Behavior Modeling Approach for Forecasting Physical Functioning of People with Multiple Sclerosis

ANINDYA DAS ANTAR, University of Michigan, USA
ANNA KRATZ, University of Michigan, USA
NIKOLA BANOVIC, University of Michigan, USA

Forecasting physical functioning of people with Multiple Sclerosis (MS) can inform timely clinical interventions and accurate "day planning" to improve their well-being. However, people's physical functioning often remains unchecked in between infrequent clinical visits, leading to numerous negative healthcare outcomes. Existing Machine Learning (ML) models trained on *in-situ* data collected outside of clinical settings (e.g., in people's homes) predict which people are currently experiencing low functioning. However, they do not forecast if and when people's symptoms and behaviors will negatively impact their functioning in the future. Here, we present a computational behavior model that formalizes clinical knowledge about MS to forecast people's end-of-day physical functioning in advance to support timely interventions. Our model outperformed existing ML baselines in a series of quantitative validation experiments. We showed that our model captured clinical knowledge about MS using qualitative visual model exploration in different "what-if" scenarios. Our work enables future behavior-aware interfaces that deliver just-in-time clinical interventions and aid in "day planning" and "activity pacing".

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; **Human computer interaction (HCI)**.

Additional Key Words and Phrases: computational behavior modeling, forecasting, symptoms, functioning, Multiple Sclerosis.

ACM Reference Format:
Anindya Das Antar, Anna Kratz, and Nikola Banovic. 2023. Behavior Modeling Approach for Forecasting Physical Functioning of People with Multiple Sclerosis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 7 (March 2023), 34 pages. https://doi.org/10.1145/3580887

1 INTRODUCTION

Highly functioning people with Multiple Sclerosis (MS) (a progressive autoimmune chronic disease of the central nervous system) [106] experience few barriers to social interactions and professional careers [105]. Functioning (inverse of disability) indicates how well people with MS can interact with their surrounding environment to perform activities of daily living (ADL) and maintain social participation [118]. However, people with MS face challenges tracking their end-of-day (EOD) physical functioning outside of clinical settings (e.g., at their homes). Despite existing medications and behavioral interventions to improve people's functioning [127], their functioning often goes unchecked in between infrequent clinical visits. This could lead to many negative consequences, including loss of employment, social disability, poor life satisfaction, and reduced physical health [91, 118].

The ability to forecast functioning of people with MS before it deteriorates could enable future user interfaces for clinically validated "just-in-time" interventions [26, 57] (e.g., "day planning" that allows people with MS to

Authors' addresses: Anindya Das Antar, adantar@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Anna Kratz, alkratz@med.umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Nikola Banovic, nbanovic@umich.edu, University of Michigan, Ann Arbor, Michigan, USA.

2474-9567/2023/3-ART7 $15.00
https://doi.org/10.1145/3580887

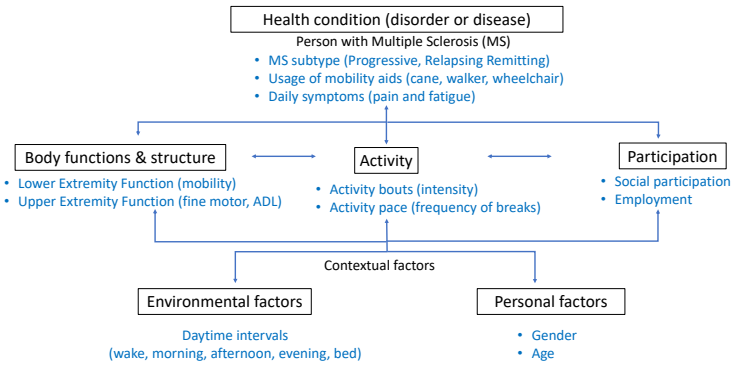


Fig. 1. Our expert-informed conceptual model of functioning and disability of people with MS based on the World Health Organization International Classification of Functioning (WHO-ICF) model [118].

plan their activities for the day in advance) [35, 45, 60, 64, 72, 103] and "activity pacing" which involves taking breaks during physical activities to manage symptoms [1–3]). For example, such interfaces could notify people with MS (or their caregivers and clinicians) and suggest relevant clinical interventions *before* their functioning worsens. However, tracking and forecasting functioning of people with MS remains challenging because their functioning (like their symptoms) is often variable, unpredictable, and fluctuate over time [104].

Recent advancements in personal, mobile, and wearable devices enable tracking of people's symptoms (e.g., fatigue, pain) and physical functioning (e.g., mobility, ADL) at their homes. Existing Machine Learning (ML) models [37, 80, 112] can use such data to predict people susceptible to or *already* experiencing low functioning. However, they do not forecast when people with MS will experience low functioning *ahead of time*.

In this paper, we present a human-centered computational modeling method that captures behaviors of people with MS to forecast their EOD functioning. We based our model on clinical knowledge from an MS disease expert (a clinical psychologist and researcher in a department of physical medicine and rehabilitation) on our team and their insights about perspectives of various stakeholders (e.g., people with MS, clinicians, caregivers). We formalized existing clinical knowledge (Fig. 1) about the relationship between people's MS condition (in particular their MS symptoms, such as pain, fatigue) and their daily physical activity (e.g., activity intensity and pace) [61–63, 70, 117], and how they both influence EOD physical functioning for people with MS [1, 3, 6, 61].

We contribute a computational modeling method (Fig. 2), which extends an existing behavior modeling approach [8] based on Inverse Reinforcement Learning (IRL) [78] and applies it to MS. We first co-engineered features with our MS domain expert; a domain-specific step that is not immediately obvious from existing work [8, 9]. We further extended the existing approach [8, 9] and replaced "off-the-shelf" ML algorithms with a Bayesian Network (BN) [18] (which we co-designed with our domain expert) to predict MS symptoms at different times of day and EOD functioning. We used the engineered features and BN to train our probabilistic generative IRL-based model to estimate conditional probabilities of different EOD functioning levels (low, medium, high) and forecast the one with maximum probability. Unlike other time series approaches (e.g., Hidden Markov Models [89]), our model differentiates between features that describe situations that people with MS find themselves in and the actions they perform in those situations to assess how each impact their EOD physical functioning.

We validated our method in a series of quantitative performance evaluation experiments and using qualitative visual model exploration. We trained and tested our model using an existing clinically validated dataset [61–63] containing data from 107 ambulatory adults with MS collected over 7 consecutive days in their homes. The data consisted of: 1) people's personal factors (age, gender), 2) health conditions (MS subtype, usage of mobility aids), 3) continuous actigraphy, 4) self-reported symptoms (pain, fatigue) five times per day, and 5) self-reported EOD

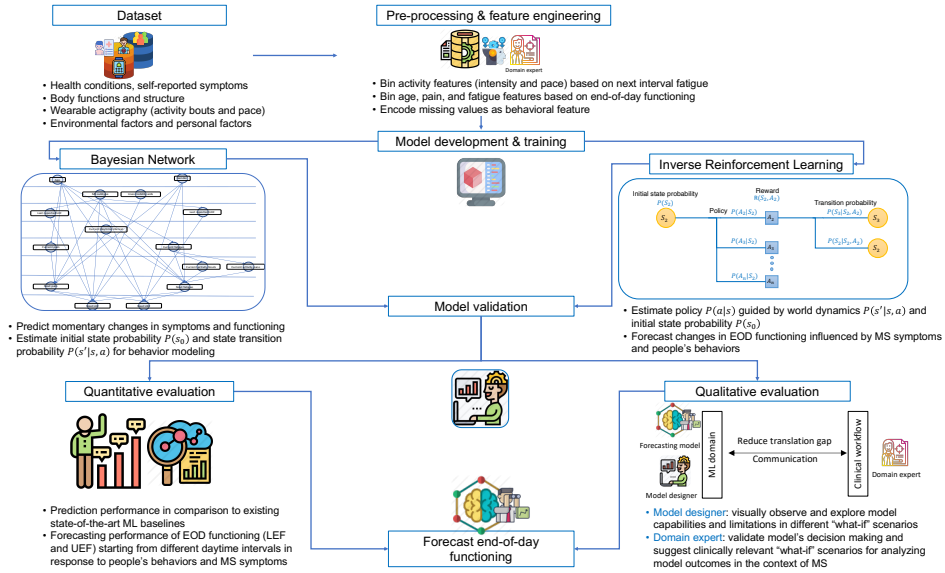


Fig. 2. Stages of our computational behavior modeling method to forecast end-of-day (EOD) physical functioning (LEF and UEF) of people with MS based on their demographics, health conditions, MS symptoms (pain and fatigue), and activities.

functioning. The dataset included two clinically validated measures of EOD physical functioning [12, 122]: 1) upper extremity functioning (UEF), and 2) lower extremity functioning (LEF). We evaluated the performance of our BN to predict people's current symptoms and EOD LEF and UEF as an intermediary step towards evaluating the performance of our IRL-based algorithm to forecast EOD LEF and UEF starting at different times of day.

Our quantitative BN model performance evaluation showed that our BN model on average outperformed existing state-of-the-art ML models [36, 37, 80, 92, 123, 125] with an improvement of 21% and 29% in the receiver operating characteristic area under the curve (ROC AUC) statistic compared to the best performing baseline when predicting EOD LEF and UEF respectively. Our baselines included models used to predict MS symptoms in the past [36, 37, 80, 92, 123, 125] and those used with the existing modeling approach that we extended [8, 9].

Our forecasting performance evaluation showed that our model is more accurate at forecasting EOD LEF and UEF compared to a baseline based on an existing Long Short-Term Memory (LSTM) architecture [25] in two experiments: 1) starting at each daytime interval and using both self-reported and passive-sensing data (with average improvements of 10% and 13% in the best ROC AUC to forecast LEF and UEF respectively), and 2) starting when participants woke up and using their self-reported data then and only passively sensed data afterwards (with average improvements of 14% and 19% in the best ROC AUC to forecast LEF and UEF respectively). Our qualitative visual model exploration in different "what-if" scenarios with two model designers and a clinical expert on our team found that our model captures meaningful relationships between people's personal and environmental factors, daily activities, MS symptoms, and EOD functioning.

Our work contributes a computational modeling method that improves on the existing ML approaches for predicting and forecasting EOD physical functioning of people with MS. Our evaluation showed evidence that this increase in performance came from our carefully engineered, expert-informed features and algorithms. We demonstrated that it is feasible to forecast physical functioning *before* it worsens (i.e., when the people with MS need it the most), with few false alarms. Our work is the first necessary step towards enabling future ubiquitous computing behavior-aware user interfaces that deliver just-in-time clinical interventions.

2 BACKGROUND AND RELATED WORK

Here, we briefly introduce MS and review existing methods for data collection and data analysis. We then outline existing approaches and algorithms for automated prediction and forecasting of MS symptoms and functioning. Note that the study of how people self-track and interpret their own data [4] is out of the scope of this work.

2.1 Multiple Sclerosis (MS) and Physical Functioning

MS is a progressive autoimmune disease of the central nervous system with a constellation of debilitating symptoms [104, 106] that affect people on biological, behavioral, and social levels [44, 120]. Clinical MS research [12, 69] studies physical functioning outcomes because they inform clinical interventions, such as "day planning" (planning activities for a day based on active time) [35, 45, 60, 64, 72, 103] and "activity pacing" (balancing periods of activities with periods of rest) [1–3]. However, forecasting physical functioning outside clinical settings remains challenging [61, 120] due to the complex associations between different factors that impact functioning, such as personal factors (e.g., age [94], gender [24, 34]), clinically identified health conditions (e.g., MS subtype [71, 82], usage of mobility aids [32, 47]), activity [3, 70], and various MS symptoms (e.g., pain, fatigue [2, 6, 83, 115]).

2.2 MS Data Collection and Analysis Methods

Researchers can gain insight into the MS symptoms and functioning of people by collecting their daily behavior and activity data in natural settings (e.g., their homes), over a long period, and at scale [127]. Research [28, 43, 59, 95, 109, 112] has already shown the feasibility of passive sensing to collect daily activity data by leveraging embedded sensors from people's mobile and wearable devices (e.g., Fitbit, Actigraph). Ecological Momentary Assessment (EMA) [62], clinically validated self-report survey instruments [88], and technology-driven empirical studies [14, 85, 111] allow measuring MS symptoms, body functioning [61], and physical and mental wellbeing [54]. Clinical researchers then apply exploratory data analysis (EDA) [114], data visualizations [33, 49], and statistical testing [62] on such data. Although such methods allow researchers to manually explore high-level relationships between factors that influence MS functioning, they are not meant for automated forecasting.

2.3 Predicting and Forecasting MS Symptoms and Functioning

Supervised ML classification [11, 37, 39, 80, 92, 123, 125] and regression [84, 90, 98, 112] algorithms can predict both subjective and objective measures of MS symptoms and functioning. Such methods can classify which people with MS are susceptible to low physical functioning, or people who are currently experiencing low functioning at the time of prediction. However, such methods are not meant for forecasting multivariate time-series MS data (e.g., forecasting in advance when people will experience low levels of functioning). Existing methods of forecasting time-series data have been applied to electronic health records (EHR) [79, 121], forest fire [27], traffic [5], etc. However, these methods can only forecast a specific value, but not multiple variables simultaneously (e.g., computing joint probabilities of multiple variables (e.g., EOD LEF and UEF)) as required in forecasting problem.

2.4 MS Symptom and Functioning Prediction and Forecast Validation and Explanation Challenges

Model designers often use aggregated performance metrics (e.g., accuracy, f1-score) to evaluate ML model performance [41]. However, solely quantitative analysis can be misleading [86] when error distributions are not uniform across subgroups of data, features, classes, and domains. Not knowing where prediction and forecasting errors originate from [23, 67, 102] can limit model usefulness [22, 55, 110]. Recent interactive ML and explainable AI (XAI) research showed that visual model exploration [9, 29, 31, 52, 66, 93, 124] could enable inspection of black-box ML models [116] to identify regions of the model producing erroneous predictions [10, 21]. Nevertheless, it remains challenging for model designers without sufficient domain knowledge to formulate situations when qualitative ML model exploration would be helpful in identifying forecasting errors and their origin [17, 50].

3 METHOD FOR FORECASTING MS FUNCTIONING

Our computational behavior model formalizes clinical knowledge and utilizes carefully engineered features to forecast the EOD functioning of people with MS. We structured our behavior model based on the WHO-ICF model of functioning and disability [118] by considering health conditions, body functions, and contextual factors of people with MS (Fig. 1). We estimated model parameters using an existing probabilistic approach [8] that captures the semantics of purposeful behaviors from different situations and actions [9] demonstrated in the data.

3.1 Dataset

We used an existing dataset [61–63] collected by MS clinicians following clinical guidelines, where the data represent the proportion of MS subtype in a population of people with MS. This Dataset contains 107 ambulatory adults (74 female and 33 male; age range 23-67, $mean = 45$, $std = 11.7$) with clinically confirmed MS. Data was collected in participants' homes over seven consecutive days and includes: 1) personal factors (age, gender), 2) health conditions (MS subtype, usage of mobility aids), 3) continuous wrist actigraphy (activity bouts and pace) from a PRO-Diary watch, 4) Ecological Momentary Assessment (EMA) of self-reported symptoms (pain and fatigue on a 0-10 scale), and 5) end-of-day (EOD) diary of clinically validated self-reported survey measures of functioning using Qualtrics online survey tool. Participants self-reported their symptoms five times each day (wake, morning (11 am), afternoon (3 pm), evening (7 pm), and bed) resulting in 3,745 records from 107 participants, and their end-of-day functioning only before bedtime (749 records). Functioning was computed using a clinically validated eight items questionnaire from the Quality of Life in Neurological Disorders (Neuro-QOL) item bank for daily assessments of lower extremity function (LEF) and upper extremity function (UEF) [75].

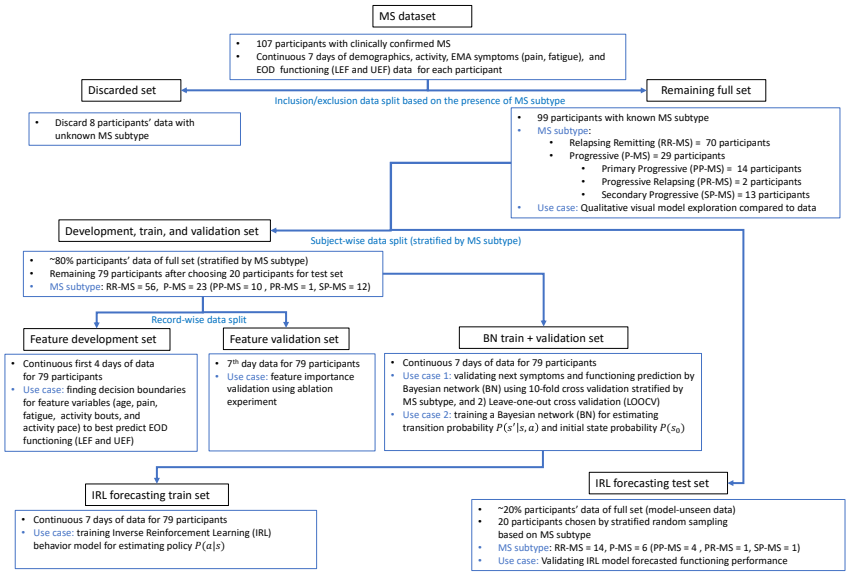


Fig. 3. Dataset split procedure.

We used this dataset for feature selection, model development, training, and validation. We applied both record-wise and subject-wise data splits to generate separate sets for each of the steps in our model development process (Fig. 3). We used stratified random sampling based on people's MS subtype (Progressive and Relapsing Remitting MS) to ensure that our training, validation, and test sets included representative samples. We also

created a hold-out test set including data from model-unseen 20 people with MS from our cohort to minimize chances of identity confounding [77] (e.g., model overfitting on the digital fingerprint of each participant).

3.2 Data Pre-processing and Feature Engineering

We used our domain expert's insights and clinical knowledge to select and engineer our features. We leveraged existing empirical evidence that behaviors and activities affect changes in daily symptoms [61–63, 70, 76, 81, 117], which in turn affect EOD physical functioning (LEF and UEF) for people with MS [1, 3, 6, 61]. We used existing categories in the data when clinically applicable (e.g., binary gender, MS subtype); when no such categories existed, we used entropy-based supervised binning [30] using data from the "feature development set" to find decision boundaries and bin other features that maximize information gain and predictive performance of our model (similar to training a decision tree with the entropy criterion using binned features to predict EOD functioning). Table 1 shows the detailed steps involved in our data processing and feature engineering.

Table 1. Data pre-processing and feature engineering in our modeling approach.

