative Care for Patients with Depression and Chronic Illnesses. *New England Journal of Medicine* 363(27):2611–2620.
- Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J (2020) Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI conference on human factors in computing systems*.
- Kim BR, Srinivasan K, Kong SH, Kim JH, Shin CS, Ram S (2023) Rolex: A novel method for interpretable machine learning using robust local explanations. *MIS Quarterly* 47(3).
- LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.
- Lee D, Manzoor E, Cheng Z (2018) Focused concept miner (fcm): Interpretable deep learning for text exploration. Available at SSRN 3304756 .
- Lemke MR, Wendorff T, Mieth B, Buhl K, Linnemann M (2000) Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. *Journal of Psychiatric Research* 34(4-5):277–283.
- Li W, Zhu W, Dorsey ER, Luo J (2020) Predicting Parkinson's Disease with Multimodal Irregularly Collected Longitudinal Smartphone Data. *2020 IEEE International Conference on Data Mining (ICDM)*.
- Lin YK, Fang X (2021) First, do no harm: Predictive analytics to reduce in-hospital adverse events. *Journal of Management Information Systems* 38(4):1122–1149.
- Liu CW, Wang W, Gao G, Agarwal R (2024) The value of virtual engagement: Evidence from a running platform. *Management Science* 70(9):6179–6201.
- Liu J, Zhang T, Han G, Gou Y (2018) TD-LSTM: Temporal Dependence-Based LSTM Networks for Marine Temperature Prediction. *Sensors* 18(11):3797.
- Liu X, Zhang B, Susarla A, Padman R (2020) Go to youtube and call me in the morning: Use of social media for chronic conditions. *MIS Quarterly* 257–283.

- Lundberg SM, Lee SI (2017) A Unified Approach to Interpreting Model Predictions.
- Luo W, Li Y, Urtasun R, Zemel R (2016) Understanding the effective receptive field in deep convolutional neural networks. *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA: Curran Associates Inc.), 4905–4913, NIPS'16, ISBN 978-1-5108-3881-9.
- Ma D, Wang Z, Xie J, Guo B, Yu Z (2020) Interpretable multivariate time series classification based on prototype learning. *Green, Pervasive, and Cloud Computing: 15th International Conference, GPC 2020, Xi'an, China, November 13–15, 2020, Proceedings 15* (Springer).
- Marsh L (2013) Depression and Parkinson's Disease: Current Knowledge. *Current neurology and neuroscience reports* 13(12):409–409.
- Mayo Clinic MC (2022) Depression (major depressive disorder) - Symptoms and causes.
- Michalak J, Troje NF, Fischer J, Vollmar P, Heidenreich T, Schulte D (2009) Embodiment of sadness and depression-gait patterns associated with dysphoric mood. *Psychosomatic Medicine* 71(5):580–587.
- Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267:1–38.
- Ming Y, Xu P, Qu H, Ren L (2019) Interpretable and Steerable Sequence Learning via Prototypes. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- Molnar C (2020) *Interpretable Machine Learning* (Lulu.com).
- Moss L, Corsar D, Shaw M, Piper I, Hawthorne C (2022) Demystifying the Black Box: The Importance of Interpretability of Predictive Models in Neurocritical Care. *Neurocritical Care* 37(2):185–191.
- Murphy KP (2022) *Probabilistic Machine Learning: An Introduction* (The MIT Press).
- Nauta M, Bree Rv, Seifert C (2021) Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*.
- NHS (2022) Symptoms - Clinical depression. *National Heath Service* .
- NIH (2022) Chronic Illness and Mental Health: Recognizing and Treating Depression. *National Institute of Mental Health* .
- Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R (2013) The grounded psychometric development and initial validation of the health literacy questionnaire (hlq). *BMC public health* 13(1):1–17.
- Oung Q, Hariharan M, Lee H, Basah S, Sarilée M, Lee C (2015) Wearable multimodal sensors for evaluation of patients with Parkinson disease. *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*.
- Padmanabhan B, Fang X, Sahoo N, Burton-Jones A (2022) Machine Learning in Information Systems Research. *MIS Quarterly* 46(1):iii–xix.

- Rai A (2017) Editor's comments: diversity of Design Science Research. *MIS Quarterly* 41(1):iii–xviii.
- Reijnders JS, Ehrt U, Weber WE, Aarsland D, Leentjens AF (2008) A systematic review of prevalence studies of depression in Parkinson's disease. *Movement Disorders* 23(2):183–189.
- Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1(5):206–215.
- Ruiz AP, Flynn M, Large J, Middlehurst M, Bagnall A (2021) The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery* 35(2):401–449.
- Rymarczyk D, Struski L, Tabor J, Zieliński B (2021) ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification.
- Shim J, van den Dam R, Aiello S, Penttinen J, Sharda R, French A (2022) The transformative effect of 5g on business and society in the age of the fourth industrial revolution. *Communications of the Association for Information Systems* 50(1):29.
- Sigcha L, Costa N, Pavón I, Costa S, Arezes P, López JM, De Arcas G (2020) Deep Learning Approaches for Detecting Freezing of Gait in Parkinson's Disease Patients through On-Body Acceleration Sensors. *Sensors 2020, Vol. 20, Page 1895* 20(7):1895–1895.
- Simchi-Levi D (2020) From the editor: diversity, equity, and inclusion in management science. *Management Science* 66(9):3802–3802.
- Singh G, Yow KC (2021) These do not Look like Those: An Interpretable Deep Learning Model for Image Recognition. *IEEE Access* 9:41482–41493.
- Sloman L, Berridge M, Homatidis S, Hunter D, Duck T (1982) Gait patterns of depressed patients and normal subjects. *American Journal of Psychiatry* 139(1):94–97.
- Statista (2022) Topic: US smartphone market.
- Taylor HL, Menachemi N, Gilbert A, Chaudhary J, Blackburn J (2023) Economic burden associated with untreated mental illness in Indiana. *JAMA Health Forum* (American Medical Association).
- Trinh L, Tsang M, Rambhatla S, Liu Y (2021) Interpretable and trustworthy deepfake detection via dynamic prototypes. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.
- Vancampfort D, Basangwa D, Kimbowa S, Firth J, Schuch F, Van Damme T, Mugisha J (2020) Test-retest reliability, validity, and correlates of the 2-min walk test in outpatients with depression. *Physiotherapy Research International* 25(2):e1821–e1821.
- Vicert (2021) Health Apps Usage Statistics — Vicert — vicert.com. <https://www.vicert.com/blog/health-apps-usage-statistics/>, [Accessed 16-08-2024].
- Wang B, Di Buccio E, Melucci M (2021a) Word2Fun: Modelling Words as Functions for Diachronic Word Representation. *Advances in Neural Information Processing Systems*.

- Wang T, Yang J, Li Y, Wang B (2021b) Partially interpretable estimators (pie): black-box-refined interpretable machine learning. *arXiv preprint arXiv:2105.02410* .
- Xie J, Chai Y, Liu X (2023) Unbox the Blackbox: Predict and Interpret YouTube Viewership Using Deep Learning. *Journal of Management Information Systems* 40(2):541–579.
- Xie J, Liu X, Zeng D, Fang X (2022) Understanding Medication Nonadherence from Social Media: A Sentiment-Enriched Deep Learning Approach. *MIS Quarterly* 46(1):341–372.
- Xie J, Zhang B, Ma J, Zeng D, Lo-Ciganic J (2021a) Readmission prediction for patients with heterogeneous medical history: A trajectory-based deep learning approach. *ACM Transactions on Management Information Systems (TMIS)* 13(2):1–27.
- Xie J, Zhang Z, Liu X, Zeng D (2021b) Unveiling the hidden truth of drug addiction: a social media approach using similarity network-based deep learning. *Journal of Management Information Systems* 38(1):166–195.
- Xu D, Ruan C, Körpeoglu E, Kumar S, Achan K (2020) Inductive representation learning on temporal graphs. *8th International Conference on Learning Representations, ICLR 2020*.
- Yu S, Chai Y, Chen H, Sherman SJ, Brown RA (2022) Wearable Sensor-based Chronic Condition Severity Assessment: An Adversarial Attention-based Deep Multisource Multitask Learning Approach. *MIS Quarterly* Forthcoming.
- Zhang D, Zhou L, Tao J, Zhu T, Gao G (2024) Ketch: A knowledge-enhanced transformer-based approach to suicidal ideation detection from social media content. *Information Systems Research* .
- Zhang H, Deng K, Li H, Albin RL, Guan Y (2020a) Deep Learning Identifies Digital Biomarkers for Self-Reported Parkinson’s Disease. *Patterns* 1(3):100042.
- Zhang W, 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(1).
- Zhang X, Gao Y, Lin J, Lu CT (2020b) Tapnet: Multivariate time series classification with attentional prototypical network. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Zhu H, Samtani S, Brown RA, Chen H (2021) A deep learning approach for recognizing activity of daily living (adl) for senior care: Exploiting interaction dependency and temporal patterns. *MIS Quarterly* 45(2):859–896.
- Zhu H, Samtani S, Chen H, Nunamaker Jr JF (2020) Human identification for activities of daily living: A deep transfer learning approach. *Journal of Management Information Systems* 37(2):457–483.

Supplementary Materials

Care for the Mind Amid Chronic Diseases: An Interpretable AI Approach Using IoT

Jiaheng Xie^{1,*,**}, Xiaohang Zhao^{2,*,**}, Xiang Liu¹, Xiao Fang¹

¹ Lerner College of Business and Economics,
University of Delaware, Newark, DE, USA

² School of Information Management & Engineering,
Shanghai University of Finance and Economics, Shanghai, China

* Corresponding Author: Xiaohang Zhao, xiaohangzhao@mail.shufe.edu.cn

** Equal Contribution

A. Recent Health Sensing Studies

Table A.1 shows the recent health sensing studies related to our study. Using walking sensor data as the input, existing studies have applied machine and deep learning methods, such as convolutional neural networks (CNNs), support vector machines (SVMs), and random forests (RFs), to detect the occurrence and severity of chronic diseases.

