IISE Transactions on Healthcare Systems Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uhse21

A Ranking-based Weakly Supervised Learning model for telemonitoring of Parkinson's disease

Dhari F. Alenezi, Hang Shi & Jing Li

To cite this article: Dhari F. Alenezi, Hang Shi & Jing Li (2022) A Ranking-based Weakly Supervised Learning model for telemonitoring of Parkinson's disease, IISE Transactions on Healthcare Systems Engineering, 12:4, 322-336, DOI: 10.1080/24725579.2022.2091065

To link to this article: https://doi.org/10.1080/24725579.2022.2091065

IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
2022, VOL. 12, NO. 4, 322–336
https://doi.org/10.1080/24725579.2022.2091065

A Ranking-based Weakly Supervised Learning model for telemonitoring of Parkinson's disease

Dhari F. Alenezi[a], Hang Shi[b], and Jing Li[a]

[a]School of Industrial and Systems Engineering, Georgia Tech, Atlanta, GA, USA; [b]Department of Neurology, Emory University, Atlanta, GA, USA

ABSTRACT

Telemonitoring is the use of electronic devices to monitor patients remotely. A model is needed to translate the data collected by a patient's mobile device into a predicted score for disease severity assessment. Labeled samples are scarce, which makes it difficult to train a supervised learning model. On the other hand, there is an abundance of samples without precise labels but whose relative rank can be known from domain knowledge. We propose a Ranking-based Weakly Supervised Learning (RWSL) model to integrate both types of data. We apply RWSL to predict Parkinson's disease severity based on mobile-collected tapping activity data of patients. RWSL achieves high predictive accuracy and outperforms competing methods.

KEYWORDS

Health care; machine learning; data mining

1. Introduction

In today's world, mobile phones have moved from being a simple communication tool to being an essential part of people's daily lives. According to App Annie, an application analytics firm, Americans spent an average of four hours per day on their phones in 2020 (Kristianto, 2021). Combining this high screen time and the increasing sensing capabilities of mobile phones, remote and continuous monitoring of an individual's health has become easier than ever, and with negligible costs. Modern smartphones are equipped with various sensors and custom-built applications that can collect user-specific health data. Monitoring such health data is commonly referred to as telemonitoring. Telemonitoring is defined as the use of technological devices to remotely monitor and transmit information related to a person's health status (Dansky et al., 2008).

In this paper we focus on the monitoring and modeling of mobile phone-collected data in patients with Parkinson's disease (PD). PD is a brain disorder that leads to aberrations in movement, including tremors, slowness of movement, rigidity, and difficulty with walking, balance, and coordination. Globally, PD affects seven to ten million people worldwide (Goetz et al., 2009). PD costs the United States $52 billion every year (Michael J. Fox Foundation, 2019). There is currently no cure for PD, but treatments are available to control and reduce symptoms. However, to effectively control symptoms, frequent monitoring of the patient's condition is required to titrate mediations, engage in PD-specific physical therapies, and adjust devices such as deep brain stimulators.

The conventional approach to assess the patient's PD condition is through clinical visits where medical experts inquire about symptoms and perform physical examination. These include questions about the patient's medical history and observations of physical signs or symptoms of PD (Rizek et al., 2016). During the physician's assessment, standardized clinical instruments/questionnaires are typically used to provide some guidance and help reduce subjectivity and misdiagnosis errors. For example, Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is one of the commonly used clinical instruments. UPDRS was developed in 1987 as a gold standard by neurologists for monitoring signs and symptoms of PD (Fahn & Elton, 1987). UPDRS remained the most widely used PD scale until an updated version, MDS-UPDRS, was commissioned by the MDS in 2007. MDS-UPDRS improves upon the original UPDRS by enhancing scale properties and including more non-motor items so that the breadth of PD manifestations is adequately captured.

In conventional clinical practice, clinical visits with a physician happen around every 4–6 months and much less frequently in resource-limited countries and regions (e.g., in years) (Dotchin et al., 2011). The lack of frequent assessment of a patient's PD condition leads to a delay in effective intervention. Telemonitoring technologies offer one possible solution to addressing this issue by making it possible to remotely and frequently assess the patient' clinical condition based on individually collected mobile health data.

To enable mobile-based telemonitoring, an application is needed to become the platform for data collection and transmission. Generally, telemonitoring systems can be divided into self-administered telemonitoring systems such as mPower (Bot et al., 2016) and wearable inertial sensors-based telemonitoring systems such as those introduced in

CONTACT Jing Li jli3175@gatech.edu School of Industrial and Systems Engineering, Georgia Tech, Atlanta, GA, USA.

Sica et al. (2021). mPower is an app installed on the user's mobile phone. Once launched, the app would guide the user to perform several pre-designed activities that measure PD symptoms, such as tapping, speaking, and walking. The data of these activities would be recorded by the mobile's built-in sensors such as accelerometers, gyroscopes, and microphones. Alternatively, wearable inertial sensors-based telemonitoring systems collect free-living movements for a prolonged period of time (Sica et al., 2021). Such an approach has been shown to reduce the well-known "Hawthorne observation effect" (Maetzler et al., 2013) since patients tend to pay less attention when performing their daily free-living activities.

Both types of aforementioned telemonitoring systems open the opportunity for more frequent and convenient assessment of the conditions for PD patients compared to the conventional clinical setting. This paper focuses on the former type of systems represented by mPower with several promising features: Firstly, mPower collects data specific to activities that a patient commonly performs in the clinic; such activities have been shown to capture symptoms relevant to assessing PD severity. On the other hand, the data collected by inertial sensors systems are less focused and not specific to activities that are commonly used to assess PD severity. Secondly, in many inertial sensors-based systems, the assessment of multiple symptoms necessitates employing a high number of wearable devices, compromising patients' comfort. In contrast, mPower does not have such issues since the telemonitoring is conducted purely through a mobile app interface. Thirdly, the ease of using a mobile phone as the data collection platform by mPower makes it possible to collect other important information for PD patients in addition to their activities, such as medication use. Such information helps provide a better assessment of the PD condition and progression.

Despite the promising features, telemonitoring systems like mPower have some limitations that need to be addressed to fully utilize the system's capabilities. First and foremost, mPower provides a data collection platform with no feedback to users (i.e., patients and physicians). Providing feedback in terms of the severity and condition of the patient's disease is critical as it would help with timely intervention. To provide such feedback, data analytics and machine learning algorithms are needed to integrate the collected activity data into predicted disease severity scores. However, current telemonitoring systems do not have adequate predictive analytics capability beyond providing a data collection platform. This paper aims to mitigate this gap.

The predictive analytics task posed above is not trivial due to some uniquely challenging properties of the datasets collected via mPower. First, there is a lack of sufficient labeled samples to train the machine learning model. Specifically, while the activity data (X) can be conveniently obtained via the telemonitoring system, it is not easy to obtain the matched observation for PD severity (Y, a.k.a. label) as this requires expert assessment. For example, using the mPower app, a patient can perform app-guided activities such as tapping, walking, and speaking anytime and anywhere; thus, the activity data can be collected as frequently as multiple times a day. However, disease severity assessment typically only happens at most every 4–6 months. As a result, the sample size with matched (X,Y) datasets from each patient is quite limited. Furthermore, because the data collection process is patient self-administered, there is missing data due to non-adherence and mistakes (Son et al., 2020), which further reduces the qualified samples that can be used to train a robust model. Lastly, as mentioned previously as a promising feature, mPower collects other crucial information such as medication use beyond just the activity data. However, how to effectively leverage this information to augment the capability of the predictive model training remains a challenging issue.

To overcome the aforementioned challenges and provide advanced analytics capability to integrate with the mPower platform, we propose a novel machine learning based predictive model called the Ranking-based Weakly Supervised Learning (RWSL) model. RWSL aims to overcome the challenge of insufficient labeled samples (X,Y) by effectively leveraging and integrating the knowledge in the abundant activity data without a matching Y, known as unlabeled data. Integrating unlabeled and labeled data to train a machine learning model is known as semi-supervised learning (SSL), belonging to the broader field of weakly-supervised learning. However, different from the existing SSL that assumes the Y for each unlabeled sample is completely unknown, we note that domain knowledge can be utilized to provide some weak labeling information of Y. Specifically, domain knowledge may allow for ranking of unlabeled samples in terms of their Y values. As mentioned previously, mPower collects other crucial patient information beyond the activity data, such as medication use. In PD, medication can lessen the severity of the disease; thus, two samples of a patient's activity data that are collected before and after medication can be ranked as $y_i \geq y_j$, where y_i and y_j are the severity levels corresponding to the two samples, even though the exact values for y_i and y_j are unknown. Our model aims to integrate insufficient labeled samples and domain knowledge based ranked samples to build a predictive model for a response variable Y (e.g., disease severity) using features X (e.g., mobile-collected activity data). Since RWSL utilizes domain knowledge to create ranked samples, the model is also related to the machine learning subfield of Information Retrieval (IR), in which learning-to-rank algorithms have been developed to determine the relevance of documents concerning a given user query and order them accordingly. However, although RWSL can incorporate ranked samples in training, the objective of RWSL is different from existing learning-to-rank algorithms in IR, as our model aims to integrate ranked samples with labeled samples to build a superior predictive model.

The remainder of this paper is organized as follows: Sec. 2 reviews the related work and points out gaps. Section 3 presents the proposed model, RWSL. Section 4 presents

simulation experiments. Section 5 presents a real data application in PD. Section 6 concludes the paper.

2. Related work

Our work exists at the intersection of three lines of research pursued by different research communities. Section 2.1 will be dedicated to reviewing research related to healthcare analytics for PD based on telemonitoring data. Section 2.2 will focus on reviewing research in two subfields of machine learning that are related to our model.

