CPSL
Cyber-Physical Systems Laboratory

Multi-Task Learning for Randomized Controlled Trials: A Case Study on Predicting Depression with Wearable Data

RUIXUAN DAI, Washington University in St. Louis, United States
THOMAS KANNAMPALLIL, Washington University in St. Louis, United States
JINGWEN ZHANG, Washington University in St. Louis, United States
NAN LV, University of Illinois Chicago, United States
JUN MA, University of Illinois Chicago, United States
CHENYANG LU, Washington University in St. Louis, United States

Washington University in St. Louis

Depression

- Depression affects more than 280 million people of all ages [1]

"An estimated **3.8%** of the population experience depression, including **5%** of adults (4% among men and 6% among women), and **5.7%** of adults older than 60 years. Approximately **280 million** people in the world have depression."

Depression treatment

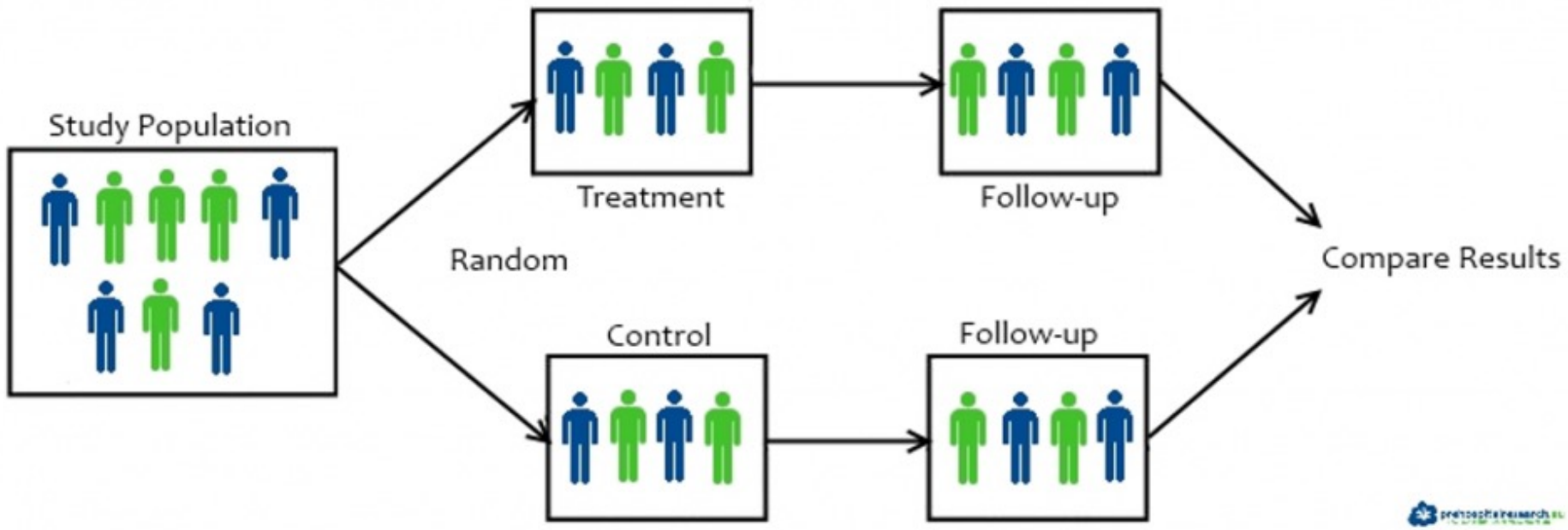
How to evaluate treatment?



[1] "Depression". *www.who.int*. Archived from the original on 26 December 2020. Retrieved 7 April 2021.

Randomized Control Trial

- A randomized controlled trial (RCT) is considered the gold standard for evaluating treatment efficacy, including in the case of mental health interventions
  - Intervention vs control group



Randomized Control Trial

- A randomized controlled trial (RCT) is considered the gold standard for evaluating treatment efficacy, including in the case of mental health interventions
  - Intervention vs control group
- Typically, RCT depends on **statistical analysis** to evaluate disparities between intervention and control groups
  - Group level instead of individual level

Goal: Personalized Depression Outcome Prediction

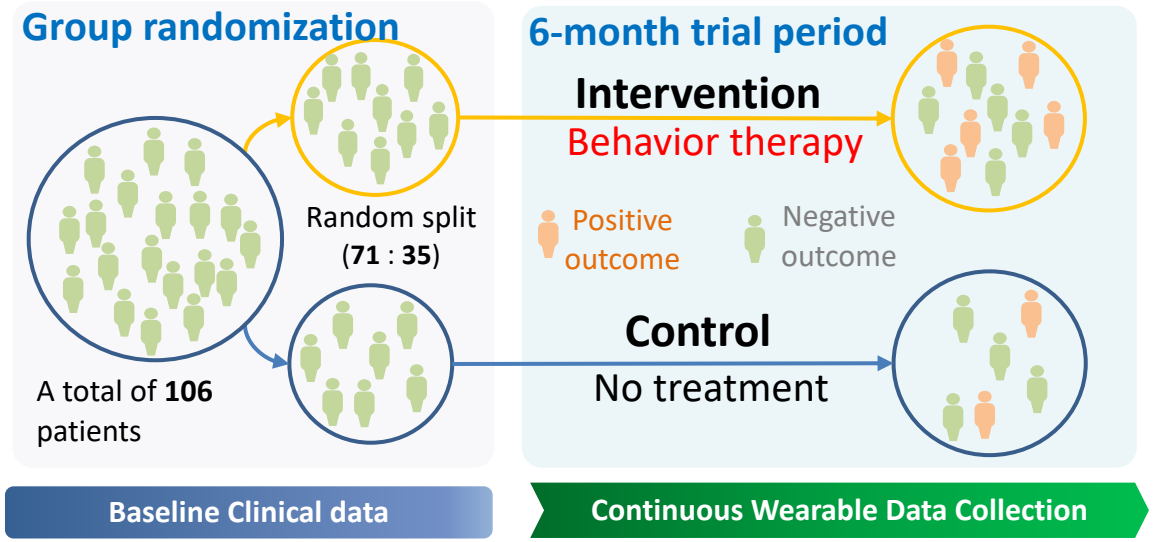
- Towards **precision medicine**: tailor treatment for individuals

Personalized prediction ⇨ Tailored depression treatment

  - Predict the depression outcome if the patient **takes the treatment** (intervention group)
  - Predict the depression outcome if the patient **does not take the treatment** (control group)