Feature Type	Raw Feature	Feature Engineering Procedure	Cutoff points	Pre-processed Feature Bins
Clinically validated self-reported functioning measures	End-of-day (EOD) lower extremity functioning (LEF) (Neuro-QoL T-s core)	G-code severity modifier scale for Neuro-QoL LEF (v1.0) to bin T-score from EOD questionnaires response (signifies mobility) [73, 74]	$c_1 = 40.5$, $c_2 = 50.9$	Low: $T - score <= 40.5$, Medium: $40.5 < T - score <= 50.9$, High: $T - score > 50.9$, Not Recorded: missing reports
	End-of-day (EOD) upper extremity functioning (UEF) (Neuro-QoL T-score)	G-code severity modifier scale for Neuro-QoL UEF (v1.0) to bin T-score from EOD questionnaires response (signifies fine motor tasks and Activities of Daily Living (ADL)) [73, 74]	$c_1 = 36.5$, $c_2 = 50$	Low: $T - score <= 36.5$, Medium: $36.5 < T - score <= 50$, High: $T - score > 50$, Not Recorded: missing reports
Health conditions	MS subtype (Progressive, Relapsing Remitting)	Feature grouping using clinical insight [108] 1. Relapsing Remitting (RR) 2. Progressive: Primary Progressive, Progressive Relapsing, and Secondary Progressive	N.A	Relapsing Remitting, Progressive
	Usage of mobility aids (e.g., cane, wheelchair)	Binary encoding	N.A	Yes, No
Personal factors	Gender (74 female, 33 male)	Binary encoding	N.A	Male, Female
	Age (mean 45, std = 11.7)	Entropy-based supervised binning Training feature: age Target class: EOD LEF (High, Medium, Low)	$c_1 = 40$, $c_2 = 55$	Younger than 40: <= 40 Between 41 to 55: $[41 - 55]$ 56 and older: >= 56
Self-reported symptoms	Current pain scale $(0 - 10)$	Entropy-based supervised binning Training feature: current pain Target class: EOD LEF (High, Medium, Low)	$c_1 = 1$, $c_2 = 4$	Low: $[0 - 1]$, Medium:$[2 - 4]$, High: $[5 - 10]$, Not Recorded: missing reports
	Current fatigue scale $(0 - 10)$	Entropy-based supervised binning Training feature: current fatigue Target class: EOD LEF (High, Medium, Low)	$c_1 = 3$, $c_2 = 6$	Low: $[0 - 3]$, Medium:$[4 - 6]$, High: $[7 - 10]$, Not Recorded: missing reports
Action features	Current activity bouts (raw data from Actigraph)	Entropy-based supervised binning Training feature: current activity bouts Target class: next fatigue (High, Medium, Low)	$c = 21.97$	Low: avg. activity bouts per epoch (15s) <= 21.97, High: avg. activity bouts per epoch (15s) > 21.97, Not Recorded: missing data
	Current activity pace (raw data from Actigraph)	Entropy-based supervised binning Training feature: current activity pace Target class: next fatigue (High, Medium, Low)	$c = 1.09$	Low: avg. pace per epoch (15s) <= 1.09, High: avg. pace per epoch (15s) > 1.09, Not Recorded: missing data

3.2.1 Clinically Validated Functioning Measures. We used clinically validated subjective measures of EOD LEF and UEF [75]. We converted raw LEF and UEF T-scores into three representative bins (low, medium, high) following clinically validated G-code severity modifiers scale cut-off points [73, 74]. We grouped the severity modifiers into three representative balanced bins (low, medium, high) for LEF and UEF (Table 1).

3.2.2 Health Conditions. Studies suggested that MS subtype [62, 94] and usage of mobility aids [32, 47] influence symptom changes and end-of-day (EOD) functional abilities for people with MS. Thus, we used these two features to represent people's health conditions. In a real-world MS population, Relapsing Remitting (RR-MS) ($\approx$ 85%) is the most common MS subtype than Progressive (P-MS) ($\approx$ 15%) [108]. Thus, most of the participants in the dataset had RR-MS subtypes (77.8%); participants with P-MS subtypes (22.2%) were few. We used clinical insights that people with different subcategories of progressive MS (e.g., Primary Progressive (PP-MS), Progressive Relapsing (PR-MS), and Secondary Progressive (SP-MS)) share similar outcomes, and thus, we combined the participants with PP-MS, PR-MS, and SP-MS subtypes into the "progressive" category (Table 1) consulting with domain expert. We used a binary feature ("Yes", "No") to indicate if participants used mobility aids (e.g., cane, walker, wheelchair).

3.2.3 Personal Factors. Age and gender directly impacts EOD functioning for people with MS [24, 61, 94]. We used the existing gender feature from the dataset which splits participants into males and females. However, modeling every age is computationally prohibitive. Thus, we used data from the "feature development set" to identify three age bins using entropy-based supervised binning [30]. We found two decision boundaries for age ($c_1 = 40$ and $c_2 = 55$) resulting in three bins (Table 1): 1) younger than 40, 2) 41 to 55, and 3) 56 and older.

3.2.4 Self-reported Symptoms. Changes in MS symptoms (e.g., pain and fatigue) throughout the day significantly influence EOD functioning for people with MS [61, 122]. Although existing ML models [16, 51, 100] used numerical symptom features (e.g., pain and fatigue on a scale from 0-10), this requires prohibitively large amount of data to model each combination of symptom levels. Thus, we used the correlation between MS symptoms and functioning to further categorize numerical symptom features into three bins (low, medium, and high) using entropy-based supervised binning [30] (Table 1 shows the cut-offs and the resulting bins).

3.2.5 Activity Bouts and Pacing. MS symptoms (e.g., fatigue) are influenced by people's activity intensity and pace [3, 70, 95, 117], which in turn impact EOD functioning for people with MS [61]. Thus, we included activity bouts (intensity) and pace as activity features. We repeated the same procedure to bin the activity bouts and pace, as with symptoms. However, the actigraphy data lacked sufficient granularity, so we could only create two representative bins (low and high) for activity bouts and pace features. Due to the lack of correlation between the actigraphy data during sleep and other daytime intervals, we discarded bedtime activity bout and pace data. Table 1 shows the resulting bins for *activity bouts* and *activity pace* features.

3.2.6 Handling Missing Values. The dataset contains missing values (activities, self-reported symptoms, and functioning) across different demographics and daytime intervals. Existing modeling approaches [19, 37, 113] adopted different strategies to deal with missing values (e.g., interpolation, mean imputation). However, tracking missing values across daytime intervals might manifest meaningful behaviors (e.g., feeling symptomatic, busy schedule, forgetting to charge the watch) relevant to functioning of people with MS. We analyzed missing value distributions across different feature variables using data from the "feature development set" and the "full set" (Table 2) and found trends in missing values (e.g., males are less likely to report symptoms and functioning than females and the majority fail to report symptoms before bedtime). Thus, we tracked missing values as behavioral features in our model to explore the underlying behaviors behind their occurrence.

Table 2. Distribution of missing/unreported feature values in our "feature development set" (first 4 days of data for 79 participants) and the "full set" (continuous 7 days of data for 99 participants).

Feature name	Feature values	Feature distribution in the data		% of missing data									
				Activity Bouts and Pace		Pain		Fatigue		LEF		UEF	
		Dev.	Full	Dev.	Full	Dev.	Full	Dev.	Full	Dev.	Full	Dev.	Full
Gender	Male	32.9%	31.3%	2.2%	6.1%	12.5%	21%	12.5%	21.2%	12.5%	22.1%	12.5%	22.1%
	Female	67.1%	68.7%	3.4%	4.9%	16%	14.8%	16.2%	14.8%	10.8%	16.2%	10.8%	16%
Age	Younger than 40	38%	41.4%	3.8%	4.8%	17%	18.7%	17%	18.7%	8.3%	13.9%	8.3%	13.9%
	Between 41 to 55	36.7%	36.4%	2.6%	7.3%	11.7%	16.6%	11.9%	16.7%	13.8%	19.8%	13.8%	19.8%
	56 and older	25.3%	22.2%	2.5%	2.9%	11.5%	13.2%	11.5%	13.4%	12.5%	22.7%	12.5%	22.1%
MS subtype	Progressive	29.1%	29.3%	2.2%	6.8%	13.7%	20.6%	13.9%	20.8%	13%	28.1%	13%	28.1%
	Relapsing Remitting	70.9%	70.7%	3.3%	4.7%	13.7%	15.1%	13.7%	15.2%	10.7%	13.9%	10.7%	13.7%
Uses mobility aids	No	72.2%	71.7%	3.8%	6.1%	14%	16.9%	14%	16.9%	12.7%	16.5%	12.7%	16.3%
	Yes	27.8%	28.3%	0.9%	3.3%	12.7%	16.3%	13%	16.5%	8%	21.9%	8%	21.9%
Daytime intervals	Wake	20%	20%	4.4%	6.6%	13%	16.7%	13%	16.7%	N.A	N.A	N.A	N.A
	Morning	20%	20%	2.8%	4.5%	11.1%	15.2%	11.1%	15.2%	N.A	N.A	N.A	N.A
	Afternoon	20%	20%	1.9%	3.8%	11.4%	13%	11.7%	13.1%	N.A	N.A	N.A	N.A
	Evening	20%	20%	2.8%	6.3%	9.8%	13.1%	9.8%	13.1%	N.A	N.A	N.A	N.A
	Bed	20%	20%	N.A	N.A	23.1%	25.7%	23.1%	26%	11.4	18%	11.4%	17.9%

3.3 Capturing Behaviors of People with MS in a Computational Model

Our computational model uses our carefully engineered features to capture the semantics of purposeful behaviors and actions in contextual situations and forecast EOD functioning of people with MS. Our model estimates the probability of changes (e.g., worsening or improvement) in MS symptoms (at different times of day) and predict EOD functioning using a Bayesian network (BN) [18]. Then, we computed the probability of people's behaviors and the probability of performing different actions preceding those symptoms and functioning using an existing behavior modeling approach [8, 9] based on Inverse Reinforcement Learning (IRL) [78]. We formalized behaviors of people with MS and defined stochastic transitions between different contextual situations as Markov Decision Process (MDP) [87] represented by a tuple:

$$\mathcal{M}_{MDP} = (S, P(s_0), A, P(a \mid s), P(s' \mid s, a)) \tag{1}$$

Here, S represents states ($s \in S$) describing different situations people can be in, $P(s_0)$ represents probability that a state initiates a behavior instance, and A denotes a set of actions ($a \in A$) people can perform in those states. The conditional probability ($P(a \mid s)$) of action a in state s captures the uncertainty of people's actions in different states. $P(s' \mid s, a)$ represents the probability of transition into the next state s' after performing action a in the state s. Since people rarely have full control over their environments, this probability distribution also encodes how different environmental conditions change between states, regardless of people's actions.

3.3.1 State and Action Features. We represented each state $s \epsilon S$ and action $a \epsilon A$ using feature vectors ($\mathcal{F}_S$ and $\mathcal{F}_A$). State features in our model included: 1) personal factors (age and gender), 2) environmental factors (current daytime interval), 3) health conditions (MS subtype, usage of mobility aids, and MS symptoms), and 4) EOD body functioning (LEF and UEF) (Table 3). Action features in our model included: 1) current activity bouts (activity intensity), 2) activity pace, and 2) if participants reported their symptoms and functioning or not (Table 4).

3.3.2 Probabilistic State Transition Model. We built a Bayesian network (BN) [18] to compute the initial state probabilities $P(s_0)$ and transition probabilities $P(s' \mid s, a)$. We defined the structure (i.e., relations between

Table 3. State features describing different situations that people with MS can be in.

State feature name	State feature description and possible values
Contextual factors: personal factors	
Gender	Gender of the person with MS {Male, Female}
Age	Age of the person with MS {Younger than 40, Between 41 to 55, 56 and Older}
Contextual factors: environmental factors	
Current daytime interval	Current time of the day {Wake, Morning, Afternoon, Evening, Bed}
Health condition (disorder or disease)	
MS subtype	Type of MS {Progressive, Relapsing Remitting} The progressive category consists of Primary Progressive (PP-MS), Progressive Relapsing (PR-MS), and Secondary Progressive (SP-MS) people with MS
Uses mobility aids	Whether a person with MS uses mobility aids (cane, walker, wheelchair, scooter, etc.) or ambulates without it {Yes, No}
Current symptom (pain)	Self-reported EMA pain score in the current interval using PRO-Diary watch {Low, Medium, High, Not Recorded}
Current symptom (fatigue)	Self-reported EMA fatigue score in the current interval using PRO-Diary watch {Low, Medium, High, Not Recorded}
Body functions and structure (last reported)	
Lower Extremity Function (LEF)	Self-reported end of day (EOD) functioning measure, which signifies mobility and daily level of difficulty completing LE motor tasks (get on and off the toilet; step-up and down curbs; get in and out of a car, etc.) {Low, Medium, High, Not Recorded}
Upper Extremity Function (UEF)	Self-reported end of day (EOD) functioning measure, which signifies fine motor, ADL, and daily level of difficulty completing UE motor tasks (turn a key in a lock; brush your teeth; make a phone call using a touch-tone keypad, etc.) {Low, Medium, High, Not Recorded}

Table 4. Action features describing various actions people with MS can perform in different situations.

Action feature name	Action feature description and possible values
Track of current activity	
Current activity bouts (activity intensity)	Activity bouts in the current interval based on the average number of activity bouts per epoch (15s) from a PRO-Diary ActiGraph watch (activity bouts during sleep has been discarded) {Low, High, Not Recorded}
Current activity pace	Activity pace signifies the number of frequent breaks in between performed activities. Pacing intensity in the current interval has been estimated from the average number of immobile bouts per epoch (15s) from a PRO-Diary ActiGraph watch with values {Low, High, Not Recorded}
Track of reported/not-reported symptoms and outcomes	
Record Pain	Status of next state EMA pain assessment {Recorded, Not Recorded}
Record Fatigue	Status of next state EMA fatigue assessment {Recorded, Not Recorded}
Record Lower Extremity Functioning (LEF)	Status of end-of-day (EOD) diary surveys of LE functional outcomes (mobility) {Recorded, Not Recorded}
Record Upper Extremity Functioning (UEF)	Status of end-of-day (EOD) diary surveys of UE functional outcomes (fine motor/ADL) {Recorded, Not Recorded}

features) of our BN (Fig. 4) based on existing literature [61–63, 70, 117] and in consultation with the domain expert on our team. We trained the BN on data from the "BN train set". We only computed the initial state probabilities $P(s_0)$ as the joint probability of all state features except for $CurrentDaytimeInterval$, which we used as evidence (i.e., $CurrentDaytimeInterval = Wake$). We set the initial state probability to 0 for all states where $CurrentDaytimeInterval \neq Wake$. We computed $P(s' \mid s, a)$ as the joint probability of pain, fatigue, LEF, and UEF in the next state given all other features from the previous state and action. Note that age, gender, MS subtype, usage of mobility aids do not change between the from and to states in each transition.

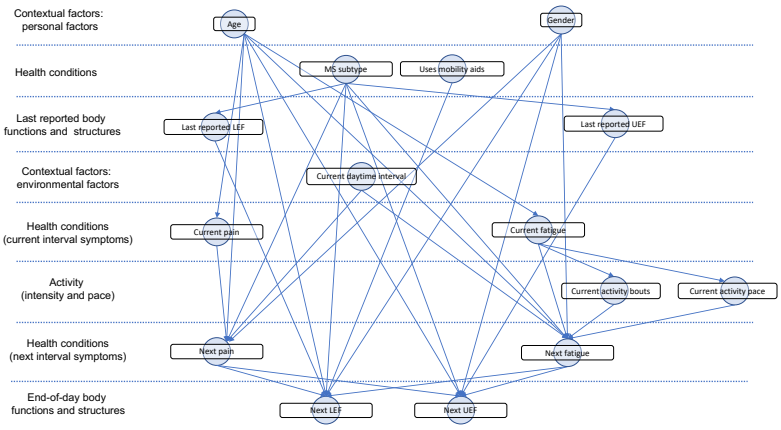


Fig. 4. Bayesian network for computing state transition probabilities and initial state probabilities.

3.3.3 IRL-based Model of Human Behavior. We computed action probabilities $P(a \mid s)$ using the MaxCausalEnt IRL [126] algorithm on the "IRL forecasting train set" and using the initial state and state transition probabilities from the above. Our final model had 30,720 possible states, 104 possible actions, 814,918 possible situation-action pairs, and 3,273,787 possible transitions between states. We performed IRL post-hoc sample size determination (SSD) [46] and showed that our training dataset (553 behavior instances from 79 participants) was adequate to accurately estimate the parameters of our IRL model (the training dataset achieved 88.47% of maximum expected information gain (EIG) corresponding to maximum resource-available sample number $n_d^{max} = 10,000$).

3.3.4 Prediction and Forecasting. To predict or forecast the value of any feature (or a combination of features) at any daytime interval, we sampled 10,000 behavior instances from our model using action probabilities $P(a \mid s)$ and state transition probabilities $P(s' \mid s, a)$ starting from a given daytime interval until the target daytime interval. For prediction, the target interval is always the next interval and the state and action features from the starting daytime interval are always fully observable, so we only use $P(s' \mid s, a)$ for sampling. For forecasting, the target daytime interval can be any daytime interval in the future. We then use the samples to compute conditional probabilities of different target feature(s) values and pick the one with the highest probability as our forecast.

4 QUANTITATIVE MODEL PERFORMANCE EVALUATION

Here, we performed a series of quantitative performance evaluation experiments: 1) an ablation experiment to investigate the importance of our features, 2) evaluation of the prediction performance of our Bayesian network (BN), and 3) evaluation of our Bayesian IRL-based algorithm to forecast EOD functioning in two experiments.

4.1 Computing Feature Importance through Ablation Experiment

We performed feature ablation study to evaluate the effect of features on BN next time interval symptoms and EOD functioning predictions. By progressively removing features (e.g., self-reports, actigraphy) from the complete set of our engineered features, we created four feature sets: 1) complete feature set, 2) discarded activities, 3) discarded activities and current symptoms, and 4) discarded activities, current symptoms, and last reported functioning. We trained four different Bayesian networks for each set of features using data from the "BN train set". We compared the BIC scores [99] for each of the four BN (Table 5), where the lowest BIC is the best.

Our results showed that removing features increased the BIC score (i.e., resulted in worse models). We also computed prediction performance metrics on the "feature validation set", and showed that the BN with the complete set of features had the best predictive performance (Table 5). Thus, this experiment confirmed that *all* of our engineered features are informative for predicting symptoms and functioning. Our findings align with existing clinical studies [53, 61] that show patient self-reports are important for capturing behaviors of people with MS, as well as reflecting what people with chronic conditions are experiencing over the course of a day.

