

# Heart2Mind: Human-Centered Contestable Psychiatric Disorder Prediction System Using Wearable ECG Monitors

**HUNG NGUYEN\***, University of New Brunswick, Canada and National Research Council Canada, Canada  
**ALIREZA RAHIMI**, University of New Brunswick, Canada  
**VERONICA WHITFORD**, University of New Brunswick, Canada  
**HÉLÈNE FOURNIER**, National Research Council Canada, Canada  
**IRINA KONDRATOVA**, National Research Council Canada, Canada  
**RENÉ RICHARD**, National Research Council Canada, Canada and University of New Brunswick, Canada  
**HUNG CAO**, University of New Brunswick, Canada

Psychiatric disorders affect millions, yet diagnosis depends on subjective assessments and uneven access to care. To address these challenges, there is a growing need for Contestable AI (CAI), a framework that extends beyond Explainable AI (XAI) by allowing clinicians to inspect, question, and revise algorithmic outputs, thereby reducing automation bias and strengthening accountability. We present Heart2Mind<sup>1</sup>, a human-centered CAI system for psychiatric disorder prediction that provides objective evidence while preserving clinical oversight. Heart2Mind collects R-R interval (RRI) time series from Polar H9/H10 wearable ECG sensors via a Cardiac Monitoring Interface and analyzes them using a Multi-Scale Temporal-Frequency Transformer (MSTFT) that combines time-domain and frequency-domain features. For contestability, the Contestable Diagnosis Interface integrates model explanations with dialogue. Self-Adversarial Explanations compare attention-based and gradient-based explanation maps to flag inconsistent predictions, and a collaboration chatbot helps users verify and challenge outputs. On the HRV-ACC dataset, MSTFT achieved 91.7% accuracy under leave-one-out cross-validation, outperforming benchmark methods. Human-centered evaluation with the Human-CAI Consensus Rate showed experts and CAI could confirm correct decisions and correct errors through readable, efficient dialogues ( $FKGL \approx 15$ , median 8.3 minutes, 4 turns). These results support low-cost wearable CAI screening with objective biomarkers, safeguards, and an interactive path for clinicians to refine recommendations.

CCS Concepts: • **Computing methodologies** → **Neural networks**; • **Applied computing** → **Health care information systems**; • **Human-centered computing** → **Natural language interfaces**.

Additional Key Words and Phrases: Psychiatric disorder prediction, human-centered contestable AI, contestable large language models, explainable AI, wearable ECG

\*Corresponding author.

<sup>1</sup>Our implementation is available at: <https://github.com/Analytics-Everywhere-Lab/heart2mind>.

Authors' Contact Information: **Hung Nguyen**, hung.ntt@unb.ca, University of New Brunswick, Fredericton, New Brunswick, Canada and National Research Council Canada, Fredericton, New Brunswick, Canada; **Alireza Rahimi**, University of New Brunswick, Fredericton, New Brunswick, Canada, alireza.rahimi@unb.ca; **Veronica Whitford**, University of New Brunswick, Fredericton, New Brunswick, Canada, veronica.whitford@unb.ca; **Hélène Fournier**, National Research Council Canada, Moncton, New Brunswick, Canada, helene.fournier@nrc.ca; **Irina Kondratova**, National Research Council Canada, Fredericton, New Brunswick, Canada, irina.kondratova@nrc-cnrc.gc.ca; **René Richard**, National Research Council Canada, Fredericton, New Brunswick, Canada and University of New Brunswick, Fredericton, New Brunswick, Canada, rene.richard@nrc-cnrc.gc.ca; **Hung Cao**, University of New Brunswick, Fredericton, New Brunswick, Canada, hcao3@unb.ca.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2476-1249/2018/8-ART111

<https://doi.org/XXXXXXX.XXXXXXX>

**ACM Reference Format:**

Hung Nguyen, Alireza Rahimi, Veronica Whitford, H       Fournier, Irina Kondratova, Ren   Richard, and Hung Cao. 2018. Heart2Mind: Human-Centered Contestable Psychiatric Disorder Prediction System Using Wearable ECG Monitors. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (August 2018), 45 pages. <https://doi.org/XXXXXXX.XXXXXXX>

**1 Introduction**

Psychiatric disorders are mental health conditions characterized by clinically significant disturbances in cognition, emotional regulation, and behavior [91]. These disorders include schizophrenia, which involves a combination of positive symptoms (e.g., delusions, hallucinations) and negative symptoms (e.g., alogia, blunted affect), and bipolar disorder, which involves extreme mood swings between manic and depressive episodes, sometimes accompanied by psychotic features. Affecting an estimated 1-3% of the global population, both schizophrenia and bipolar disorder substantially impact functional independence and quality of life [69].

Psychiatric disorders pose a significant global health burden, with profound personal and societal consequences. Approximately 90% of individuals who die by suicide have a diagnosable psychiatric disorder [12], and conditions such as schizophrenia are associated with markedly reduced life expectancy, with mortality rates increasing by 2.6% over the past decade relative to healthy populations [81]. Despite the standardized diagnostic frameworks provided by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and the International Classification of Diseases (ICD-11), clinical diagnosis continues to rely heavily on subjective self-reports and clinician-administered interviews. Even with structured tools, such as the Positive and Negative Syndrome Scale (PANSS) [5], the assessment process remains time-consuming, variable in quality, and error-prone. Blood biomarker approaches, although promising, require costly laboratory testing that is inaccessible in many regions [27]. Therefore, there is an urgent need for automated, accurate, and effective psychiatric disorder prediction. Wearable cardiography technology offers a promising alternative through heart rate variability (HRV) and R-R interval (RRI) measurements. These cardiac biomarkers reflect autonomic nervous system function, which shows dysregulation in psychiatric disorders [9, 28, 62] and correlates with symptom severity [9, 10, 93]. However, previous AI integration attempts faced limitations from focusing on less precise sensors such as photoplethysmography (PPG) [28, 42], while AI's "black-box" nature hinders clinical acceptance.