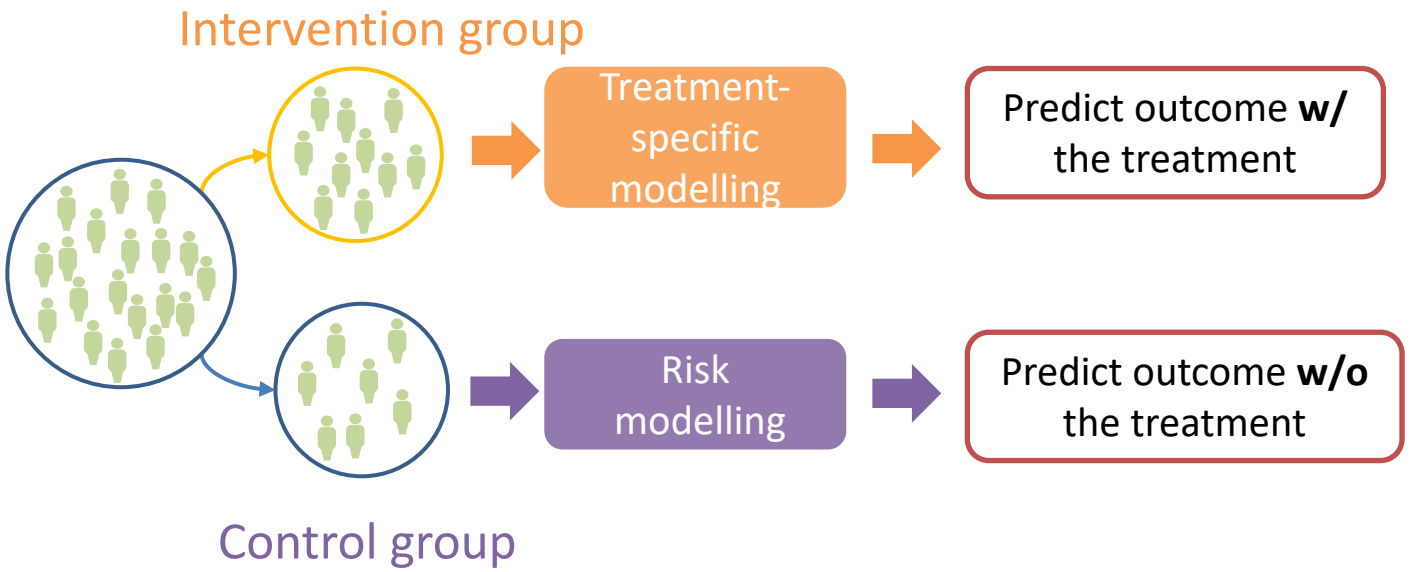
Randomized Control Trial for Depression Treatment

- Our RCT of behavior therapy for depression
  - **6-month** trial with **106** patients with depression
- Data collected:
  - Clinical features
    e.g., age, anxiety score, PTSD etc.
  - Wearable data for 6 months
    e.g., heart rate, sleep etc.
  - Outcome at the 6th month (depression remission or not)



Standard Machine Learning Methods for RCT

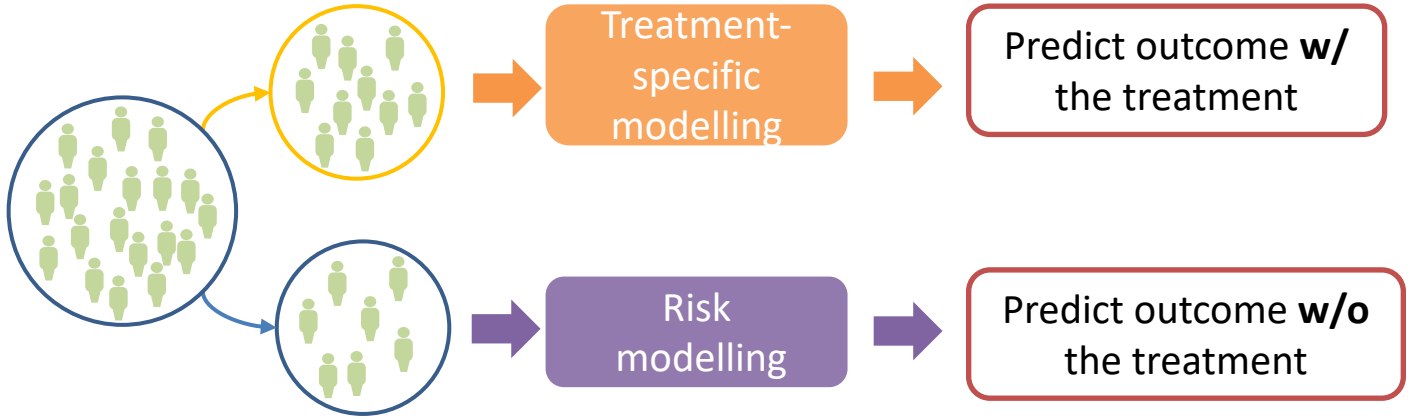
- Standard machine learning (ML) models: two **group-specific** models



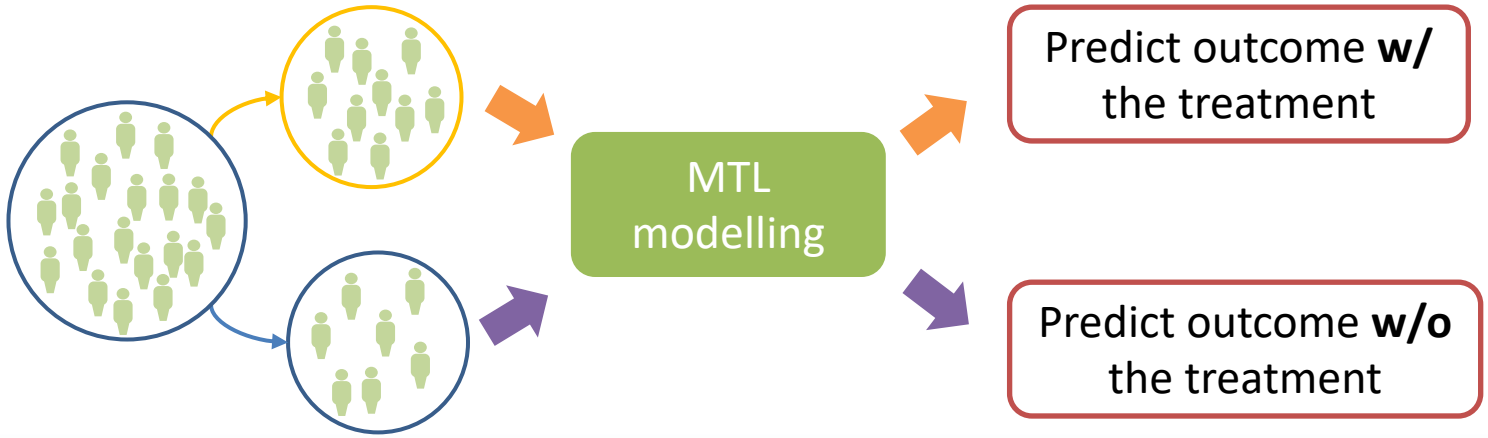
Limitation: two separate models may suffer from sample size issues and may not fully utilize the underlying characteristics of the patients.