Table 5. Comparison of feature importance in terms of BIC score and prediction performance (F1-score) of next interval symptoms (pain and fatigue) and end-of-day (EOD) functioning (LEF and UEF) using "feature validation set" data.

Feature sets	**BIC score**	**Prediction performance of Bayesian networks (criteria: F1-score)**			
		Next time interval symptoms		**EOD functioning**	
		Pain	**Fatigue**	**LEF**	**UEF**
Complete set of features	**-115,109.9**	**66.90%**	**66.03%**	**81.74%**	**87.08%**
Discarded activity features	-69,250.9	65.02%	63.70%	78.30%	86.74%
Discarded activity and current symptom features	-60,731.7	52.07%	56.85%	78.08%	80.35%
Discarded activity, current symptom, and last reported functioning features.	-28,514.8	51.23%	55.16%	63.64%	55.54%

4.2 Predicting Next Interval Symptoms and EOD Functioning of People with MS

Here, we first evaluated predicting next-time interval pain and fatigue, and EOD LEF and UEF—an intermediate step required before forecasting EOD physical functioning. We validated symptom prediction because we sample symptoms from our BN when forecasting EOD physical functioning. We evaluated the performance of our BN (Fig. 4) trained on our engineered features using 10-fold [58] and leave-one-out (LOOCV) cross-validation [119] using "BN train + validation" set (stratified ≈80% of all records to ensure all MS subtypes were represented). Note that since people reported their functioning data only during bedtime (one time per day), there are fewer LEF and UEF records than reported pain and fatigue records (reported five times per day) in our BN validation data.

We compared the performance of our model with existing supervised ML and deep learning (DL) baselines [36, 37, 80, 92, 123, 125] used to predict MS symptoms and functioning. We optimized each of the baselines using grid

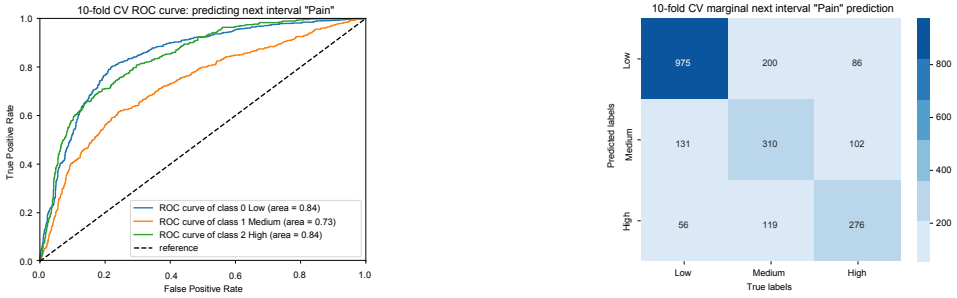
search-based inner-outer CV (Table 17). One baseline in particular (Logistic Regression) was used in the existing approach [9] that we extended. Note that existing supervised ML and DL methods required four separately trained models to predict pain, fatigue, LEF, and UEF, while our method used only a single BN for all predictions.

We reported average prediction performance metrics in Tables 6-9 along with confusion matrices and ROC-curves for 10-fold CV. We conducted LOOCV for completeness and ensured the results were comparable with 10-fold CV; detailed LOOCV results are in the appendix. Here, we focus on 10-fold CV results because uneven distribution of missing self-reports across many participants may lead to high variance in LOOCV performance [13, 42, 58]. Thus, we performed error analysis on incorrectly predicted instances using 10-fold CV results.

4.2.1 Predicting Momentary (Next Interval) Pain. When predicting pain, our BN outperformed all existing baseline models with mean accuracy of 69.28% (Table 6) and ROC AUC of 0.84 (Fig. 5a). Our error analysis showed that most prediction errors with low entropy (i.e., the model was confident about its prediction) occurred in the boundary between low and medium pain levels (Fig. 5b), which explains low ROC AUC for predicting medium pain. Also, next time interval prediction errors with high entropy (i.e., the model was least confident about its prediction) occurred when participants did not record their symptoms in the current time interval.

Table 6. Stratified 10-fold CV performance comparison of BN for predicting pain at each daytime interval.

10-fold model comparison for predicting pain	Mean Precision (95% CI)	Mean Recall (95% CI)	Mean F1-score (95% CI)	Mean Accuracy (95% CI)	Mean ROC AUC (95% CI)
Logistic Regression [37, 123, 125] with naive features	48.52% (43.65%, 53.38%)	54.09% (47.49%, 60.7%)	48.52% (43.65%, 53.38%)	47.61% (42.26%, 52.95%)	0.57 (0.52, 0.61)
Random Forest [92, 123] with naive features	46.3% (39.98%, 52.61%)	53.92% (48.34%, 59.51%)	46.3% (39.98%, 52.61%)	46.47% (41.00%, 51.93%)	0.56 (0.52, 0.60)
Support Vector Machine [37, 125] with naive features	44.57% (37.19%, 51.94%)	51.06% (40.67%, 61.46%)	44.57% (37.19%, 51.94%)	42.88% (36.23%, 49.53%)	0.50 (0.43, 0.58)
Decision Tree [80] with naive features	45.94% (40.05%, 51.84%)	53.79% (48.13%, 59.44%)	45.94% (40.05%, 51.84%)	47.16% (41.72%, 52.60%)	0.57 (0.52, 0.61)
K-nearest Neighbors [37] with naive features	44.17% (38.22%, 50.11%)	51.84% (46.27%, 57.41%)	44.17% (38.22%, 50.11%)	41.96% (36.50%, 47.43%)	0.53 (0.51, 0.56)
Elastic Net [36] with naive features	48.16% (43.19%, 53.13%)	51.35% (42.38%, 60.33%)	48.16% (43.19%, 53.13%)	47.11% (41.44%, 52.78%)	0.56 (0.52, 0.61)
Gradient Boosting [36] with naive features	48.43% (43.59%, 53.27%)	54.0% (47.38%, 60.61%)	48.43% (43.59%, 53.27%)	47.52% (42.2%, 52.84%)	0.57 (0.52, 0.61)
Ridge classifier [37] with naive features	48.91% (43.87%, 53.96%)	52.91% (43.76%, 62.06%)	48.91% (43.87%, 53.96%)	47.15% (41.68%, 52.63%)	0.54 (0.47, 0.61)
Extra Tree classifier with naive features	45.5% (38.68%, 52.31%)	52.08% (46.29%, 57.88%)	45.5% (38.68%, 52.31%)	46.03% (39.82%, 52.24%)	0.56 (0.51, 0.6)
Light GBM with naive features	45.98% (39.48%, 52.49%)	54.22% (46.86%, 61.59%)	45.98% (39.48%, 52.49%)	45.52% (39.53%, 51.5%)	0.57 (0.53, 0.6)
XGBoost with naive features	46.57% (40.92%, 52.22%)	53.62% (47.13%, 60.12%)	46.57% (40.92%, 52.22%)	46.81% (41.67%, 51.95%)	0.57 (0.53, 0.6)
AdaBoost with naive features	48.43% (44.2%, 52.65%)	53.92% (48.32%, 59.53%)	48.43% (44.2%, 52.65%)	47.32% (43.07%, 51.57%)	0.57 (0.54, 0.61)
Multi-Layer Perceptron with naive features	45.9% (39.52%, 52.27%)	53.87% (46.76%, 60.99%)	45.9% (39.52%,52.27%)	45.41% (38.69%, 52.12%)	0.57 (0.53, 0.61)
Naive Bayes with naive features	47.94% (42.48%, 53.39%)	50.74% (44.52%, 56.97%)	47.94% (42.48%, 53.39%)	46.95% (41.05%, 52.85%)	0.54 (0.51, 0.58)
Naive Bayesian Network with naive features	48.12% (37.72%, 58.52%)	52.87% (45.87%, 59.87%)	48.13% (41.50%, 54.76%)	48.85% (41.52%, 56.18%)	0.57 (0.50, 0.64)
Bayesian Network with proposed connections and proposed features	**68.25% (62.18%, 74.33%)**	**69.28% (64.19%, 74.38%)**	**68.12% (62.46%, 73.79%)**	**69.28% (64.19%, 74.38%)**	**0.73 (0.69, 0.77)**



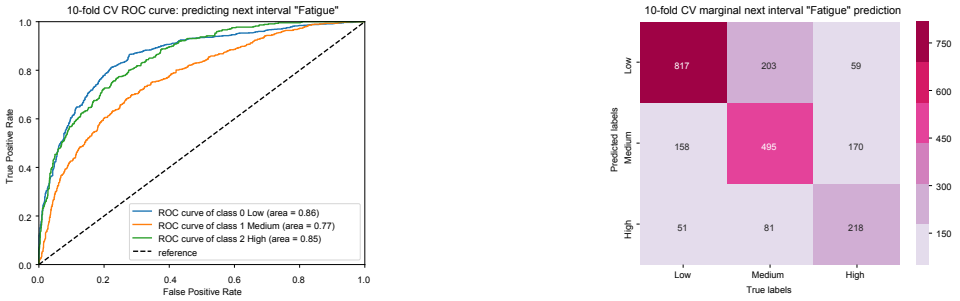
(a) (b)

Fig. 5. Results for 10-fold CV next interval pain prediction using our BN: a) ROC curve, and b) confusion matrix.

4.2.2 Predicting Momentary (Next Interval) Fatigue. When predicting fatigue, our BN outperformed all existing baseline models (Table 7) with mean accuracy of 67.92%. Our model best predicts high fatigue severity with ROC AUC of 0.86 (Fig. 6a). Most prediction errors (Fig. 6b) with low entropy occurred for participants who prefer to report their fatigue in the boundary level (e.g., medium fatigue). Whereas, predicting sudden changes in symptoms (e.g., from low to high fatigue) within an interval led to high entropy prediction errors. In such cases, we suspect that there could be other factors that influence fatigue severity (e.g., sleep disturbance, cognitive load, poor mental functioning) [7, 20, 56] that our model did not have access to from the data.

Table 7. Stratified 10-fold CV performance comparison of BN for predicting fatigue at each daytime interval.

10-fold model comparison for predicting fatigue	Mean Precision (95% CI)	Mean Recall (95% CI)	Mean F1-score (95% CI)	Mean Accuracy (95% CI)	Mean ROC AUC (95% CI)
Logistic Regression [37, 123, 125] with naive features	58.66% (53.98%, 63.34%)	58.56% (53.29%, 63.83%)	58.66% (53.98%, 63.34%)	55.0% (50.27%, 59.72%)	0.58 (0.55, 0.6)
Random Forest [92, 123] with naive features	55.46% (51.54%, 59.38%)	55.08% (50.08%, 60.08%)	55.46% (51.54%, 59.38%)	53.92% (49.47%, 58.37%)	0.58 (0.55, 0.62)
Support Vector Machine [37, 125] with naive features	58.35% (52.25%, 64.44%)	54.58% (47.53%, 61.64%)	58.35% (52.25%, 64.44%)	51.37% (45.16%, 57.57%)	0.54 (0.51, 0.56)
Decision Tree [80] with naive features	52.93% (49.68%, 56.18%)	55.16% (50.2%, 60.12%)	52.93% (49.68%, 56.18%)	52.95% (49.06%, 56.84%)	0.58 (0.56, 0.61)
K-nearest Neighbors [37] with naive features	56.35% (49.36%, 63.33%)	52.6% (45.04%, 60.17%)	56.35% (49.36%, 63.33%)	50.58% (42.74%, 58.42%)	0.55 (0.52, 0.58)
Elastic Net [36] with naive features	58.35% (53.46%, 63.23%)	58.14% (52.81%, 63.47%)	58.35% (53.46%, 63.23%)	54.67% (49.86%, 59.48%)	0.57 (0.55, 0.6)
Gradient Boosting [36] with naive features	58.75% (54.25%, 63.24%)	58.63% (53.13%, 64.13%)	58.75% (54.25%, 63.24%)	55.12% (50.6%, 59.64%)	0.58 (0.55, 0.61)
Ridge classifier [37] with naive features	58.7% (53.14%, 64.26%)	55.45% (50.36%, 60.53%)	58.7% (53.14%, 64.26%)	52.97% (47.31%, 58.64%)	0.57 (0.54, 0.6)
Extra Tree classifier with naive features	56.35% (52.87%, 59.82%)	56.66% (52.21%, 61.11%)	56.35% (52.87%, 59.82%)	55.01% (51.09%, 58.93%)	0.60 (0.56, 0.63)
Light GBM with naive features	56.57% (51.83%, 61.3%)	58.01% (52.43%, 63.58%)	56.57% (51.83%, 61.3%)	55.26% (49.85%, 60.66%)	0.60 (0.56, 0.64)
XGBoost with naive features	56.39% (52.54%, 60.24%)	58.3% (52.54%, 64.07%)	56.39% (52.54%, 60.24%)	55.87% (51.56%, 60.18%)	0.60 (0.56, 0.64)
AdaBoost with naive features	58.7% (52.81%, 64.59%)	58.18% (52.9%, 63.47%)	58.7% (52.81%, 64.59%)	55.36% (49.64%, 61.08%)	0.58 (0.56, 0.61)
Multi-Layer Perceptron with naive features	55.11% (51.6%, 58.61%)	56.95% (52.98%, 60.91%)	55.11% (51.6%, 58.61%)	54.51% (50.69%, 58.33%)	0.60 (0.56, 0.63)
Naive Bayes with naive features	55.07% (48.21%, 61.92%)	56.23% (51.11%, 61.35%)	55.07% (48.21%, 61.92%)	54.07% (48.23%, 59.91%)	0.59 (0.56, 0.61)
Naive Bayesian Network with naive features	61.72% (53.42%, 70.01%)	58.04% (52.03%, 64.04%)	53.79% (44.90%, 62.68%)	58.03% (52.03%, 64.03%)	0.59 (0.54, 0.63)
Bayesian Network with proposed connections and proposed features	**66.93% (60.79%, 73.07%)**	**67.92% (62.30%, 73.55%)**	**64.49% (57.85%, 71.14%)**	**67.92% (62.30%, 73.55%)**	**0.67 (0.62, 0.72)**



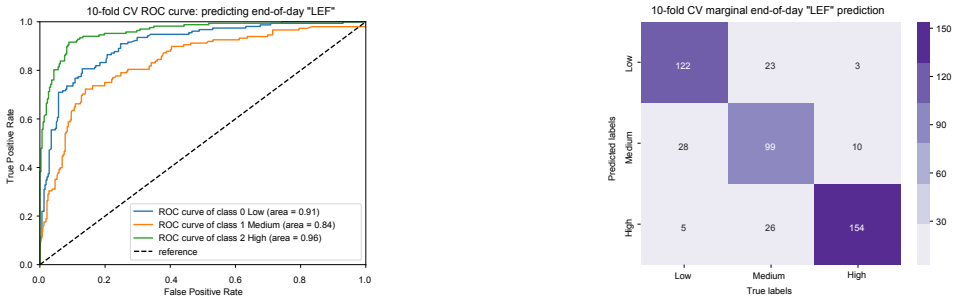
(a) (b)

Fig. 6. Results for 10-fold CV next interval fatigue prediction using our BN: a) ROC curve, and b) confusion matrix.

4.2.3 Predicting EOD LEF. When predicting EOD bedtime LEF, our BN outperformed all existing baseline models (Table 8) with mean accuracy of 80.25%. Our model better predicts high and low LEF (ROC AUC of 0.96 and 0.91 respectively) (Fig. 7a) with very few false positives. Our error analysis shows that most prediction errors occurred in the closest boundary between medium and low LEF levels (Fig. 7b). High entropy LEF prediction errors occurred primarily for participants with missing symptom values. In cases where the participants did not report their symptoms during the evening, high entropy LEF prediction errors occurred. This finding also shows

Table 8. Stratified 10-fold CV performance comparison of BN for predicting EOD bedtime Lower Extremity Functioning (LEF)

10-fold model comparison for predicting LEF	Mean Precision (95% CI)	Mean Recall (95% CI)	Mean F1-score (95% CI)	Mean Accuracy (95% CI)	Mean ROC AUC (95% CI)
Logistic Regression [37, 123, 125] with naive features	49.36% (41.84%, 56.88%)	55.23% (43.82%, 66.63%)	49.36% (41.84%, 56.88%)	48.48% (39.37%, 57.6%)	0.54 (0.46, 0.63)
Random Forest [92, 123] with naive features	54.04% (44.74%, 63.35%)	63.19% (50.47%, 75.92%)	54.04% (44.74%, 63.35%)	54.65% (44.39%, 64.90%)	0.63 (0.56, 0.71)
Support Vector Machine [37, 125] with naive features	46.6% (34.04%, 59.16%)	50.96% (34.29%, 67.62%)	46.6% (34.04%, 59.16%)	45.33% (31.72%, 58.95%)	0.53 (0.41, 0.64)
Decision Tree [80] with naive features	52.13% (43.27%, 60.98%)	59.76% (47.29%, 72.24%)	52.13% (43.27%, 60.98%)	52.19% (42.55%, 61.84%)	0.62 (0.54, 0.7)
K-nearest Neighbors [37] with naive features	45.53% (38.35%, 52.71%)	55.21% (43.29%, 67.14%)	45.53% (38.35%, 52.71%)	45.01% (37.32%, 52.70%)	0.59 (0.53, 0.64)
Elastic Net [36] with naive features	50.64% (41.18%, 60.1%)	55.8% (43.81%, 67.78%)	50.64% (41.18%, 60.10%)	49.68% (39.01%, 60.35%)	0.55 (0.46, 0.64)
Gradient Boosting [36] with naive features	49.36% (41.57%, 57.15%)	54.17% (42.08%, 66.25%)	49.36% (41.57%, 57.15%)	48.55% (39.02%, 58.08%)	0.54 (0.46, 0.63)
Ridge classifier [37] with naive features	51.06% (41.93%, 60.2%)	56.43% (44.8%, 68.06%)	51.06% (41.93%, 60.20%)	49.11% (38.7%, 59.52%)	0.55 (0.46, 0.64)
Extra Tree classifier with naive features	52.98% (45.2%, 60.75%)	60.72% (49.24%, 72.2%)	52.98% (45.2%, 60.75%)	53.22% (44.48%, 61.95%)	0.63 (0.55, 0.7)
Light GBM with naive features	50.43% (41.87%, 58.98%)	58.89% (48.17%, 69.61%)	50.43% (41.87%, 58.98%)	50.37% (41.22%, 59.51%)	0.60 (0.52, 0.68)
XGBoost with naive features	52.34% (44.64%, 60.04%)	61.84% (51.9%, 71.77%)	52.34% (44.64%, 60.04%)	52.6% (44.68%, 60.52%)	0.62 (0.55, 0.69)
AdaBoost with naive features	45.32% (40.95%, 49.69%)	54.63% (43.63%, 65.63%)	45.32% (40.95%, 49.69%)	45.91% (39.63%, 52.2%)	0.58 (0.52, 0.63)
Multi-Layer Perceptron with naive features	50.43% (40.8%, 60.05%)	57.69% (45.98%, 69.39%)	50.43% (40.8%, 60.05%)	51.24% (41.06%, 61.43%)	0.61 (0.52, 0.69)
Naive Bayes with naive features	45.32% (31.69%, 58.95%)	49.97% (31.9%, 68.05%)	45.32% (31.69%, 58.95%)	41.62% (26.09%, 57.15%)	0.54 (0.41, 0.67)
Naive Bayesian Network with naive features	58.73% (49.07%, 68.38%)	52.83% (45.51%, 60.15%)	52.73% (44.82%, 60.63%)	50.79% (43.48%, 58.10%)	0.59 (0.49, 0.69)
Bayesian Network with proposed connections and proposed features	**81.43% (76.20%, 86.65%)**	**80.25% (74.91%, 85.59%)**	**80.01% (74.61%, 85.43%)**	**80.25% (74.91%, 85.59%)**	**0.84 (0.81, 0.87)**



(a) (b)

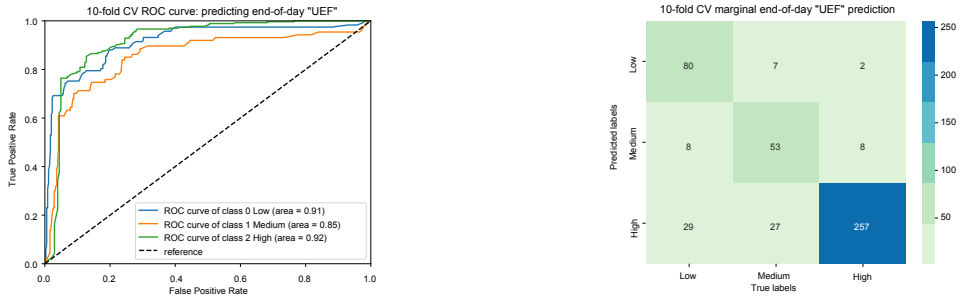
Fig. 7. Results for 10-fold CV EOD LEF prediction using our BN: a) ROC curve, and b) confusion matrix.

the necessity of analyzing meaningful behavior patterns captured by the model, when people with MS fail to self-report their functioning at the end-of-day.