In our previous work [62], we built an explainable psychiatric diagnosis system that renders model rationales as textual explanations. That contribution prioritized comprehension, helping clinicians follow the AI's reasoning, but it did not yet provide a structured path to use those insights to refine or overturn the AI's decision. Meanwhile, regulators increasingly require transparency and contestability so users can question and correct automated outputs, as reflected in General Data Protection Regulation (GDPR) [73], EU AI Act [58], Canada's Directive on Automated Decision-Making [64], Health Canada [14], UK's Medicines and Healthcare products Regulatory Agency (MHRA) [71], U.S Food and Drug Administration (FDA) policies [22], and the Montr       Declaration [66], which embed rights to explanation, human oversight, and recourse. Ethically, equipping clinicians with tools to contest AI safeguards patients and affirms that automation should support, not supplant, professional judgment. In response to these challenges, our research presents the following key contributions:

- **We propose Heart2Mind, a human-centered contestable psychiatric diagnosis system** that integrates wearable electrocardiogram (ECG) monitoring and a novel Multi-Scale Temporal-Frequency Transformer (MSTFT) to classify disorders from RRI time series by fusing multi-scale temporal and wavelet-based frequency features, enabling continuous objective assessment with short recordings.
- **We advance a Contestable AI (CAI) framework, going beyond Explainable AI (XAI)**, by integrating Self-Adversarial Explanations (SAE), which identify regions that reflect unfaithfulness in AI decisions by contrasting attention- and gradient-based explanation maps, with a Contestable Large Language Model (LLM) that supports collaboration and revision through natural language interaction, preserving clinician oversight in high-stakes care and transforming AI from passive explanation to active contestation.

- **We propose the Human–CAI Consensus Rate (HCCR)** as a human evaluation metric that quantifies whether domain experts, collaborating with CAI, achieve the correct final decision, capturing both validation of true predictions and overturning of errors.

This paper is organized as follows: Section 2 reviews related work on ECG-based psychiatric disorder prediction, including approaches based on short wearable recordings, and outlines the progression from XAI toward CAI systems in healthcare. Sections 3 and 4 provide an overview and implementation of the Heart2Mind system, including the MSTFT model and the integration of a contestable LLM. Section 5 presents experimental results on diagnostic performance and human-centered contestability evaluation. Section 6 discusses the potential of wearable ECG in psychiatric disorder prediction and future visions of CAI systems in healthcare. Section 7 concludes with key contributions and future research directions.

## 2 Related Work

This section reviews recent AI methods for predicting psychiatric disorders using wearable ECG devices, outlines the shift from XAI toward CAI, and highlights prior efforts to develop human-centered CAI systems in healthcare.

### 2.1 Psychiatric Disorder Prediction from Wearable ECG

**2.1.1 Heart–Brain Interaction and ECG Biomarkers.** Traditionally, biomarker-based psychiatric disorder prediction has centered on neurophysiological signals, most notably electroencephalography (EEG) [72], or biological data, including magnetic resonance imaging (MRI) [33], functional MRI (fMRI) [88], and genetic analysis [67]. In contrast, the diagnostic value of peripheral cardiovascular markers has only recently garnered attention in the context of predicting psychiatric disorders. One of the most crucial organs in the human body is the heart, which is tightly connected to the brain through various physiological and neurological processes. For instance, psychiatric disorders often involve physiological dysregulation that can be captured via cardiovascular biomarkers [37, 90]. Several ways of how the heart and brain communicate with each other have been explored [35]:

- **Autonomic nervous system (ANS):** Heart functions, such as heart rate (HR), blood pressure (BP), and blood flow (BF), are controlled by the ANS. BP and HR go up during the sympathetic (“fight-or-flight”) response, which is controlled by the sympathetic nervous system, whereas the same parameters go down during the parasympathetic (“rest-and-digest”) response, which is controlled by the parasympathetic nervous system. The brain controls the ANS and can influence the heart’s function through this pathway [77]. For example, in healthy individuals, higher HRV denotes greater flexibility and adaptability of the ANS. Conversely, psychological stress and emotional distress can directly suppress HRV by engaging sympathetic pathways and withdrawing parasympathetic influence [90].
- **Stress and emotions:** Stress and emotions also affect the rhythm of heartbeats. As an example, when an individual feels anxiety or fear, the brain orders the heart to speed up and increases BP [76]. Reductions in high-frequency (HF) power indicate decreased parasympathetic tone, and alterations in low-frequency (LF) power or the LF/HF ratio suggest sympathetic overactivity [36]. Similarly, sustained stress can alter the heart’s physiological properties over time, increasing the risk of heart disease [57].
- **Heart–brain feedback loop:** In a feedback loop known as the “heart–brain axis”, the heart also interacts with the brain. In this loop, neurotransmitters and hormones (e.g., serotonin, oxytocin) influence both mood and brain activity [37]. This interaction can also affect the ANS, leading to changes in HR and BP.

The relationship between the brain and the heart is intricate and crucial to human physiology. Knowing how these two organs interact with one another can help to detect psychiatric disorders, anxiety, and other neurological conditions accurately and enhance mental health. Empirical researches strongly support the role of ECG-derived measures in mental health assessment [2, 8, 15, 19, 26, 31, 35, 37, 41, 62, 81, 84, 90].

Table 1. Summary of AI techniques for ECG-based psychiatric disorder prediction. Rows in shade indicate the studies most relevant to our work. \*HC: Healthy Control, BD: Bipolar Disorder, SZ: Schizophrenia, DP: Depression.

