MuHBoost: Multi-Label Boosting for Practical Longitudinal Human Behavior Modeling

Nate Thach[1]*, Patrick Habecker[1], Anika Eisenbraun[1], Alex Mason[1], Kimberly Tyler[1], Bilal Khan[2], Hau Chan[1]

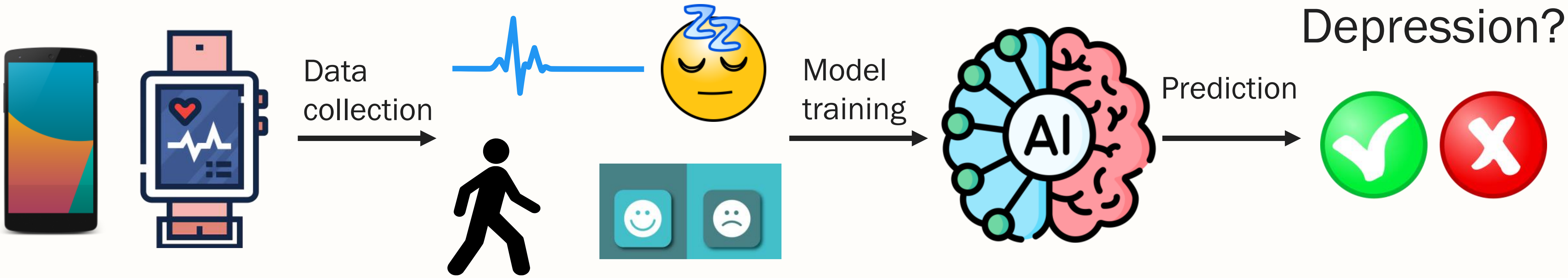
[1]University of Nebraska-Lincoln, [2]Lehigh University

ICLR 2025--Singapore

April 26, 2025

Longitudinal Human Behavior Modeling

- **Multimultidisciplinary:** psychology, human-computer interaction, ubiquitous computing, machine learning (ML)
- Collect **ubiquitous health data**, then use ML techniques to build predictive models for health/well-being outcomes



Over the last 2 weeks, how often have you been bothered by the following problems? *(circle one per question)*		Not at all	Several Days	More than half the days	Nearly every day
1	Feeling nervous, anxious, or on edge	0	1	2	3
2	Not being able to stop or control worrying	0	1	2	3
3	Little interest or pleasure in doing things	0	1	2	3
4	Feeling down, depressed, or hopeless	0	1	2	3

Ubiquitous Health Data

- Time series-like w/ rich contextual info
  - ❑ Passive sensing: resting heart rate, geolocation, phone activities
  - ❑ Self-reports: mood measures from questionnaires e.g., PANAS, PHQ-4
- Challenges
  - ► Span long time periods → long time series
  - ► High rate of missing values (sparse data)
  - ► Small sample size (~100 labeled data points or less)

(Limitations of) Related Work

- Traditional ML (e.g., SVM, random forest, boosting) and deep learning models (CNN, ResNet) yield <50% accuracy
- Most recent works employ LLMs and saw promising results, BUT:
  - ▶ Only considering standard (numerical) time series, whereas in practice:
    - ❖ Can be categorical or other types
    - ❖ Even more missing data (due to various functionalities from data collection platform)
  - ▶ Not addressing resource consumption from LLMs: (1) computing power, (2) training time, (3) price of calling APIs (e.g., GPT-3.5+)

(Limitations of) Related

- Traditional ML (e.g., SVM, random fo
  learning models (CNN, ResNet) yield
- Most recent works employ LLMs and
  - ▶ Only considering standard (numerical)
    - ❖ Can be categorical or other types
    - ❖ Even more missing data (due to various functionalities from data collection platform)
  - ▶ Not addressing resource consumption from LLMs: (1) computing power, (2) training time, (3) price of calling APIs (e.g., GPT-3.5+)

TB3 **In the past 6 months, have you smoked cigarettes?**

No (0) — SKIP TO TB5
Yes (1)

TB4 **How many cigarettes do you usually smoke in a day?**

Not at all (1)
Less than 1 cigarette a day (2)
1-5 cigarettes a day (3)
Half a pack a day (4)
A pack or more a day (5)

Figure 1: Skip Logic: If respondents answer "No" in TB3, they will automatically be directed to TB5 without being asked on TB4.

Given the following drugs, [DRUG LIST], predict whether this drug user is at-risk from using it or not. Return an array of “Yes” and “No”.

