Human-centered Explainable Psychiatric Disorder Diagnosis System using Wearable ECG Monitors

Hung Nguyen[1,2], Alireza Rahimi[1], Veronica Whitford[1], Hélène Fournier[2], Irina Kondratova[2], René Richard[2,1], Hung Cao[1]

[1] Analytics Everywhere Lab, University of New Brunswick, Canada
[2] National Research Council Canada, Canada

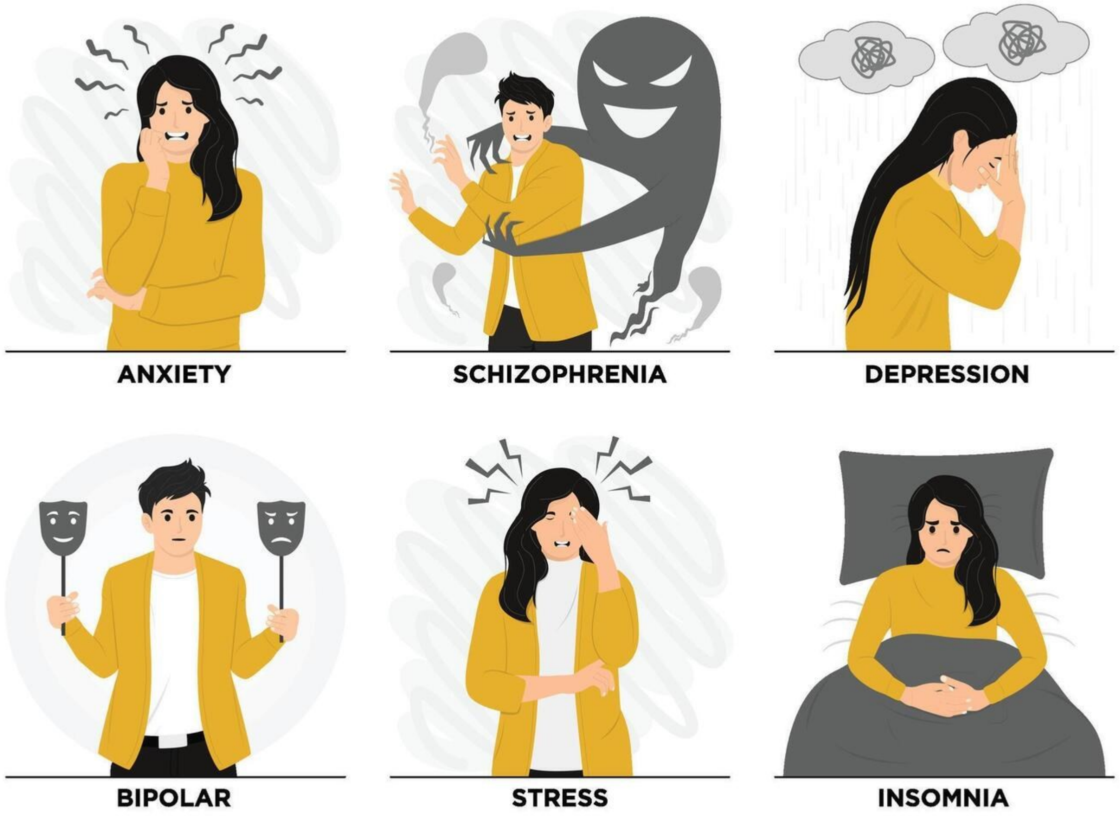
Motivation

Psychiatric Disorders

Psychiatric disorders: wide range of mental health conditions that affect the mood, thinking, and behavior.

- **Schizophrenia (SZ):** psychotic symptoms (e.g., delusions, hallucinations) and negative symptoms (e.g., alogia, blunted affect).
- **Bipolar disorder (BP):** extreme mood swings with manic and depressive episodes, sometimes accompanied by psychotic features.