Author	Targets	Subjects	Features	Methods	Devices
Valenza et al. [87]	BD	BD: 14	HRV features	SVM	Smartex PSYCHE wearable ECG
Ainunhusna et al. [2]	BD	HC: 14, BD: 18	MHR, HRV features	SVM	-
Tiryaki et al. [85]	DP	Subjects: 79 [80]	1-lead ECG	CNN	VivaLNK Continuous ECG Recorder
Zang et al. [92]	DP	HC: 37, DP: 37	2-lead ECG	CNN	-
Tasci et al. [81]	BD, DP, SZ	HC: 35, BP: 62, DP: 17, SZ: 119 [82]	12-lead ECG	ANN	Philips ECG TC20
Khare et al. [35]				Hybrid ensemble classifiers	
Telangore et al. [84]				Wavelet scattering network & Fine KNN classifier	
Corponi et al. [19]	BD	Subjects: 67	PPG features	Bayesian hierarchical model	Empatica E4 smartwatch
Buza et al. [13]	SZ/BD	HC: 30, SZ/BD: 30 [41]	RRI time series	Convolutional nearest neighbor	Polar H10
Książek et al. [40]				Ensemble of SVMs and GRU-based NN	
Nguyen et al. [62]				Time-series Convolutional Attention NN	

**2.1.2 AI Techniques for ECG-based Psychiatric Disorder Prediction.** In recent years, various computational approaches have emerged for predicting psychiatric conditions from ECG signals. Tasci et al. [81] published a 12-lead ECG dataset (Psychiatry ECG dataset [82]) and introduced a ternary pattern-based classification model using multi-level discrete wavelet transform and iterative majority voting, achieving 96.3% accuracy in distinguishing between bipolar disorder, depression, schizophrenia, and healthy controls. Using the same dataset, Khare et al. [35] developed ECGPsychNet, which employs decomposition techniques and optimizable classifier ensembles to achieve 98.2% accuracy. Building on these studies, Telangore et al. [84] combined wavelet scattering networks with Fine K-Nearest Neighbor classification, achieving 99.8% accuracy via ten-fold cross-validation.

Wearable devices have extended psychiatric monitoring beyond clinical settings [8]. Valenza et al. [87] analyzed HRV features, and nonlinear metrics collected via the PSYCHE wearable t-shirt with integrated fabric electrodes and applied an SVM classifier to forecast mood transitions between euthymic (EUT) and non-euthymic (non-EUT) states with an average accuracy of 69.0%, reaching up to 83.3% in individual cases. Tiryaki et al. [85] developed a convolutional neural network (CNN)-based method for ECG depression detection using the VivaLNK recorder, achieving over 95.0% accuracy. Cella et al. [15] employed Empatica E4 smartwatches [26] to identify HRV differences in schizophrenia patients, finding lower vagal tone correlated with symptom severity. Similarly, Corponi et al. [19] used these smartwatches to study HRV during bipolar episodes, revealing that HRV measures increased in parallel with symptom improvement. Other notable studies include Ainunhusna et al. [2]'s SVM for bipolar disorder (93.8% accuracy) and Zang et al. [92]'s 1D-CNN for depression prediction (93.9% accuracy).

While multi-day monitoring provides comprehensive data, researchers have begun investigating shorter recordings as an alternative. Inoue et al. [31] achieved high accuracy using short-time ECG measurements during structured yoga exercises. Recently, HRV-ACC dataset [41] stands out as the only publicly available resource that



provides raw beat-accurate RRI time series from a Polar H10 chest strap together with synchronized accelerometry for psychiatric assessment. RRI is appropriate for short continuous monitoring because R-peaks give precise beat timing in 5-minute windows [23] and stay more stable under light movement than PPG [28, 42]. Modeling the full RRI time series preserves brief and non-stationary autonomic patterns that summary HRV metrics often average out, enabling the model to capture dynamic heart–brain interactions in short wearable sessions. Using this dataset, Buza et al. [13] proposed a convolutional nearest neighbor approach, and Książek et al. [40] trained an ensemble of SVMs and a GRU-based neural network, achieving 80.0–83.0% classification accuracy.

These advances demonstrate that wearable ECG and HRV measurements can capture clinically meaningful alterations in psychiatric disorders, creating promising opportunities for non-invasive and continuous assessment with short recordings. However, prior studies on RRI time series have focused almost entirely on time-domain information, either operating directly on raw RRI time series or incorporating automatically derived HRV features that combine simple time- and frequency-domain statistics [13, 41, 62]. To date, no studies have introduced a learned representation of RRI that integrates temporal dynamics with frequency-domain structure in a unified model. Also, the challenge remains to develop algorithms that maintain high accuracy with minimal data while ensuring that “black-box” models are validated across diverse populations and real-world settings.

## 2.2 From XAI to Human-centered CAI Systems in Healthcare

**2.2.1 XAI Applications in Healthcare.** Despite several remarkable advances in ECG-based psychiatric disorder prediction techniques, significant challenges remain before these technologies can be successfully translated into clinical practice. The “black-box” nature of advanced AI techniques poses a barrier to clinical acceptance and implementation. Hence, in the healthcare context, XAI has grown rapidly and has proven particularly valuable, enhancing diagnostic accuracy, building trust among practitioners and patients, and ensuring ethical use of AI technologies [6, 33, 56, 59]. XAI applications provide clinicians with insights into AI-driven diagnostic decisions. Several XAI techniques have been proven valuable in clinical research: post-hoc explanations [48, 74, 75] that help identify *where* and *why* a “black-box” focuses, or counterfactual explanations [34] that enable exploration of *what-if* scenarios. Recent research shows healthcare explainability advancing through three complementary directions: (1) aligning explanations with human cognition, (2) enabling interactive clinician engagement, and (3) adapting explanations across modalities. These trends move XAI from static visualizations toward human-centered systems integrated into clinical workflows:

- **Human-cognition and psychology alignment:** Modern XAI applies cognitive psychology to align explanations with clinicians’ reasoning. Zhang and Lim [94] proposed contrastive *why A not B* narratives, while studies highlight empathy, familiarity, and causal framing as key for human-aligned explanations, favoring storytelling and domain-specific language over gradient maps [21].
- **Interactive and human-in-the-loop explanations:** Interactive explanations reflect the social nature of human explanations [55]. Static explanations are evolving into interactive systems where users can modify instance attributes [18] or create theory-driven explanations [89], or examine updated predictions by *what-if* counterfactual questions to improve decision accuracy [53].
- **Modality-adaptive explanations:** Multi-modal XAI integrates visual, textual, and interactive elements through concept-based and LLM-driven reasoning [60–62, 68], improving comprehension and supporting richer human–AI collaboration.