MuHBoost

- **M**ulti-**u**biquitous-**H**ealth **Boost**ing framework
- Idea: Use LLM within a boosting framework for **multi-label classif.**
- Extend SummaryBoost (originally for tabular data classif.):
  1. Convert each data point (time series + contextual info + label) into natural language via LLM prompting
  2. Leverage LLM to generate **weak learners** from a subset of the converted data points

     Each weak learner consists of multiple summaries, each capturing the relationship between a label vs. the provided (time series + contextual info). During inference, LLM is provided with a weak learner and asked to generate an array of labels at once.
  3. Boosting the weak learners (with AdaBoost)

MuHBoost Variants

- Due to LLM hallucinations, MuHBoost may underperform when tasked with too many labels

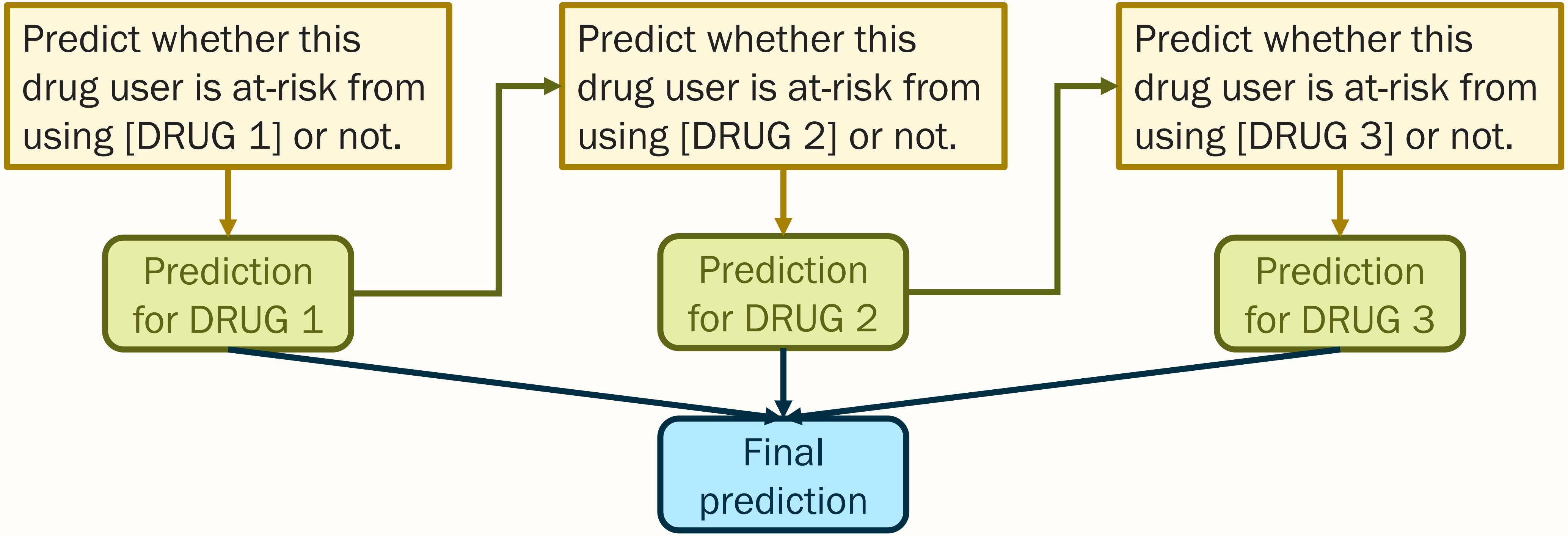
1. MuHBoost[LP+]: rephrase the prompt during inference

Given the following drugs, [DRUG LIST], predict whether this drug user is at-risk from using it or not. Return an array of “Yes” and “No”.

↓

Predict which drug(s) from the following list this drug user is at-risk from using it, or “None” otherwise: [DRUG LIST].

MuHBoost Variants



2. MuHBoost[CC]: AdaBoost + classifier chain (CC)

Boost each label independently and link the single-label predictions together in a chain

Experimental Setups

- **Datasets** (# of tasks/labels in parentheses):
  - ❑ LifeSnaps (2) & GLOBEM (3): mental health (e.g., anxiety, stress, depression)
  - ❑ CoSt (2): Undergraduates at UNL – student performance in a CS course
  - ❑ PWUD (6): Drug users across Nebraska – whether at risk of using drugs
- **Baselines:** Traditional ML models for multi-label classif., zero-shot and few-shot prompting with GPT-4
- **Evaluation metrics:** Hamming accuracy, micro-F1, macro-F1