Table A.1 Summary of Recent Health Sensing Studies

Study	Data	Number of Subjects	Task	Model
Anand and Stepp (2015)	Walking tests	25	PD detection	Regression, NB, RF
Um et al. (2017)	Walking tests	30	PD motor detection	CNN
Millor et al. (2017)	Walking tests	431	Frailty prediction	Decision tree
Watanabe et al. (2017)	Walking tests	12	Diabetes forefoot load detection	Statistical analysis
Polat (2019)	Walking tests	16	PD prediction	Regression
Coelln et al. (2019)	Walking tests	683	PD prediction	Cox
Rastegari et al. (2019)	Walking tests	43	PD diagnosis	SVM, RF, NB, AdaBoost
Nemati et al. (2020)	Cough and speech	21	Lung disease prediction	Regression, SVM, RF, MLP
Moon et al. (2020)	Walking tests	524	PD prediction	NN, SVM, KNN, decision tree, RF
Piau et al. (2020)	Walking tests	125	Fall detection	Regression

B. Recent Prototype Learning Studies

Table A.2 shows the recent prototype learning studies. Prototype learning has been extended to interpret text classification. In this vein, Ming et al. (2019) propose ProSeNet, which adds the

prototype layer after the sequence encoder (e.g., RNN). This model is able to predict the class of a sentence (e.g., positive or negative) and explain which part of the sentence (prototype) leads to such a prediction result. A number of prototype learning models have also been proposed for various tasks. For example, Rymarczyk et al. (2021) develop ProtoPShare that captures sharing property between each pair of prototypes. Nauta et al. (2021) combine decision trees and prototype learning so that the prototype reasoning process can be streamlined as a tree structure. Singh and Yow (2021) design two groups of prototypes: one group that the input looks like and the other group that the input does not look like.

Table A.2 Existing Prototype Learning Methods v.s. Our Method

Study	Novelty	Input	TP*
Chen et al. (2019)	Prototype for image classification	An image	No
Hase et al. (2019)	Hierarchical prototype	An image	No
Ming et al. (2019)	Prototype for text classification	A piece of text	No
Xu et al. (2020)	Represent attribute for zero-shot learning	An image	No
Shitole et al. (2021)	Attention maps to explain a classifier	An image	No
Rymarczyk et al. (2021)	Prototype parts share	An image	No
Nauta et al. (2021a)	Prototype and decision tree	An image	No
Wang et al. (2021)	Add embedding space using manifold	An image	No
Singh and Yow (2021)	Positive reasoning and negative reasoning	An image	No
Nauta et al. (2021b)	Generate textual info about prototypes	An image	No
Our Method	Capture temporal progression of the input	A sequence of walking segments	Yes

* TP stands for “Temporal Progression”, indicating whether a model is capable of detecting interpretable temporal progression patterns from its input and then leveraging these patterns for prediction and interpretation. Depression symptoms exhibit a temporal progression, such as in Figure 1.

C. MTSC Studies

Table A.3 shows the MTSC studies. The distance-based models usually use 1-nearest neighbor coupled with a bespoke distance function. Different from cross-sectional data, the distance between multivariate time series can be computed using dynamic time warping (DTW) (Shokoohi-Yekta et al. 2017). Shapelets are discriminatory sub-series that have practical meaning. The shapelets-based models use selected shapelets using random forests (Karlsson et al. 2016). For the histogram-based models, words in the form of unigrams and bigrams are extracted for all time series and dimensions using a sliding window for a range of window lengths. The words for each dimension and window length are concatenated into a single bag of words histogram for a series (Schäfer and Leser 2017b). Interval summarizing-based models, such as the Canonical Interval Forest is an ensemble of time series trees built using the Canonical Time-Series Characteristics features and simple summary statistics extracted from phase dependant intervals (Middlehurst et al. 2020). For deep learning-based models, various time series encoders, such as ResNet (Wang et al. 2017) and

Table A.3 Existing Multivariate Time Series Classification Methods v.s. Our Method

Study	Category	Input	Interpretable
Karlsson et al. (2016)	Shapelets	A multivariate time series of UCR data	No
Shokoohi-Yekta et al. (2017)	Distance measures	A multivariate time series of cricket umpire signals	No
Schäfer and Leser (2017a)	Histogram	A multivariate time series of UCR data	No
Schäfer and Leser (2017b)	Histogram	A multivariate time series	No
Bagnall et al. (2020)	Distance measures	A multivariate time series of UCR data	No
Middlehurst et al. (2020)	Interval summarising	A multivariate time series of UCR data	No
Dempster et al. (2020)	Interval summarising	A multivariate time series of UCR data	No
Fawaz et al. (2020)	Deep learning	A multivariate time series of synthetic data	No
Our Method	Deep learning	A time series of irregularly spaced walking segments. Each is a multivariate time series	Yes

InceptionTime (Fawaz et al. 2020), are adopted to represent the multivariate time series data. Then, multiple layers of deep learning models can be deployed for classification.

D. Prototype Learning for MTSC Methods

Table A.4 shows the prototype learning for MTSC studies. Although Ming et al. (2019) does not directly tackle the MTSC problem, its text sequence model can be adapted to process MTSC-based prototype learning problems. The learned prototype is a sentence. Ma et al. (2020) apply prototype learning to interpret the MTSC classification of vital signs. The time series vital signs are processed as an image. Multiple CNN layers and a prototype layer are used to predict Myocardial infarction. The learned prototypes are a segment of vital signs. Zhang et al. (2020c) devise TapNet for MTSC problems with high dimensionality and limited training data issues. TapNet leverages a low-dimensional feature extractor to reduce the dimension from the multivariate time series. To address the limited training data issue, the authors propose a Random Dimension Permutation as a data augmentation mechanism. As multiple data sources are concatenated and a black-box LSTM layer is deployed in the feature extractor, the prototype cannot be traced back to a local region of the input, thus hindering the model’s interpretability. Consequently, Zhang et al. (2020c) only focus on prediction. Ghosal and Abbasi-Asl (2021) develop a framework for interpretable MTSC. The multivariate time series is first separated into multiple univariate time series. For each variable, an LSTM encoder extracts a representation for its time series. A prototype layer is stacked next to learn typical patterns from each variable independently. In the end, the prototype similarities from each variable are concatenated to make the classification. In the interpretation phase, prototypes of each variable are shown independently, which are a segment of the univariate time series. Trinh et al. (2021) propose DPNet to detect deep fake videos. Videos are a special format of multivariate time series where the dimensions are the image channels. An encoder represents each video as a single

Table A.4 Existing Prototype Learning for MTSC Methods v.s. Our Method

Study	Input	Prototype	Temporal Progression of Prototype
Gee et al. (2019)	A multivariate time series of ECG	A segment of ECG signal	No
Ming et al. (2019)	A sentence	A sentence	No
Ma et al. (2020)	A multivariate time series of vital signs	A segment of vital sign signal	No
Zhang et al. (2020c)	A multivariate time series of ECG	A hidden embedding	No
Ghosal and Abbasi-Asl (2021)	A four-dimensional time series of simulated data	A segment of a univariate series	No
Trinh et al. (2021)	Videos	A clip of a video	No
Our Method	A time series of irregularly spaced multivariate time series (i.e., walking segments)	1) A region of sensor signal (symptom); 2) progression of symptom (trend)	Yes

tensor, which is compared with the prototypes. These prototypes are embeddings of typical deep fake videos. Unlike most other prototype learning studies (model-based interpretation), DPNet’s interpretation is independent of its prediction process, conducted in the post-training phase (post-hoc interpretation). The authors use Timed Quality Temporal Logic (TQTL), which is an induction method. Using the TQTL, the authors pick a clip of a video that resembles most of a fake video.

E. Deep Learning Models Considering Time Irregularity

Table A.5 shows the deep learning models considering time irregularity. Their mechanisms of incorporating time fall into three approaches. The first approach utilizes a continuous time function to model the time series data, so that the irregularly spaced temporal input can be implicitly considered (Li et al. 2020). An example of this category is Li et al. (2020) who design an ordinary differential equation to model the temporal walking physical symptoms of Parkinson’s disease. The second approach modifies the sequence feature extraction model (e.g., LSTM) and adds the time interval between consecutive states in the cell state (Baytas et al. 2017, Zhang et al. 2020a, Gao et al. 2019). Baytas et al. (2017) use this approach to predict patient subtyping using EHR. The third approach proposes new temporal embeddings and adds them as additional features to the model (Li et al. 2019, Liu et al. 2018, Mei et al. 2022). For instance, Mei et al. (2022) devise a time-varying embedding that is sensitive to time changes.

F. The Generative Process of S Characterized by Trend Prototype k

Algorithm 1 shows the generative process of S characterized by trend prototype k .

G. The Definition of Sensor Features

We explain the content of a sensor feature as mentioned in Section 3.1. At each timepoint l , the mobile sensor collects accelerometer readings $[x_l^{\mathcal{E}}, y_l^{\mathcal{E}}, z_l^{\mathcal{E}}]$ and orientation readings $[x_l^o, y_l^o, z_l^o, w_l^o]$. A graphic illustration of these sensor readings is shown in Figure A.1.