Our Approach: Multi-Task Learning (MTL) for RCT

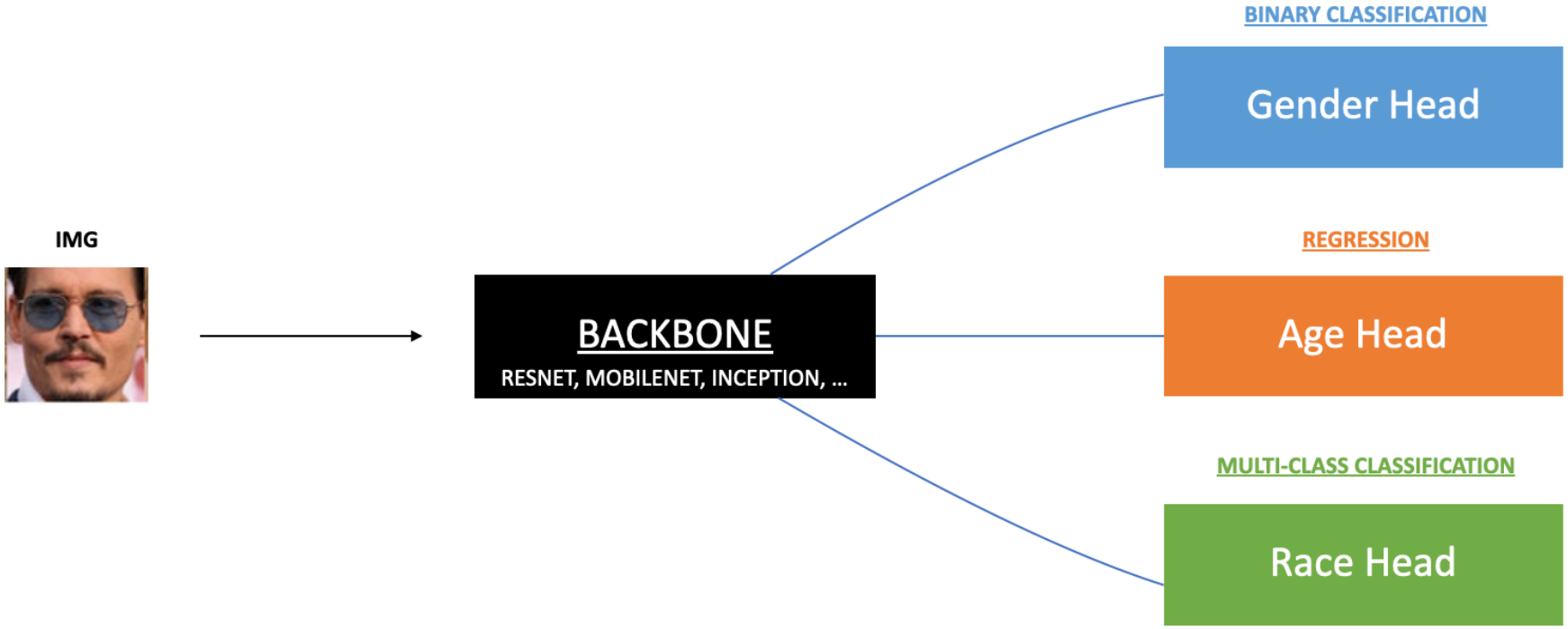
- Standard ML: build group-specific models