Affecting **1–3% of the global population.**



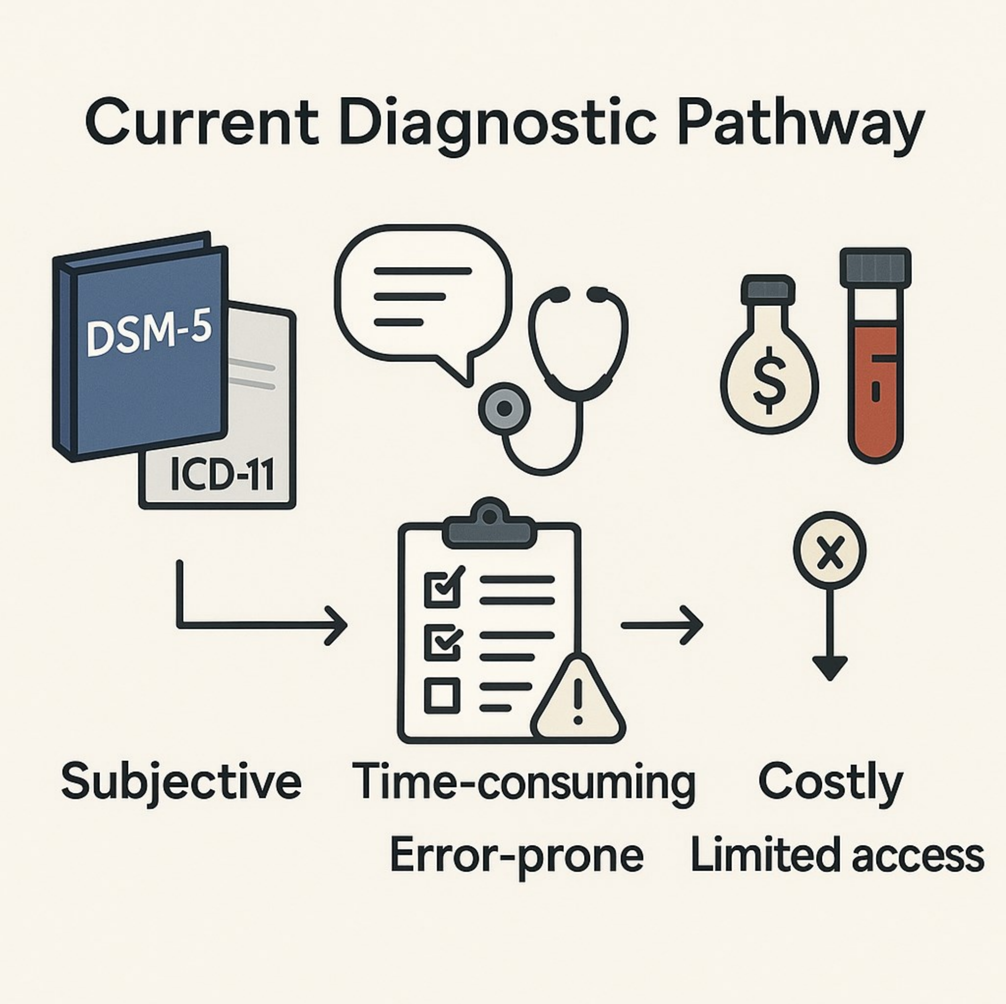
Limitations of current diagnoses

While diagnostic criteria exist:

- **Diagnostic and Statistical Manual of Mental Disorders (DSM-5)**
- **International Classification of Diseases (ICD-11)**

Current diagnosis relies heavily on subjective self-reports and clinical interviews.

- **Positive and Negative Syndrome Scale (PANSS):** Despite efforts to enhance objectivity, the process remains time-consuming and error-prone.
- **Blood biomarker** approaches require costly laboratory testing inaccessible in many regions.



Potential of Wearable Devices

Hand grip

ECG Patches

Chest strap

https://www.polar.com/ca-en/sensors/h10-heart-rate-sensor

https://www.verniercanada.ca/product/sensors/heart-rate-sensors/hand-grip-heart-rate-monitor/

https://www.vivalink.com/wearable-ecg-monitor

Smartwatches

https://www.apple.com/ca/watch/

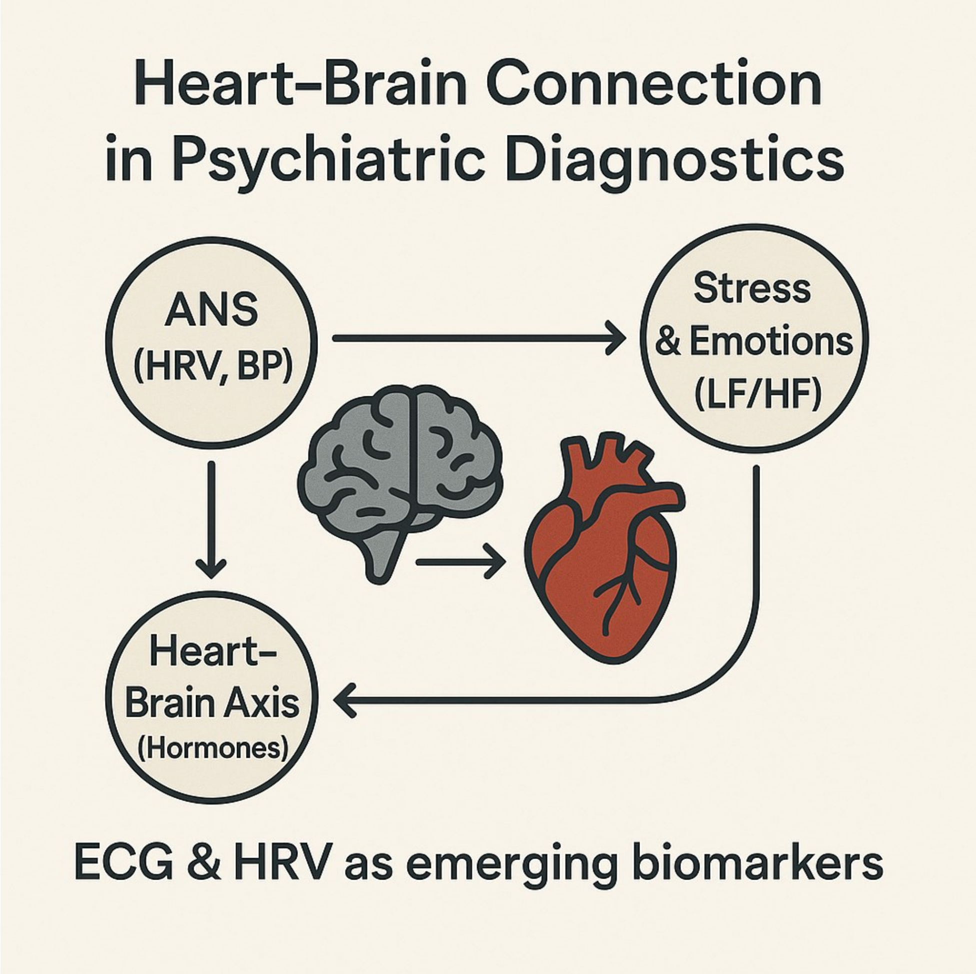
https://www.samsung.com/ca/watches/galaxy-watch/

https://www.empatica.com/research/e4/

Heart-Brain Connection

Psychiatric disorders often involve physiological dysregulation that can be captured via cardiovascular biomarkers:

- **Autonomic nervous system (ANS):** Heart functions such as heart rate (HR), blood pressure (BP).
- **Stress & Emotions:** High-frequency (HF) power or alterations in low-frequency (LF) power or the LF/HF ratio.
- **Heart-Brain Feedback Loop (Heart-Brain Axis):** Serotonin and oxytocin.



