# Supplementary Materials

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

Jiaheng Xie[1,**], Xiaohang Zhao[2,*,**], Xiang Liu[1], Xiao Fang[1]

[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 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.

**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 |

## 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 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

**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 |

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 **??**. At each timepoint $l$, the mobile sensor collects accelerometer readings $[x_l^{\mathfrak{L}}, y_l^{\mathfrak{L}}, z_l^{\mathfrak{L}}]$ 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.

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

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

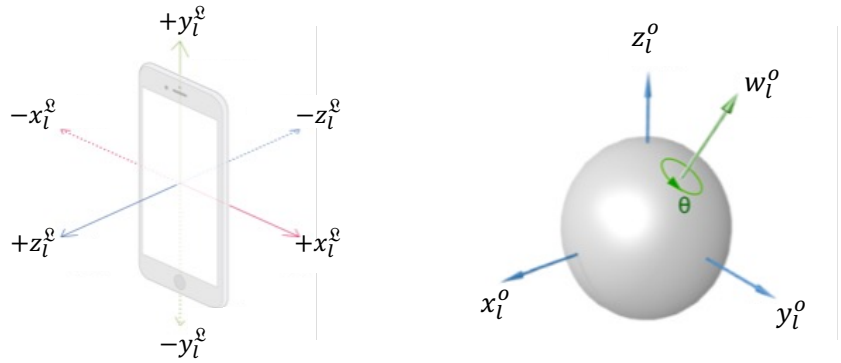
reflect the precise walking pattern in the geographic coordinate system, because it moves and rotates along with the mobile device. To address this issue and make the interpretation more meaningful in the geographic sense, we decide to work with the readings in the global reference frame. We transform 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 as[1]

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

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




Left: $[x_l^{\mathfrak{L}}, y_l^{\mathfrak{L}}, z_l^{\mathfrak{L}}]$ 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

$$a_l = [x_l^{\mathfrak{G}}, y_l^{\mathfrak{G}}, z_l^{\mathfrak{G}}]^T. \tag{EC.2}$$

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 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 $\langle \tilde{\mathcal{R}}a_1, \tilde{\mathcal{R}}a_2, \ldots, \tilde{\mathcal{R}}a_L \rangle$. Once the training stage is finished, we use the original sensor features $\langle a_1, a_2, \ldots, a_L \rangle$ 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:

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

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[.] is the matrix determinant operator, and $|.|$ 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

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

which is a diagonal matrix because $\sigma^{-1}$ 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

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

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

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

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 variety of chronic diseases, such as high blood pressure, heart disease, diabetes, chronic kidney disease, asthma, arthritis,

[2] https://www.cdc.gov/nchs/nhanes/index.htm

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 days[3]. 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.

## 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

[3] Severity Measure for Depression (American Psychiatric Association)

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 $\lambda_{\mathbb{S}} = 0.1$ and $\lambda_{\mathbb{T}} = 0.1$. 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 $H_i^{\mathbb{S}}$ 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 $3 \times 300$, 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 $128 \times 5$ 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