2.1. Healthcare analytics for PD based on telemonitoring data

Telemonitoring has enabled the generation of a tremendous amount of patient-specific data that is yet to be fully explored. Such data has made it possible for monitoring, predictive analytics, and understanding of various diseases. The importance of telemonitoring of PD lies in providing data about a patient's daily life not observed by a physician, which may help gain understanding of the complex nature of PD and develop better interventions. Generally, the healthcare analytics literature for PD can be divided based on the end goal and type of PD data utilized. For example, some work has focused on processing of the raw telemonitoring data and feature engineering to obtain proper features related to a clinician's diagnosis (Far et al., 2021; Lenain et al., 2020; Wang et al., 2020). Other research has focused on applying classification models to the extracted features related to PD to classify subjects as PD patients or healthy controls, which may help with PD screening (Abujrida et al., 2017; Arora et al., 2014; Zhang et al., 2020). Another area of research focuses on specific symptoms related to PD (Pastorino et al., 2011; Rigas et al., 2012; Yoneyama et al., 2014); for example, (Rigas et al., 2012) built a hidden Markov model for assessing tremor in PD. Lastly, one major area of research is related to the ability of using telemonitoring apps to collect samples with respect to the timing of medication; such samples have provided researchers the opportunity to study patients' responses to medication (Matarazzo et al., 2019; Zhan et al., 2016).

In summary, the existing telemonitoring data analytics research for PD primarily focuses on feature engineering, classification of PD patients from healthy controls, modeling and understanding of the dynamics of some specific symptoms, and tracking of response to medication. Little research has been done to use telemonitoring activity data to predict disease severity, which is an important task to allow for timely monitoring of disease progression and effective intervention. The present paper aims to mitigate this gap. On the other hand, it is a challenging task to train a robust machine learning model to predict disease severity using telemonitoring activity datasets. While the activity data (X) can be conveniently obtained via the telemonitoring system, it is not easy to obtain the matched observation for PD severity (Y, a.k.a. label). This results in a small labeled sample size. Furthermore, because the telemonitoring activity is patient self-administered, there is missing data due to non-adherence and mistake (Son et al., 2020), which further reduces the qualified samples that can be used to train a robust model. In this paper, we address these challenges by proposing a new machine learning model, RWSL, which integrates ranked samples and labeled samples for training a robust model to predict PD severity (Y) using activity data (X).

2.2. Weakly-supervised learning and information retrieval algorithms in machine learning

Supervised learning is a type of machine learning models that uses features $\boldsymbol{x}$ to predict or classify a response variable y. To train a supervised learning model, a training dataset is needed, which typically contains samples with both $\boldsymbol{x}$ and y available. Such samples are called labeled samples because their response variables are precisely measured. However, it may be difficult to obtain precisely measured response variable for each sample in many application domains, i.e., the label is "weak." This creates the need for weakly-supervised learning algorithms. Weakly-supervised learning can be divided into several major sub-fields: incomplete-supervision, inexact-supervision, and inaccurate-supervision (Zhou, 2018). For incomplete supervision, only a subset of the training samples have labels while the other samples are unlabeled, which is the category our proposed method falls into.

Incomplete supervision can be further divided into two sub-fields: active learning (Settles, 2009) and semi-supervised learning (Chapelle et al., 2006; Zhou & Li, 2010; Zhu, 2008). Active learning assumes that there is an 'oracle', such as a human expert, that can be queried to get labels for selected unlabeled samples (Settles, 2009). The goal of active learning algorithms is to select the most valuable unlabeled samples to query.

On the other hand, semi-supervised learning (SSL) assumes no 'oracle' intervention; instead, it aims to integrate labeled and unlabeled samples to train a model to predict or classify y using $\boldsymbol{x}$. There are different types of SSL algorithms. Some methods assume that samples have inherent cluster structure, and therefore samples falling into the same cluster should have a similar class label (Chapelle et al., 2006). Some methods treat the labels of unlabeled samples as missing values based on the assumption that labeled and unlabeled data samples are both generated from the same model (Nigam et al., 2000). Graph-based methods are popular in SSL, in which a graph is constructed with nodes corresponding to training samples and edges representing relationships between the nodes (Blum & Chawla, 2001; Fujino et al., 2006; Zhou et al., 2004; Zhu et al., 2003). Another type of popular methods assumes that low-dense regions separate labels and aims to identify a classification boundary that goes across the less-dense region while keeping the labeled data correctly classified (Chapelle & Zien, 2005; Joachims, 1999; Li et al., 2013). Despite the fact that SSL is a popular field in machine learning, to our best knowledge, there is little work on leveraging the rank information of unlabeled samples like our method.

On the other hand, we found that ranking algorithms have been mainly investigated in another machine learning field called information retrieval (IR). IR is the process of finding documents of an unstructured nature that satisfies an information/query need from extensive collections of documents (Sanderson et al., 2010). IR relies heavily on learning-to-rank algorithms since returned results may not match the search query and need to be ranked by relevance. Learning-to-rank algorithms can be divided into three categories: pointwise approaches, pairwise approaches, and listwise approaches (Liu, 2011). Pointwise approaches are the earliest in this field. The basic idea is to map the documents' feature vector from ordinal scales to numeric values, and then solve the ranking as regression, classification, and ordinal regression, respectively (Crammer & Singer, 2001; Li et al., 2007). A limitation of pointwise approaches is that the input is a single document without considering the inter-dependency between documents, and thus the position of a document in the final ranked list is invisible to the pointwise loss function (Liu, 2011). Alternatively, listwise approaches directly compare the relevance of a list of documents to a query considering the inter-dependency between documents (Cao et al., 2007; Kondor et al., 2007). Although the performance of listwise approaches has been shown to be better than pointwise and pairwise approaches in many cases, the computational complexity is very high due to the permutation-based loss function evaluation (Liu, 2011). Finally, pairwise methods compare every two documents' relevance, and then rank all the documents based on all these comparison results (Burges et al., 2005, 2006; Joachims, 2002).

Although our method can incorporate ranked samples in training, the objective of our method is different from existing learning-to-rank algorithms in IR. First, learning-to-rank algorithms are mainly aimed to rank a set of documents related to a search query. This means that the objective of the problem is to find the best set of coefficients that meets an unknown latent ranking variable. Hence, a solution of the problem may meet the ordinal requirement but not necessarily produce a good prediction performance. In contrast, our method aims to integrate ranked samples with labeled samples to build an accurate predictive model. Second, learning-to-rank algorithms start by a set of documents whose relevance to a search query is judged using a feature extraction or retrieval function. For example, decisions about ranking a certain pair of documents relative to a query may be based on how frequently a matching term appears within a specific document. This is not how features are intended to be used in our problem, while our objective is to find how features are relevant to response variable of interest.

In summary, even though our method has some connection with the fields of weakly-supervised learning and IR in machine learning, these existing fields do not address the same problem as ours, which is to integrate labeled and ranked samples to build a predictive model for a response variable y using features $\boldsymbol{x}$.

3. Ranking-based Weakly Supervised Learning (RWSL) model

3.1. Model formulation

Let $\boldsymbol{x}$ denote the feature vector, e.g., features extracted from the tapping signal recorded on a PD patient's smartphone. Let y denote the severity level of the disease, e.g., the MDS-UPDRS score of the patient. Our goal is to learn a model $f(\boldsymbol{x})$ to predict y. The uniqueness of RWSL is that it can integrate two types of data in training the predictive model: labeled samples and ranked samples. A labeled sample is one for which both the feature vector and the corresponding disease severity level are available. Suppose there are L labeled samples, $\{\boldsymbol{x}_l, y_l\}$, $l=1, \ldots, L$. Ranked samples are those for which only the feature vector of each sample is available, and there is a criterion d that allows for ranking of each pair of samples in terms of their disease severity levels. Suppose there are $|\Omega_d|$ pairs of ranked samples, $\{\boldsymbol{x}_i \succcurlyeq \boldsymbol{x}_j\}$, $(i,j) \in \Omega_d$, and the notation "$\succcurlyeq$" means that $y_i \geq y_j$, while the exact values for y_i and y_j are unknown. The criterion d is typically known by domain knowledge, which will be discussed in further detail in Sec. 3.3.

To incorporate ranked samples in model training, the problem boils down to learning the most robust function $f(\boldsymbol{x})$ that correctly predicts the ordering of any given ranked samples. Hence, the best function f is the one that will give us a positive margin such that

$$f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \geq 0, \ \ (i,j) \in \Omega_d. \tag{1}$$

There are several algorithms in the literature that attempt to solve this problem. One of the popular approaches is the SVMRank algorithm (Jankovic, 2008). SVMRank makes it possible to design an efficient algorithm for finding the function $f \in \mathcal{F}$ that maximizes the ordering constraint in (1) and generalizes well beyond the training data. The algorithm uses margin-maximization which leads to an ordering that is more robust with respect to noise in $\boldsymbol{x}$. However, SVMRank has a different objective from ours: it aims to train a model to rank samples whereas we want to train a model to predict y.

We propose the RWSL model that adopts the margin-maximization concept of SVMRank for incorporating ranked samples, while at the same time adding labeled samples through supervised learning. RWSL has the following model formulation:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\lambda_L}{2|L|}\sum_{l=1}^{L}\|y_l - f_{\boldsymbol{\beta}}(\boldsymbol{x}_l)\|_2^2 + \frac{\lambda_R}{|\Omega_d|}\sum_{(i,j)\in\Omega_d}\xi_{ij}, \\ &s.t \\ &\qquad f_{\boldsymbol{\beta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\beta}}(\boldsymbol{x}_j) \geq 1 - \xi_{ij}, \ \ \forall (i,j) \in \Omega_d, \end{aligned} \tag{2}$$

$\xi_{ij} \geq 0, \ \forall \ (i,j) \in \Omega_d$, where $\boldsymbol{\beta}$ contains the model coefficients, ξ_{ij} is a non-negative slack variable, and $\|\ \|_2^2$ is the squared L_2-norm. The objective function consists of three terms: the first term aims to achieve a good generalization on the ranked samples by maximizing the closest distance between two ranked samples defined as $\frac{1}{\|\beta\|_2}$, where it can

be shown that maximizing $\frac{1}{\|\beta\|_2}$ is equivalent to minimizing $\|\beta\|_2^2$; the second term uses a quadratic loss to encourage the predicted responses to be close to the true responses for labeled samples; the third term upper-bounds the slack variables to preserve the ordering of ranked samples, which will serve as a "soft" ordering constraint that may allow for some samples to be mis-ordered. Depending on the application, the choice for the mapping function $f_{\boldsymbol{\beta}}$ may vary. In our application, a linear function showed satisfactory results. The coefficients λ_L and λ_R control the relative degrees of emphasis on the labeled and ranked samples, respectively. The degree of emphasis on the ranked samples will depend on the noise in those samples, which is dependent on the choice of ranking threshold. In Sec. 4.3, we try to provide some guidance on the effect of ranking threshold selection in reducing noise in ranked samples.