Human-centered Explainable AI (XAI)

XAI reveals crucial insights into AI decision-making processes:

- **How can we make explanations more “human-centered” for end-users?**
- In sensitive contexts, the ability to validate or challenge AI models through explanations?

Montréal Declaration Responsible AI_

Health Canada Santé Canada

U.S. FOOD & DRUG ADMINISTRATION

AI Act

GDPR

Government of Canada Gouvernement du Canada

MENU

Canada.ca > How government works > Policies, directives, standards and guidelines

Directive on Automated Decision-Making

Medicines & Healthcare products Regulatory Agency

Our Contributions

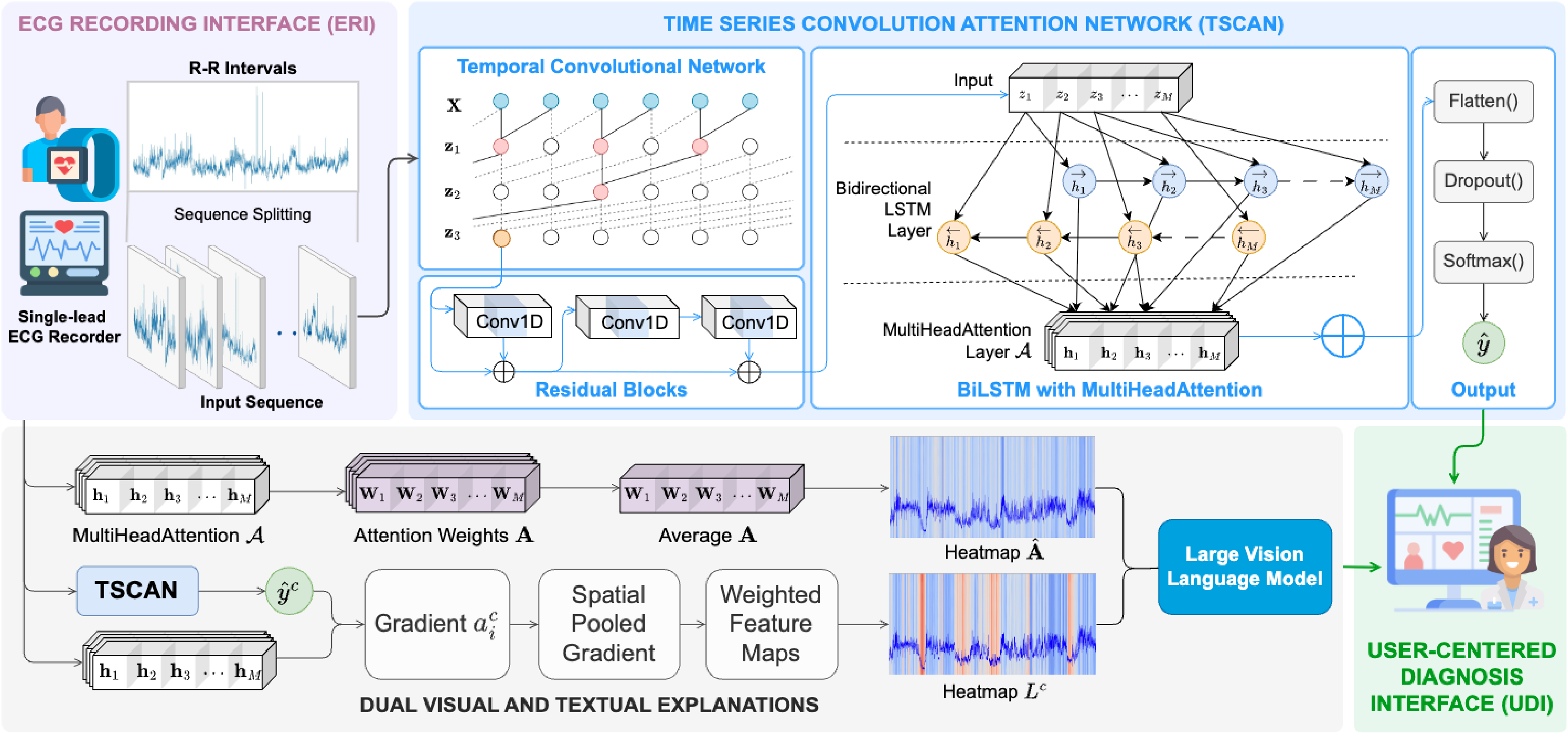
“Psychiatric Disorder Diagnosis System using Wearable ECG Monitors in a Human-centered Explainable manner?”

Our contributions are:

- **Psychiatric Disorder Diagnosis System:** Time Series Convolutional Attention Network (TSCAN) for wearable single-lead ECG to detect psychiatric disorders.
- **Dual Visual and Textual Explanations:** two visual explanation methods + large vision language models (LVLMs).
- **Evaluation and Applicability:** performance in psychiatric disorders detection and Afib detection.