**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 | $u_v = \frac{1}{L}\sum_{l=1}^{L} \lvert\lvert v_l \rvert\rvert$, $\lvert\lvert v_l \rvert\rvert = \sqrt{v_{x,l}^2 + v_{y,l}^2 + v_{z,l}^2}$ |
| St. D. of magnitude | $\sigma_v = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}(\lvert\lvert v_l \rvert\rvert - u_v)^2}$ |
| Mean x-axis jerk | $\alpha_x = \frac{1}{L-1}\sum_{l=1}^{L-1} d_{x,l}$, where $d_{x,l} = v_{x,l+1} - v_{x,l}$ |
| Mean y-axis jerk | $\alpha_y = \frac{1}{L-1}\sum_{l=1}^{L-1} d_{y,l}$, where $d_{y,l} = v_{y,l+1} - v_{y,l}$ |
| Mean z-axis jerk | $\alpha_z = \frac{1}{L-1}\sum_{l=1}^{L-1} d_{z,l}$, 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 | $\alpha_d = \frac{1}{L-1}\sum_{l=1}^{L-1} \lvert\lvert d_l \rvert\rvert$, where $\lvert\lvert d_l \rvert\rvert = \sqrt{d_{x,l}^2 + d_{y,l}^2 + d_{z,l}^2}$ |
| St. D. of jerk magnitude | $\beta_d = \sqrt{\frac{1}{L-2}\sum_{l=1}^{L-1}(\lvert\lvert d_l \rvert\rvert - \alpha_d)^2}$ |
| Stride time variability on x-axis | (1) Identify signal peaks in x-axis, $[t_1, t_2, \ldots, t_Q]$;<br>(2) Identify stride times $[dt_1, dt_2, \ldots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$;<br>(3) Compute stride time variability $V_x = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(dt_i - \overline{dt})^2}$ |
| Stride time variability on y-axis | (1) Identify signal peaks in y-axis, $[t_1, t_2, \ldots, t_Q]$;<br>(2) Identify stride times $[dt_1, dt_2, \ldots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$;<br>(3) Compute stride time variability $V_y = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(dt_i - \overline{dt})^2}$ |
| Stride time variability on z-axis | (1) Identify signal peaks in z-axis, $[t_1, t_2, \ldots, t_Q]$;<br>(2) Identify stride times $[dt_1, dt_2, \ldots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$;<br>(3) Compute stride time variability $V_z = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(dt_i - \overline{dt})^2}$ |
| Stride time vairability on magnitude | (1) Identify signal peaks in magnitude, $[t_1, t_2, \ldots, t_Q]$;<br>(2) Identify stride times $[dt_1, dt_2, \ldots, dt_{Q-1}]$, where $dt_i = t_{i+1} - t_i$;<br>(3) Compute stride time variability $V_v = \sqrt{\frac{1}{Q-2}\sum_{i=1}^{Q-1}(dt_i - \overline{dt})^2}$ |

Recall that given a walking segment, we observe a sequence of sensor signals $\langle v_1^{\mathfrak{L}}, v_2^{\mathfrak{L}}, \ldots, v_L^{\mathfrak{L}} \rangle$, 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.

architecture of the CNN layer in Table A.8. Since the BatchNorm1d layer and the 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.7 Benchmark Hyperparameter Settings**

| Model | Parameter | Values |
|---|---|---|
| TempPNet | CNN channels | $(256, 512, 256, 128)$ |
| | CNN kernel sizes | $8 \times 1$ |
| | Time encoding dimensions | 64 |
| KNN | Number of neighbors | 5 |
| SVM | Regularization parameter | 1 |
| | 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 | $3 \times 3$ |
| ProSeNet | GRU hidden units | 64 |

**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 scores[4].

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

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

[4]https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/KIQ_U_H.htm

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

| KIQ Score | F1-score | Precision | Recall |
|---|---|---|---|
| $1-2$ | $0.763 \pm 0.040$ | $0.757 \pm 0.081$ | $0.773 \pm 0.023$ |
| $3-6$ | $0.809 \pm 0.037$ | $0.807 \pm 0.061$ | $0.813 \pm 0.018$ |

# 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). 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 | $0.774 \pm 0.019$ | $0.805 \pm 0.035$ | $0.746 \pm 0.039$ |
| CNN | Raw sensor | No | $0.689 \pm 0.015$ | $0.594 \pm 0.030$ | $0.821 \pm 0.047$ |
| RNN | Raw sensor | No | $0.700 \pm 0.017$ | $0.630 \pm 0.041$ | $0.790 \pm 0.033$ |
| KNN | Features | No | $0.500 \pm 0.000$ | $0.500 \pm 0.000$ | $0.500 \pm 0.000$ |
| SVM | Features | No | $0.627 \pm 0.000$ | $0.512 \pm 0.000$ | $0.808 \pm 0.000$ |
| Random forest | Features | No | $0.577 \pm 0.041$ | $0.572 \pm 0.097$ | $0.610 \pm 0.119$ |
| AdaBoost | Features | No | $0.615 \pm 0.000$ | $0.500 \pm 0.000$ | $0.800 \pm 0.000$ |
| XGBoost | Features | No | $0.661 \pm 0.000$ | $0.580 \pm 0.000$ | $0.769 \pm 0.000$ |

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.