For psychiatric disorder prediction, various XAI approaches have been applied across modalities. Arias and Astudillo [6] applied SHAP [48] to XGBoost for schizophrenia prediction using EEG-derived features. Jimenez-Mesa et al. [33] used LIME [74] and SHAP on MRI scans for schizophrenia classification. Misgar et al. [56] applied GradCAM [75] to visualize time series model’s classification of psychiatric patients’ motor activity.

**2.2.2 Beyond Explaining: Towards Human-Centered CAI Systems in Healthcare.** While explainability provides transparency, recent regulations and clinical realities demand that healthcare AI systems go further by empowering clinicians not only to understand AI decisions but also to meaningfully challenge and correct them, leading us to explore human-centered CAI systems, as illustrated in Fig. 1.

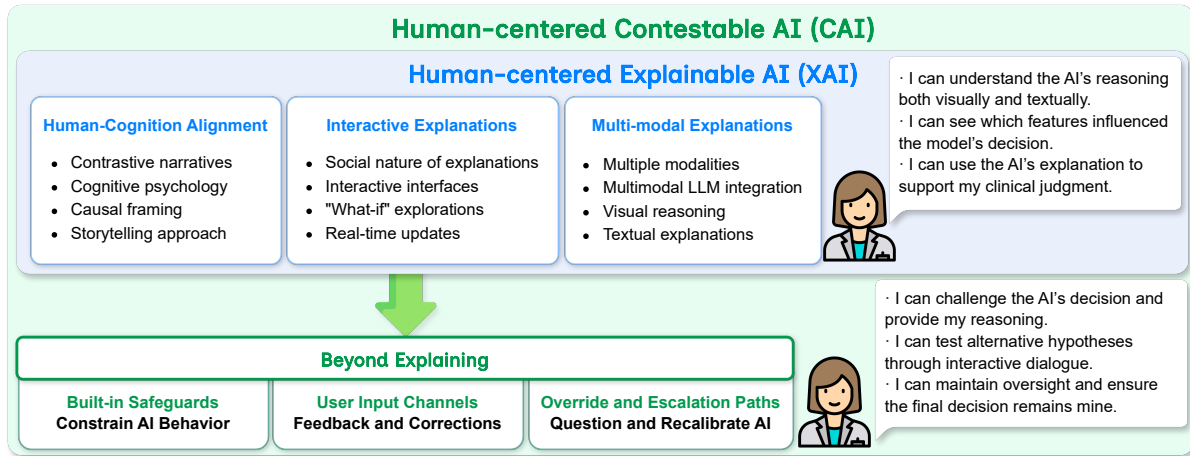


Fig. 1. Evolution from Human-centered Explainable AI (XAI) toward Contestable AI (CAI) System.

**Contestability** goes one step further than explainability: beyond merely understanding the AI, human users must be able to question, intervene in, and correct the AI's decisions. Recent regulations, such as the GDPR [73] and EU AI Act [58], emphasize the importance of model interpretability in healthcare applications. A requirement of transparency or explainability can be found in legislation. Articles 13 and 14 of the GDPR specify that if data subjects are profiled, they have a right to "meaningful information about the logic involved." This applies to the medical context as explained in the official EU guideline. GDPR's requirement for explainability should be understood as a requirement for contestability, where AI decision-making must be explainable to a degree that allows an individual to contest the system's decision. Specifically, Article 22 states that in cases where a data subject may legitimately be subjected to automated decision-making, including profiling, the data controller should safeguard the data subject's right "to express his or her point of view and to *contest* the decision" [73]. Similarly, the Canadian Directive on Automated Decision-Making [64] directly addresses contestability through specific requirements related to the recourse mechanism, where Section 6.4 explicitly states, "providing clients with any applicable recourse options that are available to them to *challenge* the administrative decision." Other regulations, such as Health Canada [14], UK's MHRA [71], U.S FDA [22], or Montréal Declaration on Responsible AI [66] highlight facets of contestability, from rights to human oversight and explanation to mechanisms for appeal, in the context of AI and automated decision-making. As a response to recent regulations that emphasize the ability to contest the AI decision and make explanations more human-oriented [55], CAI goes beyond XAI by enabling users to actively challenge and dispute those decisions based on the provided explanations, essentially giving them the ability to contest the AI's output in a meaningful way. It is not just about understanding the reasoning but also having the agency to question and potentially overturn a decision if necessary [3, 49]. Based on the regulatory frameworks presented that consistently emphasize human agency in questioning AI decisions, from GDPR's "right to contest," to Canada's requirement for "challenge" mechanisms, to various health authorities' emphasis on oversight and appeals, we define the CAI system as follows:

### Contestable AI System

An interactive computational system that allows human challenge throughout its decision-making lifecycle, maintains conscious transparency of its operational processes, and incorporates a safeguard mechanism to constrain algorithmic behavior.

In healthcare AI systems, CAI enables clinicians and patients to challenge algorithmic outputs and meaningfully influence outcomes. This aligns with clinical practice where diagnoses and treatments evolve collaboratively, with second opinion requests and initial assessments refined. AI recommendations should therefore never be considered final. While concrete examples are still emerging, promising prototypes exist. Hirsch et al. [29] developed an automated psychotherapy feedback system where therapy session evaluations are transparent and disputable. Their design provided detailed metric explanations, allowed users to trace AI reasoning to raw data, and, crucially, offered mechanisms to record disagreements or add context missed by AI. This interactive contestability keeps human experts in control, with AI assisting analysis while therapists can correct records, preventing blind reliance on imperfect algorithms. Ploug and Holm [70] advocated for patient-centric CAI diagnostics, identifying prerequisites for effective contestation: including informing users about data usage, biases, accuracy, and clinical validity, as well as early defining AI's role. Additionally, mental health AI tools should communicate their limitations (e.g., reduced reliability for older populations if trained primarily on younger subjects) so clinicians know when to question them, within workflows that allow these questions to alter decisions (e.g., triggering human review when algorithmic outputs appear questionable).