Psychiatric Disorder Detection System

Architecture



ECG Recording Interface (ERI)

Capture and process the ECG data

Polar H9 or Polar H10

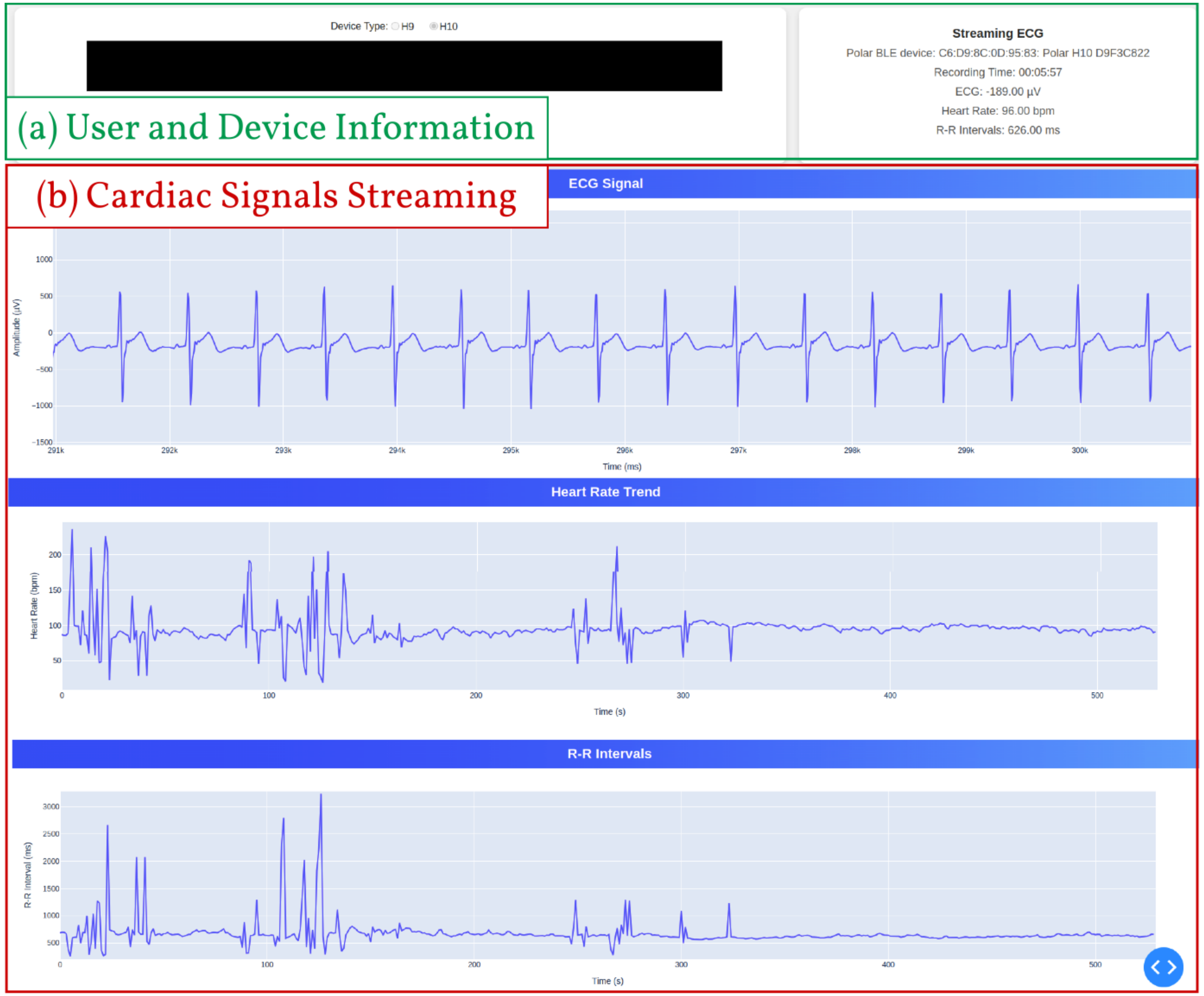
- Capture HR in BPM and R-R Intervals (RRI) in milliseconds with a 1-sec sampling rate.
- Polar H10 can record a single-lead ECG at 130 Hz with measurements in µV.

During the session (>70 minutes), users can follow light free-living protocols.

Chest strap

Hand grip

https://www.polar.com/ca-en/sensors/h10-heart-rate-sensor



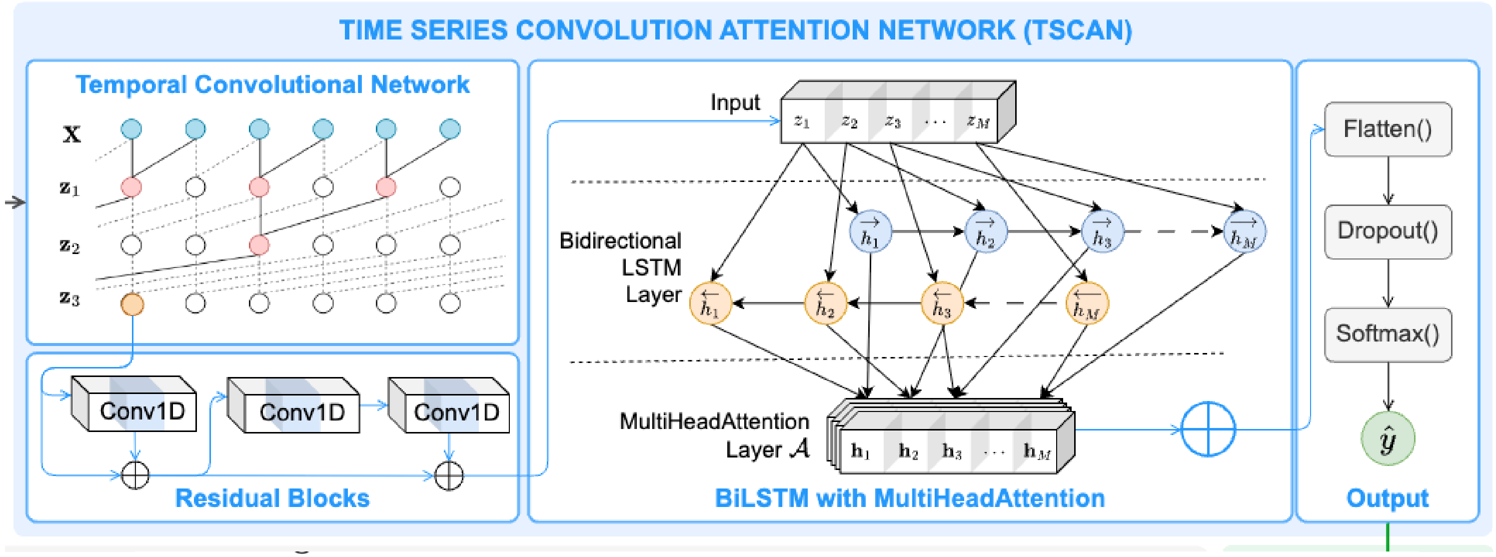
https://www.verniercanada.ca/product/sensors/heart-rate-sensors/hand-grip-heart-rate-monitor/