3.2. Optimization algorithm

The constrained optimization problem in (2) can be converted to an unconstrained optimization with a hinge loss $L_{hinge}(t) = max(0,\ 1-t)$, i.e.,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\lambda_L}{2|L|}\sum_{l=1}^{L}\|y_l - f_{\boldsymbol{\beta}}(\boldsymbol{x}_l)\|_2^2 + \frac{\lambda_R}{|\Omega_d|} \sum_{(i,j)\in\Omega_d} L_{hinge}\big(1-\big(f_{\boldsymbol{\beta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\beta}}(\boldsymbol{x}_j)\big)\big). \tag{3}$$

(3) is convex but not differentiable. To make it differentiable, the hinge loss can be replaced by another loss function that has comparable performance (Chapelle & Keerthi, 2010), a Huber loss:

$$L_{huber}(t) = \max(0,\ 1-t)^2 - \max(0,\ -t)^2 \tag{4}$$

The Huber loss combines the robustness of L_1-norm with the stability of L_2-norm. For huge errors, it is linear; for small errors, it is quadratic. The Huber loss also gives fast methods for computing gradient and performing Hessian times vector operations. The problem is now convex, unconstrained, and differentiable, and can be formulated as follows:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\lambda_L}{2|L|}\sum_{l=1}^{L}\|y_l - f_{\boldsymbol{\beta}}(\boldsymbol{x}_l)\|_2^2 + \frac{\lambda_R}{|\Omega_d|} \sum_{(i,j)\in\Omega_d} L_{huber}\big(1-\big(f_{\boldsymbol{\beta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\beta}}(\boldsymbol{x}_j)\big)\big). \tag{5}$$

By choosing a linear mapping for the f function, the problem can be further simplified to:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\lambda_L}{2|L|}\sum_{l=1}^{L}\|y_l - \boldsymbol{\beta}^T\boldsymbol{x}_l\|_2^2 + \frac{\lambda_R}{|\Omega_d|} \sum_{(i,j)\in\Omega_d} L_{huber}\Big(1-\boldsymbol{\beta}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\Big), \tag{6}$$

which can be solved using the Newton's method (Dennis & Schnabel, 1996).

3.2.1. Parameter tuning

There are two tuning parameters in RWSL, λ_L and λ_R, which control the relative degrees of emphasis on the labeled and ranked samples in training, respectively. Due to the computational efficiency of RWSL, a grid search of the optimal tuning parameters is feasible. Specifically, the dataset is divided into a training set with labeled and ranked samples, and a validation set with labeled samples. Given fixed values of the two tuning parameters, a model is trained, which is then used to predict the response variables of the samples in the validation set. The mean square error (MSE) between the predicted and true responses for the validation samples is computed. In this way, we can compute the MSEs for all possible combinations of the tuning parameter values and choose the optimal tuning parameter combination, $(\lambda_L^*, \lambda_R^*)$, as the one that yields the smallest MSE. When the dataset has limited labeled samples, a cross-validation scheme can be used to replace the training-validation split of the data. Let $\hat{\boldsymbol{\beta}}_{\lambda_L^*,\lambda_R^*}$ denote the parameter estimates by solving the optimization in (6) under $(\lambda_L^*, \lambda_R^*)$. Then, for any new sample, $\boldsymbol{x}_{new}$, the predicted response variable can be obtained by $\hat{y}_{new} = \hat{\boldsymbol{\beta}}_{\lambda_L^*,\lambda_R^*}^T \boldsymbol{x}_{new}$.

3.3. Criteria for ranking samples

The proposed RWSL model assumes the availability of ranked samples according to a criterion d. In the medical field, such criteria typically exist through domain knowledge. For example, medication can lessen the severity of a disease. Thus, two samples of a patient, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, that are collected before and after medication respectively, can be ranked as $\boldsymbol{x}_i \succcurlyeq \boldsymbol{x}_j$.

Specifically related to PD, the movement disorder of PD occurs largely due to the selective loss of neurons that results in depletion of dopamine in the striatum (Jankovic, 2008; Samii et al., 2004; Sveinbjornsdottir, 2016). Dopaminergic drugs designed to replace the action of dopamine in the deplete striatum. Generally, the clinical effect of dopaminergic drugs is noticed quickly, and may last for several hours, particularly in the early stages of the disease (Zahoor et al., 2018). As a result, it is reasonable to use medication as the criterion d to create pairs of ranked samples for each patient. It can be assumed that the disease is more severe before medication than relatively immediately after medication. Also, it is common for telemonitoring apps to collect samples with respect to the timing of medication, which naturally creates ranked samples. For example, the mPower app requests the users to perform their activities three times a day with at least one time before and one time after medication. Finally, our model is designed to address the scenario of limited labeled samples by compensating for that shortage with ranked samples; however, the model is expected to perform as good as a supervised learning model for patients without ranked samples.

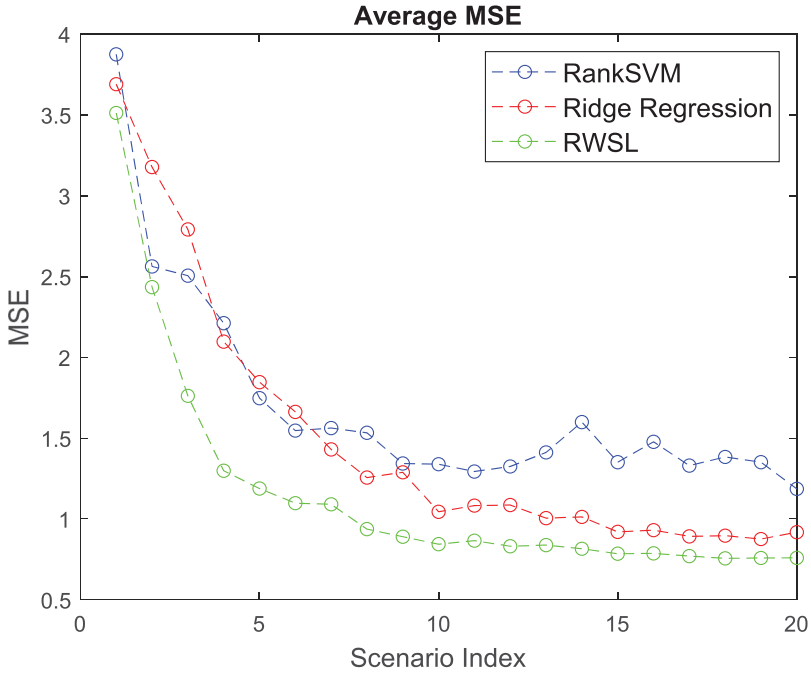

Figure 1. Average MSE comparison of three models on test data.

4. Simulation study

In this section, we use simulated data to test the performance and capabilities of our model in comparison with competing methods. Our simulation experiments in Secs. 4.1–4.4 aim to answer the following questions: How is the performance of RWSL compared to competing methods under different sample sizes for labeled and ranked data (Sec. 4.1)? How much labeled data can be saved by including ranked data in training (Sec. 4.2)? What type of ranked data should be included to achieve the best performance of RWSL (Sec. 4.3)? How robust is RWSL with respect to label noise (Sec. 4.4)?

Data generation: In all the experiments, the data generation process is similar, whereas the parameters values may vary. The data generation process starts by sampling the feature vector from a multivariate normal distribution, i.e., $\boldsymbol{x} \sim MN(0, \boldsymbol{\Sigma}_x)$, where the covariance between the i-th and j-th samples is created by having $\boldsymbol{\Sigma}_{x,i,j} = \alpha^{|i-j|}$, where $\alpha \in (0, 1]$. Then, we generate the coefficients corresponding the features by $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \sigma_\beta^2 \mathbf{I})$. All the coefficients are held fixed after they are generated and considered as the ground truth values to be compared with the estimated coefficients by a model. Furthermore, the response variable is created using an additive model as follows: $y = \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$, where ε is sampled from $N(0, \sigma_\varepsilon^2)$. σ_ε^2 is decided based on the desired signal-to-noise ratio $SNR = \frac{Var(\boldsymbol{\beta}^T \boldsymbol{x})}{\sigma_\varepsilon^2}$. In our simulations, the SNR is fixed at five unless stated otherwise. Also, the number of real features is set to be 15, while we add 15 noise features to test the robustness of the model.

Models under comparison: The proposed RWSL integrates labeled and ranked samples in training a predictive model. To our best knowledge, there is no existing model that provides this capability. Because RWSL adopts the margin-maximization concept in its formulation, which is also the basis of RankSVM, we compare RWSL with RankSVM, which can only use ranked samples in training. Additionally, we compare RWSL with supervised learning, which can only use labeled samples in training. Ridge regression is a supervised learning model that adopts a quadratic loss for labeled samples and a squared L_2 penalty for model coefficients, which are similar to the part of RWSL that accounts for labeled samples. Therefore, we choose ridge regression to compare with RWSL.