- Our proposed method: MTL with a **unified model** for two groups

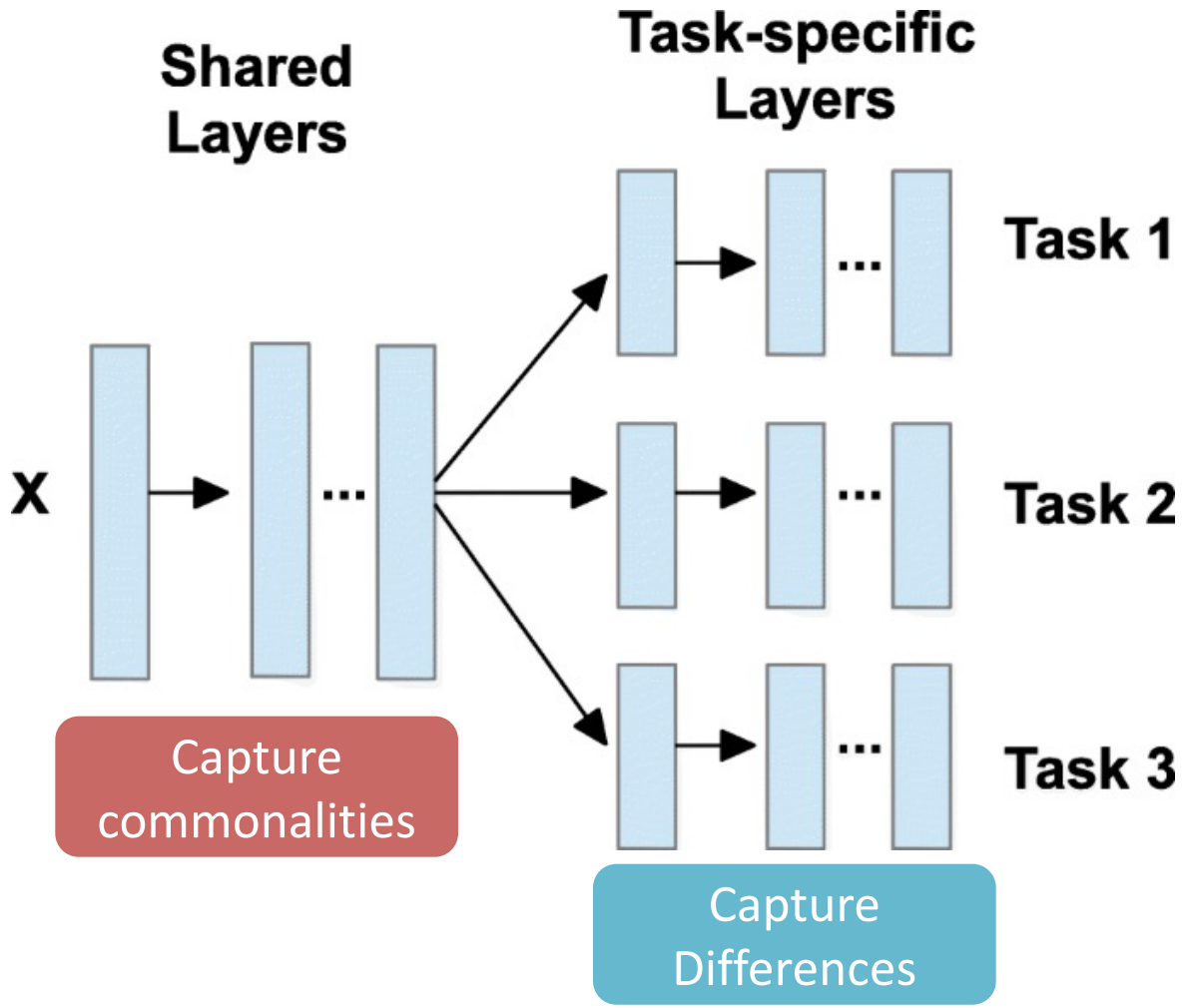


Background: MTL



1 input → 3 tasks (labels)
Utilize all the information from 3 tasks

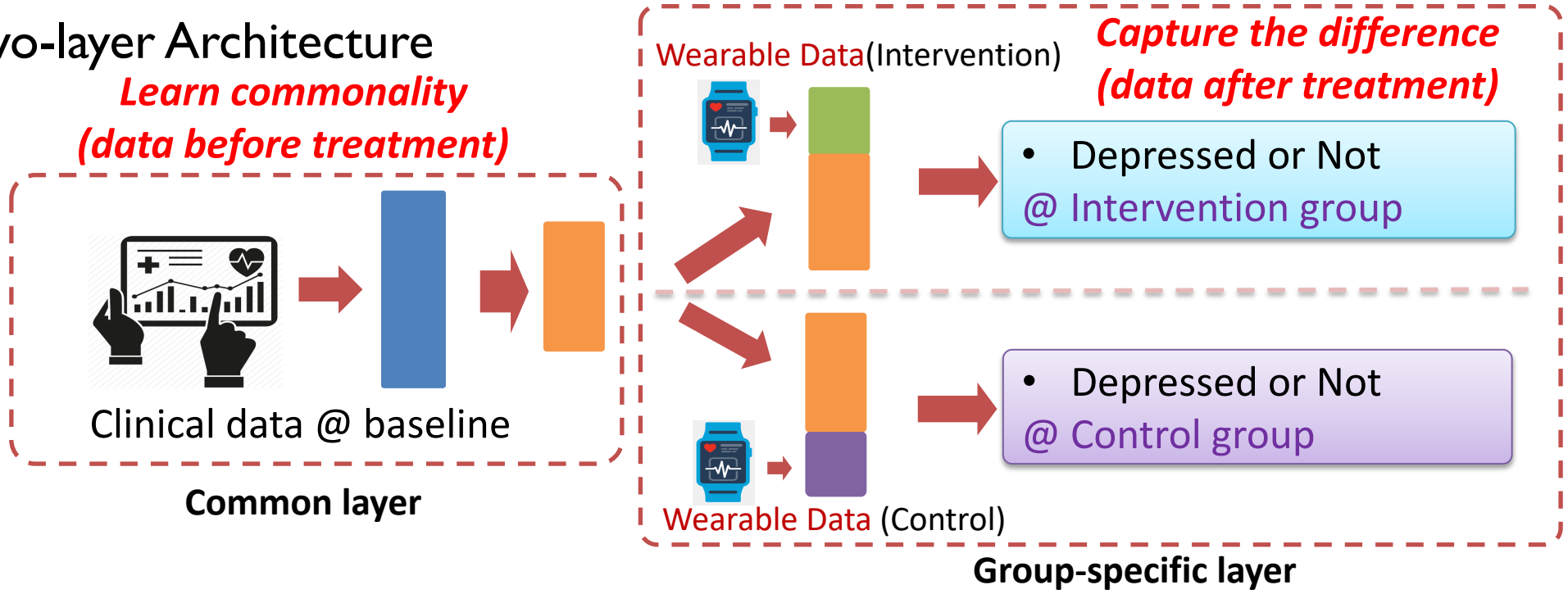
MTL Architecture



Final loss function = Weighted sum of task-specific loss functions

MTL Architectures for RCT

- Two-layer Architecture



- **Clinical data** at baseline
  - Statistically the same between 2 groups at baseline
    - E.g., age, gender, anxiety scores, etc.
- **Wearable Data** during first 2 month
  - Fitbit data is collected after therapy starts for the intervention group
    - E.g., heart rate, daily steps, sleep

Challenges

- Non-unified data
  - One label for one patient
- Task weight optimization
  - Balance weights between tasks
- Limited dataset size

Challenges and Proposed Methods

- Non-unified data
  - One label for one patient
- Task weight optimization
  - Balance weights between tasks
- Limited dataset size

- Label mask for each data sample

$$Loss_{clf} = \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} w_t \cdot mask_{n,t} \cdot BCE(y_{n,t}, \hat{y}_{n,t})$$