Results

Method	LifeSnaps	LifeSnaps+	GLOBEM	GLOBEM+	CoSt	CoSt+	PWUD	PWUD+
0-shot[BR]	[16 16 17]	[14 15 14]	[18 17 16]	[13 15 15]	[14 14 13]	[13 12 12]	[14 14 15]	[13 12 13]
0-shot[LP]	[18 18 19]	[15 14 15]	[17 18 18]	[15 16 14]	[17 18 18]	[15 16 16]	[16 18 19]	[17 16 17]
0-shot[LP+]	[17 17 16]	[12 13 13]	[16 14 17]	[10 12 11]	[18 17 17]	[16 15 15]	[18 17 16]	[15 15 14]
10-shot[BR]	[13 12 11]	[8 7 7]	[12 10 10]	[8 9 7]	[10 9 9]	[7 8 7]	[8 10 8]	[7 7 6]
10-shot[LP]	[11 10 12]	[9 9 8]	[11 11 13]	[9 8 9]	[12 13 14]	[9 11 10]	[12 13 12]	[9 11 11]
10-shot[LP+]	[10 11 10]	[7 8 9]	[14 13 12]	[7 7 8]	[11 10 11]	[8 6 8]	[11 9 10]	[10 8 9]
RF[CC]	[21 19 18]	[22 21 20]	[23 22 19]	[24 23 23]	[30 27 29]	[27 29 30]	[28 30 30]	[26 29 28]
RF[LP]	[28 29 30]	[29 27 29]	[27 26 27]	[28 29 30]	[25 24 22]	[24 23 25]	[25 19 20]	[22 24 22]
XGBoost[CC]	[19 20 23]	[20 22 21]	[22 21 20]	[20 19 24]	[23 25 27]	[26 28 24]	[24 23 27]	[27 28 24]
XGBoost[LP]	[25 30 28]	[30 26 25]	[21 24 21]	[29 28 26]	[28 30 28]	[29 26 26]	[29 27 29]	[30 26 25]
MLkNN	[24 23 22]	[26 24 24]	[19 20 22]	[30 25 28]	[21 19 20]	[19 22 23]	[23 20 18]	[21 21 23]
MLTSVM	[23 25 27]	[27 28 26]	[25 27 29]	[26 30 25]	[22 20 21]	[20 21 19]	[19 25 21]	[20 22 26]
MuHBoost	[6 6 5]	[2 2 3]	[5 6 6]	[3 3 3]	[6 7 6]	[4 5 5]	[6 6 7]	[5 5 6]
MuHBoost[LP+]	[5 4 6]	[3 3 2]	[6 4 4]	[**1** 2 2]	[5 4 4]	[2 2 2]	[3 3 4]	[**1** **1** **1**]
MuHBoost[CC]	[4 5 4]	[**1** **1** **1**]	[4 5 5]	[2 **1** **1**]	[3 3 3]	[**1** **1** **1**]	[4 4 3]	[2 2 2]

Table 1: Performance ranking (↓) of 15 methods subject to [HA miF_1 maF_1]. (+) next to the datasets' tags denotes incorporation of auxiliary data in addition to time-series data (hence there are $15 \times 2 = 30$ methods in total).

Results

MuHBoost, w/ or w/o integrating contextual info into the models, already outperforms all baselines in most cases

The 2 MuHBoost variants bring further improvements and yield the overall best performance.

MuHBoost	[6 6 5]	[2 2 3]	[5 6 6]	[3 3 3]	[6 7 6]	[4 5 5]	[6 6 7]	[5 5 6]
MuHBoost[LP+]	[5 4 6]	[3 3 2]	[6 4 4]	[1 2 2]	[5 4 4]	[2 2 2]	[3 3 4]	[1 1 1]
MuHBoost[CC]	[4 5 4]	[1 1 1]	[4 5 5]	[2 1 1]	[3 3 3]	[1 1 1]	[4 4 3]	[2 2 2]

Table 1: Performance ranking (↓) of 15 methods subject to [HA miF_1 maF_1]. (+) next to the datasets' tags denotes incorporation of auxiliary data in addition to time-series data (hence there are $15 \times 2 = 30$ methods in total).

Resource Consumption

- Computing power: negligible locally (via GPT)
- Training time and price of calling APIs (GPT-3.5):
  - ➢ Multi-label classif. helps reduce both time and cost
  - ➢ MuHBoost[CC] consumes more resources in exchange for better accuracy

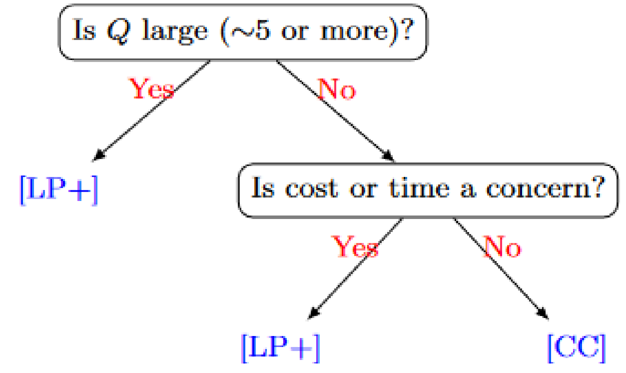


Figure S3: Decision tree for considering the two MuHBoost variants.

Conclusion

- MuHBoost tackles more practical forms of ubiquitous health data while simultaneously focusing on resource efficiency
- Help domain experts quickly develop effective personalized prevention or intervention strategies for at-risk individuals

NEBRASKA

ICLR

Thank you!

National Institutes of Health

NSF

RURAL DRUG ADDICTION RESEARCH CENTER

Nate Thach

nate.thach@huskers.unl.edu

https://openreview.net/pdf?id=BAeIAyADqn