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

**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 | $0.774 \pm 0.019$ | $0.805 \pm 0.035$ | $0.746 \pm 0.039$ |
| Chen et al. (2019) | Yes | No | $0.716 \pm 0.015$ | $0.694 \pm 0.030$ | $0.741 \pm 0.047$ |
| Ming et al. (2019) | Yes | No | $0.701 \pm 0.017$ | $0.630 \pm 0.041$ | $0.790 \pm 0.033$ |
| Gee et al. (2019) | Yes | No | $0.683 \pm 0.016$ | $0.577 \pm 0.030$ | $0.836 \pm 0.053$ |
| Ma et al. (2020) | Yes | No | $0.704 \pm 0.023$ | $0.628 \pm 0.077$ | $0.801 \pm 0.091$ |
| Fawaz et al. (2020) | No | No | $0.696 \pm 0.011$ | $0.730 \pm 0.018$ | $0.737 \pm 0.018$ |

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) | $0.774 \pm 0.019$ | $0.805 \pm 0.035$ | $0.746 \pm 0.039$ |
| TempPNet removing offset $t_0^{(k)}$ | $0.726 \pm 0.031$ | $0.678 \pm 0.101$ | $0.816 \pm 0.106$ |
| Remove trend prototype using last symptom severity | $0.741 \pm 0.012$ | $0.712 \pm 0.032$ | $0.775 \pm 0.041$ |
| Remove trend prototype using average symptom severity | $0.744 \pm 0.015$ | $0.729 \pm 0.035$ | $0.763 \pm 0.033$ |

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 | $0.774 \pm 0.019$ | $0.805 \pm 0.035$ | $0.746 \pm 0.039$ |
| 4 weeks | $0.738 \pm 0.021$ | $0.708 \pm 0.029$ | $0.772 \pm 0.039$ |
| 8 weeks | $0.730 \pm 0.027$ | $0.683 \pm 0.084$ | $0.810 \pm 0.103$ |
| 16 weeks | $0.721 \pm 0.031$ | $0.651 \pm 0.077$ | $0.828 \pm 0.087$ |

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 | $0.774 \pm 0.019$ | $0.805 \pm 0.035$ | $0.746 \pm 0.039$ |
| 20 Hz | $0.752 \pm 0.034$ | $0.674 \pm 0.075$ | $0.866 \pm 0.062$ |
| 30 Hz | $0.725 \pm 0.034$ | $0.631 \pm 0.086$ | $0.877 \pm 0.080$ |

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

### 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.

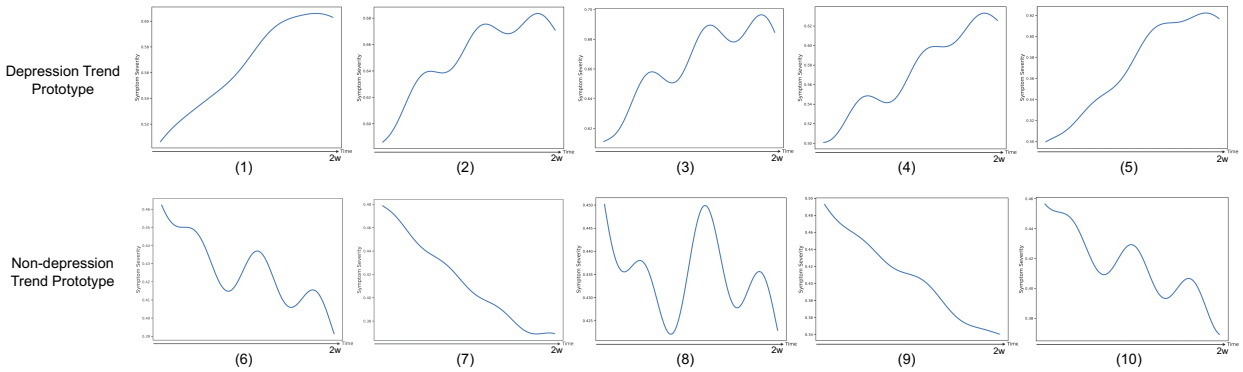
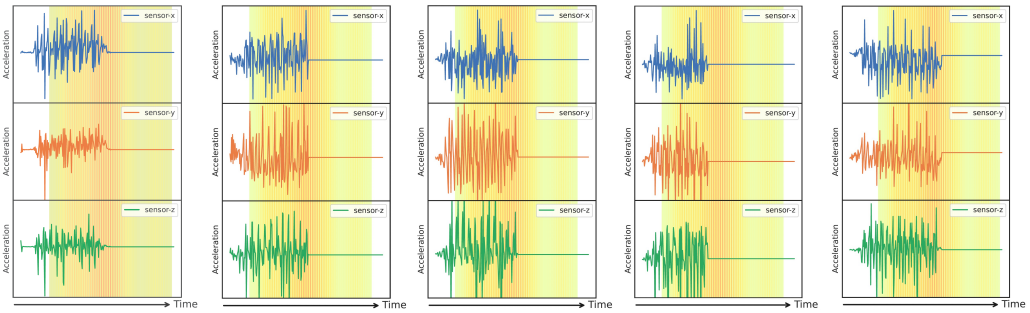