4.2.4 Predicting EOD UEF. When predicting EOD bedtime UEF, our BN outperformed all existing baseline models (Table 9) with mean accuracy of 83.06%. Our model best predicts low and high UEF with ROC AUC of 0.92 and 0.91 respectively (Fig. 8a). Our error analysis shows that UEF prediction error accumulates near the low and medium UEF boundaries (Fig. 8b). It is possible that such errors occurred because our activity features did not have enough granularity to capture fine motor tasks that are characteristic of UEF. Also, high entropy UEF prediction error occurs mostly for participants with a large number of missing actigraphy data.

Table 9. Stratified 10-fold CV performance comparison of BN for predicting EOD Upper Extremity Functioning (UEF)

10-fold model comparison for predicting UEF	Mean Precision (95% CI)	Mean Recall (95% CI)	Mean F1-score (95% CI)	Mean Accuracy (95% CI)	Mean ROC AUC (95% CI)
Logistic Regression [37, 123, 125] with naive features	57.58% (45.39%, 69.78%)	49.58% (34.2%, 64.96%)	57.58% (45.39%, 69.78%)	51.7% (37.66%, 65.74%)	0.52 (0.38, 0.66)
Random Forest [92, 123] with naive features	56.31% (44.18%, 68.43%)	52.72% (35.81%, 69.62%)	56.31% (44.18%, 68.43%)	52.3% (37.60%, 66.99%)	0.55 (0.41, 0.68)
Support Vector Machine [37, 125] with naive features	61.41% (48.71%, 74.12%)	50.62% (34.77%, 66.47%)	61.41% (48.71%, 74.12%)	54.07% (39.9%, 68.25%)	0.53 (0.4, 0.66)
Decision Tree [80] with naive features	55.67% (43.58%, 67.76%)	51.29% (34.11%, 68.48%)	55.67% (43.58%, 67.76%)	51.61% (36.89%, 66.32%)	0.53 (0.39, 0.67)
K-nearest Neighbors [37] with naive features	48.44% (34.65%, 62.22%)	46.8% (28.87%, 64.72%)	48.44% (34.65%, 62.22%)	43.01% (28.56%, 57.45%)	0.44 (0.3, 0.58)
Elastic Net [36] with naive features	57.16% (44.92%, 69.4%)	48.96% (33.52%, 64.4%)	57.16% (44.92%, 69.4%)	51.37% (37.28%, 65.47%)	0.51 (0.37, 0.65)
Gradient Boosting [36] with naive features	54.61% (40.36%, 68.85%)	48.35% (32.08%, 64.62%)	54.61% (40.36%, 68.85%)	49.68% (34.57%, 64.79%)	0.50 (0.38, 0.62)
Ridge classifier [37] with naive features	56.73% (44.15%, 69.32%)	48.59% (32.97%, 64.2%)	56.73% (44.15%, 69.32%)	51.11% (36.84%, 65.38%)	0.51 (0.37, 0.65)
Extra Tree classifier with naive features	56.95% (45.64%, 68.25%)	53.64% (37.53%, 69.75%)	56.95% (45.64%, 68.25%)	52.81% (38.63%, 66.99%)	0.56 (0.43, 0.69)
Light GBM with naive features	50.56% (35.66%, 65.46%)	45.29% (28.26%, 62.32%)	50.56% (35.66%, 65.46%)	45.8% (30.03%, 61.57%)	0.52 (0.37, 0.63)
XGBoost with naive features	54.61% (38.4%, 70.81%)	51.35% (33.05%, 69.64%)	54.61% (38.4%, 70.81%)	51.42% (34.15%, 68.7%)	0.56 (0.42, 0.7)
AdaBoost with naive features	55.25% (40.73%, 69.77%)	49.27% (33.03%, 65.51%)	55.25% (40.73%, 69.77%)	50.41% (35.25%, 65.56%)	0.55 (0.45, 0.65)
Multi-Layer Perceptron with naive features	54.18% (37.06%, 71.3%)	49.2% (29.69%, 68.71%)	54.18% (37.06%, 71.3%)	49.73% (31.22%, 68.23%)	0.54 (0.39, 0.7)
Naive Bayes with naive features	47.19% (30.72%, 63.65%)	45.29% (23.85%, 66.72%)	47.19% (30.72%, 63.65%)	43.56% (24.07%, 63.05%)	0.45 (0.26, 0.64)
Naive Bayesian Network with naive features	59.15% (51.5%, 66.80%)	55.27% (45.29%, 65.25%)	57.74% (48.51%, 66.97%)	53.99% (44.01%, 63.97%)	0.53 (0.46, 0.59)
Bayesian Network with proposed connections and proposed features	**82.60% (75.01%, 90.20%)**	**83.06% (76.46%, 89.67%)**	**81.71% (74.65%, 88.77%)**	**83.06% (76.46%, 89.67%)**	**0.85 (0.80, 0.90)**



(a) (b)

Fig. 8. Results for 10-fold CV EOD UEF prediction using our BN: a) ROC curve, and b) confusion matrix.

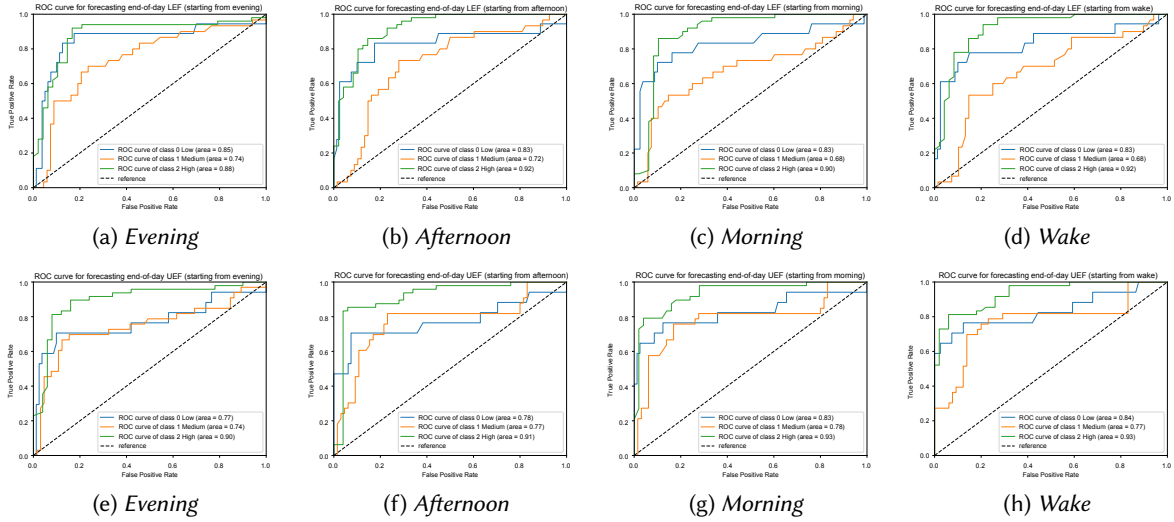
4.3 Forecasting end-of-day (EOD) Functioning of People with MS

Here, we evaluated our model's ability to accurately forecast EOD physical functioning (LEF and UEF) using model-unseen participants' data within the same cohort from the "IRL forecasting test set" (stratified ≈20% of all records). Note that when forecasting EOD functioning, prediction error compounds at each forecast step irrespective of the model choice. In addition, time granularity (e.g., start and end intervals of forecasts) as well as the information available to the model (e.g., self-reports, actigraphy) during different intervals of forecasting could influence the forecasting performance.

Thus, we evaluated our model in two scenarios: 1) with both subjective and passive sensing data (i.e., complete data) starting at each daytime interval (e.g., wake, morning, afternoon, and evening) given available data up to that starting point in time, and 2) with subjective data only collected at the wake time interval and using only passive sensing data (e.g., activity bouts, pace) for the rest of the day. We compared our model to an algorithm based on an existing LSTM architecture for sensor-based human activity recognition (HAR) [25]. We also performed an error analysis to show the feasibility of our method to forecast EOD functioning as early as when people with MS wake up to support timely interventions.

Table 10. Validation of Forecasting with Complete Data: Performance comparison of EOD bedtime LEF and UEF starting from any daytime interval given available data up to that starting point in time using model-unseen "IRL test set".

Model	Starting interval	Forecasting performance (Bedtime LEF)					Forecasting performance (Bedtime UEF)				
		Precision	Recall	F1-score	Accuracy	ROC AUC	Precision	Recall	F1-score	Accuracy	ROC AUC
LSTM [25]	Evening	64.98%	58.16%	59.62%	58.16%	0.65	59.91%	62.24%	59.20%	62.24%	0.68
	Afternoon	63.21%	55.10%	57.33%	55.10%	0.64	60.53%	61.22%	57.79%	61.22%	0.68
	Morning	62.19%	54.08%	56.31%	54.08%	0.61	56.25%	56.12%	52.91%	56.12%	0.62
	Wake	56.72%	46.93%	49.18%	46.93%	0.58	52.45%	48.01%	48.43%	47.95%	0.61
Our model	Evening	**69.83%**	**65.31%**	**66.26%**	**65.31%**	**0.75**	74.96%	71.43%	72.36%	71.43%	0.79
	Afternoon	68.16%	61.22%	62.5%	61.22%	0.72	75.86%	72.45%	73.49%	72.45%	0.78
	Morning	66.05%	57.14%	58.61%	57.14%	0.69	**78.61%**	**74.49%**	**75.51%**	**74.49%**	**0.81**
	Wake	65.98%	57.14%	58.24%	57.14%	0.70	77.04%	71.43%	72.55%	71.43%	0.79



(a) *Evening* (b) *Afternoon* (c) *Morning* (d) *Wake*

(e) *Evening* (f) *Afternoon* (g) *Morning* (h) *Wake*

Fig. 9. Validation of Forecasting with Complete Data: ROC curves when forecasting EOD bedtime LEF (a, b, c, d) and UEF (e, f, g, h) starting from any daytime intervals.

4.3.1 Validation of Forecasting with Complete Data. Our results showed that our IRL-based forecasting algorithm (given only start interval self-reports) outperformed LSTM (given available data up to that starting point in time to look back) in forecasting end-of-day (bedtime) functioning with the best ROC AUC of 0.75 (LEF) and 0.81 (UEF) (Table 10). This experiment showed that our algorithm can forecast EOD functioning in advance starting from any daytime interval (e.g., wake, morning, afternoon, and evening) given only start-interval self-report data, which reduces the need for repeated self-reports for the rest of the day.

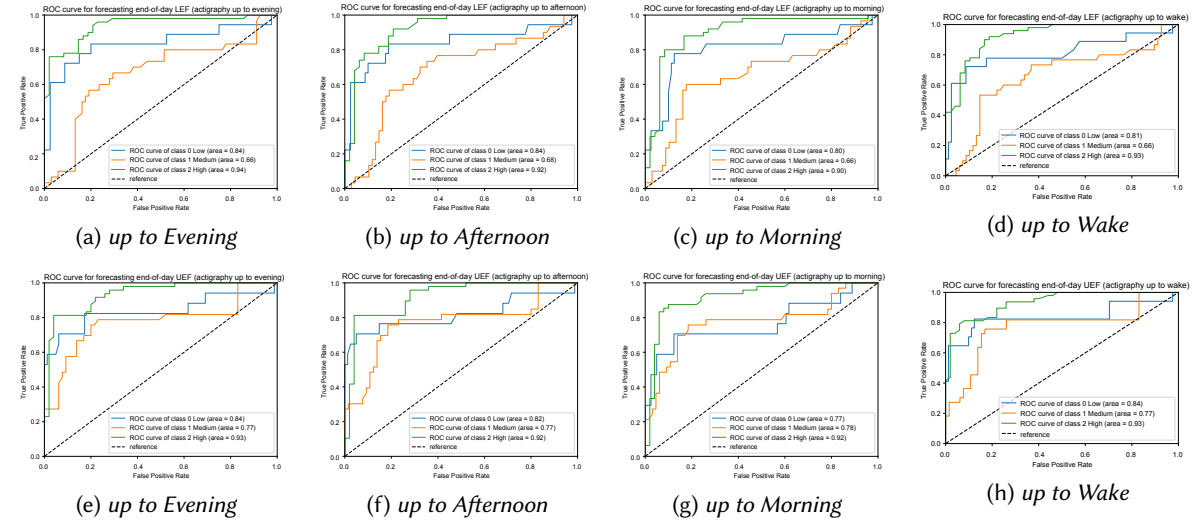
Our summary statistics show that LEF forecasting performance gradually decreased the further we started from the bedtime interval with the highest ROC AUC of 0.75 for evening as the start interval (Table 10). Thus, LEF appears to be associated with symptoms and behaviors that occur closer to bedtime. ROC curves (Fig. 9) show that our model makes most errors when true EOD LEF was medium. Forecasting performance for low and high LEF was relatively high regardless of which daytime interval we started our forecast in. Our error analysis found that most forecasting errors occurred when participants failed to report their symptoms at the start of our forecast. Our model performance was stable for each successive step of forecasting UEF with the highest ROC AUC of 0.81. Thus, UEF appears to be less influenced by symptom changes across daytime intervals and depends more on the previous day's UEF.

4.3.2 Validation of Forecasting with Only Passive Sensing Data. Our results showed that our model outperformed LSTM in forecasting end-of-day (bedtime) functioning with the best ROC AUC of 0.78 (LEF) and 0.82 (UEF) (Table 11). This experiment showed that our model can forecast EOD functioning as early as when people with MS wake up given only passive actigraphy data without asking for any self-reports throughout the day. Our results show that our model performance improved for both LEF and UEF when passive actigraphy data was available unlike our previous forecasting experiment, where only initial self-reports were given to our model.

Our summary statistics and ROC curves (Fig. 10) show that LEF forecasting performance started decreasing when afternoon and evening actigraphy data were not given. Thus, LEF appears to be more associated with actigraphy data closer to bedtime (e.g., afternoon and evening). Similar to our previous forecasting experiment, our model performance was stable for UEF with the highest ROC AUC of 0.82. Our results show that the availability of passive actigraphy data throughout the day improves UEF forecasting performance.

Table 11. Validation of Forecasting with Only Passive Sensing Data: Performance comparison of EOD bedtime LEF and UEF starting from wake interval with subjective data only collected at the wake time interval and using only passive sensing data (e.g., activity bouts, pace) for the rest of the day up to different daytime intervals from model-unseen "IRL test set".

Model	Actigraphy data up to	Forecasting performance (Bedtime LEF)					Forecasting performance (Bedtime UEF)				
		Precision	Recall	F1-score	Accuracy	ROC AUC	Precision	Recall	F1-score	Accuracy	ROC AUC
LSTM [25]	Evening	61.02%	56.22%	57.22%	56.22%	0.64	54.89%	57.14%	52.43%	57.14%	0.63
	Afternoon	59.28%	56.12%	57.19%	56.12%	0.61	50.06%	52.04%	48.09%	52.04%	0.61
	Morning	59.50%	54.08%	55.40%	54.08%	0.60	49.50%	51.02%	48.01%	51.02%	0.59
	Wake	51.22%	41.84%	43.64%	41.83%	0.56	49.38%	46.93%	46.96%	46.93%	0.56
Our model	Evening	**68.53%**	**64.86%**	**63.27%**	**71.63%**	**0.78**	74.49%	75.42%	74.49%	77.51%	**0.82**
	Afternoon	61.22%	62.51%	61.22%	69.45%	0.74	73.47%	74.52%	73.47%	78.06%	0.82
	Morning	60.20%	62.14%	60.20%	68.52%	0.71	**76.53%**	**77.46%**	**76.53%**	**79.79%**	0.80
	Wake	63.27	59.46%	58.16%	68.52%	0.71	75.51%	76.49%	75.51%	79.19%	0.81



(a) *up to Evening* (b) *up to Afternoon* (c) *up to Morning* (d) *up to Wake*

(e) *up to Evening* (f) *up to Afternoon* (g) *up to Morning* (h) *up to Wake*

Fig. 10. Validation of Forecasting with only Passive Sensing Data: ROC curves when forecasting EOD bedtime LEF (a, b, c, d) and UEF (e, f, g, h) starting from wake interval given actigraphy data up to a daytime interval.

5 QUALITATIVE MODEL PERFORMANCE EVALUATION

Here, we present further error analysis of our forecasting method to understand where forecasting errors from our evaluation originated. Although our quantitative results showed high performance compared to existing

baselines, solely quantitative analysis can be misleading [41, 86]. Thus, to further debug our model, we conducted qualitative visual explorations of model-simulated behaviors with the model designers on our team. In addition, our domain expert assisted with formulating contextual scenarios to focus our exploration on relevant parts of the large state space of possible situations people with MS can find themselves in. This allowed us to explore how well the model captures actual behaviors of people with MS, in addition to simple error analysis of instances where our model made the wrong forecasts.