Table A.5 Deep Learning Models Considering Time Irregularity

Study	Context	Mechanism of Incorporating Time	Purpose	Interpretable
Li et al. (2019)	Times series prediction	Transfer different attention weights to different timesteps	Diversify the attention for time series prediction	No
Gao et al. (2019)	Cancer detection	Consider time intervals in LSTM	Encode irregular timepoint input to prediction	No
Li et al. (2020)	Parkinson's detection	Ordinary differential equations to model time series	Encode time dimension to prediction	No
Baytas et al. (2017)	Patient subtyping	Add time interval in LSTM cell state	Encode irregular timepoint input to prediction	No
Liu et al. (2018)	Temperature prediction	Design closeness, period, and trend as additional features	Encode temporal input to prediction	No
Mei et al. (2022)	Time series prediction	Design time-varying embedding as additional features	Encode time dimension to prediction	No
Zhang et al. (2020a)	Health state detection	Add time interval in LSTM cell state	Encode irregular timepoint input to prediction	No
Ours	Depression detection	Design continuous temporal prototypes	Model a time series of irregularly spaced multivariate time series to improve interpretation	Yes

Algorithm 1 The Generative Process of S Characterized by Trend Prototype k

- 1: Compute $t_0^{(k)}$ via Equation 10.
- 2: Compute $t_i^{(k)} = t_i - t_0^{(k)}$ for $i = 1, 2, \dots, N$.
- 3: Compute $\tilde{\mathcal{G}}^{(k)}(t_i^{(k)})$ for $i = 1, 2, \dots, N$ using Equation 7.
- 4: Draw $S_i \sim \mathcal{LN}(\tilde{\mathcal{G}}^{(k)}(t_i^{(k)}), I)$ for $i = 1, 2, \dots, N$.

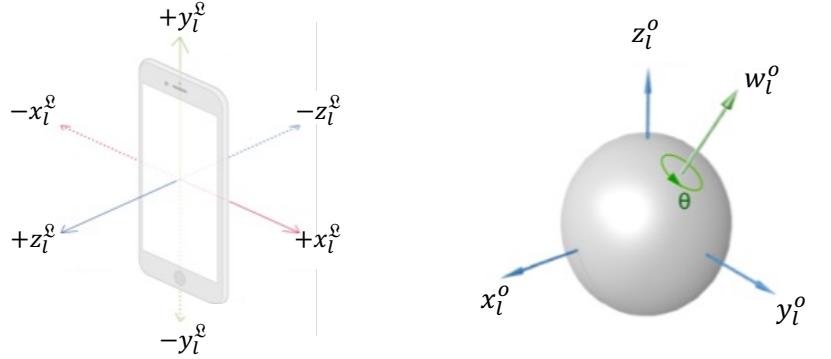
The accelerometer readings are in the local reference frame, which most existing studies rely on (Piau et al. 2020, Yu et al. 2022, Zhu et al. 2021). However, these local reference readings do not reflect the precise walking pattern in the geographic coordinate system, because it moves and rotates along with the mobile device. To address this issue and make the interpretation more meaningful in the geographic sense, we decide to work with the readings in the global reference frame. We transform the local reference frame accelerometer vector $v_l^{\mathfrak{L}} = [x_l^{\mathfrak{L}}, y_l^{\mathfrak{L}}, z_l^{\mathfrak{L}}]^T$ to the global reference frame as $v_l^{\mathfrak{G}} = [x_l^{\mathfrak{G}}, y_l^{\mathfrak{G}}, z_l^{\mathfrak{G}}]^T$ via the quaternion rotation

$$v_l^{\mathfrak{G}} = \mathcal{R}_l v_l^{\mathfrak{L}}$$

where \mathcal{R}_l is the rotation matrix derived from quaternion $[x_l^o, y_l^o, z_l^o, w_l^o]$ as follows. In general, given a quaternion $[x, y, z, w]$, the corresponding rotation matrix \mathcal{R} is defined as¹

$$\mathcal{R} = \begin{bmatrix} w^2 + x^2 - y^2 - z^2 & 2xy - 2wz & 2xz + 2wy \\ 2xy + 2wz & w^2 - x^2 + y^2 - z^2 & 2yz - 2wx \\ 2xz - 2wy & 2yz + 2wx & w^2 - x^2 - y^2 + z^2 \end{bmatrix}. \quad (\text{EC.1})$$

¹https://en.wikipedia.org/wiki/Quaternions_and_spatial_rotation



Left: $[x_l^g, y_l^g, z_l^g]$ measures the acceleration readings along the x, y, z axes in the local reference frame.

Right: $[x_l^o, y_l^o, z_l^o, w_l^o]$ measures the movement and rotation of the local reference frame relative to the global reference frame. The local reference frame is fixed to the mobile device, and moves and rotates along with the device. The local reference frame is the coordinate system of the mobile device. The axis of local reference frame changes relative to the earth when the device's orientation changes. The global reference frame is the coordinate system when the device is placed horizontally and the device's x axis points to magnetic north. Therefore, the global reference frame is fixed to the earth regardless the device moves or not.

Figure A.1 Sensor Readings (Apple Inc 2022)

Then, we define the sensor feature a_l as

$$a_l = [x_l^g, y_l^g, z_l^g]^T. \quad (\text{EC.2})$$

Following the best practice of data augmentation for mobile sensor data (Um et al. 2017, Zhang et al. 2020b), during the training stage, we further transform the input sensor features with random rotations to improve the generalization ability of our model. To do this, we first sample a quaternion using Algorithm 2 where $\text{Uniform}(b_1, b_2)$ means the uniform distribution on the interval $[b_1, b_2]$, and then plug the obtained quaternion into Equation EC.1 to construct a random rotation matrix. Within each training epoch and for each walking test, we construct a random rotation matrix $\tilde{\mathcal{R}}$ as described, and then use it to transform all sensor features of the walking test as $\langle \tilde{\mathcal{R}}a_1, \tilde{\mathcal{R}}a_2, \dots, \tilde{\mathcal{R}}a_L \rangle$. Once the training stage is finished, we use the original sensor features $\langle a_1, a_2, \dots, a_L \rangle$ to do inference.

Algorithm 2 Sample a Quaternion

- 1: Draw $x \sim \text{Uniform}(0, 1)$, $y \sim \text{Uniform}(0, 1)$, $z \sim \text{Uniform}(0, 1)$.
 - 2: Let $\text{norm} = \sqrt{x^2 + y^2 + z^2}$, and set $x = x/\text{norm}$, $y = y/\text{norm}$, $z = z/\text{norm}$.
 - 3: Draw $\theta \sim \text{Uniform}(0, 2\pi)$.
 - 4: Set $w = \cos(\theta/2)$, $x = x \sin(\theta/2)$, $y = y \sin(\theta/2)$, $z = z \sin(\theta/2)$
 - 5: Return $[x, y, z, w]$
-

H. The Density Function of the Logistic-Normal Distribution

We employ the change of variables formula (Murphy 2022) to derive Equation 8, which in general states that if a random vector $x \in R^M$ is mapped to another random vector $z \in R^M$ by an invertible function f , i.e., $z = f(x)$, then the density function of z , denoted by $p_z(z)$, is related to the density function of x , denoted by $p_x(x)$, by the following relationship:

$$p_z(z) = p_x(g(z)) |\det[J_g(z)]| \quad (\text{EC.3})$$

where g is the inverse function of f , $J_g(z) \in R^{M \times M}$ is the Jacobian matrix of g evaluated at z , $\det[\cdot]$ is the matrix determinant operator, and $|\cdot|$ is the absolute value operator.

In our case, $z = \sigma(x)$, which means that $f = \sigma$ and $g = \sigma^{-1}$. Using the fact that $\partial x_m / \partial z_m = 1/(z_m(1-z_m))$, the corresponding Jacobian matrix can be computed as

$$J_g(z) = \begin{pmatrix} \frac{1}{z_1(1-z_1)} & 0 & \dots & 0 \\ 0 & \frac{1}{z_2(1-z_2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{z_M(1-z_M)} \end{pmatrix}, \quad (\text{EC.4})$$

which is a diagonal matrix because σ^{-1} has been element-wisely applied on z . Given that $0 < z_m < 1$ for $m = 1, 2, \dots, M$, all the diagonal elements are positive, and thus we have

$$|\det[J_g(z)]| = \frac{1}{\prod_{m=1}^M z_m(1-z_m)}. \quad (\text{EC.5})$$

Recall that $x \sim \mathcal{N}(\mu, \Sigma)$, then the corresponding density function is given by

$$p_x(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^M \det[\Sigma]}}. \quad (\text{EC.6})$$

Plugging Equation EC.5 and EC.6 into Equation EC.3, and setting $x = \sigma^{-1}(z)$, we obtain Equation 8.

I. NHANES Dataset

NHANES is designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations for a nationally representative sample of all ages. To produce reliable statistics, NHANES over-samples persons 60 and older, African Americans, and Hispanics. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.² These medical measurements offer accurate labels for a

²<https://www.cdc.gov/nchs/nhanes/index.htm>

variety of chronic diseases, such as high blood pressure, heart disease, diabetes, chronic kidney disease, asthma, arthritis, stroke, and cancer. Among the health measurements, depression is one of the prioritized diseases of NHANES. NHANES deploys PHQ-9 to collect patients' depression diagnoses, making it a clinically accurate and relevant dataset for depression research (Yu et al. 2020, Vallance et al. 2011). Because of such a wide range of chronic disease coverage as well as precise measurement of depression, this dataset is an ideal testbed for our study.

In order to invite research to examine the impact of physical activities on chronic diseases, NHANES added wearable sensor data collection in conjunction with depression and other chronic disease diagnoses in the 2013 cohort. For the wearable sensor data collection, participants were first mailed a sensor device and wore it for continuous seven days. Timestamps were captured to ensure the sensor data were indeed recorded during this study window. This is sufficiently long to depict the severity change of patients' depression, as the depression severity assessment window recommended by the American Psychiatric Association is seven days³. The device was the ActiGraph model GT3X+, which measured acceleration every 1/80th of a second (80 Hz). Because of such fine-grained wearable sensor data, accurate clinical depression diagnoses from PHQ-9, and the large and representative participant base, we choose the NHANES dataset for all the following empirical analyses.

To construct our input data, we utilize the accelerometer data. The data were collected continuously. However, not all segments are usable. Patients may not wear the device from time to time, take a shower, sleep, sit while working, and so on. These time segments are not walking data and are not applicable to our study. Therefore, we use the state-of-the-art walking activity detector to filter the walking segments (Czech and Patel 2019). We acknowledge that other activities other than walking, such as sleep quality and work productivity, may also indicate depression. Measuring these activities requires different sensors other than accelerometers. From the practical point of view, imposing such hardware and additional data collection requirements results in additional consensus from users and excludes a large portion of low-income users whose devices are not equipped with those types of sensors. Those device restrictions, hardware deficiency, increased cost, and consensus issues will eventually impede the successful implementation and equitability of the resulting detection model. Our walking-based detection, on the other hand, is the easiest to implement as such functions can be deployed to existing m-health apps where location consensus is already obtained and accelerometers are readily installed in most mobile devices.