**Figure A.2 Trend Prototypes**

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.

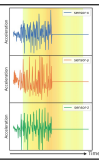
Leveraging the above-learned trend prototypes and symptom prototypes, our model can interpret the prediction of depression for every patient. We randomly select two patients (one depressed and one non-depressed) and showcase TempPNet's interpretation for them. Figure A.4 shows the interpretation of the depressed patient, and Figure A.5 shows the interpretation of the non-depressed

| 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**

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 package[5] 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).

| | Walking Symptom Sensor Signal Visualization | Walking Speed | Stride | Speed of Lifting Motion of Leg |
|---|---|---|---|---|
| Depression |  | 0.24 ↓ | 0.42 ↓ | 0.79 ↓ |

Note: ↓ lower than usual; − same as usual; ↑ higher than usual

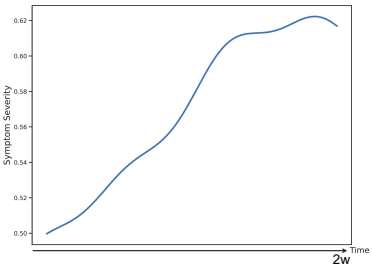



**Figure A.4 Interpretation of A Depressed Patient**

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 speed[6], shorter stride, and slower lifting motion

[5] https://pypi.org/project/gaitpy/

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

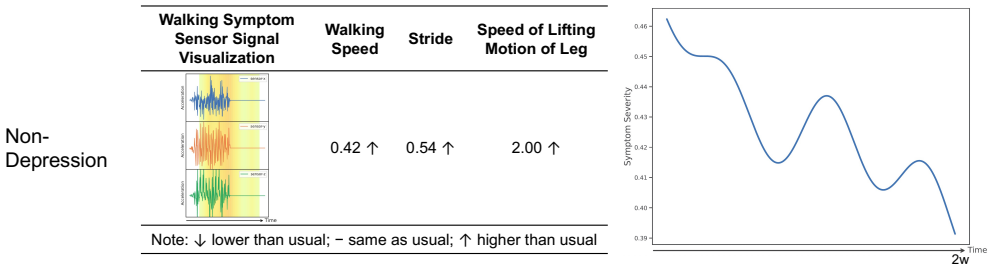



**Figure A.5 Interpretation of A Non-depressed Patient**

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.