Performance metrics: Several performance metrics are adopted to compare the models: (1) Mean Squared Error (MSE) measures the average squared difference between predicted and true responses. The MSE is aimed to evaluate the model's ability to quantify the disease severity accurately. (2) Predictive correlation, i.e., the Pearson correlation between predicted and true responses, which complements MSE as it is bounded between -1 and 1. The predictive correlation is aimed to evaluate if the prediction index is reasonably correlated with the true disease severity, to identify if the patient's condition is getting better or worse. (3) Feature selection (FS) accuracy. As our simulation data purposely includes noise features, we define the following metric to evaluate the capability of a model for identifying the true features:

$$FS\ accuracy = \frac{\sum_{i \in K} |\hat{\beta}_i|}{\sum_{i \in K} |\hat{\beta}_i| + \sum_{j \in N} |\hat{\beta}_j|},$$

where K and N are index sets of true and noise features, respectively.

4.1. Model performance under different sample sizes

This experiment aims to evaluate and compare the performances of models under different sample sizes. We create 20

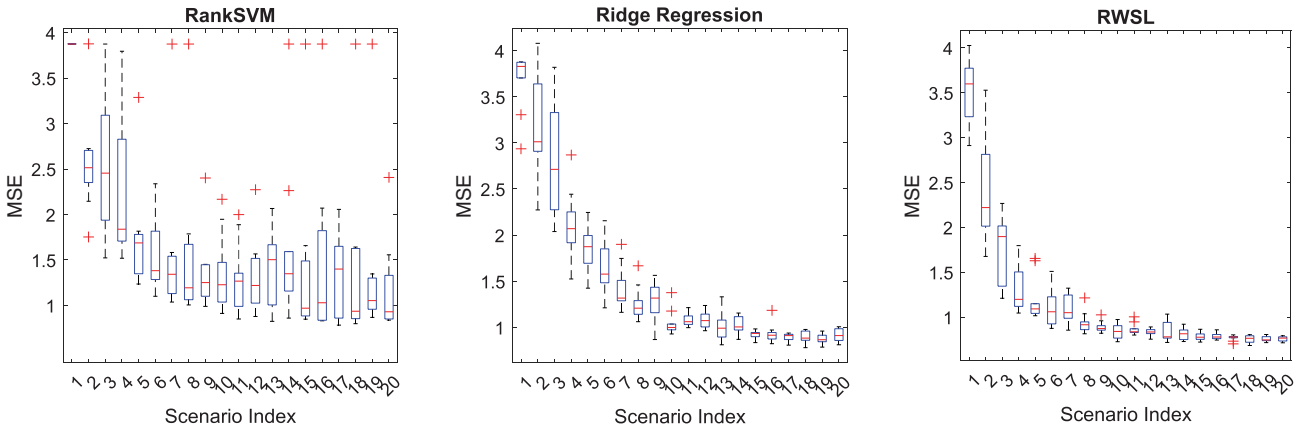

Figure 2. Boxplots of MSE on test data for three models.

different scenarios, starting from a small-sample setting with 10 labeled samples and 10 ranked pairs (i.e., scenario 1), and gradually increasing the numbers of labeled samples and ranked pairs in a fixed step size in each scenario until reaching a large-sample setting of 200 labeled samples and 1000 ranked pairs (i.e., scenario 20). In each scenario, 70% of the labeled samples and ranked pairs are used for training and the rest for validation. Additionally, we generate a blind test set of 300 labeled samples to evaluate and compare the prediction accuracy of different model.

In each scenario, RWSL is trained based on both labeled samples and ranked pairs. The two tuning parameters of RWSL are selected by minimizing the MSE on the validation data. Ridge regression can only include labeled samples in training and parameter tuning of ridge regression is done similarly to RWSL by minimizing the validation MSE. RankSVM can only include ranked pairs in training and the model is to rank samples not to predict a response variable y. Thus, parameter tuning of RankSVM cannot use validation MSE. Instead, we tune RankSVM by maximizing the percentage of pairs ordered correctly (POC) among the total number of ranked pairs included in the validation set. In the experiments, we note that the POC criterion suffers from the fact that there could be several tunning parameters' values that give the same maximum POC. Therefore, a selection rule has to be in place to deal with this situation. We follow a rule that chooses the smallest parameter value, which showed the most robust performance in our experiments.

The training/validation/testing process is repeated ten times for each model and the average MSE on the test set is plotted in Figure 1. We can observe that the average MSE achieved by RWSL is consistently lower than ridge regression and RankSVM. To assess the statistical significance of the performance difference, we conduct one-sided hypothesis testing to compare RWSL and each of the two competing models. The result shows that the mean MSE of RWSL is significantly lower than the competing models for all scenarios except scenario 1 that is the most challenging small-sample setting. Furthermore, comparing RWSL with ridge regression, we can see that the gap between the two models is most significant in scenarios 2–6, which are the scenarios with limited labeled samples. This shows the advantage of RWSL due to its capability of incorporating ranked pairs to augment labeled samples in training. Moreover, comparing RWSL with RankSVM, we can observe that the gap between the two models gets bigger in later scenarios. This indicates that RankSVM has a performance limit even with a large number of ranked samples used in training, whereas RWSL does not suffer from this issue due to the inclusion of labeled samples.

Furthermore, in order to compare not only the average but also the variability of MSE over different runs, we show the boxplot of MSE at each scenario for the three models in Figure 2. We can observe a faster drop in both the mean and variance of MSE for RWSL. We can also notice that RankSVM experiences higher variance and mean fluctuations. This is because RankSVM, by design, is a model to rank samples, instead of predicting a response variable y. When being used as a predictive model, the performance of RankSVM is less satisfying and is unstable.

Additionally, we compare the three models on other performance metrics such as predictive correlation and feature selection accuracy. Specifically, Figures 3 and 4 show average curve and boxplots of the correlation between true and predicted responses on the test set. Figures 5 and 6 show the average curve and boxplots of the feature selection accuracy. Similar observations to MSE can be drawn.

4.2. Saving of labeled samples by RWSL

The findings in the previous section showed that RWSL can help us improve performance under a limited number of labeled samples compared with using a supervised learning model such as ridge regression. The improvement is achieved by incorporating ranked samples in training, which are typically easier to obtain than labeled samples. This result gives rise to an essential question of whether RWSL can achieve a desired performance by just increasing the number of ranked samples and fixing the labeled samples to a low level. In other words, it is desirable to learn how many labeled samples can be saved by adopting RWSL. To answer this question, we focus on scenario 8 in Sec. 4.1, in which RWSL achieved an average MSE of 0.937 with a training set of 80 labeled samples and 375 ranked pairs. Ridge regression trained only based on 80 labeled samples achieved an average MSE of 1.256. To identify the number of additional labeled samples needed for

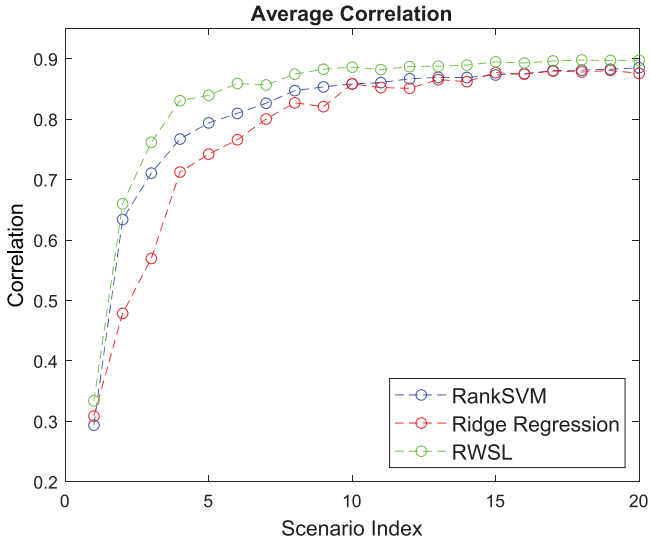

Figure 3. Average predictive correlation comparison of three models on test data.

ridge regression to reach the same MSE as RWSL, we vary the labeled samples from 80 to 150 with a step size of 10 and train ridge regression under each sample size. Figure 7 shows the MSE of ridge regression (red curve) in comparison with RWSL (blue curve).

We can observe that the MSE of ridge regression drops to be lower than that of RWSL with 150 labeled samples in training. To account for statistical significance in comparing models, we conduct one-sided hypothesis testing to see when ridge regression has statistically equivalent MSE to RWSL. We find that the first time when the two models have statistically equivalent performance is when ridge regression includes 130 samples in training. This indicates that an increment of 130–80 = 50 labeled samples in ridge regression is equivalent to 375 ranked pairs in RWSL. If labeled samples are expensive to collect while ranked pairs cost substantially less, the saving of the cost in data collection by RWSL is a clear benefit.

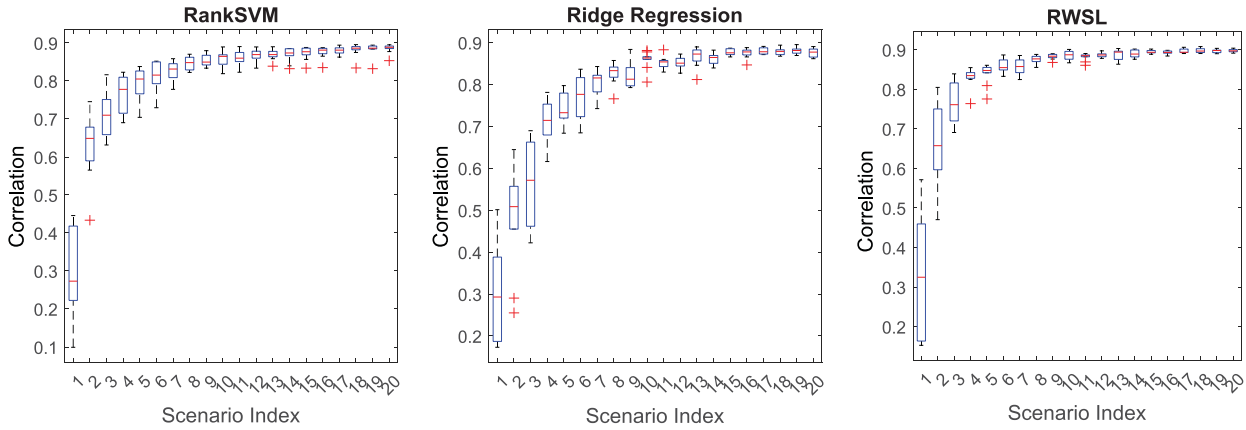

Figure 4. Boxplots of predictive correlation on test data for three models.