Time Series Convolutional Attention Network (TSCAN)

TSCAN integrates:

- a Temporal Convolutional Network (TCN)
- residual blocks
- bidirectional LSTMs (BiLSTMs) with multi-head attention

to capture both local and global temporal dependencies in RRI sequences.



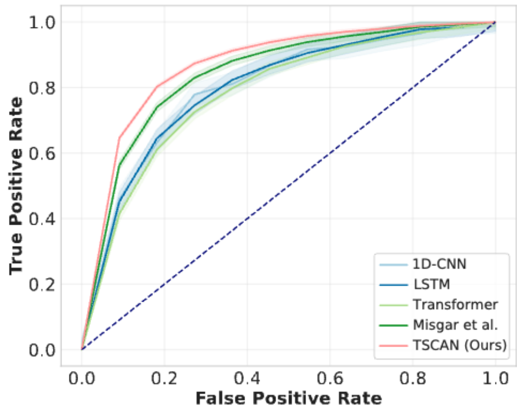
Model	Precision	Recall	F1 Score	AUC	Accuracy
(a) 5-fold cross-validation					
1D-CNN	0.747 ± 0.070	0.826 ± 0.070	0.784 ± 0.070	0.895 ± 0.035	0.750 ± 0.070
LSTM	0.760 ± 0.056	0.583 ± 0.056	0.659 ± 0.056	0.889 ± 0.028	0.667 ± 0.056
Transformer	0.710 ± 0.042	0.833 ± 0.042	0.767 ± 0.042	0.875 ± 0.021	0.750 ± 0.042
Misgar et al. [15]	0.833 ± 0.028	0.667 ± 0.028	0.741 ± 0.028	0.928 ± 0.014	0.766 ± 0.028
TSCAN (Ours)	**0.858 ± 0.014**	**0.896 ± 0.014**	**0.876 ± 0.014**	**0.948 ± 0.007**	**0.866 ± 0.014**
(b) Leave-one-out cross-validation					
1D-CNN	0.781	0.833	0.806	0.826	0.8
LSTM	0.727	0.8	0.762	0.799	0.75
Transformer	0.92	0.767	0.836	0.85	0.85
Misgar et al. [15]	0.884	0.767	0.821	0.906	0.833
Buza et al. [4]	-	-	-	0.910	0.833
TSCAN (Ours)	**0.962**	**0.833**	**0.893**	**0.933**	**0.900**

Time Series Convolutional Attention Network (TSCAN) Performance

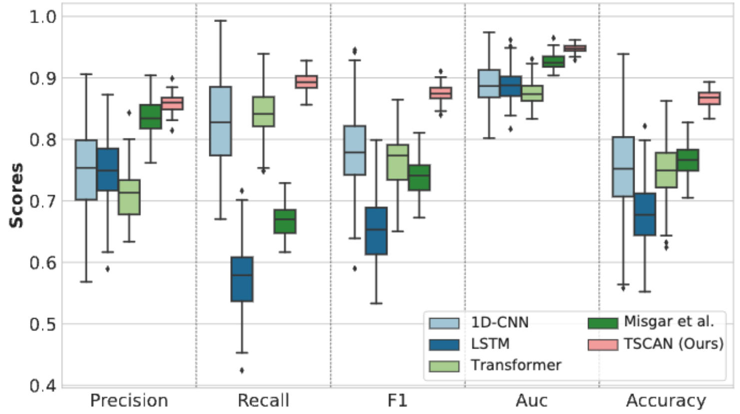
We evaluate on the HRV-ACC dataset:

- Contains physiological data of: 30 SZ/BP, 30 HC.
- 1.5–2h with Polar H10.
- Overall, the results highlight the effectiveness of TSCAN in accurately and robustly classifying instances with strong performance of its components.