**Table A.17 Summary Statistics (Categorical)**

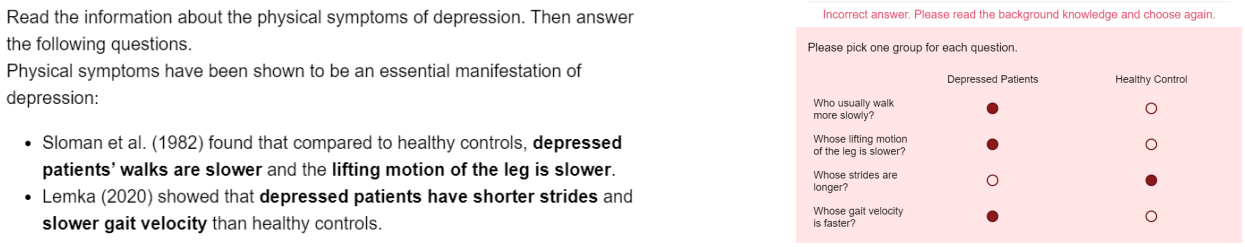
| Variable | Category | Count | Variable | Category | Count |
|---|---|---|---|---|---|
| Age | 18 and lower | 1 | Education | College freshman | 2 |
| | 18-24 | 32 | | College junior | 1 |
| | 25-34 | 30 | | College senior | 18 |
| | 35-44 | 2 | | Master | 31 |
| | 45-54 | 1 | | Doctorate | 14 |
| Gender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



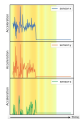

(a) Knowledge Training Reading (b) Knowledge Training Test

**Figure A.6 Depression Knowledge Training**

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

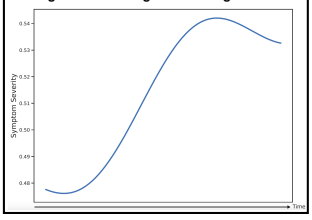
**Model A's interpretation of this prediction includes two parts:**
1. As shown in Figure A1: This user is predicted as depressed because he/she presents Figure A1's walking pattern, whose sensor signal visualization and walking features are also presented. The walking features are indicated whether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

**Figure A1. Walking Pattern**

| Walking Pattern Sensor Signal Visualization | Walking Speed | Stride | Speed of Lifting Motion of Leg |
|---|---|---|---|
|  | 0.44 ↓ | 0.95 ↓ | 1.67 ↓ |

Note: ↓ lower than usual; − same as usual; ↑ higher than usual

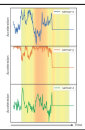
**Figure A2. Walking Pattern Progression**




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

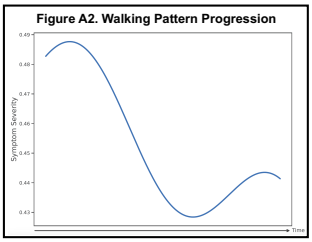
**Model A's interpretation of this prediction includes two parts:**
1. As shown in Figure A1: This user is predicted as non-depressed because he/she presents Figure A1's walking pattern, whose sensor signal visualization and walking features are also presented. The walking features are indicated whether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

**Figure A1. Walking Pattern**

| Walking Pattern Sensor Signal Visualization | Walking Speed | Stride | Speed of Lifting Motion of Leg |
|---|---|---|---|
|  | 0.87 ↑ | 1.15 ↑ | 3.38 ↓ |

Note: ↓ lower than usual; − same as usual; ↑ higher than usual

**Figure A2. Walking Pattern Progression**

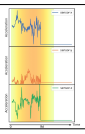



(a) Group TempPNet

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

**Model A's interpretation of this prediction includes one part**:
1. 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.

**Figure A. Walking Pattern**

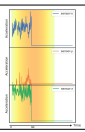
| Walking Pattern Sensor Signal Visualization | Walking Speed | Stride | Speed of Lifting Motion of Leg |
|---|---|---|---|
|  | 0.46 ↓ | 1.02 ↓ | 1.72 ↓ |

Note: ↓ lower than usual; − same as usual; ↑ higher than usual

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

**Model A's interpretation of this prediction includes one part**:
1. 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.

**Figure A. Walking Pattern**

| Walking Pattern Sensor Signal Visualization | Walking Speed | Stride | Speed of Lifting Motion of Leg |
|---|---|---|---|
|  | 1.42 ↑ | 1.99 ↑ | 7.96 ↑ |

Note: ↓ lower than usual; − same as usual; ↑ higher than usual

(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, Sarillee 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, Kulić 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.