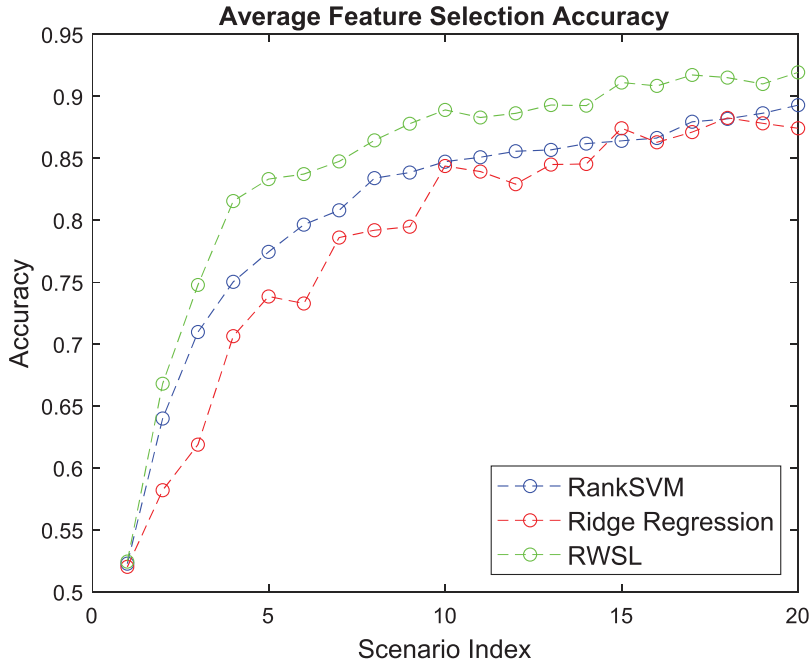

Figure 5. Average feature selection accuracy comparison of three models on test data.

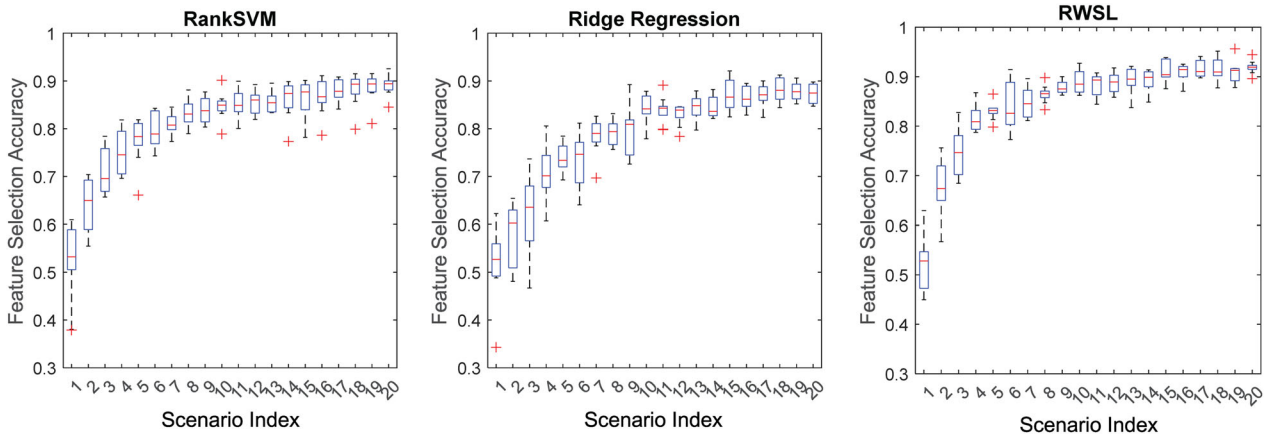

Figure 6. Boxplots of feature selection accuracy on test data for three models.

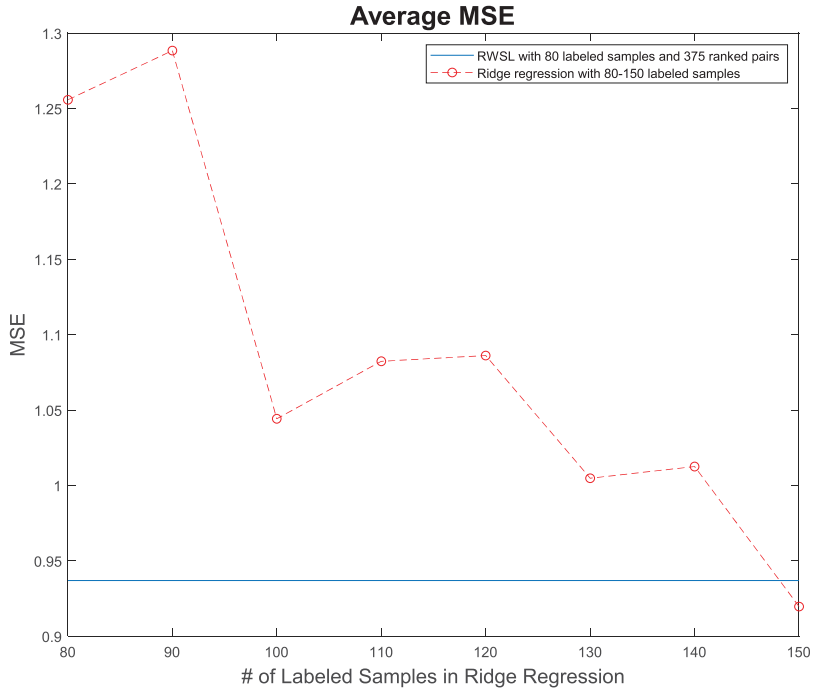

Figure 7. MSE performance of ridge regression with an increasing labeled sample size.

4.3. Performance sensitivity with respect to ranked data difference

Ranked samples play an important role in training RWSL. However, two samples satisfying $y_i \geq y_j$ could have slightly or drastically different response variables. This experiment aims to evaluate how the difference between ranked samples, i.e., $|y_i - y_j|$, affects the performance of RWSL. Let $\Delta y = |y_i - y_j|$ denote a pre-selected threshold. Only ranked samples exceeding Δy are included in model training. We compare three settings with small, medium, and large Δy. The three settings are chosen relative to the value of the noise level in the data. The small Δy is smaller than 1σ of the noise, the medium Δy is around 3σ of the noise, and the large Δy is just above 6σ of the noise. Under each setting, ranked samples of a size between 150 and 900 are simulated, creating 15 scenarios, while the labeled samples are kept at 50 for all scenarios. Figure 8 shows the average MSE on test data under the three settings of Δy. We can see that the MSE performance with medium Δy is the best. The value of the medium Δy is around 3σ of the noise in the response variable. The practical implication of this finding is to provide some guidance on the effect of threshold selection. Underestimating the threshold will result in incorrectly ranked samples due to noise. On the other hand, overestimating the threshold will result in ignoring many valid ranked samples. Hence, the choice of the threshold will affect the type of ranked samples included in training RWSL and eventually affect the performance of the model.

4.4. Performance sensitivity with respect to labeled data noise

In many applications, it may be difficult to measure the response variable precisely. Thus, measurement errors or label noise in the training data are inevitable. Supervised learning models like ridge regression completely rely on labeled samples in training. As a result, they are susceptible

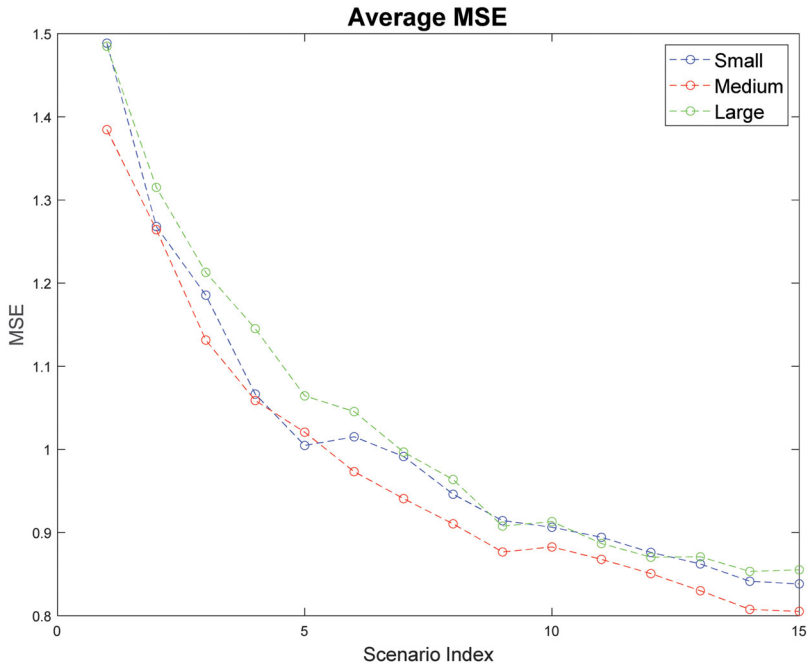

Figure 8. Average MSE comparison of RWSL under small, medium, and large ranked data difference.

to label noise. On the other hand, RWSL may be more robust as it can additionally include ranked data in training that is less susceptible to label noise. To verify this, we add label noise as a percentage of the variance of the response variable when simulating the training data. The test set is kept without noise for performance evaluation. We set the label noise percentage at 0%, 20%, 40%, and 60%. Under each setting, we compare the performance of RWTL with ridge regression. The results are shown in Figure 9. As the noise percentage increases, we can observe that the performance gap between RWSL and ridge regression increases. Such a behavior can be explained by the sole reliance of ridge regression on labeled samples in training. On the other hand, RWSL draws its performance from both labeled and ranked data, which results in high robustness against noise. Hence, under high noise scenarios, RWSL offers the ability to tolerate perturbations that might affect prediction performance.