Model	Precision	Recall	F1 Score	AUC	Accuracy
(a) 5-fold cross-validation					
1D-CNN	0.747 ± 0.070	0.826 ± 0.070	0.784 ± 0.070	0.895 ± 0.035	0.750 ± 0.070
LSTM	0.760 ± 0.056	0.583 ± 0.056	0.659 ± 0.056	0.889 ± 0.028	0.667 ± 0.056
Transformer	0.710 ± 0.042	0.833 ± 0.042	0.767 ± 0.042	0.875 ± 0.021	0.750 ± 0.042
Misgar et al. [15]	0.833 ± 0.028	0.667 ± 0.028	0.741 ± 0.028	0.928 ± 0.014	0.766 ± 0.028
TSCAN (Ours)	**0.858 ± 0.014**	**0.896 ± 0.014**	**0.876 ± 0.014**	**0.948 ± 0.007**	**0.866 ± 0.014**
(b) Leave-one-out cross-validation					
1D-CNN	0.781	0.833	0.806	0.826	0.8
LSTM	0.727	0.8	0.762	0.799	0.75
Transformer	0.92	0.767	0.836	0.85	0.85
Misgar et al. [15]	0.884	0.767	0.821	0.906	0.833
Buza et al. [4]	-	-	-	0.910	0.833
TSCAN (Ours)	**0.962**	**0.833**	**0.893**	**0.933**	**0.900**



(a) AUC-ROC curves



(b) Metrics

Dual Visual and Textual Explanations

Visual Explanations (Two Methods):

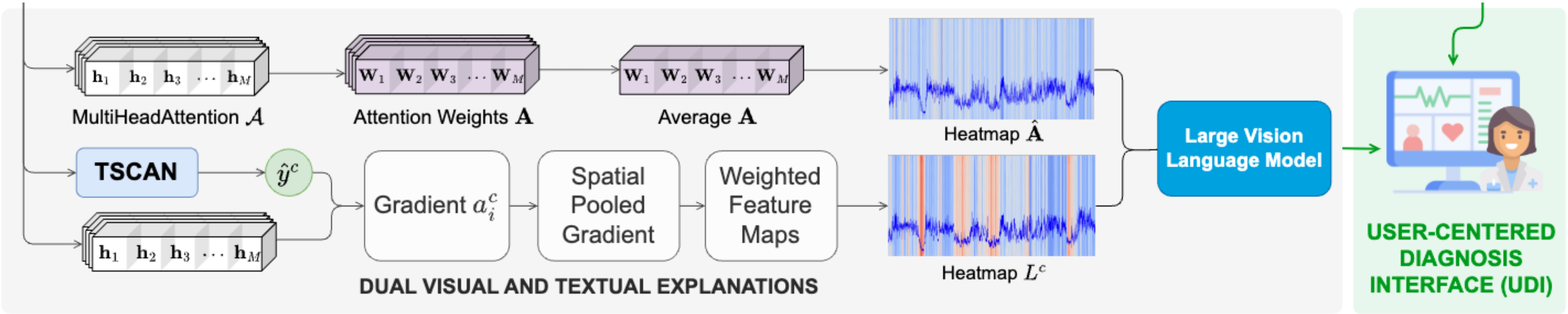
- **Attention-based**: Heatmaps showing which time periods the model focuses on most
- **Gradient-based**: Highlights input regions with highest impact on predictions

Textual Explanations:

- Large Vision Language Model (GPT-4o) processes visual heatmaps + RRI data
- Generates human-readable clinical interpretations

Combined Output:

- Visual heatmaps for pattern identification
- Natural language summaries for clinical understanding



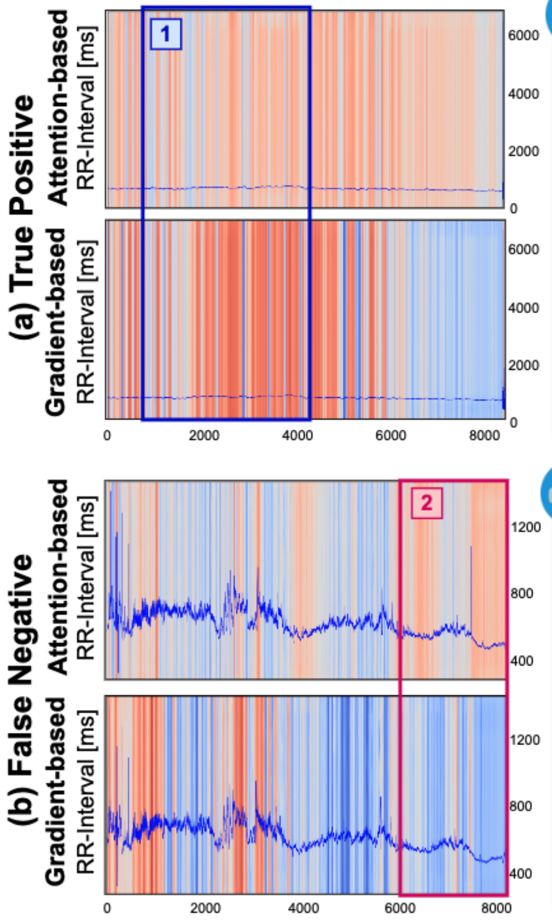
Explanations for Psychiatric Disorders

True Positive Case:

- Visual explanations show **consistent attention regions**
- Both heatmaps highlight same important RRI patterns
- LVLM confirms reliable model decision

False Negative Case:

- **Disagreement** between attention-based and gradient-based maps
- Inconsistencies in latter part of RRI sequence
- LVLM identifies potential model uncertainty areas



The saliency maps effectively highlight critical RRI regions, with the attention map covering broader intervals (1000-3000) and the gradient map focusing on localized fluctuations (2000-2500). Both maps show consistency but differ in granularity. HRV metrics (SDNN: 99.66 ms, RMSSD: 93.79 ms, pNN50: 0.70%) reveal low variability, especially with pNN50. Temporal dynamics are well-captured, though the gradient map may overfit narrow intervals. The underemphasis on spikes (>6000) is a limitation. Dual visual explanations provide reasonable support for the AI model's prediction, suggesting it is likely accurate and trustworthy within the analyzed regions, especially in the highlighted ranges (500-4000).