- Utilize task uncertainty for weights during training

- Besides MTL, use regularization and feature engineering approach

MTL with Uncertainty

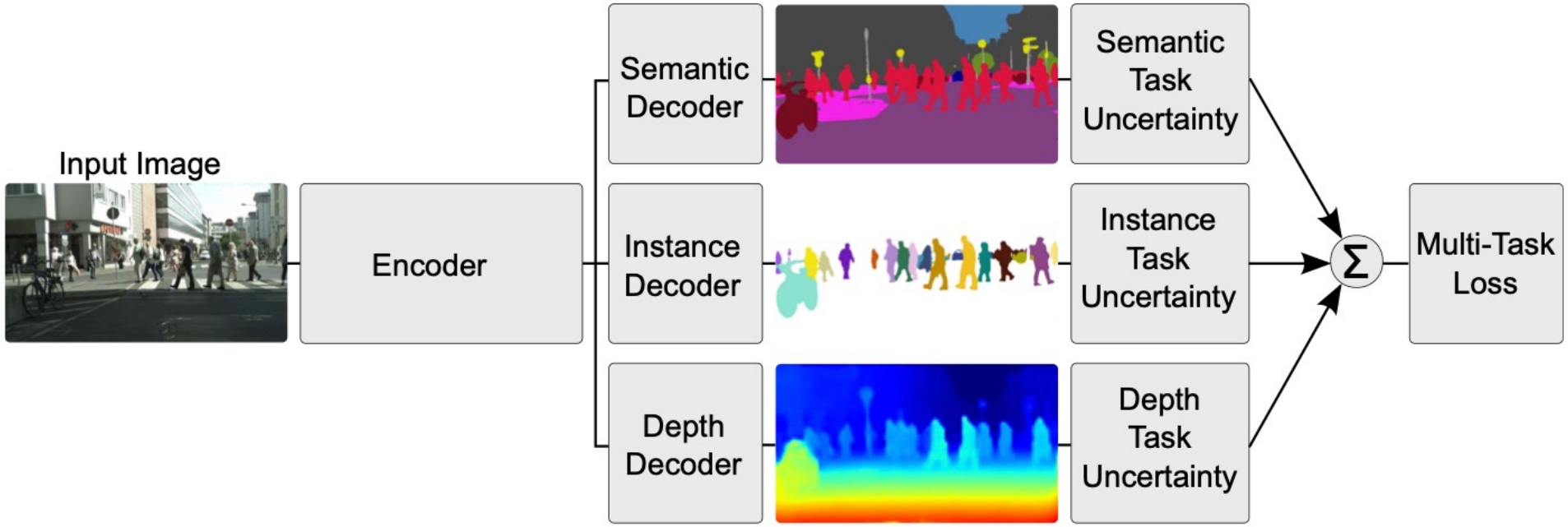


Figure 1: **Multi-task deep learning.** We derive a principled way of combining multiple regression and classification loss functions for multi-task learning. Our architecture takes a single monocular RGB image as input and produces a pixel-wise classification, an instance semantic segmentation and an estimate of per pixel depth. Multi-task learning can improve accuracy over separately trained models because cues from one task, such as depth, are used to regularize and improve the generalization of another domain, such as segmentation.

Higher uncertainty → Lower weight

[1]A. Kendall, Y. Gal, R. Cipolla, Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, Cvpr. (2018) 7482–7491.

Dynamic Task Weights for MTL

$$Loss_{MTL} = w_{INT} \cdot Loss_{INT} + w_{con} \cdot Loss_{CON}$$

- **Weights** control the contributions of each group to the overall model
  - Hyperparameters: fixed during training
  - Fixed weights could lead to negative transfers [1]
- Instead, use **uncertainty** to **automatically** optimize the weights [1]

$$Loss'_{MTL} \approx \frac{1}{\sigma_{INT}^2} \cdot Loss_{INT} + \frac{1}{\sigma_{CON}^2} \cdot Loss_{CON} + \log(\sigma_{INT}) + \log(\sigma_{CON})$$

"Pseudo weights" — regularization terms

  - σ_t is trainable and **dynamic changing** to balance the contributions

[1]A. Kendall, Y. Gal, R. Cipolla, Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, Cvpr. (2018) 7482–7491.

Trainable Uncertainty Factor

- A trainable uncertainty factor σ_t (for each task) is added to the sigmoid function to mimic the temperature in the Boltzmann distribution

$$y'_{n,t} = p(y_{n,t} = 1 | f_t(x_n), \sigma_t) = Sigmoid(\frac{1}{\sigma_t^2} f_t(x_n)) = \frac{e^{f_t(x_n)/\sigma_t^2}}{1 + e^{f_t(x_n)/\sigma_t^2}}$$

$$Loss'_{clf} = \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} mask_{n,t} \cdot BCE(y'_{n,t}, \hat{y}_{n,t})$$

$$\approx \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} \frac{1}{\sigma_t^2} \cdot mask_{n,t} \cdot BCE(y_{n,t}, \hat{y}_{n,t}) + \sum_{t=1}^{T} \log(\sigma_t)$$

Our RCT for Depression Treatment

- **106** patients with depression
- Remove the patients whose total yield of wearable data is less than 10%
  - **59** patients in intervention group
  - **30** patients in control group
- **58** clinical features
- **10** daily wearable features
  - Sedentary minutes, lightly active minutes, minutes of heart rate zone in cardio, minutes awake in main sleep…

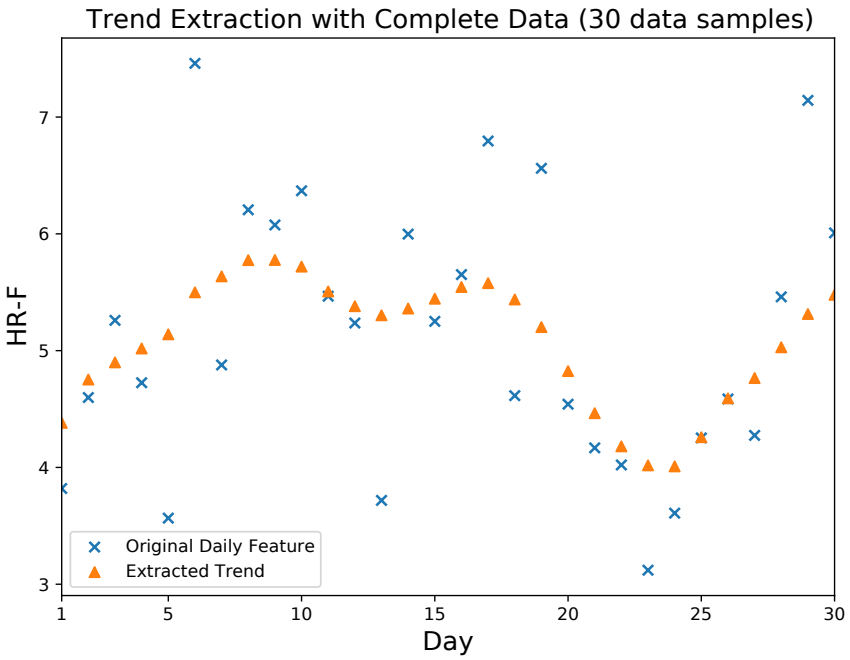
Variable	Intervention (n=59)	Control (n=30)	*p* value
Age, year	46.86 ± 12.11	47.47 ± 12.88	0.829
Female, %	76.2%	76.6%	0.824
Race/Ethnicity, %			0.191
Non-Hispanic White	18.6%	13.3%	
Non-Hispanic Black	55.9%	50%	
Asian/Pacific Islander	3.3%	0%	
Hispanic	13.5%	33.3%	
Other Race	8.4%	3.3%	
Education, %			0.497
High school/GED or less	6.7%	16.7%	
Some college	42.3%	33.3%	
College graduate	28.8%	30%	
Post college	22%	20%	
Weight, kg	100.73 ± 15.46	99.54 ± 14.19	0.726
SBP, mmHg	122.95 ± 14.96	120.63 ± 15.96	0.5
DBP, mmHg	77.25 ± 8.69	75.83 ± 10.45	0.498
Leisure, MET mins/week [43]	654.58 ± 819.72	953.20 ± 1131.43	0.158
Work, MET mins/week [43]	336.27 ± 937.05	217.33 ± 999.95	0.581
Energy expenditure, kcal/kg/d [52]	33.52 ± 2.23	33.64 ± 2.32	0.807
SPSI-R:S raw score [16]	12.88 ± 2.41	13.13 ± 2.40	0.639
PPO raw score [16]	10.59 ± 4.54	12.58 ± 3.33	0.037
NPO raw score [16]	7.07 ± 3.48	8.62 ± 3.52	0.051
RPS raw score [16]	10.46 ± 4.48	11.29 ± 3.67	0.382
ICS raw score [16]	4.05 ± 3.65	4.74 ± 3.22	0.382
AS raw score [16]	5.54 ± 4.54	4.84 ± 3.46	0.461
PROMIS sleep disturbance t score [78]	57.96 ± 7.44	57.08 ± 8.05	0.611
PROMIS sleep impairment t score [78]	56.95 ± 8.80	54.83 ± 9.09	0.291
SCL-20 score [24]	1.21 ± 0.67	1.15 ± 0.59	0.695
GAD-7 score [70]	7.10 ± 5.12	6.63 ± 3.77	0.659
PTSD severity score [4]	36.12 ± 14.57	33.77 ± 10.25	0.432
SF-8 physical component score [73]	45.00 ± 8.48	47.38 ± 8.30	0.21
SF-8 mental component score [73]	39.65 ± 11.37	42.57 ± 9.00	0.225
COPE total scores [9]	57.02 ± 12.09	57.40 ± 14.63	0.896
BRISC total scores [77]	30.12 ± 6.23	31.07 ± 6.06	0.495