5.1 Probability Sampling, Contextual Scenarios, and Visualization Properties

The domain expert on our team suggested contextual scenarios for visualizing and understanding the influence of people's behavior on their MS symptoms and end-of-day functioning. Here, we present a subset of most relevant scenarios (Table 12) to illustrate the approach and our findings. For each scenario, we used our generative model to sample N behavior instances throughout the day based on the possible combinations of feature variables (n) that minimized the absolute error ($\epsilon = 0.05$) of our sample joint probability distribution estimates under fixed confidence ($1 - \delta = 0.8$) [65]. Then, we filtered behavior instances from model samples and data (containing model-unseen populations) based on each "what-if" condition.

Table 12. Experimental setup for generating and simulating behavior instances in contextual scenarios suggested by the domain expert. We selected only EOD LEF as the outcome feature for scenario 3 and 4 as symptom onsets at the end-of-day mostly impacts people's lower extremity functioning (LEF).

Scenario	Selected features			Feature value combinations number (n)	Sample size (N)		Absolute error (ϵ)		Confidence ($1 - delta$)		Figure	
	Time independent	Time dependent	Outcome		Data	Model	Data	Model	Data	Model	Data	Model
1. Influence of personal factors on end-of-day functioning	1. Gender 2. Age	N.A	End-of-day (EOD) lower extremity functioning (LEF)	24	693	3251	0.11	0.05	0.8	0.8	Fig. 11a	Fig. 11b
			End-of-day (EOD) upper extremity functioning (LEF)	24	693	3251	0.11	0.05	0.8	0.8	Fig. 12a	Fig. 12b
2. Influence of health conditions on end-of-day functioning	1. MS subtype 2. Usage of mobility aids	N.A	EOD LEF	16	693	2474	0.09	0.05	0.8	0.8	Fig. 13a	Fig. 13b
			EOD UEF	16	693	2474	0.09	0.05	0.8	0.8	Fig. 14a	Fig. 14b
3. What-if people start their day feeling high fatigue during wake and morning interval	1. MS subtype = Relapsing Remitting 2. Usage of mobility aids = No	1. Fatigue (high at wake and morning) 2. Activity bouts	EOD LEF	46	17	4718	0.83	0.05	0.8	0.8	Fig. 15a	Fig. 15b
4. What-if people start their day feeling low fatigue during wake and morning interval	1. MS subtype = Relapsing Remitting 2. Usage of mobility aids = No	1. Fatigue (low at wake and morning) 2. Activity bouts	EOD LEF	46	188	4718	0.25	0.05	0.8	0.8	Fig. 16a	Fig. 16b

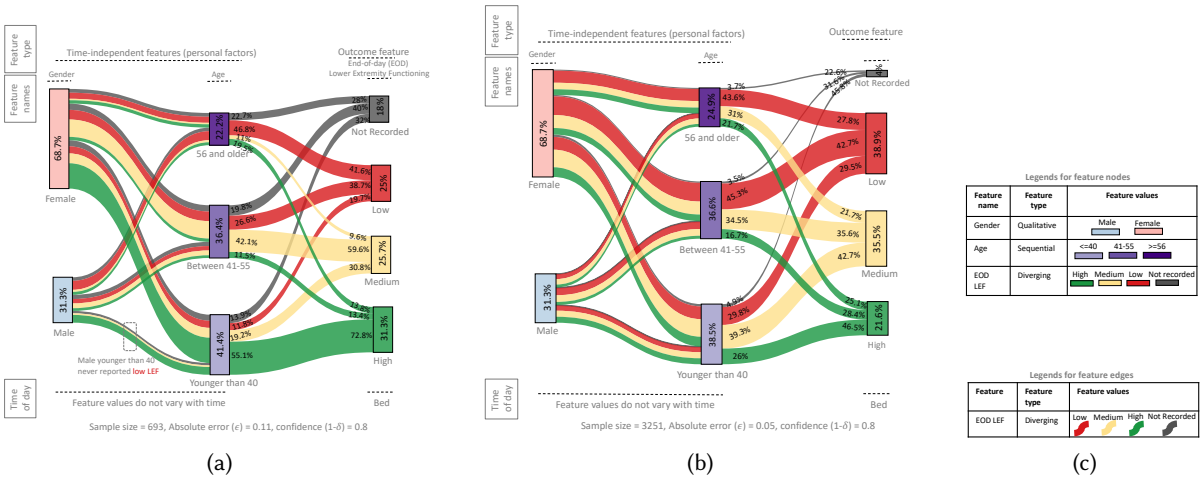
For visualizing each scenario, we selected features from three categories: 1) Time-independent features (e.g., gender, age, MS subtype, usage of mobility aids), 2) Time-dependent features (e.g., pain, fatigue, activity bouts, pace), and 3) outcome features (e.g., end-of-day lower extremity functioning (EOD LEF) and upper extremity functioning (EOD UEF)). Our visualizations (Fig. 11- 16) display the relationship between these features and end-of-day functioning at different times of the day using Sankey plots, where rectangular node heights represent the distribution of each feature value and arc-shaped edge widths represent the source of outcome feature values between two neighboring sources and target feature nodes from left to right.

We display the time information (e.g., wake, morning, afternoon, evening, and bed) in the X-axis of the Sankey plot for each type of feature value. Furthermore, we selected: 1) sequential color palettes for the nodes corresponding to age, fatigue, and activity bouts, 2) qualitative palettes for gender and MS subtype, and 3)

diverging palettes for the EOD LEF and UEF. We choose red, yellow, green, and grey colored edges to represent low, medium, high, and not recorded functioning outcomes across different feature values. The two authors who designed and implemented the models presented and discussed the scenarios and corresponding visualizations with the author with clinical expertise who evaluated whether the model captured meaningful behaviors and existing clinical knowledge. Here, we document and explain the four most relevant scenarios from the data ("full set": 99 participants' data with 693 behavior instances) and the model recommended by the domain expert.

5.2 Scenario 1: Influence of Personal Factors on end-of-day (EOD) Functioning

Despite MS not being directly inherited or contagious, epidemiologists (who study MS disease patterns) identified that personal factors (e.g., gender, age) could impact end-of-day functioning of people with MS [107]. Therefore, we validated that our model captured the influence of gender and age on people's end-of-day (EOD) lower extremity functioning (LEF) (Fig. 11) and upper extremity functioning (UEF) (Fig. 12) compared to the data it was trained on.



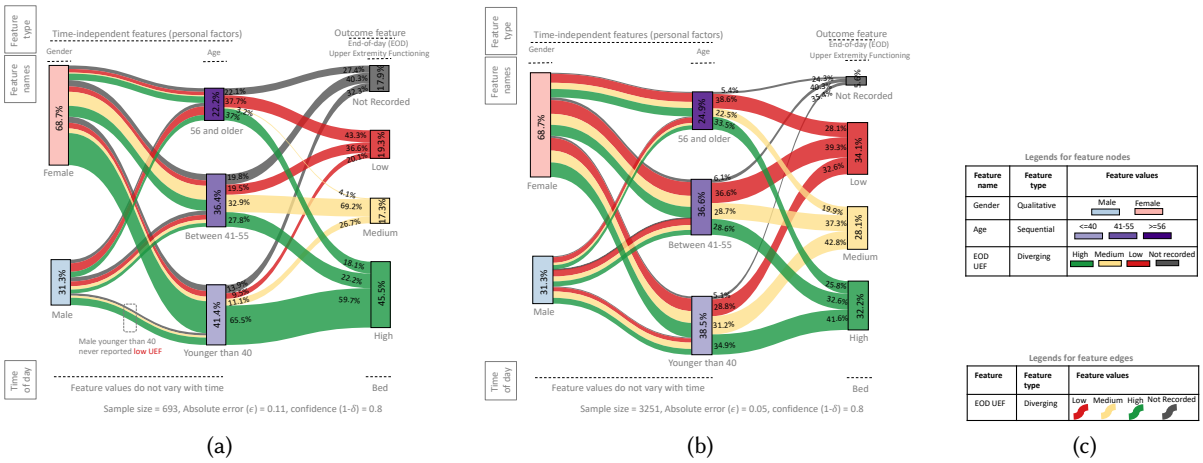
(a) (b) (c)

Fig. 11. Scenario 1 for EOD LEF: Visualizing the influence of personal factors (gender, age) on end-of-day lower extremity functioning (EOD LEF) for: a) 693 behavior instances from the data, and b) 3,251 behavior instances sampled from the model.

Our first observation was that the model correctly captured the proportion of gender at birth and ages in the data. Behavior instances from the data and model-generated samples indicate that older people are more likely to report low functioning, indicating a lower susceptibility to MS among older adults [107]. Also, we found an overestimation of behavior instances with low EOD functioning in the model compared to the data, which likely resulted from our model trying to make a prediction for functioning values that were not recorded in the data. Also, our model could sample from parts of the state space unrepresented in the data. For example, in the data, males younger than 40 never reported low EOD LEF and UEF; our model can sample what happens when they do.

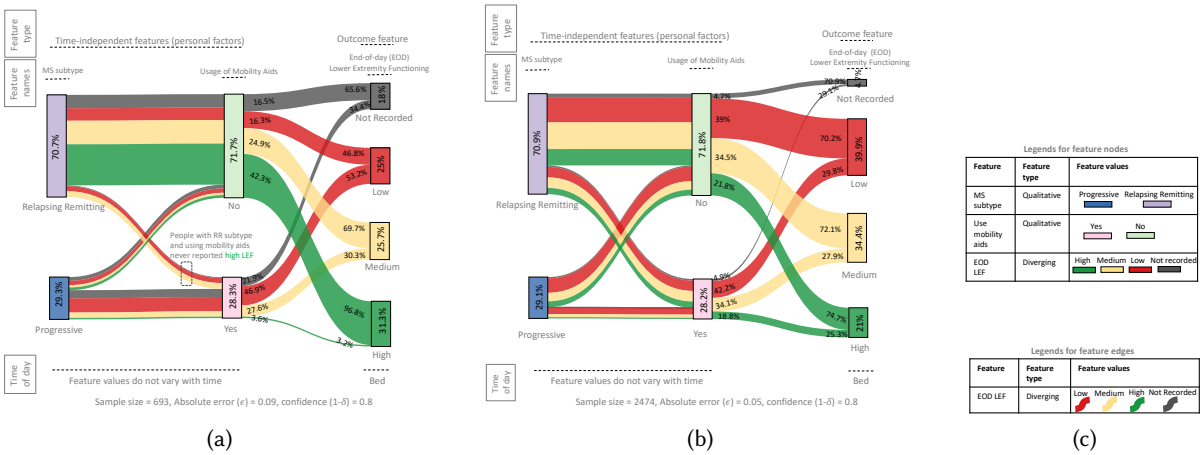
5.3 Scenario 2: Influence of Health Conditions on end-of-day (EOD) Functioning

During clinical visits, health conditions (e.g., MS subtype, usage of mobility aids) constitute the most perceptible features that clinicians use to determine the pattern and progression of MS. Therefore, we compared visualizations from the model and data to understand whether our model captured the influence of MS subtype and mobility aids (e.g., wheelchair, cane) on people's EOD functioning (Fig. 13 and Fig. 14).



(a) (b) (c)

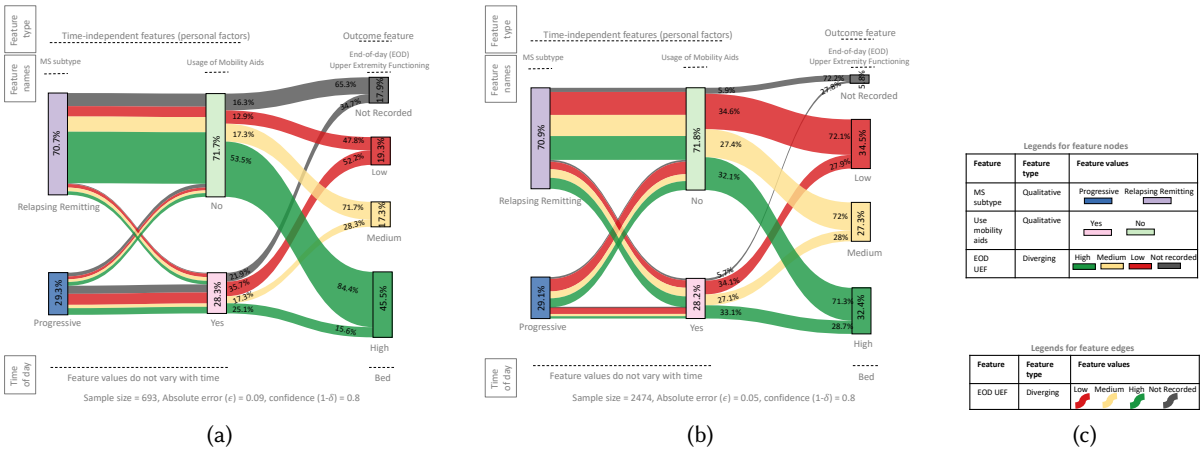
Fig. 12. Scenario 1 for EOD UEF: Visualizing the influence of personal factors (gender, age) on end-of-day upper extremity functioning (EOD UEF) for: a) 693 behavior instances from the data, and b) 3,251 behavior instances sampled from the model.



(a) (b) (c)

Fig. 13. Scenario 2 for EOD LEF: Visualizing the influence of health condition (MS subtype, usage of mobility aids) on end-of-day lower extremity functioning (EOD LEF) for: a) 693 behavior instances from the data, and b) 2,474 behavior instances sampled from the model.

We observed that our model correctly captured the distributions of MS subtype and usage of mobility aids in the data. Data and model both indicate an elevated distribution of high UEF values relative to high LEF values, indicating people using mobility aids face more challenges performing lower-extremity activities (e.g., mobility). Contrary to the data, our model forecasted that people with progressive MS are less likely to use mobility aids. This error could originate from a missing connection between MS subtype and usage of mobility aids features in our Bayesian network. However, this error did not affect our model's prediction and forecasting abilities as we set these two feature values from the data as initial state variables. Nevertheless, our visualizations helped us identify errors in the model that were otherwise difficult to spot using quantitative performance evaluation.

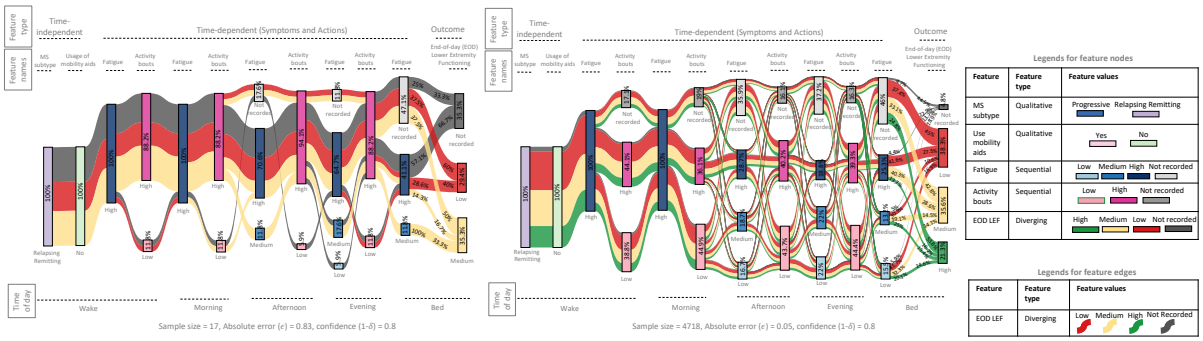


(a) (b) (c)

Fig. 14. Scenario 2 for EOD UEF: Visualizing the influence of health condition (MS subtype, usage of mobility aids) on end-of-day upper extremity functioning (EOD UEF) for: a) 693 behavior instances from the data, and b) 2,474 behavior instances sampled from the model.

5.4 Scenario 3: What-if People Start their Day Feeling High Fatigue

Existing research showed that changes in MS symptoms (e.g., fatigue [104]) throughout the day could significantly interfere with people's daily functional ability at home and work [61, 122]. Clinical research [105] suggests that managing activity intensity and pacing after symptom onsets help reduce the probability of low functioning. Therefore, we analyzed the influence of activity bouts (intensity) and fatigue levels on end-of-day (EOD) lower extremity functioning (LEF) when people start their day (wake and morning) feeling high fatigue (Fig. 15). We already established that people who use mobility aids are more likely to report lower LEF. Therefore, to avoid any confounds from the model predicting that people with progressive MS are disproportionately likely to not use mobility aids, we scoped our further analysis to behaviors from people with the Relapsing-Remitting subtype who do not use mobility aids.



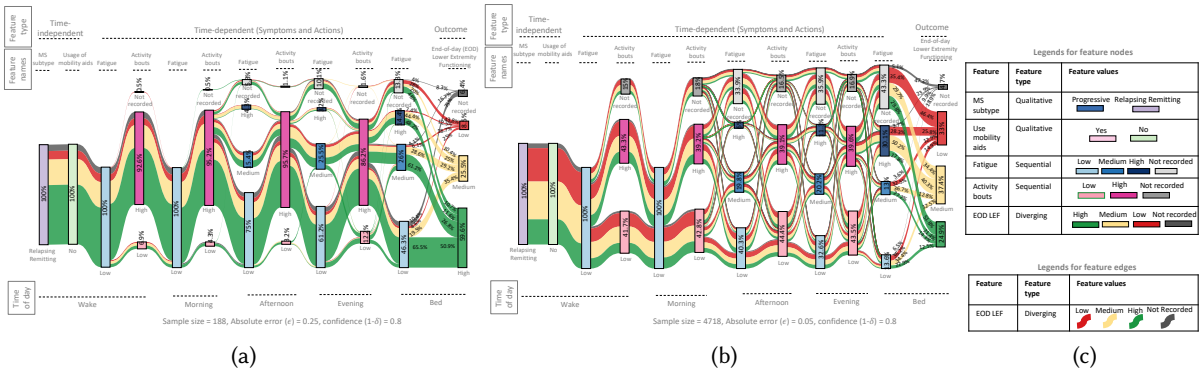
(a) (b) (c)

Fig. 15. Scenario 3: Visualizing the influence of fatigue and activity bouts on end-of-day lower extremity functioning (EOD LEF) when people start their day feeling symptomatic (high fatigue during wake and morning interval). Visualizations are shown for: a) 17 behavior instances from the data (out of 693), and b) 4,718 behavior instances sampled from the model.