The saliency maps show potential inconsistency in focus, especially towards the end of the sequence (~6000-8000). The attention-based map broadly highlights regions of moderate RRI fluctuation (~500-3000) but diminishes focus near the end, while the gradient-based map inconsistently shifts attention to narrow peaks in the later segments. HRV metrics (SDNN: 78.68 ms, RMSSD: 29.19 ms, pNN50: 1.59%) reflect lower variability, which may explain reduced attention to abrupt changes. The misalignment in temporal focus between maps at the end suggests possible model uncertainty in capturing critical late-stage transitions. These inconsistencies warrant further review for robust explanations.

Fig. 4: The dual visual and textual explanations of model's detection of PDs.

Explanations for A

System Adaptability Test

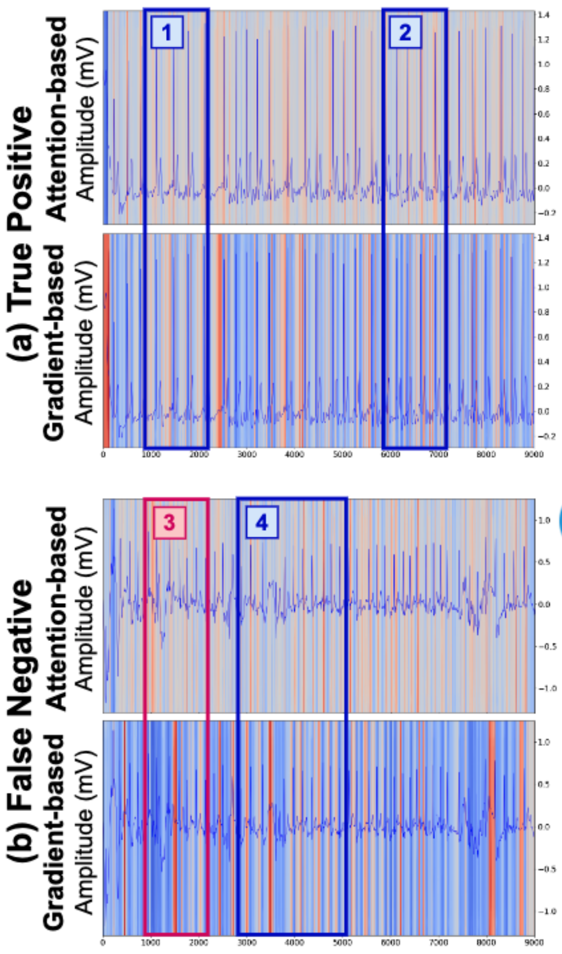
Analytics Everywhere Lab AT THE UNIVERSITY OF NEW BRUNSWICK

Canada NRC·CNRC

Adaptability: Atrial Fibrillation (AF) Detection 2017 PhysioNet Challenge dataset.

Key Insights:

- Explanations reveal **feature importance patterns**
- Highlight **model agreement/disagreement** areas
- Enable **error analysis** for clinical review
- Improve model interpretability and trust



The attention-based and gradient-based saliency maps highlight R-peaks and surrounding regions, focusing on critical intervals like 1000–2000 and 6000–7000, which are clinically relevant for AFib classification. While slight disagreements appear in the early timesteps, likely due to noise, the maps converge later, emphasizing similar features. Metrics derived from the ECG signal—SDNN (225.01 ms), RMSSD (241.46 ms), and pNN50 (78.79%)—confirm significant variability and frequent irregularities, strongly indicative of AFib. These combined findings suggest that the AI model's prediction is accurate and trustworthy, with the maps effectively capturing relevant temporal features despite initial noise.