³Severity Measure for Depression (American Psychiatric Association)

J. Benchmark Methods and Hyperparameter Settings

According to our literature review, we select three groups of benchmarks. The first group is black-box deep learning models, CNN and RNN. They have been commonly used in prior motion sensor-based predictions (Yu et al. 2022, Zhu et al. 2021). The second group uses manually crafted features, such as mean, variance, and standard deviation of sensor signals, as the input (Oung et al. 2015, Yu et al. 2022). These features are shown in Table A.6. The benchmarks in the second group and their features are in line with Yu et al. (2022), which includes k-nearest neighbors (KNN), support vector machine (SVM), random forest, AdaBoost, and XGBoost. The third group includes the state-of-the-art and the most widely recognized MTSC and prototype learning for MTSC models. From the MTSC studies, we select the most state-of-the-art study (Fawaz et al. 2020) as a benchmark, since it is also deep learning-based, thus being able to learn representation from raw sensor data. It also achieves better performance than other MTSC studies. The other MTSC studies are traditional machine learning-based which require feature engineering. However, Fawaz et al. (2020) is not interpretable. From prototype learning for MTSC studies, we select Ma et al. (2020), Gee et al. (2019), Ming et al. (2019), Chen et al. (2019) as benchmarks, because they are the most state-of-the-art, and their input data format is the closest to sensor data. These are the most related benchmarks to this study. Compared to our model, these benchmarks cannot model or interpret the temporal progressions of prototypes. The hyperparameters of the benchmarks are summarized in Table A.7. These hyperparameters are fine-tuned for each benchmark after large-scale experiments. The following evaluation results report the fine-tuned performances for the benchmarks.

To implement our model, we adopt the following hyperparameter setting. We set the embedding dimension of symptom prototypes as $n_e = 128$, the time embedding dimension as $n_d = 64$, the regularization weights as $\lambda_S = 0.1$ and $\lambda_T = 0.1$. We train our model with the Adam optimizer (Kingma and Ba 2015) with a learning rate of 0.001 and a batch size of 32. The specification of the CNN layer is as follows:

Recall that X_i denotes the sequence of sensor features of the i th walking segment performed by a focal patient. In this section, we focus on extracting the feature matrix H_i^S from a single walking segment, and therefore drop the subscript i to simplify the notation. In general, different walking segments have different lengths. In what follows, we assume that X has been downsampled with the frequency of 10 Hz, while the treatment of other sampling frequencies should be adjusted proportionally.

To facilitate batch training, we reshape each segment into a matrix of size 3×300 , by either padding it with zero columns if its length is smaller than 300, or discarding the extra columns if its length is larger than 300. After the reshaping step, we treat each segment X_j as an input of 3 channels and length 300, and then use the same CNN layer to extract a feature matrix of size 128×5

Table A.6 Features for Conventional Machine Learning Models

Feature Name	Formula
Mean x-axis values	$u_x = \frac{1}{L} \sum_{l=1}^L v_{x,l}$
Mean y-axis values	$u_y = \frac{1}{L} \sum_{l=1}^L v_{y,l}$
Mean z-axis values	$u_z = \frac{1}{L} \sum_{l=1}^L v_{z,l}$
St. D. of x-axis values	$\sigma_x = \sqrt{\frac{1}{L-1} \sum_{l=1}^L (v_{x,l} - u_x)^2}$
St. D. of y-axis values	$\sigma_y = \sqrt{\frac{1}{L-1} \sum_{l=1}^L (v_{y,l} - u_y)^2}$
St. D. of z-axis values	$\sigma_z = \sqrt{\frac{1}{L-1} \sum_{l=1}^L (v_{z,l} - u_z)^2}$
Mean magnitude	$u_v = \frac{1}{L} \sum_{l=1}^L \ v_l\ , \ v_l\ = \sqrt{v_{x,l}^2 + v_{y,l}^2 + v_{z,l}^2}$
St. D. of magnitude	$\sigma_v = \sqrt{\frac{1}{L-1} \sum_{l=1}^L (\ v_l\ - u_v)^2}$
Mean x-axis jerk	$\alpha_x = \frac{1}{L-1} \sum_{l=1}^{L-1} d_{x,l}, \text{ where } d_{x,l} = v_{x,l+1} - v_{x,l}$
Mean y-axis jerk	$\alpha_y = \frac{1}{L-1} \sum_{l=1}^{L-1} d_{y,l}, \text{ where } d_{y,l} = v_{y,l+1} - v_{y,l}$
Mean z-axis jerk	$\alpha_z = \frac{1}{L-1} \sum_{l=1}^{L-1} d_{z,l}, \text{ where } d_{z,l} = v_{z,l+1} - v_{z,l}$
St. D. of x-axis jerk	$\beta_x = \sqrt{\frac{1}{L-2} \sum_{l=1}^{L-1} (d_{x,l} - \alpha_x)^2}$
St. D. of y-axis jerk	$\beta_y = \sqrt{\frac{1}{L-2} \sum_{l=1}^{L-1} (d_{y,l} - \alpha_y)^2}$
St. D. of z-axis jerk	$\beta_z = \sqrt{\frac{1}{L-2} \sum_{l=1}^{L-1} (d_{z,l} - \alpha_z)^2}$
Mean jerk magnitude	$\alpha_d = \frac{1}{L-1} \sum_{l=1}^{L-1} \ d_l\ , \text{ where } \ d_l\ = \sqrt{d_{x,l}^2 + d_{y,l}^2 + d_{z,l}^2}$
St. D. of jerk magnitude	$\beta_d = \sqrt{\frac{1}{L-2} \sum_{l=1}^{L-1} (\ d_l\ - \alpha_d)^2}$
Stride time variability on x-axis	(1) Identify signal peaks in x-axis, $[t_1, t_2, \dots, t_Q]$; (2) Identify stride times $[dt_1, dt_2, \dots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_x = \sqrt{\frac{1}{Q-2} \sum_{i=1}^{Q-1} (dt_i - \bar{dt})^2}$
Stride time variability on y-axis	(1) Identify signal peaks in y-axis, $[t_1, t_2, \dots, t_Q]$; (2) Identify stride times $[dt_1, dt_2, \dots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_y = \sqrt{\frac{1}{Q-2} \sum_{i=1}^{Q-1} (dt_i - \bar{dt})^2}$
Stride time variability on z-axis	(1) Identify signal peaks in z-axis, $[t_1, t_2, \dots, t_Q]$; (2) Identify stride times $[dt_1, dt_2, \dots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_z = \sqrt{\frac{1}{Q-2} \sum_{i=1}^{Q-1} (dt_i - \bar{dt})^2}$
Stride time variability on magnitude	(1) Identify signal peaks in magnitude, $[t_1, t_2, \dots, t_Q]$; (2) Identify stride times $[dt_1, dt_2, \dots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$; (3) Compute stride time variability $V_v = \sqrt{\frac{1}{Q-2} \sum_{i=1}^{Q-1} (dt_i - \bar{dt})^2}$

Recall that given a walking segment, we observe a sequence of sensor signals $\langle v_1^\mathfrak{L}, v_2^\mathfrak{L}, \dots, v_L^\mathfrak{L} \rangle$, where $v_l^\mathfrak{L}$ is the accelerometer readings recorded in the local reference frame at timepoint l , as explained in Appendix G. To simplify notation, we drop the superscript \mathfrak{L} , and write $v_l = [v_{x,l}, v_{y,l}, v_{z,l}]^T$. Following Yu et al. (2022), we define the following features for each given walking segment, shown in Table A.6.

from X_j . The CNN layer is composed by a sequence of one-dimensional convolution (Conv1d) layers, each followed in order by a one-dimensional batch normalization layer (BatchNorm1d), a max pooling layer (MaxPool1d), and lastly a leaky ReLU layer (LeakyReLU) with slope 0.01 for non-linear activation (Goodfellow et al. 2016). Following the style of Zhu et al. (2021), we report the detailed architecture of the CNN layer in Table A.8. Since the BatchNorm1d layer and the

Table A.7 Benchmark Hyperparameter Settings

Model	Parameter	Values
TempPNet	CNN channels	(256, 512, 256, 128)
	CNN kernel sizes	8×1
	Time encoding dimensions	64
	Number of neighbors	5
KNN	Regularization parameter	1
SVM	Kernel coefficient	0.001
Random Forest	Number of estimators	100
AdaBoost	Number of estimators	50
XGBoost	Number of estimators	100
	Minimum loss reduction for further partition	1.5
	Subsample	0.6
ProtoPNet	CNN channels	(32, 64, 128, 256)
	CNN kernel sizes	3×3
ProSeNet	GRU hidden units	64

LeakyReLU layer do not change the input shape, we do not list them in Table A.8. After the last Conv1d layer in Table A.8, we do not add the MaxPool1d layer nor the LeakyReLU layer.

Table A.8 The Specification of the CNN Layer

	Kernel Size	Stride	Output Channel	Output Shape
Conv1d	8	1	256	(256, 293)
MaxPool1d	2	2		(256, 146)
Conv1d	8	1	512	(512, 139)
MaxPool1d	2	2		(512, 69)
Conv1d	8	1	256	(256, 62)
MaxPool1d	2	2		(256, 31)
Conv1d	8	1	128	(128, 24)
MaxPool1d	2	2		(128, 12)
Conv1d	8	1	128	(128, 5)

K. Evaluations Conditioned on Pre-existing Chronic Disease Severity (NHANES)

Since our dataset contains numerous pre-existing chronic diseases, we select two of them (diabetes and kidney disease) to showcase our model's performances when conditioned on specific disease severities, reported in Tables A.9 and A.10. Diabetes severity is determined based on HgVA1c levels (Care 2018, King and Xiang 2019). Kidney disease severity is based on KIQ scores⁴.