Abbreviations: AS, avoidance style; COPE, COPE Inventory survey, including 14 components [9]; BRISC, BRISC questionnaire of emotional resilience and self-efficacy Survey, including 3 components [77]; DBP, diastolic blood pressure; GAD-7, generalized anxiety disorder scale-7; ICS, impulsivity/carelessness style; MET, metabolic equivalent task; NPO, negative problem orientation; PPO, positive problem orientation; PROMIS, Patient-Reported Outcomes Measurement Information System; PTSD, post-traumatic stress disorder; RPS, rational problem solving; SBP, systolic blood pressure; SCL-20, Symptom Checklist-20; SF-8, Short Form 8 Health Survey; SPSI-R:S, Social Problem Solving Inventor -Revised: Short Form.

Data Preprocessing

- For wearable features, singular spectrum analysis is used to extract the trend (10 features)
- Maximum, minimum, median, slope, and intercept are computed from the extracted trend (10*5)
- Feature selection for all features
  - ANOVA for continuous features
  - Chi2 for categorical features

  (10 clinical features + 10 wearable features)



Selected Features

Selected clinical features				Selected wearable features			
Features	Ave. stat	Ave. p	Times	Features	Ave. stat	Ave. p	Times
PROMIS sleep disturbance t score	15.68	2.95E-04	100	distanceTotal_median	5.54	3.55E-02	99
PROMIS sleep impairment t score	12.43	1.24E-03	100	restlessCounts_intercept	2.61	1.23E-01	99
SCL20 baseline score	8.88	5.81E-03	100	restlessCounts_min	2.75	1.19E-01	98
NPO raw score	5.41	3.03E-02	100	minutesAwake_intercept	2.31	1.45E-01	97
COPE acceptance score	5.20	3.36E-02	99	minutesAwake_min	2.94	1.17E-01	91
PTSD severity score	4.70	4.49E-02	99	distanceTotal_intercept	2.80	1.30E-01	86
Work, MET mins/week	5.06	3.10E-02	97	restlessCounts_slope	3.47	9.76E-02	84
Sex	4.27	4.77E-02	90	lightlyActiveMinutes_min	2.11	1.78E-01	75
COPE denial score	4.36	4.89E-02	89	minutesAwake_slope	2.87	1.27E-01	64
BRISC skill scores	4.16	5.21E-02	65	HRzoneFatBurnMinutes_min	1.92	1.90E-01	57
SF-8 mental component score	3.81	6.33E-02	40	restlessCounts_max	2.16	1.70E-01	42
SF-8 physical component score	3.94	5.84E-02	12	distanceTotal_max	2.00	1.95E-01	31
GAD-7 score	3.70	6.52E-02	4	HRzoneFatBurnMinutes_slope	2.37	1.72E-01	30
DBP	3.09	8.34E-02	2	sedentaryMinutes_median	1.98	1.84E-01	22
COPE plan score	3.04	8.57E-02	1	sedentaryMinutes_min	1.76	2.13E-01	12
COPE active score	2.22	1.41E-01	1	HRzoneFatBurnMinutes_median	2.56	1.69E-01	4
Energy expenditure	2.15	1.47E-01	1	lightlyActiveMinutes_intercept	1.98	1.79E-01	3
				sedentaryMinutes_slope	1.68	2.12E-01	2
				minutesAwake_max	1.21	2.77E-01	2
				minutesAwake_median	1.49	2.27E-01	1
				lightlyActiveMinutes_max	1.28	2.61E-01	1

*Ave. stat: average statistic of either ANOVA test or χ^2 test from different CV splits.
*Ave. p: average p−value of either ANOVA test or χ^2 test from different CV splits.