Behavior instances from data (Fig. 15a) show that a high level of fatigue at the beginning will *always* end the day with low or medium functioning for people with MS. However, it is challenging to generate insights based on only 17 behavior instances (out of 693) from the data without any example of high EOD LEF. Even in this underrepresented scenario, our model-generated samples (Fig. 15b) provided insights into the possibility of recovery (e.g., high functioning) by showing some salient paths (thick edges) of high LEF for people who start their day with high fatigue. Model samples showed that people are more likely to have high fatigue during bedtime and low EOD functioning if they continuously perform high activity bouts. In contrast, few salient paths from the model indicated that people who manage and lower their activity bouts throughout the day could experience lower fatigue during bedtime and achieve high EOD LEF. Also, our model detected an increase in missing fatigue values before bedtime, following a similar trend in the data, and demonstrated people with end-of-day missing fatigue values are likely to end their day with low EOD LEF. This highlights a use-case where our model could be used to detect high fatigue during wake and morning intervals and send people with MS or their caregivers reminders about "just-in-time clinical" interventions to help improve EOD functioning.

5.5 Scenario 4: What-if People Start their Day Feeling Low Fatigue

According to our domain expert, minimal symptom levels during the morning boost people's endurance and active time [104]. However, controlling energy expenditure based on daily activity intensity is necessary to manage MS episodes and exacerbation. Therefore, we assessed the effect of activity bouts (intensity) and fatigue levels on EOD LEF when people start their day (wake and morning) with low fatigue levels (Fig. 16).



(a) (b) (c)

Fig. 16. Scenario 4: Visualizing the influence of fatigue and activity bouts on end-of-day lower extremity functioning (EOD LEF) when people start their day with minimal symptoms (low fatigue during wake and morning interval). Visualizations are shown for: a) 188 behavior instances from the data (out of 693), and b) 4,718 behavior instances sampled from the model.

Behavior instances from data (Fig. 16a) show that people are more likely (188 behavior instances out of 693) to start their day with low fatigue and mostly report high EOD LEF. However, our model-generated samples (Fig. 16b) show that people can still have low EOD LEF despite starting the day with minimal fatigue. In both data and the model, we observed that high activity bouts and fatigue lead the salient paths (thick edges) toward low EOD LEF. Conversely, the salient pathways leading to high EOD LEF involve episodes of low activity bouts and low fatigue. The domain expert on our team pointed out that people may overdo high activity bouts if they start the day with minimal fatigue. Thus, continuous high activity bouts can exacerbate the progressions of fatigue, which could explain our model overestimating low EOD LEF in this scenario. The ability of our model to forecast low EOD LEF when people overdo activities informs a use-case where the model could be used to detect when to encourage people with MS to manage their activity intensity and pace.

6 DISCUSSION

Our work showed the feasibility of formalizing the behaviors of people with MS in a computational model to enable forecasting their end-of-day (EOD) physical functioning as early as when they wake up. Our results are the first step towards more accurate forecasting of the functioning of people with MS that can be used to deploy real-world just-in-time interventions. In particular, our results point to additional features that could further improve the performance of our algorithm. Our feature ablation experiment established that demographics and contextual factors alone cannot predict MS symptoms and functioning. Instead, our results showed the value of self-reported symptoms and functioning as well as passively sensed activity data.

However, despite its value, collecting self-reported symptoms five times per day could be tedious and burdensome. Our results showed that our method can forecast EOD physical functioning without the need to repeatedly collect self-reported symptoms. Our method samples possible symptoms in next time intervals using our BN. Although the performance of our BN when predicting the highest probability next time interval pain and fatigue has room for improvement, our forecasting evaluation showed that the forecasting performance of our algorithm was not affected by it. This is because our forecasting algorithm does not use only the most likely symptom level—it uses the probability distribution over different symptom levels. This is supported by our findings that most of the symptom prediction errors happened in the boundary between symptom values. In such cases, the next most probable symptom value was almost always the correct one.

Our model performance did not degrade in our carefully designed evaluation experiments (LOOCV, 10-fold CV) following the best practices from applied ML, thus showing the stability and feasibility of our algorithm. We argued that uneven distributions of missing self-reported values across such a large number of participants in the data may lead to a high variance in model performance estimated using LOOCV, even when such results are as good as ours. There are potential future scopes to tackle this problem by improving the data collection protocol to reduce the number of missing values and generate synthetic data by the model that captures behavior characteristics behind missing value generation.

Nevertheless, our results point to ways to improve symptom prediction. Our error analysis showed that most prediction errors occurred in the symptom value boundaries (e.g., between low and medium, medium and high) and during sudden shifts between low and high symptom levels. We suspect that in such cases, other factors (such as sleep disturbances, cognitive load, and poor mental functioning) [7, 20, 56] that were not present in our dataset could affect symptom severity. Such improvement is particularly important because our evaluation showed that it is difficult to predict the bedtime EOD functioning of participants who have a higher percentage of missing self-reported data.

Our findings showed the value of using passively collected activity intensity and pacing data to forecast physical functioning. However, our evaluation also showed that the activity features contributed only a small increase in forecasting performance of LEF and UEF despite clinical research suggesting a stronger relationship [3, 61, 95]. Instead, in our forecasts, UEF depended more on the previous day's UEF, and LEF depended more on self-reported symptoms data closer to bedtime. One way to improve forecasting performance could be to collect actigraphy features at a finer granularity (e.g., sampling at a higher frequency) combined with activity recognition (e.g., detecting Activities of Daily Living that participants performed at different times of day).

Our forecasting algorithm was most accurate at forecasting high and low levels of LEF and UEF (with AUC ROC ranging from 0.8 to over 0.9). Accurately forecasting those two functioning values is the most important prerequisite for deciding to recommend just-in-time interventions when people with MS need them the most (e.g., when they are at risk of reduced functioning) with few false alarms (e.g., notifying them to reduce their activity intensity or to pace themselves unnecessarily).

We further qualitatively evaluated our model using visual model exploration, which was meant to aid model designers on our team to detect errors by visually exploring patterns of behavior captured in our model. Such

errors could otherwise be difficult to spot using solely quantitative performance evaluation. We did not aim to generate new medical insight from our visual explorations of contextual scenarios. Instead, we aimed to validate (from the model designer's perspective) whether our model provided reasonable functioning level predictions that reflect *already known* clinical insights.

Despite our careful feature engineering, principled human-centered behavior modeling, and comprehensive evaluation, our approach comes with some limitations. For example, we trained and evaluated our model on a single cohort of 99 people with MS. Although we used one of the largest available datasets, more data will almost always lead to better model performance and a lower chance of overfitting. However, such limitation is likely to persist in practice too, since it is exceptionally challenging and time-consuming to collect data in the wild for people with clinically confirmed chronic conditions, such as MS. As a result, data availability and the size needed to train and test models that forecast functioning of people with MS will always be a challenge.

To account for the lack of other datasets, we followed existing applied machine learning best practices (record-wise and subject-wise data splitting) to generate separate sets for feature selection, model development, training, and validation to reduce the effects of overfitting. We ensured that our training, validation, and test sets had coverage across labels using stratification based on MS subtype. To test for the generalizability of our model to other people with MS within our clinical population, we used data from the model-unseen hold-out test set to minimize the chances of identity confounding problems [77].

Although our method is general enough to generalize to modeling other cohorts of people with MS and potentially apply to people with other similar chronic conditions, we make no attempt to show any evidence for the transfer of knowledge to any other population (e.g., people with MS from another clinical center). In precision health, translating the predictive knowledge of a behavior model across various cohorts of patients with different environmental conditions from various healthcare institutions (e.g., hospitals, medical centers) is a known challenge [101]. For example, a model that can predict the mortality rate in ICU for one hospital may not predict the same with similar accuracy in other hospitals [38, 96]. There is scope for improving the ML model's ability to transfer clinical knowledge across populations in the future for ensuring broader adoption of ML.

7 CONCLUSION AND FUTURE WORK

In this paper, we presented our human-centered method for modeling behaviors of people with MS to forecast their physical functioning. Although our model is designed specifically to forecast the functioning of people with MS, the potential applications of our method to forecast the functioning of people with similar neurological and chronic conditions are evident. Our comprehensive model evaluation, which included quantitative and qualitative visual model exploration has implications for human-centered applied ML approaches to support future ubiquitous computing applications (e.g., activity-based ubiquitous computing [68], human-centered machine learning (HCML) [40, 97], eXplainable Activity Recognition (XAR) [15]).

Our findings motivate future research that should explore and identify features that could further improve our forecasting performance (e.g., sleep quality, traces of fine upper extremity movement, ADLs). This has implications not only for what data to include in future data collection but also for future activity sensing and smart ubiquitous tracking technology used to collect the data. For example, we identified that coarse actigraphy data sampled at low frequencies provide poor physical activity features to predict upper extremity functioning (UEF) associated with fine motor skills. Future research could explore if higher frequency actigraphy combined with cost-efficient privacy-preserving ubiquitous sensing [48] could generate data that improve our forecasting performance. However, it is also important to consider the computational costs associated with additional features that influence each other (e.g., activities, symptoms, and functioning).

Our work currently targets model developers by providing limited interaction (i.e., interpreting static visualizations), which can help them debug the model compared to data. Such detailed error analysis would not have been

possible using only quantitative methods or would have been otherwise very tedious. However, our work has implications for designing future user interfaces that will target clinicians and end-users (e.g., people with MS). For example, our qualitative validation through static visualizations could inform the design of future interactive visualization tools that allow clinicians to interact with the model, explain model decisions, and identify the effects of such models on various stakeholders. However, designing such interfaces is not trivial and it will require iterative design and evaluation of different prototypes at different levels of fidelity.

Our research informs the design of behavior-aware user interfaces that deliver just-in-time clinical interventions to help people with MS manage their symptoms and physical functioning before they worsen. Thus, future research could explore how to design interfaces to deliver interventions, such as "day planning" [35, 45, 60, 64, 72, 103] and "activity pacing" [1–3], based on our model forecasts and validate such interventions in a clinical setting. Our work could inform the design and development of ubiquitous computing systems and interactive behavior-aware technologies to enable more people with MS to access healthcare and self-manage their functioning.

ACKNOWLEDGMENTS

Toyota Research Institute (TRI) provided funds to assist the authors with their research, but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. We thank the reviewers, associate editors, and members of Computational HCI Lab and Interactive Sensing and Computing Lab for their insightful comments that helped us improve this paper. Anindya Das Antar would like to thank his wife Purba.

REFERENCES

[1] Ulric S Abonie, Andrew M Edwards, and Florentina J Hettinga. 2020. Optimising activity pacing to promote a physically active lifestyle in medical settings: A narrative review informed by clinical and sports pacing research. *Journal of sports sciences* 38, 5 (2020), 590–596.

[2] Ulric S Abonie and Florentina J Hettinga. 2021. Effect of a tailored activity pacing intervention on fatigue and physical activity behaviours in adults with multiple sclerosis. *International Journal of Environmental Research and Public Health* 18, 1 (2021), 17.

[3] Ulric S Abonie, Femke Hoekstra, Bregje L Seves, Lucas HV van der Woude, Rienk Dekker, and Florentina J Hettinga. 2020. Associations between Activity Pacing, Fatigue, and Physical Activity in Adults with Multiple Sclerosis: A Cross Sectional Study. *Journal of Functional Morphology and Kinesiology* 5, 2 (2020), 43.

[4] Parastoo Abtahi, Victoria Ding, Anna C. Yang, Tommy Bruzzese, Alyssa B. Romanos, Elizabeth L. Murnane, Sean Follmer, and James A. Landay. 2020. Understanding Physical Practices and the Role of Technology in Manual Self-Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 115 (Dec. 2020), 24 pages. https://doi.org/10.1145/3432236

[5] Imad Alawe, Adlen Ksentini, Yassine Hadjadj-Aoul, and Philippe Bertin. 2018. Improving traffic forecasting for 5G core network scalability: A machine learning approach. *IEEE Network* 32, 6 (2018), 42–49.

[6] Deborah Antcliff, Malcolm Campbell, Steve Woby, and Philip Keeley. 2015. Assessing the psychometric properties of an activity pacing questionnaire for chronic pain and fatigue. *Physical Therapy* 95, 9 (2015), 1274–1286.

[7] AM Bamer, KL Johnson, D Amtmann, and GH Kraft. 2008. Prevalence of sleep problems in individuals with multiple sclerosis. *Multiple Sclerosis Journal* 14, 8 (2008), 1127–1130.

[8] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K. Dey. 2016. Modeling and Understanding Human Routine Behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 248–260. https://doi.org/10.1145/2858036.2858557

[9] Nikola Banovic, Anqi Wang, Yanfeng Jin, Christie Chang, Julian Ramos, Anind Dey, and Jennifer Mankoff. 2017. Leveraging Human Routine Models to Detect and Generate Human Behaviors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6683–6694. https://doi.org/10.1145/3025453.3025571

[10] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[11] Liliana Barrios, Pietro Oldrati, Marc Hilty, David Lindlbauer, Christian Holz, and Andreas Lutterotti. 2021. Smartphone-Based Tapping Frequency as a Surrogate for Perceived Fatigue: An in-the-Wild Feasibility Study in Multiple Sclerosis Patients. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 89 (Sept. 2021), 30 pages. https://doi.org/10.1145/3478098

[12] Ralph HB Benedict, Roee Holtzer, Robert W Motl, Frederick W Foley, Sukhmit Kaur, David Hojnacki, and Bianca Weinstock-Guttman. 2011. Upper and lower extremity motor function and cognitive impairment in multiple sclerosis. *Journal of the International*

Neuropsychological Society 17, 4 (2011), 643–653.
[13] Yoshua Bengio and Yves Grandvalet. 2003. No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems* 16 (2003).
[14] Jason A Berard, Andra M Smith, and Lisa AS Walker. 2019. Predictive models of cognitive fatigue in multiple sclerosis. *Archives of Clinical Neuropsychology* 34, 1 (2019), 31–38.
[15] Claudio Bettini, Gabriele Civitarese, and Michele Fiori. 2021. Explainable Activity Recognition over Interpretable Models. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 32–37.
[16] Aman Bhandari and Todd Wagner. 2006. Self-reported utilization of health care services: improving measurement and accuracy. *Medical Care Research and Review* 63, 2 (2006), 217–235.
[17] Karthik S Bhat and Neha Kumar. 2020. Sociocultural Dimensions of Tracking Health and Taking Care. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.
[18] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.
[19] Anne C Black, Ofer Harel, and Gregory Matthews. 2012. Techniques for analyzing intensive longitudinal data with missing values. (2012).
[20] Tiffany J Braley, Anna L Kratz, Neeraj Kaplish, and Ronald D Chervin. 2016. Sleep and cognitive function in multiple sclerosis. *Sleep* 39, 8 (2016), 1525–1533.
[21] Fraser S Brown, Stella A Glasmacher, Patrick KA Kearns, Niall MacDougall, David Hunt, Peter Connick, and Siddharthan Chandran. 2020. Systematic review of prediction models in relapsing remitting multiple sclerosis. *PloS one* 15, 5 (2020), e0233575.
[22] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.
[23] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*. 161–168.
[24] Ilaria Casetta, Trond Riise, Monica Wamme Nortvedt, Nicola Tiberio Economou, Riccardo De Gennaro, Patrik Fazio, Edward Cesnik, Vittorio Govoni, and Enrico Granieri. 2009. Gender differences in health-related quality of life in multiple sclerosis. *Multiple Sclerosis Journal* 15, 11 (2009), 1339–1346.
[25] Yuwen Chen, Kunhua Zhong, Ju Zhang, Qilong Sun, and Xueliang Zhao. 2016. LSTM networks for mobile human activity recognition. In *2016 International conference on artificial intelligence: technologies and applications*. Atlantis Press, 50–53.
[26] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. 2017. Semi-automated tracking: a balanced approach for self-monitoring applications. *IEEE Pervasive Computing* 16, 1 (2017), 74–84.
[27] Ehsan H Chowdhury and Quazi K Hassan. 2015. Operational perspective of remote sensing-based forest fire danger forecasting systems. *ISPRS Journal of Photogrammetry and Remote Sensing* 104 (2015), 224–236.
[28] Gloria Dalla-Costa, Marta Radaelli, Simona Maida, Francesca Sangalli, Bruno Colombo, Lucia Moiola, Giancarlo Comi, and Vittorio Martinelli. 2017. Smart watch, smarter EDSS: Improving disability assessment in multiple sclerosis clinical practice. *Journal of the neurological sciences* 383 (2017), 166–168.
[29] Anind K Dey, Raffay Hamid, Chris Beckmann, Ian Li, and Daniel Hsu. 2004. a CAPpella: programming by demonstration of context-aware applications. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 33–40.
[30] James Dougherty, Ron Kohavi, and Mehran Sahami. 1995. Supervised and Unsupervised Discretization of Continuous Features. In *Machine Learning Proceedings 1995*, Armand Prieditis and Stuart Russell (Eds.). Morgan Kaufmann, San Francisco (CA), 194–202. https://doi.org/10.1016/B978-1-55860-377-6.50032-3
[31] Nel Escher and Nikola Banovic. 2020. Exposing Error in Poverty Management Technology: A Method for Auditing Government Benefits Screening Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 064 (May 2020), 20 pages. https://doi.org/10.1145/3392874
[32] Brain T Fay and Michael L Boninger. 2002. The science behind mobility devices for individuals with multiple sclerosis. *Medical engineering & physics* 24, 6 (2002), 375–383.
[33] Jean-Daniel Fekete, Jarke J. van Wijk, John T. Stasko, and Chris North. 2008. *The Value of Information Visualization*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–18. https://doi.org/10.1007/978-3-540-70956-5_1
[34] Oscar Fernández, Karine Baumstarck-Barrau, Marie-Claude Simeoni, Pascal Auquier, and MusiQoL Study Group. 2011. Patient characteristics and determinants of quality of life in an international population with multiple sclerosis: assessment using the MusiQoL and SF-36 questionnaires. *Multiple Sclerosis Journal* 17, 10 (2011), 1238–1249.
[35] Marcia Finlayson. 2005. Pilot study of an energy conservation education program delivered by telephone conference call to people with multiple sclerosis. *NeuroRehabilitation* 20, 4 (2005), 267–277.
[36] Samuele Fiorini, Alessandro Verri, Annalisa Barla, Andrea Tacchino, and Giampaolo Brichetto. 2017. Temporal prediction of multiple sclerosis evolution from patient-centered outcomes. In *Machine Learning for Healthcare Conference*. 112–125.

[37] Samuele Fiorini, Alessandro Verri, Andrea Tacchino, Michela Ponzio, Giampaolo Brichetto, and Annalisa Barla. 2015. A machine learning pipeline for multiple sclerosis course detection from clinical scales and patient reported outcomes. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4443–4446.

[38] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet. Digital health* 2, 9 (Sep 2020), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2 32864600[pmid].