L. Second Dataset (mPower) Results

L.1. Data Collection and Preprocessing

For generalizability considerations, we have obtained the second dataset: mPower, a smartphone-based study that collects daily motion sensor signals for chronic disease patients (Bot et al. 2016).

⁴https://www.cdc.gov/Nchs/Nhanes/2013-2014/KIQ_U_H.htm

Table A.9 Evaluations Conditioned on Diabetes Severity (NHANES)

HgbA1c	Diabetes Severity	F1-score	Precision	Recall
< 5.7	No diabetes	0.770 ± 0.022	0.750 ± 0.040	0.793 ± 0.016
5.7 – 6.4	Pre-diabetes	0.796 ± 0.022	0.792 ± 0.046	0.802 ± 0.020
6.5 – 9	Diabetes	0.762 ± 0.047	0.734 ± 0.054	0.795 ± 0.068
> 9	Severe diabetes	0.789 ± 0.091	0.801 ± 0.025	0.797 ± 0.181

Table A.10 Evaluations Conditioned on Kidney Disease Severity (NHANES)

KIQ Score	F1-score	Precision	Recall
1 – 2	0.763 ± 0.040	0.757 ± 0.081	0.773 ± 0.023
3 – 6	0.809 ± 0.037	0.807 ± 0.061	0.813 ± 0.018

To acquire the depression label, we leverage the MDS-UPDRS survey from this dataset. The MDS-UPDRS survey is originally used to evaluate Parkinson’s disease severity. Part of its questions overlaps with the PHQ-9 depression assessment questionnaire. We select those overlapped questions to measure depression status, including MDS-UPDRS 1.3-1.5 and 1.7-1.8, whose total score is 20. In clinical practice, patients with a PHQ-9 score over 4 (total score is 27) are diagnosed as depressed ([Patient 2022](#)). Similarly, we label patients whose MDS-UPDRS score is over $20 \times 4 / 27 \approx 3$ as depressed and the remaining as non-depressed. Since the MDS-UPDRS is a crude depression screening measure, this is to predict depressed mood rather than diagnose depression. Our data usage (depression detection using MDS-UPDRS and sensor) has been approved by Synapse and our institute’s IRB. The walking tests in the mPower dataset were done individually and unsupervised at home. The participants downloaded an app on their mobile phones. The app gives participants instructions to walk. The app automatically records the walking data. Only the mobile phone was used. No other equipment was necessary. This test setting is the norm in sensor-based disease monitoring, as is widely adopted by the health sensing studies in Table A.1.

To construct our input data, we utilize the accelerometer data from the mPower dataset. These data are collected from walking tests – each test is composed of walking 20 steps in a straight line (outbound), turning around and standing for 30 seconds (rest), and walking 20 steps back (return). In the walking tests, the accelerometer records a tri-axial acceleration reading sampled at a frequency of 100 Hz. To reduce noise and prevent overfitting, we follow the standard sensor data preprocessing technique ([Sigcha et al. 2020](#)) to resample the readings at a frequency of 10 Hz. For each patient, we select a window of two weeks and utilize the accelerometer data in this time window as the sensor input for this patient. Unlike the 7-day time window in the NHANES analysis which is recommended by American Psychiatric Association as the depression severity assessment window, we choose the two-week time window in the mPower dataset, because the walking tests are conducted voluntarily by users and thus not as dense as the continuously recorded

NHANES dataset. We use a relatively longer time window to include sufficient walking tests for model training. This time window is also not too long to be irrelevant to the current depression status. We only select the patients with at least one chronic disease based on their answers to two questions in the demographic survey: “Have you been diagnosed by a medical professional with Parkinson disease?” and “Has a doctor ever told you that you have any of the following conditions? Please check all that apply.” Among them, we also remove the patients who did not participate in the walking experiments (no walking sensor data). In the end, we generated a dataset of 3,154 walking tests, encompassing 916 chronic disease patients (496 depressed and 420 non-depressed). Each walking test includes a sequence of motion sensor readings. Due to the complexity and high budget of sensor data collection, our data size is in line with or larger than most sensor-based disease prediction studies (Zhu et al. 2021, Jacobson and Chung 2020, Farhan et al. 2016, Moon et al. 2020, Coelln et al. 2019). We split this dataset into 60% for training, 20% for validation, and 20% for test.

L.2. Depression Prediction Evaluation

We first compare with the commonly used machine and deep learning models in sensing studies. Compared to the best deep learning model (RNN), our model increases F1-score by 0.074. This increase is attributed to our model’s capability of capturing temporal symptom progression and depression symptoms. Compared to the leading feature-based ML model (XGBoost), TempPNet boosts F1-score by 0.113. This performance enhancement is due to our model’s ability to learn effective features from the raw sensor signal.

Table A.11 Prediction Performance Comparison with Machine and Deep Learning Methods (mPower)

Model	Input	Interpretable	F1-score	Precision	Recall
TempPNet (Ours)	Raw sensor	Yes	0.774 ± 0.019	0.805 ± 0.035	0.746 ± 0.039
CNN	Raw sensor	No	0.689 ± 0.015	0.594 ± 0.030	0.821 ± 0.047
RNN	Raw sensor	No	0.700 ± 0.017	0.630 ± 0.041	0.790 ± 0.033
KNN	Features	No	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000
SVM	Features	No	0.627 ± 0.000	0.512 ± 0.000	0.808 ± 0.000
Random forest	Features	No	0.577 ± 0.041	0.572 ± 0.097	0.610 ± 0.119
AdaBoost	Features	No	0.615 ± 0.000	0.500 ± 0.000	0.800 ± 0.000
XGBoost	Features	No	0.661 ± 0.000	0.580 ± 0.000	0.769 ± 0.000

Compared to regular MTSC models without temporal progressions of prototypes (Fawaz et al. 2020), our model increases F1-score by 0.078. This result proves that capturing the prototypes and their temporal progressions assists in prediction performance. Compared to the best-performing prototype learning for MTSC model (Chen et al. 2019), TempPNet improves F1-score by 0.058. Such a significant performance gain indicates that capturing the temporal progressions of prototypes greatly contributes to depression prediction.

Table A.12 Prediction Performance Comparison with MTSC and Prototype Learning for MTSC (mPower)

Model	Interpretable	Progression of Prototype	F1-score	Precision	Recall
TempPNet (Ours)	Yes	Yes	0.774 ± 0.019	0.805 ± 0.035	0.746 ± 0.039
Chen et al. (2019)	Yes	No	0.716 ± 0.015	0.694 ± 0.030	0.741 ± 0.047
Ming et al. (2019)	Yes	No	0.701 ± 0.017	0.630 ± 0.041	0.790 ± 0.033
Gee et al. (2019)	Yes	No	0.683 ± 0.016	0.577 ± 0.030	0.836 ± 0.053
Ma et al. (2020)	Yes	No	0.704 ± 0.023	0.628 ± 0.077	0.801 ± 0.091
Fawaz et al. (2020)	No	No	0.696 ± 0.011	0.730 ± 0.018	0.737 ± 0.018

Since our model consists of multiple critical design components, we further perform ablation studies to show their effectiveness, as reported in Table A.13. We first remove the latent trend starting time design ($t_0^{(k)}$). We also remove the trend prototype design. After removing the trend prototype, the model loses the capability of detecting temporal symptom progression. Consequently, we test two options: using the last symptom severity to predict depression and using the average symptom severity over time to predict depression. Table A.13 suggests that removing any design component will significantly hamper the prediction accuracy, proving that our design choice is optimal.

Table A.13 Ablation Studies (mPower)

Model	F1-score	Precision	Recall
TempPNet (Ours)	0.774 ± 0.019	0.805 ± 0.035	0.746 ± 0.039
TempPNet removing offset $t_0^{(k)}$	0.726 ± 0.031	0.678 ± 0.101	0.816 ± 0.106
Remove trend prototype using last symptom severity	0.741 ± 0.012	0.712 ± 0.032	0.775 ± 0.041
Remove trend prototype using average symptom severity	0.744 ± 0.015	0.729 ± 0.035	0.763 ± 0.033

As our model takes the sensor data from an observation window as the input, we analyze how the length of the observation window influences the prediction accuracy. We show the results of the 2-week, 4-week, 8-week, and 16-week observation windows in Table A.14. Beyond two weeks, patients may have depressive and non-depressive episodes from time to time. Thus, noisy observations arise. Therefore, we use the 2-week observation window for all the other analyses.

Table A.14 Analysis of Observation Window (Signal Frequency = 10 Hz; mPower)

Observation Window	F1-score	Precision	Recall
2 weeks	0.774 ± 0.019	0.805 ± 0.035	0.746 ± 0.039
4 weeks	0.738 ± 0.021	0.708 ± 0.029	0.772 ± 0.039
8 weeks	0.730 ± 0.027	0.683 ± 0.084	0.810 ± 0.103
16 weeks	0.721 ± 0.031	0.651 ± 0.077	0.828 ± 0.087

To reduce noise in the sensor data and avoid overfitting, sensor-based prediction studies usually downsample the sensor signals (Sigcha et al. 2020). We test the effect of different sample rates in

Table A.15: 10 Hz, 20 Hz, and 30 Hz. The results suggest that 10 Hz signal frequency achieves the best performance. Therefore, we use the 10 Hz signal frequency for all the other analyses.