20 runs of 5-fold cross validation

Regularization to Avoid Overfitting

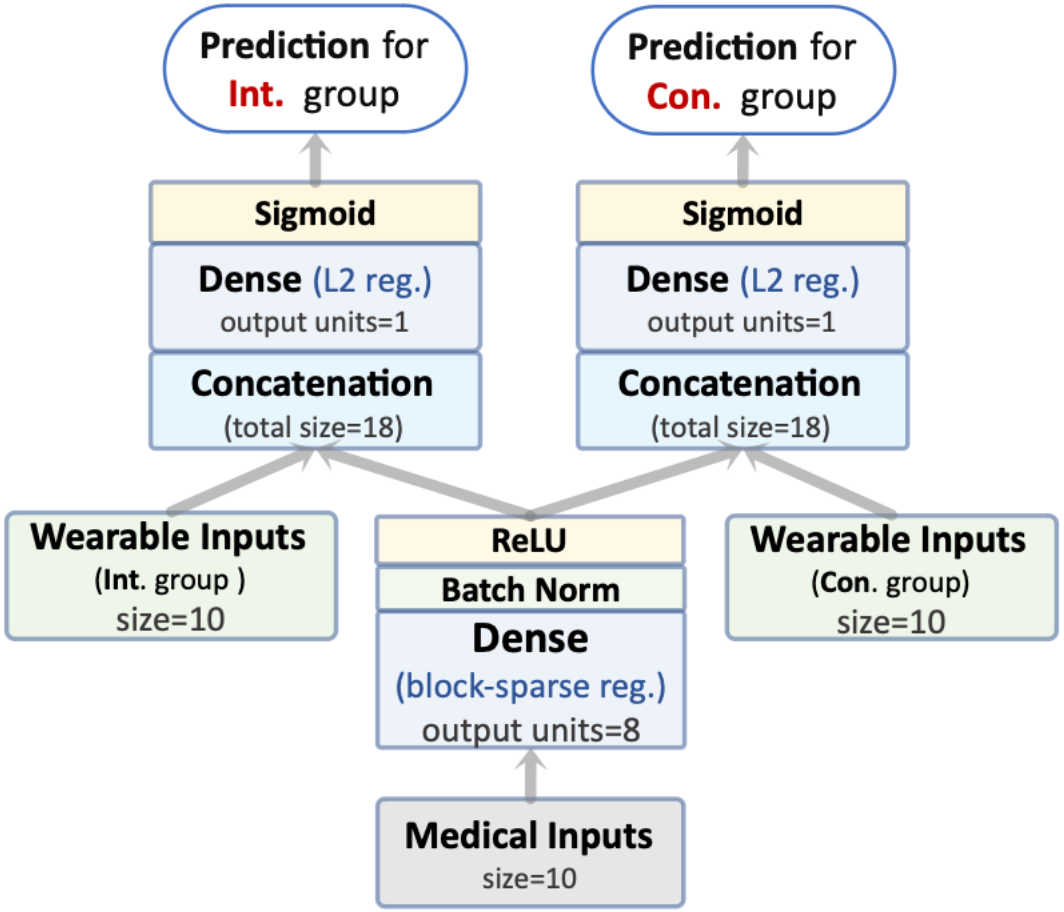
- Block-sparse regularization for shared layer

$$R_{blk} = \| |a_1|_2, |a_2|_2, ..., |a_n|_2 \|_1$$

- L2 regularization for task-specific layer

$$R_{tsk,j} = \|b_j\|_2$$

Key idea: less parameters to avoid overfitting



Evaluation Setting

- **Single task learning (STL):** train separate models for each group
  - Logistic regression (LR), Support Vector Machine (SVM), Gradient Boosting(GBDT), Random Forest (RF), AdaBoost (Ada), Shallow deep neural network
- **MTL** with different settings
  - **MTL-1**: MTL trained on a **single** group separately -> reduced to group-specific
  - **MTL-fixed**: MTL trained on the **combined** group with **fixed** weights
    - Using **grid-search** to find the best weights for each group
      - We found the fixed weights cannot achieve the best for both
  - **MTL-dynamic** (ours) **:** MTL trained on the **combined** group with **dynamic** weights

Evaluations

		Intervention		Control	
Category	Model	AUROC	AUPRC	AUROC	AUPRC
STL-separate	SVM	0.607(0.053)	0.556(0.052)	0.734(0.063)	0.561(0.112)
	RF	0.667(0.040)	0.601(0.046)	0.755(0.066)	0.538(0.089)
	Ada	0.615(0.042)	0.567(0.045)	0.681(0.076)	0.433(0.086)
	GBDT	0.657(0.061)	0.582(0.060)	0.724(0.098)	0.487(0.072)
	ANN	0.659(0.040)	0.581(0.051)	0.754(0.081)	0.548(0.136)
	LR	**0.697(0.050)**	**0.636(0.063)**	**0.794(0.067)**	**0.601(0.093)**
STL-unified	SVM	0.533(0.062)	0.477(0.057)	0.683(0.107)	0.458(0.091)
	RF	0.649(0.053)	0.587(0.063)	0.704(0.077)	0.474(0.086)
	Ada	0.609(0.056)	0.562(0.062)	0.619(0.093)	0.340(0.064)
	GBDT	**0.650(0.060)**	**0.596(0.067)**	0.555(0.089)	0.332(0.076)
	ANN	0.629(0.071)	0.561(0.054)	0.734(0.045)	0.528(0.056)
	LR	0.639(0.049)	0.571(0.059)	**0.759(0.047)**	**0.569(0.084)**
MTL	MTL-1[1]	0.695(0.032)	0.641(0.051)	0.784(0.049)	0.589(0.073)
	MTL-fixed[2]	0.707(0.052)	0.653(0.046)	0.807(0.063)	0.615(0.083)
	MTL-dynamic	**0.725(0.059)**	**0.668(0.068)**	**0.813(0.077)**	**0.637(0.061)**

[1]MTL-1 is the single task learning but with the same architecture of our proposed MTL model.
[2]MTL-fixed is trained on the combined group, but the performances of two tasks are from separate models.

- Difference between two tasks matters
- MTL trained on the combined groups improves performance
- MTL with dynamic weights → the best performance for both groups

Evaluations(2): fixed weights vs dynamic weights

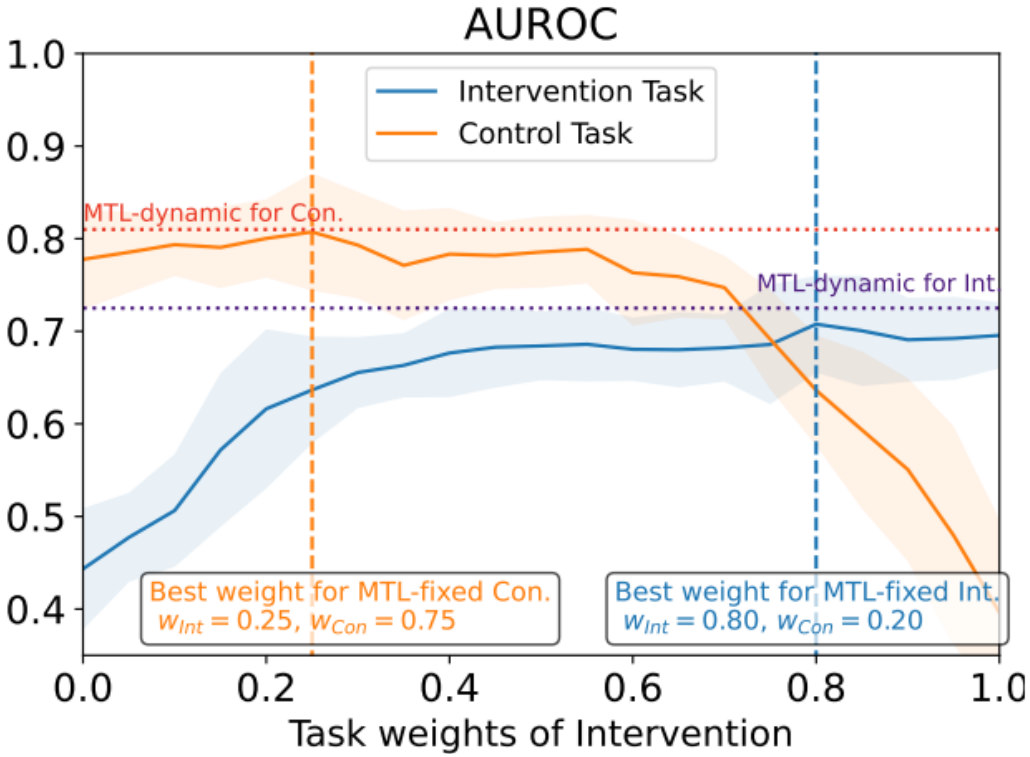
- MTL-fixed: grid search for best weights