5. Real data application

In this section, we will present an application of RWSL for predicting the MDS-UPDRS of PD patients using the tapping data collected by their mobile phones.

5.1. Data description

The data was collected by the mPower study (Bot et al., 2016). For each PD patient enrolled in the study, the patient was asked to use the mPower app installed on their mobile phone to perform a tapping activity three times a day: before medication, after medication, and another time. The tapping activity is pre-designed, which requires the patient to use two fingers from the same hand to alternatively tap two stationary points on the screen for 20 seconds. Figure 10 shows the interface on a mobile phone taken from the mPower app (Bot et al., 2016). The tapping activity aims to measure dexterity as assessed by speed, precision, and steadiness (Bot et al., 2016). Raw sensor data collected during a single session of performing the tapping activity is in the form of time series of the screen x-y coordinates on each tap. From the raw time series data, 43 features were extracted, such as the number of taps and the mean inter-tapping interval based on findings from previous studies (Arora et al., 2015; Kassavetis et al., 2016; Tavares et al. 2005). The 43 features are included in the RWSL model and histograms of some of those features can be seen in Figure 11.

In addition to providing tapping data, each participant of the mPower study was also requested to fill out the MDS-UPDRS questionnaire, which was collected on a less frequent basis (usually monthly). The MDS-UPDRS total score is used as the response variable for representing PD severity. MDS-UPDRS score ranges between 0 and 64, with 0 denoting a healthy response, 64 indicating complete disability.

Among the PD participants in the mPower study, males comprise a higher proportion than females accounting for 65.8% of cases. Although an imbalance exists between sexes, the participants are well age-matched with respect to the ages at which males report disease onset (56.6 ± 9.6 years), showing no significant difference to that of females (56.2 ± 9.0 years) (Prince et al., 2018). Moreover, regarding the correlation between gender and tapping activity performance, no significant correlation was found in baseline performance between males (135.4 ± 61.3 taps) and females (133.9 ± 58.0 taps). Regarding the correlation between years since diagnosis and tapping performances, no significant correlations were found regarding baseline and steady-state performance (Prince et al., 2018). Lastly, the correlation

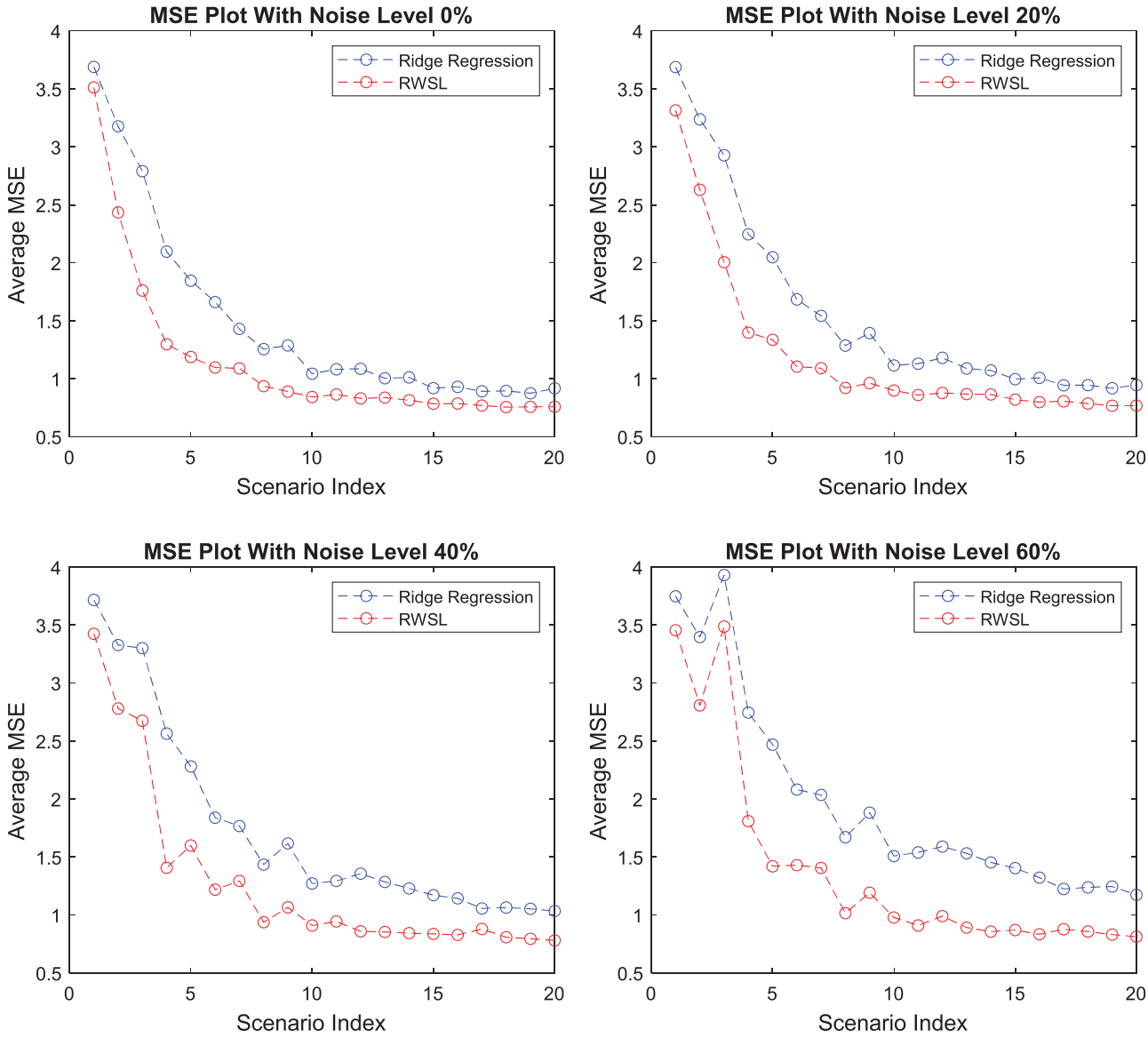

Figure 9. Average MSE comparison between RWSL and ridge regression under different levels of label noise.

between the age of the patient and the clinical severity assessed by the UPDRS scale was found to be weak and statistically insignificant ($p > 0.05$) (Taravari et al., 2014).

5.2. Composition of ranked and labeled datasets

Since mPower is a mobile-based study, the number of participants is quite large. However, because the data is self-collected instead of being collected in a controlled environment, quality of the data collected from each patient varies significantly from one patient to another due to patient commitment and compliance. Therefore, to ensure a more reliable dataset, patient selection needs to be conducted. Our analysis is restricted to patients who performed at least 30 tapping tasks before medication, and 30 or more tasks after medication, and we have 57 such patients. Among those patients, we chose 16 patients who showed strong statistical evidence for medication response based on hypothesis testing suggested by the literature (Chaibub Neto et al., 2016). From these patients, there are a total of 3711 tapping activity records before and after medication. Finally, the ranked dataset is composed of before- and after- medication information within a 3-day window, producing 1284 valid ranked pairs.

Labeled samples are those with MDS-UPDRS measurements available. There is substantial missing data of MDS-UPDRS due to lack of commitment from participants. A possible solution to this problem is to interpolate the data with missing values. According to the PD literature (Tsanas et al., 2010), a linear trend of MDS-UPDRS as PD progresses is the most plausible trend. Hence, we linearly interpolate the MDS-UPDRS to create weekly measurements, ultimately obtaining 168 labeled samples.

5.3. Model training and performance evaluation

We train RWSL, RankSVM, and ridge regression using 5-fold cross-validation (CV) and the CV accuracy of each model is reported in Table 1 using two metrics: MSE and predictive correlation. We can see that RWSL, in bold,

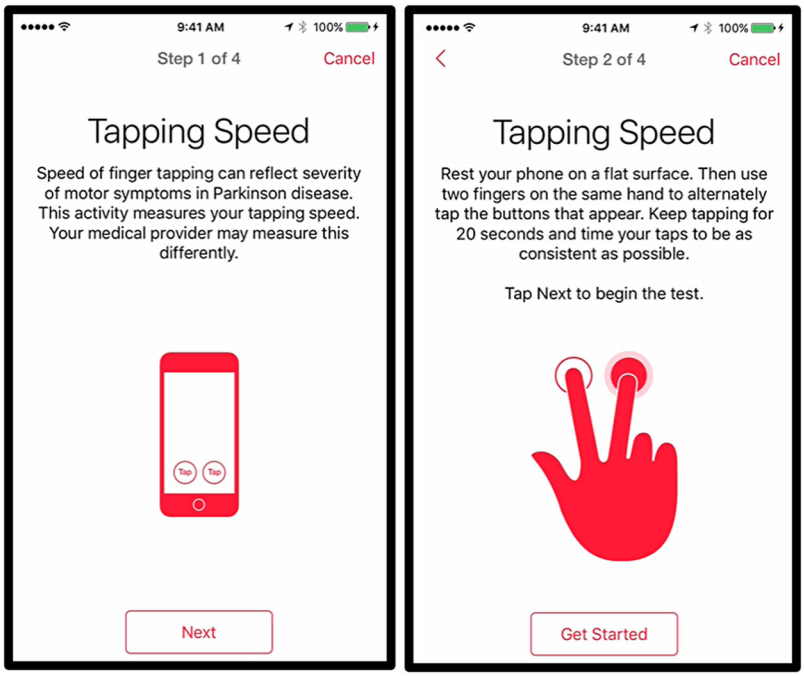

Figure 10. The mPower app instructs a patient to perform a tapping activity (Bot et al., 2016).