Table A.15 Analysis of Signal Frequency (mPower)			
Signal Frequency	F1-score	Precision	Recall
10 Hz	0.774 ± 0.019	0.805 ± 0.035	0.746 ± 0.039
20 Hz	0.752 ± 0.034	0.674 ± 0.075	0.866 ± 0.062
30 Hz	0.725 ± 0.034	0.631 ± 0.086	0.877 ± 0.080

PD and depression have a certain correlation because they share similar walking symptoms. To make sure that our model is actually predicting depression instead of PD severity, we perform evaluations that are conditioned on the PD severity. We divide the patients into groups based on their PD severity score (the summation of the MDS-UPDRS questions (Goetz et al. 2008)). Conditioned on each PD severity score, we report our model’s depression prediction performance. To make sure there is sufficient data points in a group to train and test our model, we only select the groups where there are at least 20 patients. This is also in line with the health sensing studies in Table A.1, where most studies have more than 20 subjects. Groups smaller than that do not have enough statistical power, thus inappropriate to perform reliable evaluations. If our model indeed predicts depression, we expect that, conditioned on each PD severity score, our model’s performance should remain consistently high. If our model only predicts PD severity, conditioned on a PD severity score, the performance should be very low because in this group the model has not seen different values of the outcome, thus unable to update parameters well. Table A.16’s results prove that given any PD severity score, our model is able to accurately predict depression consistently. Therefore, our model indeed predicts depression rather than PD severity.

Table A.16 Evaluations Conditioned on PD Severity (mPower)

PD Severity Score	F1-score	Precision	Recall
6	0.796 ± 0.042	0.944 ± 0.027	0.692 ± 0.064
7	0.768 ± 0.095	0.729 ± 0.112	0.817 ± 0.084
8	0.774 ± 0.045	0.772 ± 0.032	0.778 ± 0.060
9	0.837 ± 0.044	0.936 ± 0.036	0.757 ± 0.050
10	0.814 ± 0.098	0.780 ± 0.116	0.853 ± 0.076
11	0.752 ± 0.068	0.637 ± 0.042	0.917 ± 0.126
12	0.848 ± 0.035	0.779 ± 0.012	0.932 ± 0.066
13	0.844 ± 0.038	0.748 ± 0.030	0.969 ± 0.056
14	0.779 ± 0.027	0.858 ± 0.037	0.713 ± 0.035
15	0.780 ± 0.060	0.848 ± 0.034	0.722 ± 0.094
16	0.810 ± 0.062	0.931 ± 0.108	0.720 ± 0.046
17	0.845 ± 0.137	0.879 ± 0.085	0.821 ± 0.180
18	0.743 ± 0.098	0.686 ± 0.092	0.812 ± 0.114

L.3. Interpretation of Depression Prediction

Beyond depression prediction, TempPNet is capable of interpreting why a patient is classified as depressed by presenting the contributing temporal symptom progression (trend prototype) and the corresponding walking symptom (symptom prototype). Figure A.2 shows the most salient trend prototypes that our model learned. These trend prototypes are the prototypical depression or non-depression trend. For each picture, the x-axis is time, and the y-axis is symptom severity.

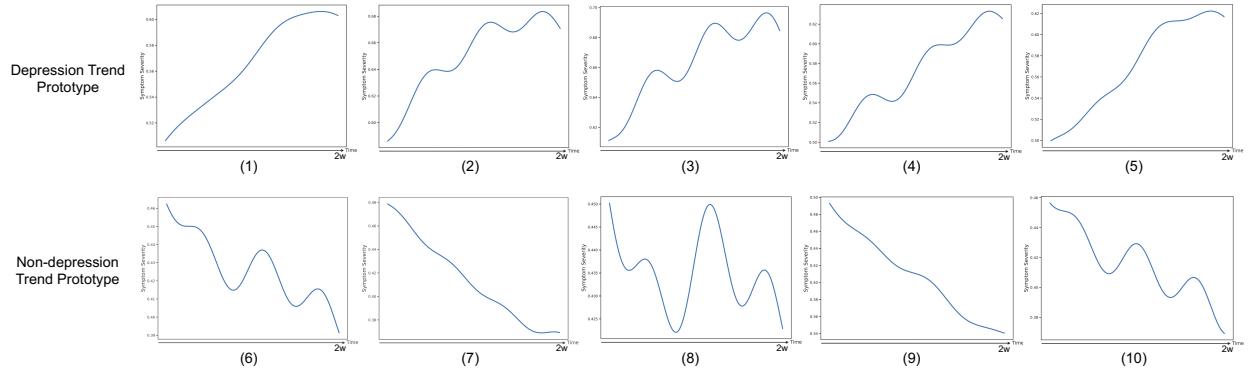


Figure A.2 Trend Prototypes

Trend prototypes (1)-(5) are depression trend prototypes. They represent the severities of depression symptoms trending up. Some of them have deviations from time to time in the upward trend, such as (2)-(4), representing temporary symptom relief and deterioration of depression. This conforms with the typical depression trend (Bockting et al. 2015). Trend prototypes (6)-(10) are non-depression trend prototypes, where (6), (7), (9), and (10) represent trending down and (8) represents fluctuating with no trend. These are not typical depression trends. Each trend prototype is coupled with underlying symptoms. Figure A.3 shows the symptom prototypes that our model learned. The trend prototypes are learned using all the patients' data rather than relying on a single patient's data. Each patient's observed walking test could be at different stages of a trend — some at the rising stage, some at the stable stage, among others. Together they depict a complete trend. Multiple patients could also share the same trend prototype if their symptom severity levels are at the same stage (e.g., all on the rise).

The prototype visualization in Figure A.3 shows the sensor signal of the symptoms. These symptom prototypes are learned by our method across all patients. Each patient may present none, one, or more of these symptoms in their walking patterns. Prior literature suggests that depression walking symptoms can be reflected in gait features, such as walking speed, stride, and speed of the lifting motion of the leg (Sloman et al. 1982, Lemke et al. 2000, NHS 2022). When interpreting the prediction of a patient, these gait features can be computed for the symptom prototypes.

Symptom Prototype	(1)	(2)	(3)	(4)	(5)
Prototype Visualization					
Walking Speed	0.24	0.24	0.42	0.41	0.24
Stride	0.50	0.42	0.54	0.49	0.44
Speed of Lifting Motion of the Leg	0.83	0.79	2.00	2.30	0.85

Note: Each of these symptom prototype represents a typical depression or non-depression walking pattern. Many gait features can be derived from a walking signal. Take three features that are frequently referenced in the depression literature as an example: walking speed, stride, and lifting. In our study population, the average walking speed is 0.31. The average stride is 0.47. The average lifting motion is 1.38. For example, symptom prototype 1 shows much slower walk than usual and much slower lifting motion than usual. This is indicative of depression. Symptom prototype 4 shows much faster walk than usual, normal stride, and much faster lifting motion. This is indicative of non-depression.

Figure A.3 Symptom Prototypes

Leveraging the above-learned trend prototypes and symptom prototypes, our model can interpret the prediction of depression for every patient. We randomly select two patients (one depressed and one non-depressed) and showcase TempPNet’s interpretation for them. Figure A.4 shows the interpretation of the depressed patient, and Figure A.5 shows the interpretation of the non-depressed patient. For simplicity, we only show the trend prototype with the highest existing strength and the corresponding symptom prototype with the highest existing strength in these examples. For the symptom prototype, we also compute the gait features using the GaitPy package⁵ to explain the encoded information from the visualization. The arrows after the gait features denote whether a feature is higher or lower than an average human. They do not imply any trend information (neither go up nor go down).

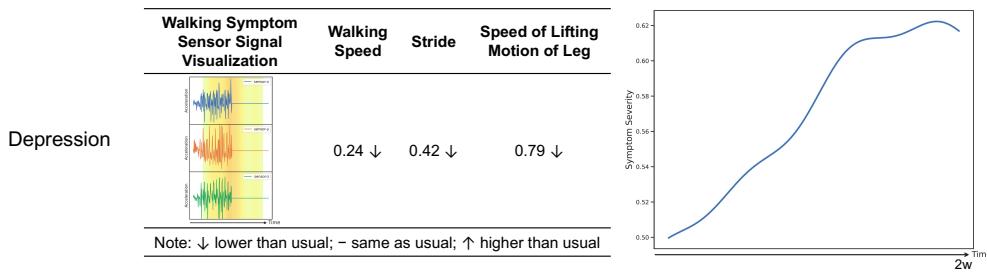


Figure A.4 Interpretation of A Depressed Patient

⁵<https://pypi.org/project/gaitpy/>

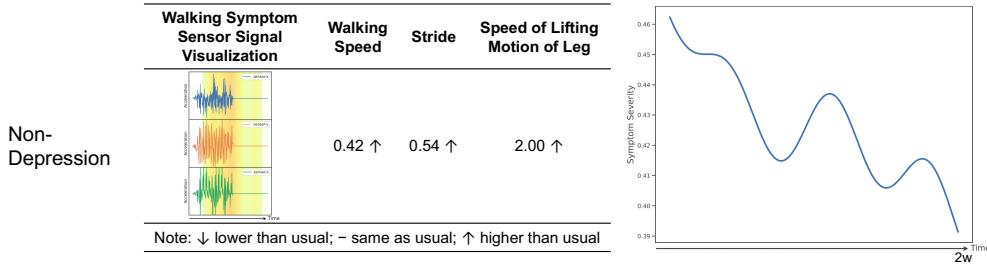


Figure A.5 Interpretation of A Non-depressed Patient

TempPNet predicts the patient in Figure A.4 as depressed for two reasons. First, this patient's walking patterns strongly present a walking symptom like in the left part of Figure A.4. This walking symptom is manifested as slower-than-usual walking speed⁶, shorter stride, and slower lifting motion of the leg. This symptom conforms with the depression physical symptoms in the literature (Sloman et al. 1982, Lemke et al. 2000, NHS 2022). Second, the severity of the previously mentioned symptom presents a temporal progression pattern like the right part of Figure A.4. TempPNet believes this temporal symptom progression pattern resembles a typical depression progression pattern. According to the depression progression literature (Bockting et al. 2015, Dattani et al. 2021), this judgment makes sense — this patient's depression walking symptom first worsens rapidly and then peaks, similar to the onset and acute phases in Figure 1.

TempPNet predicts the patient in Figure A.5 as non-depressed for two reasons. First, this patient's walking patterns strongly present a walking symptom like in the left part of Figure A.5. This walking symptom is manifested as faster-than-usual walking speed, longer stride, and faster lifting motion of the leg. This symptom does not resemble the typical depression walking symptoms in the related literature (Sloman et al. 1982, Lemke et al. 2000, NHS 2022). Second, the severity of the previously mentioned symptom presents a temporal progression pattern like the right part of Figure A.5. The symptom severity trends down and has fluctuations in the middle. This trend does not resemble a typical depression trend.