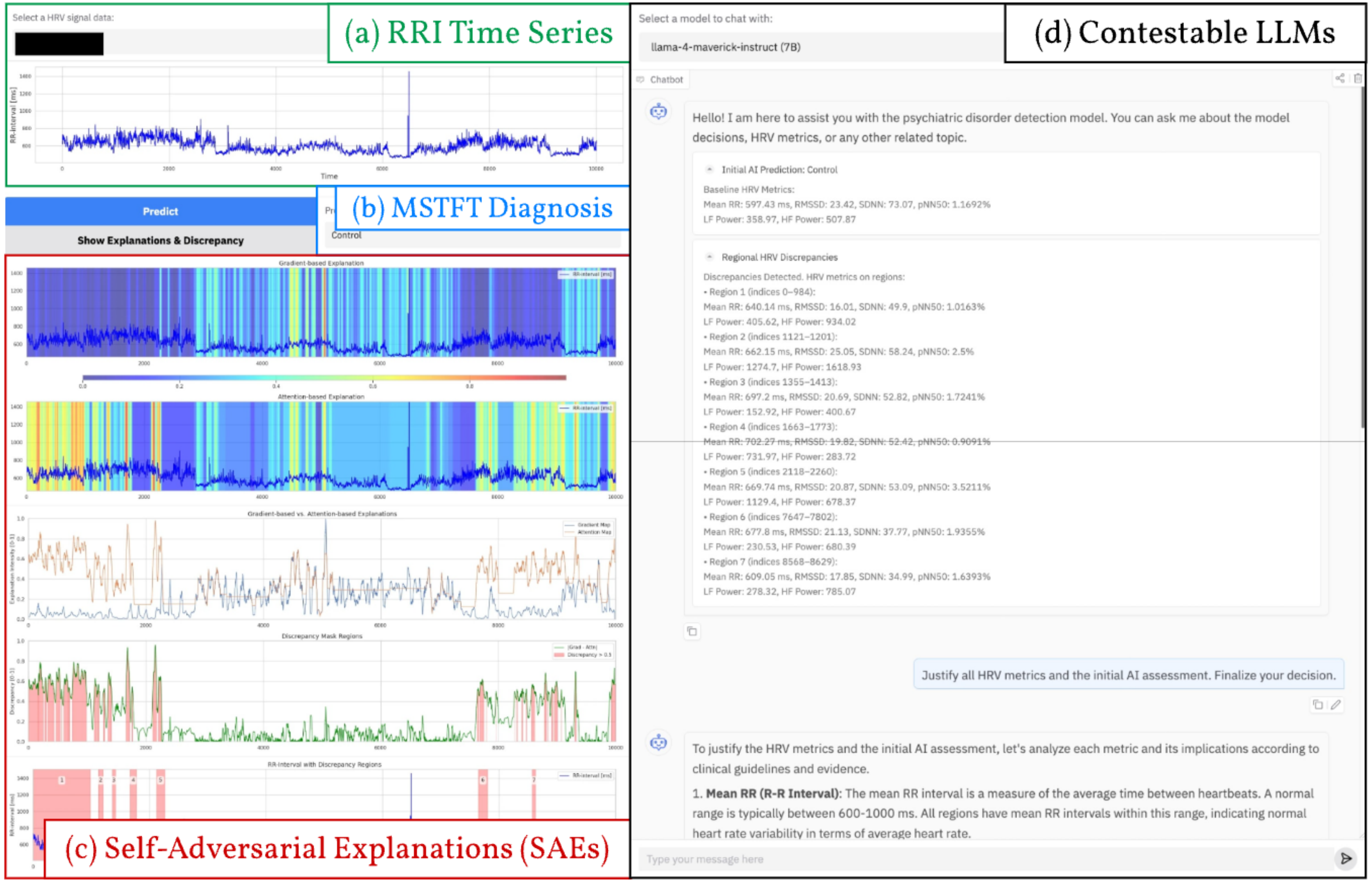
The attention-based and gradient-based maps show agreement in key regions, such as 3000–5000, focusing on R-peaks. However, disagreements in earlier intervals (e.g., 1000–2000) suggest noise or missed features. The maps lack the broader dispersion typical of AFib, focusing too narrowly on periodic features rather than chaotic irregularities. Metrics—SDNN (107.59 ms), RMSSD (143.04 ms), and pNN50 (62.87%)—indicate moderate variability but do not exhibit the extreme fluctuations associated with AFib. These factors suggest the AI model likely made a wrong detection, as its explanations fail to capture the characteristic irregularities of AFib.

Fig. 5: The dual visual and textual explanations of model's detection of AF.

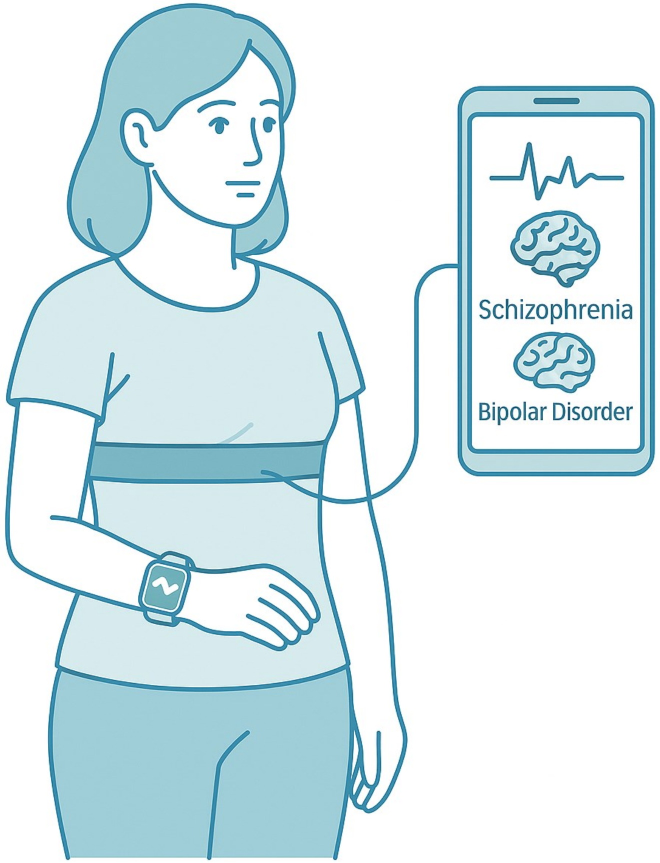
Future Works & Conclusion

Future Works

- A self-adversarial explanation module.
- A contestable LLMs chatbot.
- **Multimodal data integration** (EEG, activity, etc.).
- Comprehensive cognitive clinical evaluations.



Conclusion



- **Heart-brain connection** opens new diagnostic pathways
- **Objective biomarker** replacing subjective diagnostic methods
- **Earlier detection** potential through continuous monitoring
- **Dual explanations** (visual + textual) increase transparency + **User-centered interface** for clinician workflow
- Our implementation is at: github.com/Analytics-Everywhere-Lab/heart2mind

Our mission

We're recruiting MSc. and PhD students!

If you are interested, contact hcao3@unb.ca

Hung Cao, PhD

Assistant Professor, Lab Director
Analytics Everywhere Lab
University Of New Brunswick, Canada
hcao3@unb.ca

Affiliated Faculty

Francis Palma, PhD

Monica Wachowicz, PhD

Trevor Hanson, PhD

Rene Richard, MSc

Students

Hung Nguyen

Asfia Kawnine

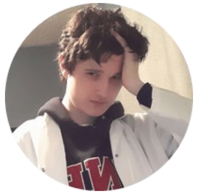
Atah Nuh Mih

Alireza Rahimi

Pavi P

…

Krishno Dey



Connor McLenaghan

Ishan Randeniya

Simran Dadhich

Bohdan Savchuk

…