In summary, building XAI and CAI systems in the healthcare context, particularly in mental health, is now recognized as vital for the safe, ethical, and effective deployment of these systems. Explainability provides transparency and insight, enabling clinicians to integrate AI recommendations into their decision-making process with appropriate confidence. Contestability ensures that clinicians and patients remain empowered, so that the AI is a tool at their service rather than an opaque authority. Together, a human-centered CAI system should emphasize features such as:

- (1) **Human-centered explanations** to ground any contestation, because users can only know when to contest a decision if they understand the rationale behind models, even without or with very little knowledge about AI and XAI. Explanations should adapt to match user expertise and context, providing simplified content for less experienced users while offering technical depth for human experts. The system should assess user needs and adjust explanation complexity, style, and focus accordingly. Effective explanations optimize utility metrics, including usefulness, readability, intractability, plausibility, faithfulness, and fairness, incorporating user feedback to refine communication strategies [63, 70].
- (2) **Built-in safeguards** to constrain the behavior of AI systems, where procedural safeguards restrict unilateral AI systems' decisions. One type of safeguard involves creating a self-adversarial decision-making process, such as using different XAI [63] or reliability techniques, including computational argumentation [24], uncertainty scores, and domain-specific rule checks. Another can be achieved by introducing a second automated system external to the controlling organization, through which machine decisions are made. When inconsistencies arise between primary system and verification mechanisms, the case can either be escalated for human expert evaluation or handled through built-in conflict resolution protocols.
- (3) **User input channels** for feedback and corrections allow users to provide alternative suggestions, with the system acknowledging input by updating model or logging disagreements. These channels must integrate seamlessly into workflows, creating audit trails that support accountability and enable continuous learning from human expertise, transforming AI from a static tool into a collaborative partner.
- (4) **Override and escalation paths** enable users to ask follow-up questions, validate or correct input data, and observe how outputs change. Reject/override mechanisms either recalibrate recommendations or re-route

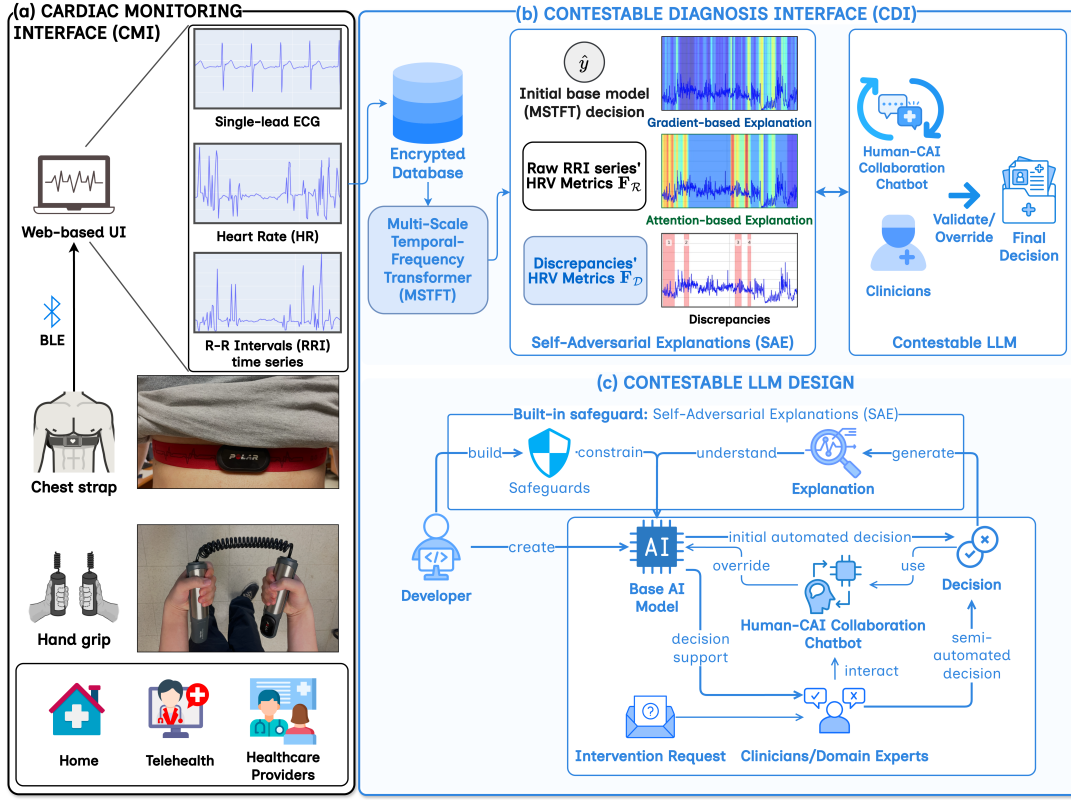


Fig. 2. Overview of the Heart2Mind system: (a) Cardiac Monitoring Interface (CMI), (b) Contestable Diagnosis Interface (CDI) encompasses (c) Contestable LLM with Built-in Safeguard and Human-CAI Collaboration Chatbot as core modules.

decisions to human supervisors for review. Multi-model agreement, where multiple models or algorithms cross-validate interpretations, strengthens reliability. Governance mechanisms should monitor for drifts or biases, with safety constraints requiring human confirmation for high-stakes decisions. These paths operationalize regulatory guidelines by ensuring transparent audit trails and meaningful human oversight throughout the decision-making process.