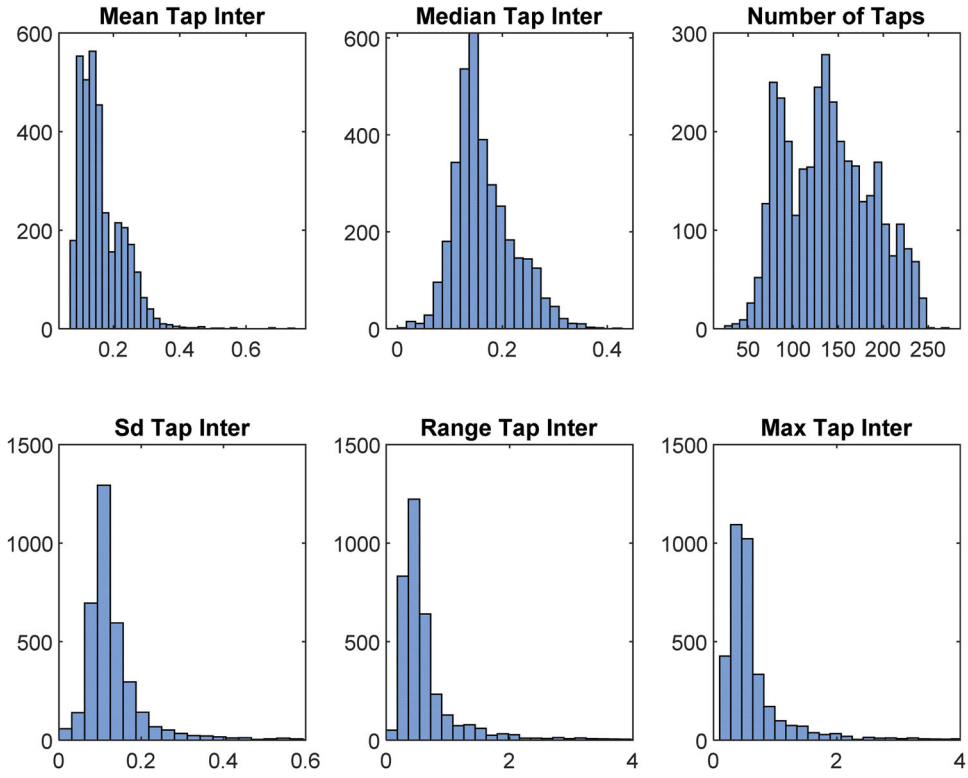

Figure 11. Histograms of some of the tapping features.

Table 1. CV accuracy of MDS-UPDRS prediction based on tapping features.

Model	RankSVM	Ridge regression	RWSL
Predictive correlation: Mean (std)	0.7335 (0.141)	0.6913 (0.158)	**0.7579 (0.083)**
MSE: Mean (std)	0.0384 (0.016)	0.0342 (0.016)	**0.0276 (0.013)**

produces the smallest MSE and the highest predictive correlation, and the smallest standard deviation in both metrics demonstrating good stability of the algorithm.

Among the 43 features included in training the RWSL, the coefficients of two features, median tapping interval and mean tapping interval, rank the top in magnitude. The absolute values of model coefficients are commonly interpreted in the regularized regression literature as crude feature importance scores (Tibshirani, 1996). The sum of absolute coefficients of these two features takes 37% of the sum of absolute total coefficients where we can observe their estimates in Table 2. This is consistent with the literature in

Table 2. Top coefficient estimates of RWSL.

Feature	Mean	Std
Median tap inter	0.6791	0.1490
Mean tap inter	0.7494	0.1572
Button freq	−0.1244	0.0149
Sd tap inter	0.1539	0.0263
Skew tap inter	−0.1497	0.0219
Sd drift right	0.2667	0.0618

which PD patients have been found to have a shorter inter-tap interval of finger tapping due to a lack of control in fine motor capabilities (Roalf et al., 2018; Tavares et al., 2005).

5.4. Practical implications

Telemonitoring is an emerging health care platform enabled by smartphones and wearables, which produces a tremendous amount of patient-specific health data. This paper addresses the data science challenges in leveraging the telemonitoring platform to benefit patient care. Specifically, the proposed RWSL allows for both labeled and ranked samples to be integrated in training a robust model to predict the disease severity of patients with PD based upon mobile-collected tapping activity data. The model provides a step toward continuous disease monitoring of PD patients, thereby providing a tool for physicians to get insight into patients' real-life functioning, deliver timely interventions, reduce clinic visits, and facilitate improving of patients' life quality. Those benefits are of extreme interest, especially to health care practitioners and patients within less privileged healthcare systems under limited resources.

6. Conclusion

We proposed a new RWSL model that allows for both labeled and ranked samples to be integrated in training a predictive model. Simulation experiments showed that RWSL was superior to ridge regression and RankSVM, especially in scenarios of scarce labeled samples. RWSL was applied to the tapping activity data of patients with PD collected by their mobile phones and demonstrated good performance in predicting their disease severity.

Despite the promising features of our model, some limitations need to be addressed and open the doors for future research. First, PD condition/severity is a latent construct, it cannot be perfectly known no matter which clinical instrument is used or even by more comprehensive clinical assessment. In this paper, we chose to use the MDS-UPDRS score as a surrogate measure for PD condition/severity because it has a long history of being a gold standard by neurologists for monitoring signs and symptoms of PD, and it has shown excellent internal consistency across multiple studies and across stages of disease severity as measured by the Hoehn and Yahr staging system (Goetz et al., 2008; Martínez-Martín et al., 1994; Louis et al., 1996). On the other hand, we acknowledge that MDS-UPDRS may not perfectly capture all aspects of the disease. It is expected that a more comprehensive PD severity scale will be developed by the medical society in near future driven by the advances of diagnostic instrument and medical sciences. By that time, RWSL can be re-trained to predict the new severity scale using mobile activity data to allow for more accurate and reliable prediction of PD severity for each patient.

Furthermore, the telemonitoring data collection process is patient self-administered. Hence, there is missing data due to non-adherence and mistakes. This raises questions that are worth exploring in the future about the best data imputation and pre-processing techniques that can mitigate the effect of these occurrences in the data. Also, related to the specific structure of our model, RWSL relies only on medication as the criteria to create ranked samples, and it is worth studying for including other PD-related criteria to create ranked samples.

Finally, it is worth mentioning that there are practical issues that need to be addressed before our model can be implemented in clinical practice. For example, the mPower dataset we used in this paper was created by a research study, whose quality is relatively higher than the data from the general patient population. It remains a challenge about how to improve user compliance for using the app and generate quality-assured tapping activity data for each patient. Furthermore, RWSL predicts disease severity of each patient based on their mobile-collected datasets. While the information may be useful to support clinical decision, this tool needs to gain physicians' trust before they are willing to adopt it. To this end, extensive clinical validations are needed. Also, since the data is collected by patients' mobile phones, substantial efforts would be needed to ensure cyber security so that the data is not maliciously altered, misplaced, or misused.

Consent and approval

This study has been exempted from the requirement for approval by an institutional review board. We have only used publicly available data from the mPower Public Researcher Portal. https://www.synapse.org/#!Synapse:syn4993293.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is partially supported by National Science Foundation (NSF DMS2053170).

References

Abujrida, H., Agu, E., & Pahlavan, K. (2017, November). *Smartphone-based gait assessment to infer Parkinson's disease severity using crowdsourced data*. 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT) (pp. 208–211). IEEE.

Arora, S., Venkataraman, V., Donohue, S., Biglan, K. M., Dorsey, E. R., & Little, M. A. (2014, May). *High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones*. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3641–3644). IEEE.

Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K. M., Dorsey, E. R., & Little, M. A. (2015). Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Parkinsonism & Related Disorders*, *21*(6), 650–653.

Blum, A., & Chawla, S. (2001). *Learning from labeled and unlabeled data using graph mincuts*. Proceedings of the Eighteenth International Conference on Machine Learning (pp. 19–26).

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., … Trister, A. D. (2016). The mPower study, Parkinson disease mobile data collected using. *ResearchKit. Scientific Data*, *3*(1), 1–9.

Burges, C., Ragno, R., & Le, Q. (2006). Learning to rank with non-smooth cost functions. *Advances in Neural Information Processing Systems*, *19*, 193–200.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). *Learning to rank using gradient descent*. Proceedings of the 22nd International Conference on Machine Learning, August (pp. 89–96). https://doi.org/10.1145/1102351.1102363

Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F., & Li, H. (2007, June). *Learning to rank: from pairwise approach to listwise approach*. Proceedings of the 24th International Conference on Machine Learning (pp. 129–136).

Chaibub Neto, E. L. I. A. S., Bot, B. M., Perumal, T., Omberg, L., Guinney, J., Kellen, M., … Trister, A. D. (2016). *Personalized hypothesis tests for detecting medication response in Parkinson disease patients using iPhone Sensor data*. Biocomputing 2016: Proceedings of the Pacific Symposium (pp. 273–284).

Chapelle, O., & Keerthi, S. S. (2010). Efficient algorithms for ranking with SVMs. *Information Retrieval*, *13*(3), 201–215. https://doi.org/10.1007/s10791-009-9109-9

Chapelle, O., Schölkopf, B., & Zien, A. (2006). A discussion of semi-supervised learning and transduction. In O. Chapelle, B. Schölkopf, & A. Zien (Eds.), *Semi-supervised learning* (pp. 473–478). MIT Press.

Chapelle, O., & Zien, A. (2005). Semi-supervised classification by low density separation. *Aistats*, *2005*, 57–64.

Crammer, K., & Singer, Y. (2001). Pranking with ranking. *Nips*, *1*, 641–647.

Dansky, K. H., Vasey, J., & Bowles, K. (2008). Impact of telehealth on clinical outcomes in patients with heart failure. *Clinical Nursing Research*, *17*(3), 182–199.

Dennis Jr., J. E., & Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. Society for Industrial and Applied Mathematics.