M. Summary Statistics and Designs of Human Evaluations

The summary statistics and randomization *p*-values of the user study are reported in Tables A.17, A.18, and A.19. The knowledge training in the user study is shown in Figure A.6. The user study groups are shown in Figure A.7.

⁶The usual case is computed as the mean of each gait feature among the non-depressed participants in the dataset.

Table A.17 Summary Statistics (Categorical)

Variable	Category	Count	Variable	Category	Count
Age	18 and lower	1	Education	College freshman	2
	18-24	32		College junior	1
	25-34	30		College senior	18
	35-44	2		Master	31
	45-54	1		Doctorate	14
Gender	Female	39			
	Male	27			

Table A.18 Summary Statistics (Continuous)

Statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Trust in AI	1.000	2.000	3.000	2.530	3.000	4.000
Health Literacy	1.500	2.750	3.000	2.966	3.188	4.000

Table A.19 Randomization Checks

	Age	Education	Gender	Trust in AI	Health Literacy
P-value	0.776	0.632	0.749	0.243	0.127

Read the information about the physical symptoms of depression. Then answer the following questions.

Physical symptoms have been shown to be an essential manifestation of depression:

- Sloman et al. (1982) found that compared to healthy controls, **depressed patients' walks are slower** and the **lifting motion of the leg is slower**.
- Lemka (2020) showed that **depressed patients have shorter strides** and **slower gait velocity** than healthy controls.

(a) Knowledge Training Reading

Incorrect answer. Please read the background knowledge and choose again.

Please pick one group for each question.	
Depressed Patients	Healthy Control
<input checked="" type="radio"/>	<input type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input checked="" type="radio"/>
<input checked="" type="radio"/>	<input type="radio"/>

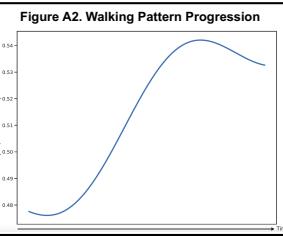
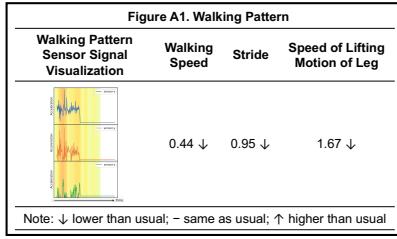
(b) Knowledge Training Test

Figure A.6 Depression Knowledge Training

Input: walking sensor signals of a user
Model A's prediction outcome: depressed

Model A's interpretation of this prediction includes two parts:

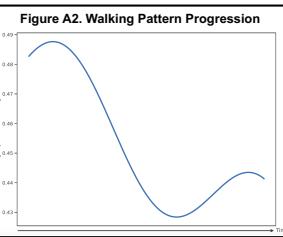
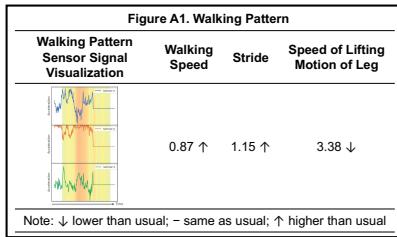
- As shown in Figure A1: This user is predicted as depressed because he/she presents Figure A1's walking pattern, whose sensor signal visualization and walking features are also presented. The walking features are indicated whether they are higher or lower than usual using an arrow sign after the number.
- As shown in Figure A2: The severity of the walking pattern in Figure A1 also progresses over time like in Figure A2.



Input: walking sensor signals of a user
Model A's prediction outcome: non-depressed

Model A's interpretation of this prediction includes two parts:

- As shown in Figure A1: This user is predicted as non-depressed because he/she presents Figure A1's walking pattern, whose sensor signal visualization and walking features are also presented. The walking features are indicated whether they are higher or lower than usual using an arrow sign after the number.
- As shown in Figure A2: The severity of the walking pattern in Figure A1 also progresses over time like in Figure A2.

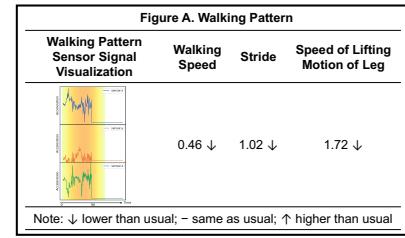


(a) Group TempPNet

Input: walking sensor signals of a user
Model A's prediction outcome: depressed

Model A's interpretation of this prediction includes one part:

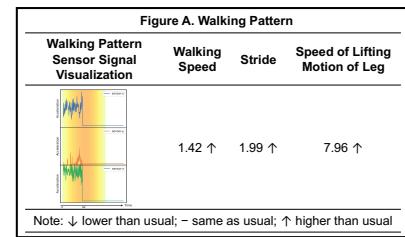
- As shown in Figure A: This user is predicted as depressed because he/she presents Figure A's walking pattern, whose sensor signal visualization and walking features are also presented. The walking features are indicated whether they are higher or lower than usual using an arrow sign after the number.



Input: walking sensor signals of a user
Model A's prediction outcome: non-depressed

Model A's interpretation of this prediction includes one part:

- As shown in Figure A: This user is predicted as non-depressed because he/she presents Figure A's walking pattern, whose sensor signal visualization and walking features are also presented. The walking features are indicated whether they are higher or lower than usual using an arrow sign after the number.



(b) Group Baseline

Figure A.7 User Study Groups

N. Implications for Other Business Areas

Mobile analytics: The advent of IoT and mobile apps has enabled innovative approaches to collect granular and real-time data to assess various human behaviors (Chen et al. 2012). To harness such data volume and granularity, our method allows mobile business analytics researchers to systematically extract interpretable patterns for dynamic physical activities and assess user behaviors based on those patterns. For instance, driving behavior is of great interest to auto insurance companies to personalize insurance premiums. They have been using sensing technology in conjunction with mobile apps to detect careless drivers. Examples include Geico's DriveEasy, Progressive's Snapshot, and StateFarm's Drive Safe & Save. Our method can help them detect the risk of each driver while providing an interpretation of what driving behavior and its temporal trend attribute to such driving risks. With such understanding, auto insurance companies can offer not only their customers reasons for the premium increase or decrease but also recommendations for correcting specific driving behaviors.

Health information technology: Our proposed method is particularly useful for those models that rely on snapshot information to make a prediction but neglect the temporal changes of such information. For example, one closely related HIT area that TempPNet can be generalized to is Parkinson's disease (PD) management. Similar to our study, wearable sensor data can be collected to reflect the motion symptoms of PD. The symptom prototype of our method can detect typical PD walking symptoms, such as smaller steps, slower speed, less trunk movement, and a narrow base of support.⁷ PD patients' symptoms may also form a trend. The trend prototype of our method is able to detect such a trend, predict PD severity, and interpret such a prediction. This new capability enables health organizations and startups to work together to manage PD symptoms timely. Consider another HIT area, as sensor data naturally resemble image data, TempPNet can be used to process time-series images and videos. For example, imaging results, such as X-ray and MRI, from routine doctor's visits and physicals can be processed by our model. We can pinpoint the abnormal patterns from each imaging result as well as the risky trend over time. Recent HIT studies also examine patient engagement in health education YouTube videos (Liu et al. 2020). The symptom prototype in our method can be adapted to recognize typical objects in each video frame, and the trend prototype can be used to capture the temporal changes of these objects, such as shape and angle changes, location moves, and context shifts. This information is essential to understand what type of information is more effective to engage patients on video platforms.

Investment portfolio choice: In finance and accounting, portfolio managers often rely on expected return and volatility (e.g., Sharpe Ratio) to select stocks. Our method is able to supplement this process. The time-series 10-K and 10-Q documents reveal a company's financial health, business environment, and strategic development. Apart from the commonly used accounting measures, these temporal textual data can be utilized to predict the return and volatility of stocks as well. Our method can also disclose the typical text contents that appear in these documents (e.g., certain business foci, investment areas, and competitor dynamics) and its temporal progression that attribute to a low return and high volatility prediction.

Social media analytics: Recent social media analytics studies in IS have investigated user behaviors such as medication nonadherence (Xie et al. 2022) and emotions (Chau et al. 2020). These social media data naturally form a temporal pattern where our method can play a pivotal role. For instance, a user's historical social media posts can be fed into our model to detect their emotional distress. The symptom prototype in our method can discover the typical phrases (e.g., major life events) in their posts that are mostly related to their emotional distress. The trend prototype can further depict the temporal progress of such events as well as how it contributes to the decision.

⁷<https://www.parkinson.org/understanding-parkinsons/symptoms/movement-symptoms/trouble-moving>

References

- Anand S, Stepp CE (2015) Listener Perception of Monopitch, Naturalness, and Intelligibility for Speakers With Parkinson's Disease. *Journal of Speech, Language, and Hearing Research* 58(4):1134–1144.
- Apple Inc (2022) Understanding Reference Frames and Device Attitude | Apple Developer Documentation. *Apple Developer*.
- Bagnall A, Flynn M, Large J, Lines J, Middlehurst M (2020) On the usage and performance of hive-cote v1. 0. *Proceedings of the 5th workshop on advances analytics and learning on temporal data, lecture notes in artificial intelligence*.
- Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J (2017) Patient Subtyping via Time-Aware LSTM Networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Bockting CL, Hollon SD, Jarrett RB, Kuyken W, Dobson K (2015) A lifetime approach to major depressive disorder: The contributions of psychological interventions in preventing relapse and recurrence. *Clinical Psychology Review* 41:16–26.
- Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Doerr M, Pratap A, Wilbanks J, Dorsey ER, Friend SH, Trister AD (2016) The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data* 3(1):1–9.
- Care I (2018) Standards of medical care in diabetes—2018 abridged for primary care providers .
- Chau M, Li TM, Wong PW, Xu JJ, Yip PS, Chen H (2020) Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS quarterly* 44(2).
- Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C (2019) This Looks Like That: Deep Learning for Interpretable Image Recognition. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.
- Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: From big data to big impact. *MIS quarterly* 1165–1188.
- Coelln Rv, Dawe RJ, Leurgans SE, Curran TA, Truty T, Yu L, Barnes LL, Shulman JM, Shulman LM, Bennett DA, Hausdorff JM, Buchman AS (2019) Quantitative mobility metrics from a wearable sensor predict incident parkinsonism in older adults. *Parkinsonism & Related Disorders* 65:190–196.
- Czech MD, Patel S (2019) Gaitpy: An open-source python package for gait analysis using an accelerometer on the lower back. *Journal of Open Source Software* 4(43):1778.
- Dattani S, Ritchie H, Roser M (2021) Mental Health. *Our World in Data* .
- Dempster A, Petitjean F, Webb GI (2020) Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery* 34(5):1454–1495.

