



Augmented Risk Prediction for the Onset of Alzheimer's Disease from Electronic Health Records with Large Language Models

Jiankun Wang¹, Sumyeong Ahn², Taykhoom Dalal³, Xiaodan Zhang², Weishen Pan³, Qiannan Zhang³, Junyuan Hong⁴, Bin Chen², Hiroko H. Dodge⁴, Fei Wang³, Jiayu Zhou¹

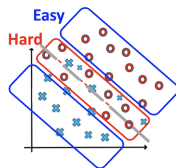
¹ University of Michigan, ² Michigan State University, ³ Weill Cornell Medicine, Cornell University, ⁴ Massachusetts General Hospital, Harvard Medical School



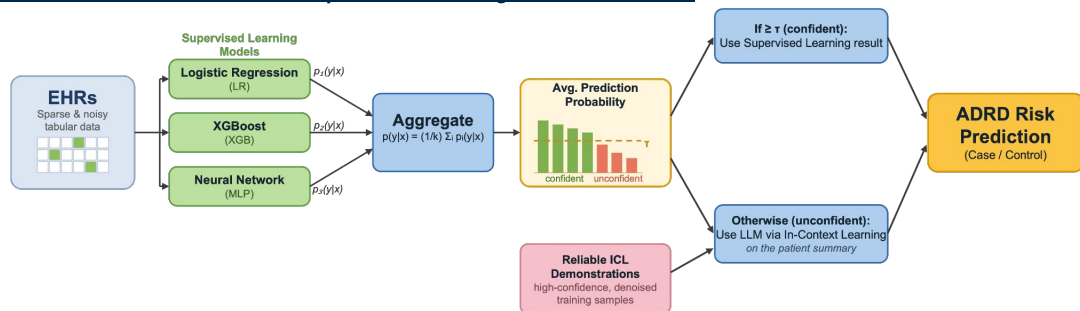
Full Paper

Alzheimer's Disease Risk Prediction

- Traditional Supervised Learning**
 - Pros:** strong learning capability to extract common pattern on **easy** samples
 - Cons:** struggle to handle hard samples with sparse or noisy features
- LLMs (In-context Learning)**
 - Pros:** strong reasoning power to handle **hard** samples given the specific context
 - Cons:** struggle to capture common pattern with limited context



Confidence-driven Collaboration between Supervised Learning Models and LLMs



1 Train SLs; average their prediction probabilities.

2 Confident ($\geq \tau$): trust the supervised models.

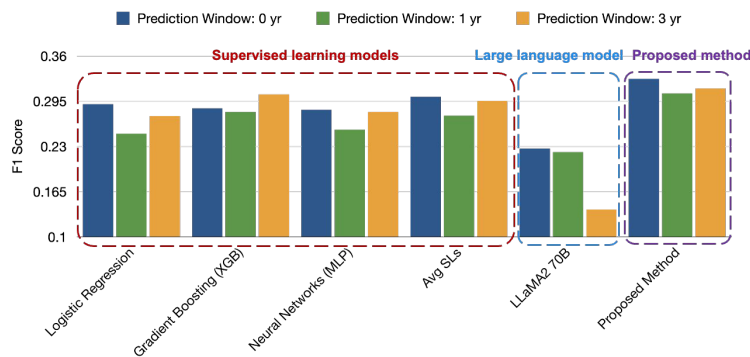
3 Unconfident ($< \tau$): defer to the LLM via cleaned ICL.

Performance of **easy/hard** samples

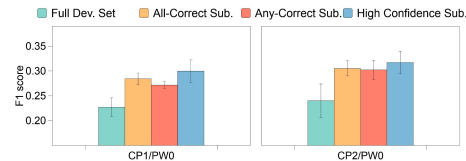
Dataset	Model	Precision	Recall	F1 score
CP1 / PW1	LR	.168 ±.027	.616 ±.088	.262 ±.034
	XGB	.233 ±.018	.392 ±.035	.292 ±.024
	MLP	.209 ±.037	.443 ±.032	.283 ±.041
	Avg. SLs	.230 ±.028	.444 ±.029	.302 ±.029
	LLaMA2 70B	.150 ±.009	.394 ±.045	.217 ±.015
Dataset	Model	Precision	Recall	F1 score
CP1 / PW1	LR	.110 ±.026	.822 ±.100	.192 ±.038
	XGB	.148 ±.054	.233 ±.127	.178 ±.077
	MLP	.072 ±.037	.487 ±.271	.125 ±.065
	Avg. SLs	.109 ±.038	.519 ±.217	.180 ±.064
	LLaMA2 70B	.160 ±.030	.542 ±.070	.247 ±.041

* Data from Oregon Health & Science University (OHSU) Hospital

Detection of ADRD Using Different Prediction Windows

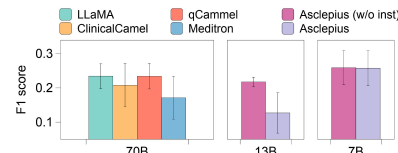


Study on Different Denoising Strategies in ICL



Denoising strategies **significantly enhance** the performance of ICL

Impact of LLM Model Size/Medical Data Fine-tuning



Fine-tuning **does not always improve** the target task's performance. Better performance can be achieved **even with small models**

Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant IIS-2212174, IIS-1749940, Office of Naval Research N00014-24-1-2168, and National Institute on Aging (NIA) RF1AG072449.

Ours: the superiority in F1 score, demonstrating the effectiveness of **combining the strengths of traditional SLs and LLMs** in risk prediction.