Dotchin, C., Jusabani, A., & Walker, R. (2011). Three year follow up of levodopa plus carbidopa treatment in a prevalent cohort of patients with Parkinson's disease in Hai, Tanzania. *Journal of Neurology*, *258*(9), 1649–1656.

Fahn, S., Elton, R. & Members of the UP DRS Development Committee. (1987). The unified Parkinson's disease rating scale. In S. Fahn, C. D. Marsden, D. B. Calne, & M. Goldstein (Eds.), *Recent developments in Parkinson's disease*, Vol. 2 (pp. 153–163). Florham Park: McMellam Health Care Information.

Far, M. S., Eickhoff, S. B., Goni, M., & Dukart, J. (2021). Exploring test-retest reliability and longitudinal stability of digital biomarkers for Parkinson disease in the m-power data set: Cohort study. *Journal of Medical Internet Research*, *23*(9), e26608.

Fujino, A., Ueda, N., & Saito, K. (2006). A hybrid generative/discriminative classifier design for semi-supervised learing. *Transactions of the Japanese Society for Artificial Intelligence*, *21*, 301–309.

Goetz, C. G., Stebbins, G. T., Wolff, D., DeLeeuw, W., Bronte-Stewart, H., Elble, R., Hallett, M., Nutt, J., Ramig, L., Sanger, T., Wu, A. D., Kraus, P. H., Blasucci, L. M., Shamim, E. A., Sethi, K. D., Spielman, J., Kubota, K., Grove, A. S., Dishman, E., & Taylor, C. B. (2009). Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device. *Movement Disorders*, *24*(4), 551–556.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., … Movement Disorder Society UPDRS Revision Task Force. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, *79*(4), 368–376. https://doi.org/10.1136/jnnp.2007.131045

Joachims, T. (1999). *Transductive inference for text classification using support vector machines*. ICML, June (Vol. 99, pp. 200–209).

Joachims, T. (2002). *Optimizing search engines using clickthrough data*. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July (pp. 133–142). https://doi.org/10.1145/775047.775067

Kassavetis, P., Saifee, T. A., Roussos, G., Drougkas, L., Kojovic, M., Rothwell, J. C., Edwards, M. J., & Bhatia, K. P. (2016). Developing a tool for remote digital assessment of Parkinson's disease. *Movement Disorders Clinical Practice*, *3*(1), 59–64.

Kondor, R., Howard, A., & Jebara, T. (2007). Multi-object tracking with representations of the symmetric group. In *Artificial intelligence and statistics* (pp. 211–218). PMLR.

Kristianto, D. (2021). Winning the Attention War: Consumers in nine major markets now spend more than four hours a day in Apps. *App Annie*.

Lenain, R., Weston, J., Shivkumar, A., & Fristed, E. (2020). Surfboard: Audio Feature Extraction for Modern Machine Learning. Interspeech 2020. doi:10.21437/interspeech.2020-2879.

Li, P., Wu, Q., & Burges, C. (2007). Mcrank: Learning to rank using multiple classification and gradient boosting. *Advances in Neural Information Processing Systems*, *20*, 897–904.

Li, Y. F., Tsang, I. W., Kwok, J. T., & Zhou, Z. H. (2013). Convex and scalable weakly labeled SVMs. *Journal of Machine Learning Research*, *14*(29), 2151–2188.

Liu, T. Y. (2011). Applications of learning to rank. In *Learning to rank for information retrieval*. Springer. https://doi.org/10.1007/978-3-642-14267-3_14.

Louis, E. D., Lynch, T., Marder, K., & Fahn, S. (1996). Reliability of patient completion of the historical section of the Unified Parkinson's Disease Rating Scale. *Movement Disorders : official Journal of the Movement Disorder Society*, *11*(2), 185–192.

Maetzler, W., Domingos, J., Srulijes, K., Ferreira, J. J., & Bloem, B. R. (2013). Quantitative wearable sensors for objective assessment of Parkinson's disease. *Movement Disorders*, *28*(12), 1628–1637.

Martínez-Martín, P., Gil-Nagel, A., Gracia, L. M., Gómez, J. B., Martínez-Sarriés, J., & Bermejo, F. (1994). Unified Parkinson's disease rating scale characteristics and structure. *Movement Disorders*, *9*(1), 76–83.

Matarazzo, M., Arroyo-Gallego, T., Montero, P., Puertas-Martín, V., Butterworth, I., Mendoza, C. S., Ledesma-Carbayo, M. J., Catalán, M. J., Molina, J. A., Bermejo-Pareja, F., Martínez-Castrillo, J. C., López-Manzanares, L., Alonso-Cánovas, A., Rodríguez, J. H., Obeso, I., Martínez-Martín, P., Martínez-Ávila, J. C., de la Cámara, A. G., Gray, M., … Sánchez-Ferro, Á. Á (2019). Remote monitoring of treatment response in Parkinson's disease: the habit of typing on a computer. *Movement Disorders*, *34*(10), 1488–1495.

Michael J. Fox Foundation. (2019). Parkinson's disease economic burden on patients, families and the federal government is $52 billion, doubling previous estimates.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*(2/3), 103–134. https://doi.org/10.1023/A:1007692713085

Pastorino, M., Cancela, J., Arredondo, M. T., Pansera, M., Pastor-Sanz, L., Villagra, F., … Martin, J. A. (2011). *Assessment of bradykinesia in Parkinson's disease patients through a multi-parametric system*. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August (pp. 1810–1813). IEEE.

Prince, J., Arora, S., & de Vos, M. (2018). Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. *Physiological Measurement*, *39*(4), 044005.

Rigas, G., Tzallas, A. T., Tsipouras, M. G., Bougia, P., Tripoliti, E. E., Baga, D., Fotiadis, D. I., Tsouli, S. G., & Konitsiotis, S. (2012). Assessment of tremor activity in the Parkinson's disease using a set of wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, *16*(3), 478–487.

Rizek, P., Kumar, N., & Jog, M. S. (2016). An update on the diagnosis and treatment of Parkinson disease. *CMAJ : Canadian Medical Association Journal = Journal de L'Association Medicale Canadienne*, *188*(16), 1157–1165. https://doi.org/10.1503/cmaj.151179

Roalf, D. R., Rupert, P., Mechanic-Hamilton, D., Brennan, L., Duda, J. E., Weintraub, D., Trojanowski, J. Q., Wolk, D., & Moberg, P. J. (2018). Quantitative assessment of finger tapping characteristics in mild cognitive impairment, Alzheimer's disease, and Parkinson's disease. *Journal of Neurology*, *265*(6), 1365–1375.

Samii, A., Nutt, J. G., & Ransom, B. R. (2004). Parkinson's disease. *Lancet (London, England)*, *363*(9423), 1783–1793. https://doi.org/10.1016/S0140-6736(04)16305-8

Sanderson, M., Christopher, D., & Manning, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100–103. https://doi.org/10.1017/S1351324909005129

Settles, B. (2009). *Active learning literature survey*. Computer Sciences Technical Report 1648. University of Wisconsin-Madison.

Sica, M., Tedesco, S., Crowe, C., Kenny, L., Moore, K., Timmons, S., Barton, J., O'Flynn, B., & Komaris, D.-S. (2021). Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review. *PLoS One*, *16*(2), e0246528.

Son, J., Flatley Brennan, P., & Zhou, S. (2020). A data analytics framework for smart asthma management based on remote health information systems with bluetooth-enabled personal inhalers. *MIS Quarterly*, *44*(1), 285–303.

Sveinbjornsdottir, S. (2016). The clinical symptoms of Parkinson's disease. *Journal of Neurochemistry*, *139*, 318–324. https://doi.org/10.1111/jnc.13691

Taravari, A., Medziti, F., Grunevska, B., Adili, F., Ademi, B., Miftari, V., & Haliti, G. (2014). Correlation of age and severity of clinical manifestation assessed by UPDRS in patients with idiopathic Parkinson's disease. *Medical Archives (Sarajevo, Bosnia And Herzegovina)*, *68*(1), 44–46.

Tavares, T., Jefferis, A. L., Koop, G. S., Hill, M., Hastie, B. C., Heit, T., Bronte, G., & Stewart, H. M. (2005). Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. *Movement Disorders*, *20*(10), 1286–1298. https://doi.org/10.1002/mds.20556

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178 https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Bio-Medical Engineering*, *57*(4), 884–893.

Wang, M., Ge, W., Apthorp, D., & Suominen, H. (2020). Robust feature engineering for Parkinson disease diagnosis: New machine learning techniques. *JMIR Biomedical Engineering*, *5*(1), e13611. https://doi.org/10.2196/13611

Yoneyama, M., Kurihara, Y., Watanabe, K., & Mitoma, H. (2014). Accelerometry-based gait analysis and its application to Parkinson's disease assessment—part 1: detection of stride event. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(3), 613–622.

Zahoor, I., Shafi, A., & Haq, E. (2018). Pharmacological treatment of Parkinson's disease. In T. B. Stoker, & J. C. Greenland, (Eds.), *Parkinson's disease: pathogenesis and clinical aspects* [Internet]. Brisbane (AU): Codon Publications.

Zhan, A., Little, M. A., Harris, D. A., Abiola, S. O., Dorsey, E., Saria, S., & Terzis, A. (2016). High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection. *ArXiv preprint arXiv:1601.00960*.

Zhang, H., Deng, K., Li, H., Albin, R. L., & Guan, Y. (2020). Deep learning identifies digital biomarkers for self-reported Parkinson's disease. *Patterns (New York, N.Y.)*, *1*(3), 100042.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in neural information processing systems* (pp. 321–328).

Zhou, Z. H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44–53. https://doi.org/10.1093/nsr/nwx106

Zhou, Z. H., & Li, M. (2010). Semi-supervised learning by disagreement. *Knowledge and Information Systems*, *24*(3), 415–439. https://doi.org/10.1007/s10115-009-0209-z

Zhu, X. (2008). Semi-supervised learning literature survey contents. *SciencesNew York*, *10*(1530), 10.

Zhu, X., Ghahramani, Z., & Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 912–919.