- Farhan AA, Yue C, Morillo R, Ware S, Lu J, Bi J, Kamath J, Russell A, Bamis A, Wang B (2016) Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. *2016 IEEE Wireless Health (WH)*.
- Fawaz H, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, Webb GI, Idoumghar L, Muller PA, Petitjean F (2020) Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* 34(6):1936–1962.
- Gao R, Huo Y, Bao S, Tang Y, Antic SL, Epstein ES, Balar AB, Deppen S, Paulson AB, Sandler KL, Massion PP, Landman BA (2019) Distanced LSTM: Time-Distanced Gates in Long Short-Term Memory Models for Lung Cancer Detection. Suk HI, Liu M, Yan P, Lian C, eds., *Machine Learning in Medical Imaging*.
- Gee AH, Garcia-Olano D, Ghosh J, Paydarfar D (2019) Explaining deep classification of time-series data with learned prototypes. *CEUR workshop proceedings* (NIH Public Access).
- Ghosal GR, Abbasi-Asl R (2021) Multi-modal prototype learning for interpretable multivariable time series classification. *arXiv preprint arXiv:2106.09636* .
- Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stern MB, Dodel R, et al. (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):2129–2170.
- Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (The MIT Press).
- Hase P, Chen C, Li O, Rudin C (2019) Interpretable Image Recognition with Hierarchical Prototypes. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7:32–40.
- Jacobson NC, Chung YJ (2020) Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. *Sensors* 20(12):3572.
- Karlsson I, Papapetrou P, Boström H (2016) Generalized random shapelet forests. *Data mining and knowledge discovery* 30:1053–1085.
- King DE, Xiang J (2019) The dietary inflammatory index is associated with diabetes severity. *The Journal of the American Board of Family Medicine* 32(6):801–806.
- Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. *International Conference on Learning Representations*, URL <http://arxiv.org/abs/1412.6980>.
- Lemke MR, Wendorff T, Mieth B, Buhl K, Linnemann M (2000) Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. *Journal of Psychiatric Research* 34(4-5):277–283.
- Li W, Zhu W, Dorsey ER, Luo J (2020) Predicting Parkinson's Disease with Multimodal Irregularly Collected Longitudinal Smartphone Data. *2020 IEEE International Conference on Data Mining (ICDM)*.
- Li Y, Zhu Z, Kong D, Han H, Zhao Y (2019) EA-LSTM: Evolutionary attention-based LSTM for time series prediction. *Knowledge-Based Systems* 181:104785.

- Liu J, Zhang T, Han G, Gou Y (2018) TD-LSTM: Temporal Dependence-Based LSTM Networks for Marine Temperature Prediction. *Sensors* 18(11):3797.
- Liu X, Zhang B, Susarla A, Padman R (2020) Go to youtube and call me in the morning: Use of social media for chronic conditions. *MIS Quarterly* 257–283.
- Ma D, Wang Z, Xie J, Guo B, Yu Z (2020) Interpretable multivariate time series classification based on prototype learning. *Green, Pervasive, and Cloud Computing: 15th International Conference, GPC 2020, Xi'an, China, November 13–15, 2020, Proceedings* 15 (Springer).
- Mei H, Yang C, Eisner J (2022) Transformer Embeddings of Irregularly Spaced Events and Their Participants.
- Middlehurst M, Large J, Bagnall A (2020) The canonical interval forest (cif) classifier for time series classification. *2020 IEEE international conference on big data (big data)* (IEEE).
- Millor N, Lecumberri P, Gómez M, Martinez A, Martinikorena J, Rodríguez-Mañas L, García-García FJ, 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):2018–2025.
- Ming Y, Xu P, Qu H, Ren L (2019) Interpretable and Steerable Sequence Learning via Prototypes. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- Moon S, Song HJ, Sharma VD, Lyons KE, Pahwa R, Akinwuntan AE, 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):125.
- Murphy KP (2022) *Probabilistic Machine Learning: An Introduction* (The MIT Press).
- Nauta M, Bree Rv, Seifert C (2021a) Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*.
- Nauta M, Jutte A, Provoost J, Seifert C (2021b) This Looks Like That, Because ... Explaining Prototypes for Interpretable Image Recognition. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.
- Nemati E, Rahman MJ, Blackstock E, Nathan V, Rahman MM, Vatanparvar K, Kuang J (2020) Estimation of the Lung Function Using Acoustic Features of the Voluntary Cough. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*.
- NHS (2022) Symptoms - Clinical depression. *National Heath Service* .
- Oung Q, Hariharan M, Lee H, Basah S, Sariltee M, Lee C (2015) Wearable multimodal sensors for evaluation of patients with Parkinson disease. *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*.
- Patient (2022) PHQ-9 Depression Test Questionnaire. *patient.info* .

- Piau A, Mattek N, Crissey R, Beattie Z, Dodge H, Kaye J (2020) When Will My Patient Fall? Sensor-Based In-Home Walking Speed Identifies Future Falls in Older Adults. *The Journals of Gerontology: Series A* 75(5):968–973.
- Polat K (2019) A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests. *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*.
- Rastegari E, Azizian S, Ali H (2019) Machine Learning and Similarity Network Approaches to Support Automatic Classification of Parkinson's Diseases Using Accelerometer-based Gait Analysis. *Hawaii International Conference on System Sciences 2019 (HICSS-52)* .
- Rymarczyk D, Struski L, Tabor J, Zieliński B (2021) ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification.
- Schäfer P, Leser U (2017a) Fast and accurate time series classification with weasel. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.
- Schäfer P, Leser U (2017b) Multivariate time series classification with weasel+ muse. *arXiv preprint arXiv:1711.11343* .
- Shitole V, Li F, Kahng M, Tadepalli P, Fern A (2021) One Explanation is Not Enough: Structured Attention Graphs for Image Classification. *Advances in Neural Information Processing Systems*.
- Shokoohi-Yekta M, Hu B, Jin H, Wang J, Keogh E (2017) Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery* 31:1–31.
- Sigcha L, Costa N, Pavón I, Costa S, Arezes P, López JM, De Arcas G (2020) Deep Learning Approaches for Detecting Freezing of Gait in Parkinson's Disease Patients through On-Body Acceleration Sensors. *Sensors 2020, Vol. 20, Page 1895* 20(7):1895–1895.
- Singh G, Yow KC (2021) These do not Look like Those: An Interpretable Deep Learning Model for Image Recognition. *IEEE Access* 9:41482–41493.
- Sloman L, Berridge M, Homatidis S, Hunter D, Duck T (1982) Gait patterns of depressed patients and normal subjects. *American Journal of Psychiatry* 139(1):94–97.
- Trinh L, Tsang M, Rambhatla S, Liu Y (2021) Interpretable and trustworthy deepfake detection via dynamic prototypes. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.
- Um TT, Pfister FMJ, Pichler D, Endo S, Lang M, Hirche S, Fietzek U, Kulic D (2017) Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.
- Vallance JK, Winkler EA, Gardiner PA, Healy GN, Lynch BM, Owen N (2011) Associations of objectively-assessed physical activity and sedentary time with depression: Nhanes (2005–2006). *Preventive medicine* 53(4-5):284–288.

- Wang J, Liu H, Wang X, Jing L (2021) Interpretable Image Recognition by Constructing Transparent Embedding Space.
- Wang Z, Yan W, Oates T (2017) Time series classification from scratch with deep neural networks: A strong baseline. *2017 International joint conference on neural networks (IJCNN)* (IEEE).
- Watanabe A, Noguchi H, Oe M, Sanada H, 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:e5350616.
- Xie J, Liu X, Zeng D, Fang X (2022) Understanding Medication Nonadherence from Social Media: A Sentiment-Enriched Deep Learning Approach. *MIS Quarterly* 46(1):341–372.
- Xu W, Xian Y, Wang J, Schiele B, Akata Z (2020) Attribute Prototype Network for Zero-Shot Learning. *Advances in Neural Information Processing Systems*.
- Yu B, Zhang X, Wang C, Sun M, Jin L, Liu X (2020) Trends in depression among adults in the united states, nhanes 2005–2016. *Journal of Affective Disorders* 263:609–620.
- Yu S, Chai Y, Chen H, Sherman SJ, Brown RA (2022) Wearable Sensor-based Chronic Condition Severity Assessment: An Adversarial Attention-based Deep Multisource Multitask Learning Approach. *MIS Quarterly* Forthcoming.
- Zhang D, Thadajarassiri J, Sen C, Rundensteiner E (2020a) Time-Aware Transformer-based Network for Clinical Notes Series Prediction. *Proceedings of the 5th Machine Learning for Healthcare Conference*.
- Zhang H, Deng K, Li H, Albin RL, Guan Y (2020b) Deep Learning Identifies Digital Biomarkers for Self-Reported Parkinson’s Disease. *Patterns* 1(3):100042.
- Zhang X, Gao Y, Lin J, Lu CT (2020c) Tapnet: Multivariate time series classification with attentional prototypical network. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Zhu H, Samtani S, Brown RA, Chen H (2021) A deep learning approach for recognizing activity of daily living (adl) for senior care: Exploiting interaction dependency and temporal patterns. *MIS Quarterly* 45(2):859–896.