### 3 Heart2Mind System Overview

This section introduces the Heart2Mind system, a human-centered CAI system designed for psychiatric disorder prediction to ensure transparency and preserve human oversight in clinical decision support. As illustrated in Fig. 2, the system consists of two primary interfaces: (1) Cardiac Monitoring Interface (CMI) and (2) Contestable Diagnosis Interface (CDI), where the CDI incorporates both built-in safeguards and a contestable LLM system to operationalize human-AI collaboration, detailed as follows:

- (1) **Cardiac Monitoring Interface (CMI)** (Fig. 2a): CMI is a web-based platform that captures real-time cardiac signals (ECG, HR, and RRI) from wearable Polar H9 and H10 sensors via Bluetooth Low Energy (BLE). The interface preprocesses, encrypts, and stores the signals in a secure clinical database. Sessions

can occur in supervised clinical settings, remote telehealth environments, or independent home use. This design enables continuous, privacy-compliant monitoring using accessible wearable devices.

- (2) **Contestable Diagnosis Interface (CDI)** (Fig. 2b-c): CDI functions as the main decision-support environment that presents model predictions, generates visual explanations, and enables clinicians to contest or refine AI outputs. It comprises three integrated modules:
  - (a) **Base Prediction Model: Multi-Scale Temporal-Frequency Transformer (MSTFT)** forms the diagnostic foundation, processing RRI time series to classify each case as “control” or “treatment.” It captures both temporal rhythm patterns and spectral information using multi-scale temporal convolutions and wavelet-based frequency transforms, which are fused through cross-attention and self-attention mechanisms, enabling the robust classification of psychiatric conditions from short wearable sessions.
  - (b) **Built-in Safeguards: Self-Adversarial Explanations (SAE)** enhances reliability by comparing attention-based and gradient-based visual explanations of the MSTFT’s reasoning. It highlights regions where the two explanation methods disagree, signaling potential unfaithfulness in model focus. When these discrepancies exceed a threshold, the case is automatically flagged as uncertain and prioritized for human review. SAE thus serves as a built-in safeguard that constrains unilateral model decision-making and supports transparent diagnostic reasoning.
  - (c) **Contestable LLM** integrates an LLM to transform explainability into interactive contestability. It interprets MSTFT predictions, SAE discrepancies to generate contextual natural language explanations that clinicians can interrogate, verify, or dispute. Contestable LLM system is structured around two complementary mechanisms:
    - (i) **Built-in Safeguard Integration:** Contestable LLM assesses SAE discrepancy regions together with the base model prediction to decide whether to retain the original output or propose a new decision. Also, if an intervention request is made, a collaboration process with the human is started to incorporate human knowledge and to finalize the decision.
    - (ii) **Human-CAI Collaboration Chatbot:** The chatbot provides an interface where the human can review the prediction, inspect explanation maps and discrepancy regions, and assess the model reasoning while requesting deeper analysis or additional information through dialogue. Each session becomes a collaboration between algorithmic evidence and clinical expertise, ensuring that human oversight remains decisive in the final outcome.

**Data Pipeline.** The system leverages RRI as the primary input for psychiatric disorder prediction due to its beat-level precision and stability during short continuous monitoring sessions. The MSTFT model processes raw RRI time series to capture dynamic autonomic patterns that standard summary statistics average out. To support the contestable LLM and facilitate clinician understanding, we compute standard HRV metrics (time-domain and frequency-domain features) from detected discrepancy regions identified by SAE. These metrics provide clinically familiar, interpretable summaries that enhance transparency and enable effective human-AI collaboration.

## 4 Implementation

This section describes Heart2Mind’s implementation: (1) CMI serves as the foundational interface for seamless continuous monitoring of high-quality cardiac signals from wearable devices, and (2) CDI enables clinicians to view initial base MSTFT model’s psychiatric disorder predictions, examine explanations and discrepancies through the built-in safeguard, and collaborate with the contestable LLM to finalize decisions.

### 4.1 Cardiac Monitoring Interface (CMI)

**4.1.1 Wearable ECG Monitor.** Our system uses Polar H9/H10 sensors, selected for their high precision in single-lead ECG signal acquisition and HRV measurements [45]. Both devices capture HR in beats per minute (BPM)



and RRI in milliseconds (ms) with a 1-second sampling rate, while the H10 additionally records single-lead ECG at 130 Hz with measurements in microvolts ( $\mu V$ ), providing richer physiological data for the analysis.

**4.1.2 Interface.** Building upon the BleakHeart library [79], we developed the monitoring interface that integrates with wearable devices to capture real-time cardiac signals and store them securely in an encrypted database. Appendix A details the recording workflow and modular architecture. The recording process begins with users self-entering their personal information (name, age, and sex), which is encrypted using a unique ID key for privacy compliance, then selecting their device type (H9/H10) to initiate recording. CMI automatically scans and connects to Polar devices via Bluetooth Low Energy (BLE), configuring appropriate data streams based on device capabilities (ECG/HR/RRI for H10; HR/RRI for H9). During the session, users can follow light, free-living protocols that include short corridor walks interleaved with seated rest periods, during which they can sit down and attend to daily business tasks. Raw sensor data undergoes real-time processing with timestamps synchronized to the host system clock, ensuring temporal accuracy crucial for HRV analysis. The processed cardiac signals are simultaneously streamed to the web-based user interface, as shown in Fig. 3, which displays ECG waveforms, HR, and RRI time series in an intuitive format for clinical observation. Upon session completion, typically after a predetermined duration (at least 70 minutes), the interface prompts users to stop recording. The collected cardiac signals are then automatically saved in a structured format and securely transmitted to an encrypted database accessible to authorized clinicians. This comprehensive data capture and storage approach ensures both the integrity of the physiological measurements and compliance with healthcare data protection standards.