[39] Soudeh Ghafouri-Fard, Mohammad Taheri, Mir Davood Omrani, Amir Daaee, and Hossein Mohammad-Rahimi. 2020. Application of artificial neural network for prediction of risk of multiple sclerosis based on single nucleotide polymorphism genotypes. *Journal of Molecular Neuroscience* 70, 7 (2020), 1081–1087.

[40] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 3558–3565.

[41] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. (2021).

[42] Quentin F Gronau and Eric-Jan Wagenmakers. 2019. Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational brain & behavior* 2, 1 (2019), 1–11.

[43] Gabriel Guo, Hanbin Zhang, Liuyi Yao, Huining Li, Chenhan Xu, Zhengxiong Li, and Wenyao Xu. 2022. MSLife: Digital Behavioral Phenotyping of Multiple Sclerosis Symptoms in the Wild Using Wearables and Graph-Based Statistical Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 158 (dec 2022), 35 pages. https://doi.org/10.1145/3494970

[44] Martin Heine, Ingrid van de Port, Marc B Rietberg, Erwin EH van Wegen, and Gert Kwakkel. 2015. Exercise therapy for fatigue in multiple sclerosis. *Cochrane Database of Systematic Reviews* 9 (2015).

[45] Christa Holberg and Marcia Finlayson. 2007. Factors influencing the use of energy conservation strategies by persons with multiple sclerosis. *American Journal of Occupational Therapy* 61, 1 (2007), 96–107.

[46] Tahera Hossain, Wanggang Shen, Anindya Das Antar, Snehal Prabhudesai, Sozo Inoue, Xun Huan, and Nikola Banovic. 2022. A Bayesian Approach for Quantifying Data Scarcity When Modeling Human Behavior via Inverse Reinforcement Learning. *ACM Trans. Comput.-Hum. Interact.* (may 2022). https://doi.org/10.1145/3551388 Just Accepted.

[47] Lisa I Iezzoni, Sowmya R Rao, and R Philip Kinkel. 2009. Patterns of mobility aid use among working-age persons with multiple sclerosis living in the community in the United States. *Disability and health journal* 2, 2 (2009), 67–76.

[48] Yasha Iravantchi, Karan Ahuja, Mayank Goel, Chris Harrison, and Alanson Sample. 2021. PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[49] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. *Visual Analytics: Definition, Process, and Challenges*. Springer Berlin Heidelberg, Berlin, Heidelberg, 154–175. https://doi.org/10.1007/978-3-540-70956-5_7

[50] Daniel Kerrigan, Jessica Hullman, and Enrico Bertini. 2021. A Survey of Domain Knowledge Elicitation in Applied Machine Learning. *Multimodal Technologies and Interaction* 5, 12 (2021). https://doi.org/10.3390/mti5120073

[51] Edward Kim, Jesus Lovera, Laura Schaben, Jorge Melara, Dennis Bourdette, and Ruth Whitham. 2010. Novel method for measurement of fatigue in multiple sclerosis: Real-Time Digital Fatigue Score. *Journal of Rehabilitation Research & Development* 47, 5 (2010).

[52] Seungjun Kim, Dan Tasse, and Anind K Dey. 2017. Making Machine-Learning Applications for Time-Series Sensor Data Graphical and Interactive. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 2 (2017), 1–30.

[53] Sonya Kim, Vance Zemon, Joseph F Rath, MaryAnn Picone, Elizabeth S Gromisch, Heather Glubo, Lucia Smith-Wexler, and Frederick W Foley. 2017. Screening instruments for the early detection of cognitive impairment in patients with multiple sclerosis. *International journal of MS care* 19, 1 (2017), 1–10.

[54] Annet M Kleiboer, Roeline G Kuijer, Joop J Hox, Peter JH Jongen, Stephan TFM Frequin, and Jozien M Bensing. 2007. Daily negative interactions and mood among patients and partners dealing with multiple sclerosis (MS): The moderating effects of emotional support. *Social Science & Medicine* 64, 2 (2007), 389–400.

[55] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

[56] Lindsey M Knowles, Anne Arewasikporn, Anna L Kratz, Aaron P Turner, Kevin N Alschuler, and Dawn M Ehde. 2021. Early Treatment Improvements in Depression Are Associated With Overall Improvements in Fatigue Impact and Pain Interference in Adults With Multiple Sclerosis. *Annals of Behavioral Medicine* 55, 9 (2021), 833–843.

[57] Elena D Koch, Talar R Moukhtarian, Caroline Skirrow, Natali Bozhilova, Philip Asherson, and Ulrich W Ebner-Priemer. 2021. Using e-diaries to investigate ADHD–State-of-the-art and the promising feature of just-in-time-adaptive interventions. *Neuroscience & Biobehavioral Reviews* (2021).

[58] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.
[59] Daphne Kos, Guy Nagels, Marie B D'Hooghe, William Duquet, Stephan Ilsbroukx, Stijn Delbeke, and Eric Kerckhofs. 2007. Measuring activity patterns using actigraphy in multiple sclerosis. *Chronobiology international* 24, 2 (2007), 345–356.
[60] Anna L Kratz, Kevin N Alschuler, David A Williams, and Dawn M Ehde. 2021. Development and pilot testing of a web-based symptom management program for multiple sclerosis: My MS toolkit. *Rehabilitation Psychology* (2021).
[61] Anna L Kratz, Tiffany J Braley, Emily Foxen-Craft, Eric Scott, John F Murphy III, and Susan L Murphy. 2017. How do pain, fatigue, depressive, and cognitive symptoms relate to well-being and social and physical functioning in the daily lives of individuals with multiple sclerosis? *Archives of physical medicine and rehabilitation* 98, 11 (2017), 2160–2166.
[62] Anna L Kratz, Susan L Murphy, and Tiffany J Braley. 2017. Ecological momentary assessment of pain, fatigue, depressive, and cognitive symptoms reveals significant daily variability in multiple sclerosis. *Archives of physical medicine and rehabilitation* 98, 11 (2017), 2142–2150.
[63] Anna L Kratz, Susan L Murphy, and Tiffany J Braley. 2017. Pain, fatigue, and cognitive symptoms are temporally associated within but not across days in multiple sclerosis. *Archives of physical medicine and rehabilitation* 98, 11 (2017), 2151–2159.
[64] Anna L Kratz, Susan L Murphy, Tiffany J Braley, Neil Basu, Shubhangi Kulkarni, Jenna Russell, and Noelle E Carlozzi. 2019. Development of a person-centered conceptual model of perceived fatigability. *Quality of Life Research* 28, 5 (2019), 1337–1347.
[65] Andreas Krause and Carlos Guestrin. 2005. Near-Optimal Nonmyopic Value of Information in Graphical Models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence* (Edinburgh, Scotland) *(UAI'05)*. AUAI Press, Arlington, Virginia, USA, 324–331.
[66] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5686–5697. https://doi.org/10.1145/2858036.2858529
[67] Himabindu Lakkaraju and Osbert Bastani. 2020. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833
[68] Yang Li and James A Landay. 2006. Exploring activity-based ubiquitous computing: interaction styles, models and tool support. *computing* 11 (2006), 13.
[69] JP Lord, MM Portwood, WM Fowler, JS Lieberman, and R Carson. 1987. Upper vs lower extremity functional loss in neuromuscular disease. *Archives of physical medicine and rehabilitation* 68, 1 (1987), 8–9.
[70] RA Lorenz, S Auerbach, J Koedbangkham, T Loux, E Shacham, and FP Thomas. 2018. The Relationship between Actigraphy Measured Sleep and Fatigue among Adults with Multiple Sclerosis. *Sleep* 41, suppl_1 (2018), A384–A384.
[71] Fred D Lublin, Stephen C Reingold, et al. 1996. Defining the clinical course of multiple sclerosis: results of an international survey. *Neurology* 46, 4 (1996), 907–911.
[72] Kathleen Matuska, Virgil Mathiowetz, and Marcia Finlayson. 2007. Use and perceived effectiveness of energy conservation strategies for managing multiple sclerosis fatigue. *American Journal of Occupational Therapy* 61, 1 (2007), 62–69.
[73] Health Measures. 2021. *G-code Severity Modifiers for Neuro-QoL™*. https://www.healthmeasures.net/score-and-interpret/interpret-scores/neuro-qol/g-code-severity-modifiers
[74] Health Measures. 2021. *Neuro-QoLTM score ranges associated with individual G-code severity modifiers for Neuro-QoL Physical Function Scores*. https://www.healthmeasures.net/images/scoring/G-code_severity_modifiers/G_Codes_for_PROMIS_and_Neuro-QoL_Mar12_2019.pdf
[75] Deborah M Miller, Francois Bethoux, David Victorson, Cindy J Nowinski, Sarah Buono, Jin-Shei Lai, Katy Wortman, James L Burns, Claudia Moy, and David Cella. 2016. Validating Neuro-QoL short forms and targeted scales with people who have multiple sclerosis. *Multiple Sclerosis Journal* 22, 6 (2016), 830–841.
[76] Roger J Mills and Carolyn A Young. 2011. The relationship between fatigue and other clinical features of multiple sclerosis. *Multiple Sclerosis Journal* 17, 5 (2011), 604–612.
[77] Elias Chaibub Neto, Abhishek Pratap, Thanneer M Perumal, Meghasyam Tummalacherla, Phil Snyder, Brian M Bot, Andrew D Trister, Stephen H Friend, Lara Mangravite, and Larsson Omberg. 2019. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ digital medicine* 2, 1 (2019), 1–6.
[78] Andrew Y. Ng and Stuart J. Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 663–670.
[79] Beau Norgeot, Benjamin S Glicksberg, Laura Trupin, Dmytro Lituiev, Milena Gianfrancesco, Boris Oskotsky, Gabriela Schmajuk, Jinoos Yazdany, and Atul J Butte. 2019. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA network open* 2, 3 (2019), e190606–e190606.
[80] Diana Ohanian, Abigail Brown, Madison Sunnquist, Jacob Furst, Laura Nicholson, Lauren Klebek, and Leonard A Jason. 2016. Identifying key symptoms differentiating myalgic encephalomyelitis and chronic fatigue syndrome from multiple sclerosis. *Neurology (E-Cronicon)*

4, 2 (2016), 41.
[81] World Health Organization et al. 2002. Towards a common language for functioning, disability, and health: ICF. *The international classification of functioning, disability and health* (2002).
[82] K O'Connell, D Langdon, N Tubridy, M Hutchinson, and C McGuigan. 2015. A preliminary validation of the brief international cognitive assessment for multiple sclerosis (BICAMS) tool in an Irish population with multiple sclerosis (MS). *Multiple sclerosis and related disorders* 4, 6 (2015), 521–525.
[83] N Pfaffenberger, K-P Pfeiffer, M Deibl, S Höfer, V Günther, and H Ulmer. 2006. Association of factors influencing health-related quality of life in MS. *Acta neurologica scandinavica* 114, 2 (2006), 102–108.
[84] Lorraine J Phillips and Alexa K Stuifbergen. 2006. Predicting Continued Employment in Persons with Multiple Sclerosis. *Journal of Rehabilitation* 72, 1 (2006).
[85] Daniel JH Powell, Christina Liossi, Wolff Schlotz, and Rona Moss-Morris. 2017. Tracking daily fatigue fluctuations in multiple sclerosis: ecological momentary assessment provides unique insights. *Journal of behavioral medicine* 40, 5 (2017), 772–783.
[86] Snehal Prabhudesai, Nicholas Chandler Wang, Vinayak Ahluwalia, Xun Huan, Jayapalli Rajiv Bapuraj, Nikola Banovic, and Arvind Rao. 2021. Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation. *Frontiers in Neuroscience* 15 (2021), 1236. https://doi.org/10.3389/fnins.2021.740353
[87] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
[88] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. 385–394.
[89] Lawrence Rabiner and Biinghwang Juang. 1986. An introduction to hidden Markov models. *ieee assp magazine* 3, 1 (1986), 4–16.
[90] A Riazi, AJ Thompson, and JC Hobart. 2004. Self-efficacy predicts self-reported health status in multiple sclerosis. *Multiple Sclerosis Journal* 10, 1 (2004), 61–66.
[91] Daniela Rodriguez Rincon, Brandi Leach, Jack Pollard, Sarah Parkinson, Evangelos Gkousis, Catherine A Lichten, Jon Sussex, and Catriona Manville. 2019. *Exploring the Societal Burden of Multiple Sclerosis: A Study Into the Non-clinical Impact of the Disease, Including Changes with Progression.* RAND.
[92] P Roca, A Attye, L Colas, A Tucholka, P Rubini, S Cackowski, J Ding, J-F Budzik, F Renard, S Doyle, et al. 2020. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagnostic and Interventional Imaging* 101, 12 (2020), 795–802.
[93] Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L. Glassman, and Finale Doshi-Velez. 2021. Evaluating the Interpretability of Generative Models by Interactive Reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 80, 15 pages. https://doi.org/10.1145/3411764.3445296
[94] Luis Ruano, Emilio Portaccio, Benedetta Goretti, Claudia Niccolai, Milton Severo, Francesco Patti, Sabina Cilia, Paolo Gallo, Paola Grossi, Angelo Ghezzi, et al. 2017. Age and disability drive cognitive impairment in multiple sclerosis across disease subtypes. *Multiple Sclerosis Journal* 23, 9 (2017), 1258–1267.
[95] Jennifer M Ryan, Jennifer Fortune, Andrea Stennett, Cherry Kilbride, Nana Anokye, Christina Victor, Wendy Hendrie, Mohamed Abdul, Lorraine DeSouza, Grace Lavelle, et al. 2017. Changing physical activity behaviour for people with multiple sclerosis: protocol of a randomised controlled feasibility trial (iStep-MS). *BMJ open* 7, 11 (2017).
[96] Alexander J. Ryu, Santiago Romero-Brufau, Ray Qian, Heather A. Heaton, David M. Nestler, Shant Ayanian, and Thomas C. Kingsley. 2022. Assessing the Generalizability of a Clinical Machine Learning Model Across Multiple Emergency Departments. *Mayo Clinic Proceedings: Innovations, Quality Outcomes* 6, 3 (2022), 193–199. https://doi.org/10.1016/j.mayocpiqo.2022.03.003
[97] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. 2016. Human-centered machine learning through interactive visualization. ESANN.
[98] Margaret M Schmitt, Yael Goverover, John DeLuca, and Nancy Chiaravalloti. 2014. Self-efficacy as a predictor of self-reported physical, cognitive, and social functioning in multiple sclerosis. *Rehabilitation psychology* 59, 1 (2014), 27.
[99] Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics* (1978), 461–464.
[100] Shahnaz Shahrbanian. 2017. What does pain in MS feel like? -A multicenter cross-sectional study of pain descriptors and characteristics. *Internal Medicine and Care (IMC)* 4, 1 (2017).
[101] Eli Sherman, Hitinder Gurm, Ulysses Balis, Scott Owens, and Jenna Wiens. 2018. Leveraging Clinical Time-Series Data for Prediction: A Cautionary Tale. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2017 (16 Apr 2018), 1571–1580. https://pubmed.ncbi.nlm.nih.gov/29854227 29854227[pmid].
[102] Galit Shmueli et al. 2010. To explain or to predict? *Statistical science* 25, 3 (2010), 289–310.
[103] Arielle M Silverman, Aimee M Verrall, Kevin N Alschuler, Amanda E Smith, and Dawn M Ehde. 2017. Bouncing back again, and again: a qualitative study of resilience in people with multiple sclerosis. *Disability and Rehabilitation* 39, 1 (2017), 14–22.
[104] National MS Society. 2021. *MS Signs and Symptoms.* https://www.nationalmssociety.org/Symptoms-Diagnosis/MS-Symptoms
[105] National MS Society. 2021. *Treating MS?* https://www.nationalmssociety.org/Treating-MS
[106] National MS Society. 2021. *What is MS?* https://www.nationalmssociety.org/What-is-MS

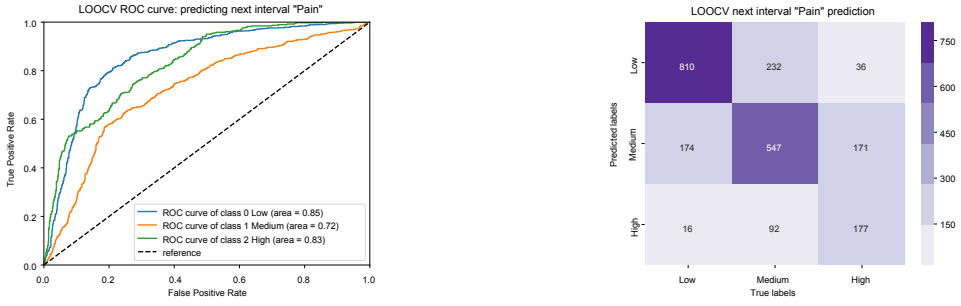
[107] National MS Society. 2021. *Who gets MS?* https://www.nationalmssociety.org/What-is-MS/Who-Gets-MS
[108] National MS Society. 2022. *Types of MS.* https://www.nationalmssociety.org/What-is-MS/Types-of-MS
[109] Jacob J Sosnoff, Michael J Socie, Morgan K Boes, Brian M Sandroff, and Robert W Motl. 2012. Does a waist-worn ActiGraph accelerometer quantify community ambulation in persons with multiple sclerosis? *Age (yr)* 46, 11.7 (2012), 23.
[110] David F Steiner, Po-Hsuan Cameron Chen, and Craig H Mermel. 2020. Closing the translation gap: AI applications in digital pathology. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* (2020), 188452.
[111] Merrill J Turpin B Occ Thy, Miho Asano, and Marcia Finlayson. 2015. Combining qualitative and quantitative data collection and analysis methods in understanding multiple sclerosis fatigue management. *International Journal of Qualitative Methods* 14, 2 (2015), 53–68.
[112] Catherine Tong, Matthew Craner, Matthieu Vegreville, and Nicholas D. Lane. 2019. Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Connnected Wellness Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 106 (Sept. 2019), 19 pages. https://doi.org/10.1145/3351264
[113] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 6 (2001), 520–525.
[114] John W. Tukey. 1977. *Exploratory data analysis.* Addison-Wesley. http://www.worldcat.org/oclc/03058187
[115] REA Van Emmerik, JG Remelius, MB Johnson, LH Chung, and JA Kent-Braun. 2010. Postural control in women with multiple sclerosis: effects of task, vision and symptomatic fatigue. *Gait & posture* 32, 4 (2010), 608–614.
[116] Alfredo Vellido. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications* 32, 24 (2020), 18069–18083.
[117] Richard Warner. 2006. Understanding fatigue in multiple sclerosis: A case study inquiry. *British Journal of Neuroscience Nursing* 2, 9 (2006), 462–469.
[118] World Health Organization (WHO). 2021. *International Classification of Functioning, Disability and Health (ICF).* https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health
[119] Tzu-Tsung Wong. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* 48, 9 (2015), 2839–2846.
[120] K Wynia, B Middel, JP Van Dijk, JHA De Keyser, and SA Reijneveld. 2008. The impact of disabilities on quality of life in people with multiple sclerosis. *Multiple Sclerosis Journal* 14, 7 (2008), 972–980.
[121] Shihao Yang, Mauricio Santillana, John S Brownstein, Josh Gray, Stewart Richardson, and SC Kou. 2017. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC infectious diseases* 17, 1 (2017), 1–9.
[122] Nuray Yozbatıran, Ferdi Baskurt, Zeliha Baskurt, Serkan Ozakbas, and Egemen Idiman. 2006. Motor assessment of upper extremity function and its relation with fatigue, cognitive function and quality of life in multiple sclerosis patients. *Journal of the neurological sciences* 246, 1-2 (2006), 117–122.
[123] Jan Yperman, Thijs Becker, Dirk Valkenborg, Veronica Popescu, Niels Hellings, Bart Van Wijmeersch, and Liesbet M. Peeters. 2020. Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis. *BMC Neurology* 20, 1 (21 Mar 2020), 105. https://doi.org/10.1186/s12883-020-01672-w
[124] Enhao Zhang and Nikola Banovic. 2021. Method for Exploring Generative Adversarial Networks (GANs) via Automatically Generated Image Galleries. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 76, 15 pages. https://doi.org/10.1145/3411764.3445714
[125] Yijun Zhao, Brian C Healy, Dalia Rotstein, Charles RG Guttmann, Rohit Bakshi, Howard L Weiner, Carla E Brodley, and Tanuja Chitnis. 2017. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PloS one* 12, 4 (2017), e0174866.
[126] Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. 2010. Modeling Interaction via the Principle of Maximum Causal Entropy. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Haifa, Israel) *(ICML'10)*. Omnipress, USA, 1255–1262. http://dl.acm.org/citation.cfm?id=3104322.3104481
[127] Tjalf Ziemssen, Raimar Kern, and Katja Thomas. 2016. Multiple sclerosis: clinical profiling and data collection as prerequisite for personalized medicine approach. *BMC neurology* 16, 1 (2016), 1–10.