$$Loss_{MTL} = w_{INT} \cdot Loss_{INT} + w_{con} \cdot Loss_{CON}$$

  - Assume: $\boldsymbol{w_{INT} + w_{CON} = 1}$

- MTL-dynamic: dynamically changing the weights

$$Loss \approx \frac{\mathbf{1}}{\boldsymbol{\sigma_{INT}^2}} \cdot Loss_{INT} + \frac{\mathbf{1}}{\boldsymbol{\sigma_{CON}^2}} \cdot Loss_{CON} + \log(\boldsymbol{\sigma_{INT}}) + \log(\boldsymbol{\sigma_{CON}})$$

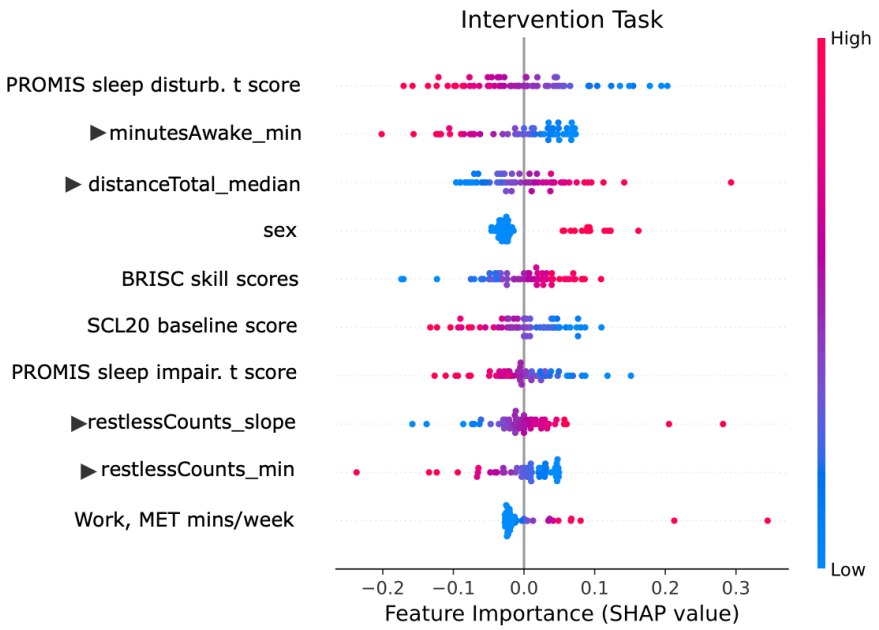


Contribution of Wearable Data

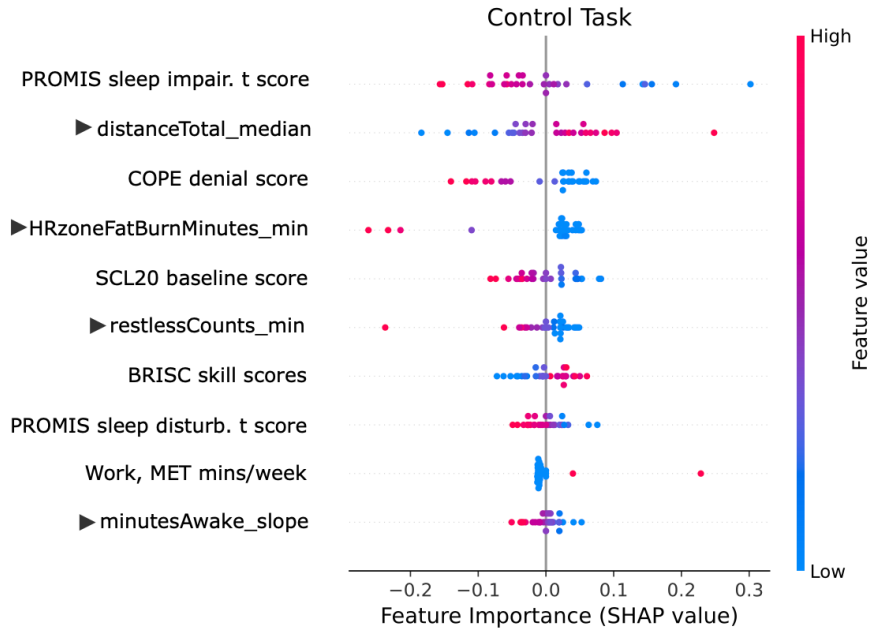
	Intervention		Control	
	AUROC	AUPRC	AUROC	AUPRC
W/ wearable data	**0.725(0.059)**	**0.668(0.068)**	**0.813(0.077)**	**0.637(0.061)**
W/O wearable data	0.652(0.041)	0.601(0.056)	0.727(0.081)	0.513(0.088)

Feature Importance with SHAP

- SHAP (SHapley Additive exPlanations) is used to compute feature importance
  - Game theory
  - Evaluate the contribution of each feature



(a) Shap summary plot for the intervention task



(b) Shap summary plot for the control task

Conclusions

- Proposed **multi-task learning** framework for **randomized control trail**
  - The dynamic weights in MTL balance task contributions
- Presented a **case study** of a real RCT for depression treatment
  - Combining wearable data and clinic data
  - Our proposed MTL method shows the best performance

Limitation

- Assuming the treatment does not change after group split and the model needs to be trained retrospectively on groups of patients
- Limited sample size
- No evaluation about the impact of different lengths of wearable data