## 4.2 Contestable Diagnosis Interface (CDI)

**4.2.1 Base Prediction Model: Multi-Scale Temporal-Frequency Transformer (MSTFT).** MSTFT processes RRI time series acquired from CMI. As shown in Fig. 5, it integrates multi-scale temporal and frequency-domain wavelet features through cross-attention fusion and self-attention to classify as “control” or “treatment”.

**Input Preprocessing and Positional Encoding.** Our model begins by processing the raw RRI time series input  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T) \in \mathbb{R}^{T \times 1}$  of length  $T$ . Recognizing cardio signals recorded from wearable devices often contain noise artifacts, we enhance model robustness by introducing controlled Gaussian noise  $\mathcal{N}(0, \lambda^2)$  to the input

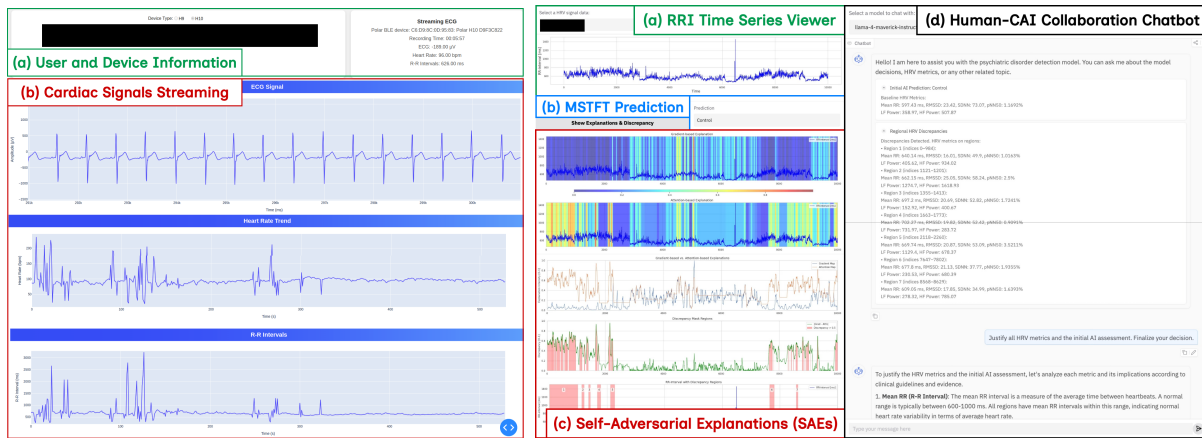


Fig. 3. CMI including: (a) User controls and device status indicators, (b) Cardiac signals streaming: ECG, HR, and RRI time series. Fig. 4. CDI including: (a) RRI time series viewer, (b) initial psychiatric disorder prediction by base MSTFT model, (c) SAE built-in safeguard, (d) Human-CAI collaboration chatbot.

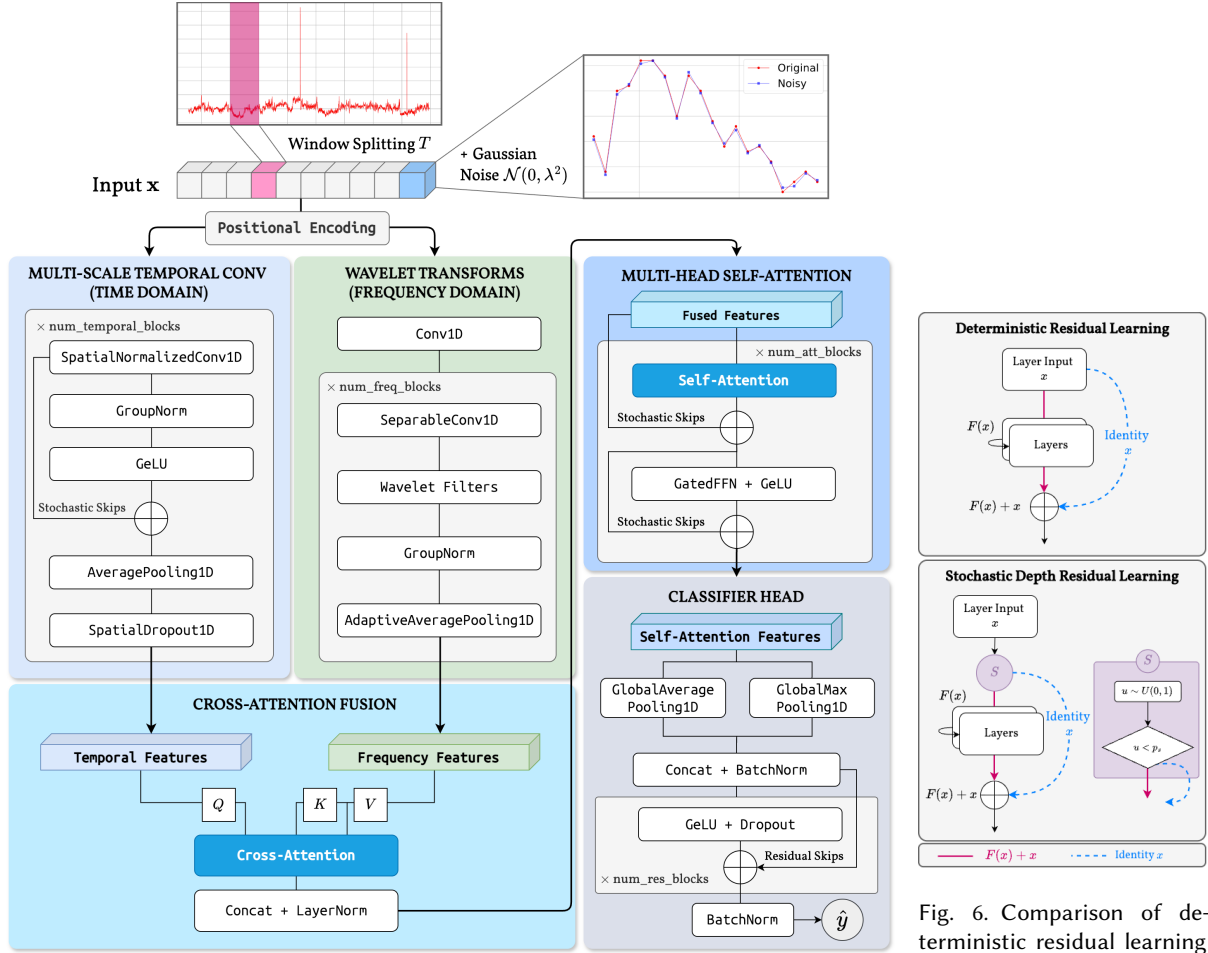


Fig. 5. Architecture overview of the Multi-Scale Temporal-Frequency Transformer (MSTFT).