A APPENDIX AND SUPPLEMENTARY MATERIALS

A.1 Leave-one-out Cross Validation for Next interval Pain Prediction

Table 13. Leave-one-out cross-validation (LOOCV) performance comparison of Bayesian Network with the proposed feature engineering approach to predict next interval pain for people with MS across different daytime intervals.

LOOCV model comparison for predicting pain	Mean Precision	Mean Recall	Mean F1-score	Mean Accuracy	Mean ROC AUC
Logistic Regression with naive features	54.91%	48.78%	51.40%	48.78%	0.56
Random Forest with naive features	48.86%	45.23%	46.82%	45.23%	0.54
Support Vector Machine with naive features	50.36%	45.48%	47.59%	45.48%	0.54
Decision Tree with naive features	48.44%	45.03%	46.54%	45.03%	0.54
K-nearest Neighbors with naive features	50.41%	44.55%	46.84%	44.55%	0.53
Elastic Net with naive features	51.63%	44.90%	47.46%	44.90%	0.53
Gradient Boosting with naive features	50.21%	45.06%	47.10%	45.06%	0.54
Ridge classifier with naive features	51.29%	45.73%	47.94%	45.73%	0.54
Extra Tree classifier with naive features	48.50%	55.08%	45.50%	49.03%	0.56
Light GBM with naive features	49.98%	54.27%	46.88%	46.52%	0.57
XGBoost with naive features	49.57%	56.62%	49.57%	49.81%	0.57
AdaBoost with naive features	52.43%	53.12%	53.34%	53.12%	0.57
Multi-Layer Perceptron with naive features	48.80%	53.87%	48.80%	48.41%	0.57
Naive Bayes with naive features	51.63%	45.13%	47.64%	45.13%	0.53
Naive Bayesian Network with naive features	51.12%	53.87%	52.13%	50.35%	0.56
Bayesian Network with proposed connections and proposed features	**69.34%**	**68.03%**	**68.50%**	**68.02%**	**0.73**



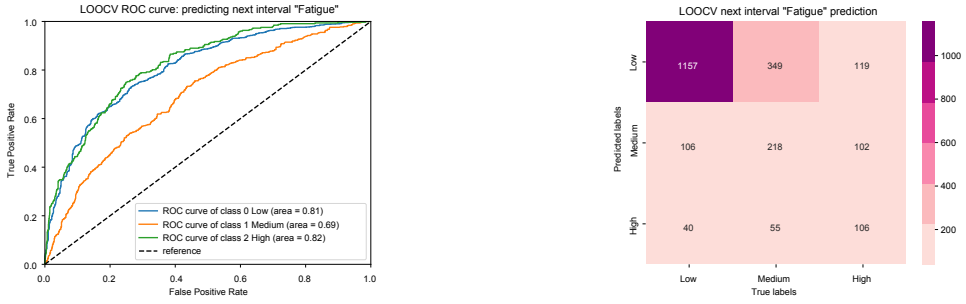
(a) (b)

Fig. 17. Results for Leave-one-out cross validation (LOOCV) next interval pain prediction using our BN: a) ROC curve, and b) confusion matrix.

A.2 Leave-one-out Cross Validation for Next interval Fatigue Prediction

Table 14. Leave-one-out cross-validation (LOOCV) performance comparison of Bayesian Network with the proposed feature engineering approach to predict fatigue for people with MS across different daytime intervals.

LOOCV model comparison for predicting fatigue	**Mean Precision**	**Mean Recall**	**Mean F1-score**	**Mean Accuracy**	**Mean ROC AUC**
Logistic Regression with naive features	64.47%	57.82%	67.19%	57.82%	0.53
Random Forest with naive features	65.03%	52.66%	57.38%	52.66%	0.52
Support Vector Machine with naive features	63.27%	53.92%	61.03%	53.92%	0.51
Decision Tree with naive features	64.22%	51.27%	56.20%	51.27%	0.53
K-nearest Neighbors with naive features	62.16%	50.55%	54.96%	50.55%	0.51
Elastic Net with naive features	64.27%	51.77%	57.54%	51.77%	0.51
Gradient Boosting with naive features	64.03%	51.81%	57.09%	51.81%	0.51
Ridge classifier with naive features	65.34%	52.56%	58.37%	52.56%	0.51
Extra Tree classifier with naive features	64.35%	56.56%	56.43%	53.01%	0.53
Light GBM with naive features	63.56%	53.01%	53.17%	54.36%	0.55
XGBoost with naive features	61.36%	53.30%	53.33%	54.35%	0.53
AdaBoost with naive features	64.80%	52.38%	53.70%	52.47%	0.55
Multi-Layer Perceptron with naive features	65.71%	53.57%	54.44%	54.51%	0.51
Naive Bayes with naive features	66.76%	52.70%	57.90%	52.70%	0.53
Naive Bayesian Network with naive features	63.72%	58.14%	55.79%	58.14%	0.59
Bayesian Network with proposed connections and proposed features	**73.60%**	**65.76%**	**68.48%**	**65.76%**	**0.65**



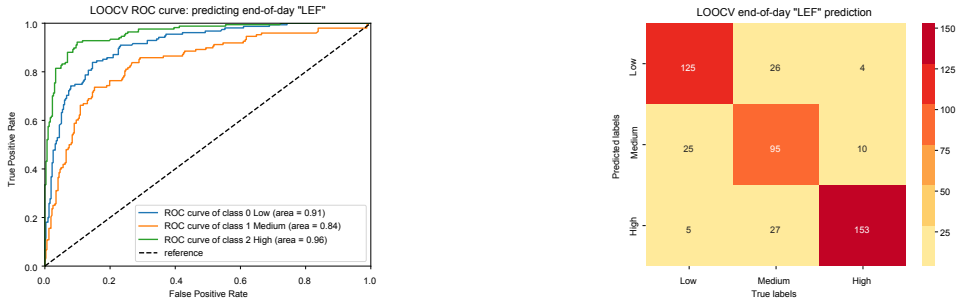
(a) (b)

Fig. 18. Results for Leave-one-out cross validation (LOOCV) next interval fatigue prediction using our BN: a) ROC curve, and b) confusion matrix.

A.3 Leave-one-out Cross Validation for EOD LEF Prediction

Table 15. Leave-one-out cross-validation (LOOCV) performance comparison of Bayesian Network with the proposed feature engineering approach to predict the end-of-day Lower Extremity Functioning (EOD LEF) for people with MS.

LOOCV model comparison for predicting LEF	Mean Precision	Mean Recall	Mean F1-score	Mean Accuracy	Mean ROC AUC
Logistic Regression with naive features	56.82%	53.40%	55.35%	53.40%	0.64
Random Forest with naive features	54.99%	52.77%	53.35%	52.77%	0.64
Support Vector Machine with naive features	55.81%	52.41%	53.26%	52.41%	0.64
Decision Tree with naive features	52.58%	49.79%	50.58%	49.79%	0.62
K-nearest Neighbors with naive features	48.28%	45.96%	46.65%	45.96%	0.59
Elastic Net with naive features	50.52%	44.01%	45.55%	44.01%	0.57
Gradient Boosting with naive features	47.64%	42.67%	43.98%	42.67%	0.56
Ridge classifier with naive features	48.69%	43.99%	45.22%	43.99%	0.57
Extra Tree classifier with naive features	53.47%	52.72%	52.48%	53.72%	0.63
Light GBM with naive features	53.43%	55.53%	53.34%	54.12%	0.62
XGBoost with naive features	54.34%	56.48%	55.18%	54.65%	0.61
AdaBoost with naive features	47.32%	53.32%	47.32%	47.91%	0.57
Multi-Layer Perceptron with naive features	52.43%	59.93%	53.16%	54.14%	0.60
Naive Bayes with naive features	50.05%	44.52%	46.01%	44.52%	0.58
Naive Bayesian Network with naive features	58.73%	53.83%	55.73%	53.79%	0.59
Bayesian Network with proposed connections and proposed features	**80.41%**	**79.36%**	**79.71%**	**79.36%**	**0.84**



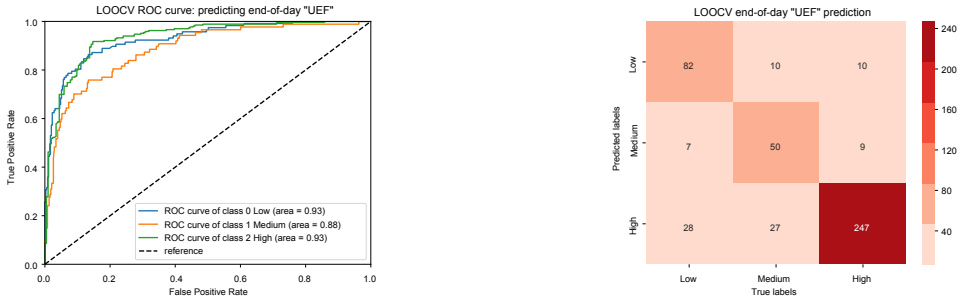
(a) (b)

Fig. 19. Results for Leave-one-out cross-validation (LOOCV) EOD LEF prediction using our BN: a) ROC curve, and b) confusion matrix.

A.4 Leave-one-out Cross Validation for Next interval UEF Prediction

Table 16. Leave-one-out cross-validation (LOOCV) performance comparison of Bayesian Network with the proposed feature engineering approach to predict the end-of-day Upper Extremity Functioning (EOD UEF) for people with MS.

LOOCV model comparison for predicting UEF	Mean Precision	Mean Recall	Mean F1-score	Mean Accuracy	Mean ROC AUC
Logistic Regression with naive features	67.77%	57.23%	64.20%	57.23%	0.56
Random Forest with naive features	62.68%	53.94%	57.20%	53.94%	0.56
Support Vector Machine with naive features	67.91%	55.82%	60.28%	55.82%	0.56
Decision Tree with naive features	61.91%	53.67%	56.84%	53.67%	0.56
K-nearest Neighbors with naive features	60.74%	52.00%	55.33%	52.00%	0.54
Elastic Net with naive features	65.40%	52.77%	57.37%	52.77%	0.53
Gradient Boosting with naive features	61.81%	51.43%	55.33%	51.43%	0.53
Ridge classifier with naive features	63.00%	52.02%	56.15%	52.02%	0.53
Extra Tree classifier with naive features	61.85%	55.71%	56.95%	55.13%	0.56
Light GBM with naive features	60.56%	54.90%	57.56%	53.8%	0.52
XGBoost with naive features	61.81%	55.35%	57.11%	52.51%	0.54
AdaBoost with naive features	63.25%	52.71%	56.34%	52.14%	0.56
Multi-Layer Perceptron with naive features	65.28%	52.31%	55.23%	51.43%	0.56
Naive Bayes with naive features	60.6%	51.56%	55.05%	51.56%	0.54
Naive Bayesian Network with naive features	61.15%	59.27%	61.74%	59.99%	0.57
Bayesian Network with proposed connections and proposed features	**82.95%**	**80.64%**	**81.31%**	**80.64%**	**0.81**



(a) (b)

Fig. 20. Results for Leave-one-out cross validation (LOOCV) EOD UEF prediction using our BN: a) ROC curve, and b) confusion matrix.

A.5 Model Hyperparameters

We used the Grid Search Cross-validation method to tune the model hyperparameters.

Table 17. Model hyperparameters using grid search cross-validation method for the prediction tasks.

Model	Hyperparameter value			
	Pain prediction	**Fatigue prediction**	**LEF prediction**	**UEF prediction**
Logistic Regression	{'C': 1.0, 'penalty': 'l2'}	{'C': 0.1, 'penalty': 'l2'}	{'C': 0.001, 'penalty': 'l2'}	{'C': 0.01, 'penalty': 'l2'}
Random Forest	{'max_depth': 50, 'min_samples_split': 2, 'n_estimators': 400, 'criterion': 'entropy'}	{'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 600, 'criterion': 'entropy'}	{'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200, 'criterion': 'entropy'}	{'max_depth': 100, 'min_samples_split': 10, 'n_estimators': 400, 'criterion': 'entropy'}
Support Vector Machine	{'C': 0.5, 'kernel': 'linear', 'tol': 0.001}	{'C': 0.1, 'kernel': 'rbf', 'tol': 0.001}	{'C': 0.5, 'kernel': 'sigmoid', 'tol': 0.01}	{'C': 0.1, 'kernel': 'linear', 'tol': 0.001}
Decision Tree	{'criterion': 'entropy', 'max_depth': 50, 'min_samples_split': 50}	{'criterion': 'gini', 'max_depth': 50, 'min_samples_split': 10}	{'criterion': 'entropy', 'max_depth': 100, 'min_samples_split': 10}	{'criterion': 'entropy', 'max_depth': 100, 'min_samples_split': 2}
K-nearest Neighbors	{'n_neighbors': 7, 'weights': 'distance', 'leaf_size': 30}	{'n_neighbors': 7, 'weights': 'uniform', 'leaf_size': 30}	{'n_neighbors': 5, 'weights': 'distance', 'leaf_size': 30}	{'n_neighbors': 7, 'weights': 'distance', 'leaf_size': 30}
Elastic Net	{'penalty' = 'elasticnet', 'l1_ratio' = 0.9, 'solver' = 'saga'}	{'penalty' = 'elasticnet', 'l1_ratio' = 0.9, 'solver' = 'saga'}	{'penalty' = 'elasticnet', 'l1_ratio' = 0.9, 'solver' = 'saga'}	{'penalty' = 'elasticnet', 'l1_ratio' = 0.9, 'solver' = 'saga'}
Gradient Boosting	{'loss': 'log_loss', 'learning_rate': 0.5, 'n_estimators': 500}	{'loss': 'log_loss', 'learning_rate': 0.1, 'n_estimators': 200}	{'loss': 'log_loss', 'learning_rate': 0.5, 'n_estimators': 500}	{'loss': 'log_loss', 'learning_rate': 0.5, 'n_estimators': 100}
Ridge classifier	{'alpha': 0.5, 'tol': 0.01}	{'alpha': 1, 'tol': 0.01}	{'alpha': 0.5, 'tol': 0.01}	{'alpha': 0.5, 'tol': 0.01}
Extra Tree classifier	{'criterion': 'entropy', 'min_samples_leaf': 2, 'min_samples_split': 5}	{'criterion': 'entropy', 'min_samples_leaf': 2, 'min_samples_split': 5}	{'criterion': 'entropy', 'min_samples_leaf': 2, 'min_samples_split': 2}	{'criterion': 'gini', 'min_samples_leaf': 1, 'min_samples_split': 2}
Light GBM	{'boosting_type' = 'gbdt', 'num_leaves' = 31, n_estimators = 100}	{'boosting_type' = 'gbdt', 'num_leaves' = 50, n_estimators = 500}	{'boosting_type' = 'gbdt', 'num_leaves' = 50, n_estimators = 100}	{'boosting_type' = 'gbdt', 'num_leaves' = 50, n_estimators = 200}
XGBoost	{'booster': 'gbtree', 'max_depth': 5, 'gamma': 0.01}	{'booster': 'dart', 'max_depth': 6, 'gamma': 0.01}	{'booster': 'gbtree', 'max_depth': 5, 'gamma': 0}	{'booster': 'gbtree', 'max_depth': 6, 'gamma': 0.01}
AdaBoost	{'algorithm': 'SAMME', 'n_estimators': 100}	{'algorithm': 'SAMME', 'n_estimators': 200}	{'algorithm': 'SAMME', 'n_estimators': 500}	{'algorithm': 'SAMME', 'n_estimators': 300}
Multi-Layer Perceptron	{'activation': 'logistic', 'hidden_layer_sizes': (100,), 'solver': 'adam', 'L2 strength': 0.0001, 'learning_rate': 0.001}	{'activation': 'logistic', 'hidden_layer_sizes': (300,), 'solver': 'adam', 'L2 strength': 0.0001, 'learning_rate': 0.001}	{'activation': 'relu', 'hidden_layer_sizes': (200,), 'solver': 'adam', 'L2 strength': 0.0001, 'learning_rate': 0.001}	{'activation': 'logistic', 'hidden_layer_sizes': (100,), 'solver': 'adam', 'L2 strength': 0.0001, 'learning_rate': 0.001}
Naive Bayes	{'var_smoothing': 0.01}	'solver': 'adam', 'L2 strength': 0.0001,	{'var_smoothing': 0.5}	{'var_smoothing': 0.5}
Our Bayesian Network	{'estimator' = 'Bayesian estimator', 'inference' = 'maximum weighted likelihood sampling'} feature nodes and connections were made following findings from existing research and consulting with a domain expert			