sequence  $x$ , providing regularization against noise sensitivity during the training phase. Then, the positional encoding  $P \in \mathbb{R}^{T \times d}$  projects the input to higher-dimension space as follows:

$$\tilde{x} = C(x + \epsilon) + P, \text{ where } \epsilon \sim \mathcal{N}(0, \lambda^2); \quad P_{(t, 2k)} = \sin\left(\frac{t}{\tau^{2k/d}}\right), \quad P_{(t, 2k+1)} = \cos\left(\frac{t}{\tau^{2k/d}}\right), \quad (1)$$

where  $\tau$  controls wavelength progression,  $C$  denotes the embedding 1D convolutional layer that transforms the input to dimension  $d$ .

*Multi-Scale Temporal Convolutions Block (Time Domain).* The multi-scale temporal convolutions block captures cardiac rhythm patterns across multiple scales through dilated convolutions. Starting with the encoded input  $Z_t^{(0)} = \tilde{x}$ , each layer applies convolutions with exponentially increasing dilation rates:

$$F^{(i)} = \text{GrpNorm}\left[\sigma_g\left(C_D^{d=2^i}(Z_t^{(i)})\right)\right], \quad Z_t^{(i+1)} = S(Z_t^{(i)}, F^{(i)}), \quad \forall i \in \{0, \dots, n_t - 1\}, \quad (2)$$

where  $C_{\mathcal{D}}^{(d=2^i)}$  represents causal dilated convolutions with dilation rate  $2^i$ ,  $\text{GrpNorm}(\cdot)$  provides group layer normalization,  $\sigma_g(\cdot)$  is the GELU activation function,  $n_t$  is the total number of temporal blocks, and  $S$  is the stochastic depth residual learning block. Each subsequent block reduces the number of filters by half to ensure multi-scale representation learning. Finally, the temporal features  $\mathbf{Z}_t$  are achieved.

*Stochastic Depth.* We adopt stochastic depth (see Fig. 6) instead of deterministic residual connection because it improves generalization and test accuracy in deep models, enhances robustness through path diversity, and reduces expected training compute by executing fewer layers per step [30]. Formally, we implement a stochastic depth module that introduces probabilistic regularization:  $S(x, F(x)) = x + \mathbf{1}\{u < p_s\}F(x)$ , where  $p_s$  is the survival probability of the residual path and  $u$  is a uniform random variable on  $[0, 1]$ . During training, randomly bypassing residual connections lowers the expected depth and shortens backpropagation paths, which eases optimization. When the skip survives, the identity mapping preserves gradient flow.

**4.2.2 Wavelet Transforms Block (Frequency Domain).** In addition to the temporal analysis, the frequency domain block extracts spectral features via learnable wavelet transforms modeled as separable convolutions:

$$\mathbf{Z}_f^{(0)} = C_f(\tilde{\mathbf{x}}), \quad \mathbf{Z}_f^{(j+1)} = \text{GrpNorm} \left[ \sigma_g \left( C_S(\mathbf{Z}_f^{(j)}) \right) \right], \quad \forall j \in \{0, \dots, n_f - 1\}, \quad (3)$$

where  $C_f$  initializes frequency-specific 1D convolutional embeddings,  $C_S$  is the 1D separable convolutional layer,  $n_f$  is the total frequency blocks. Adaptive average pooling adjusts frequency-domain features to match temporal feature dimensionality to achieve the frequency features  $\mathbf{Z}_f$ .

*Cross-Attention Fusion Block.* This block serves as a critical component for integrating information from the temporal and frequency domains. This block employs an efficient cross-attention mechanism with learned projections into a shared embedding space, enabling linear-time computation complexity rather than quadratic complexity [17] while ensuring balanced representation of both temporal and frequency domains and preventing single-domain dominance in the fused features [44]. Given the temporal features  $\mathbf{Z}_t$  and frequency features  $\mathbf{Z}_f$ , we first project them to a common embedding space with dimension  $d_p$  using linear transformations  $\mathbf{H}_t = \mathcal{W}_t \mathbf{Z}_t, \mathbf{H}_f = \mathcal{W}_f \mathbf{Z}_f$ , where  $\mathcal{W}_t$  and  $\mathcal{W}_f$  are learnable projection matrices. The cross-attention mechanism then operates by treating the temporal features as queries, and the frequency features as keys and values:  $\mathbf{Q} = \mathbf{H}_t \mathbf{W}^Q, \mathbf{K} = \mathbf{H}_f \mathbf{W}^K, \mathbf{V} = \mathbf{H}_f \mathbf{W}^V$ , where  $\mathbf{W}^Q, \mathbf{W}^K$ , and  $\mathbf{W}^V$  are learnable weight matrices for query, key, and value projections, respectively. Using these projections, each attention head computes:

$$\mathbf{h}_i = \text{softmax} \left( \frac{\mathbf{Q} \mathbf{W}_i^Q (\mathbf{K} \mathbf{W}_i^K)^T}{\sqrt{d_k}} \right) \mathbf{V} \mathbf{W}_i^V, \quad \mathbf{F}_{\text{multi}} = \text{Concat}(\mathbf{h}_1, \dots, \mathbf{h}_h) \mathbf{W}^O, \quad (4)$$

where  $d_k$  is the key dimension. The fused representation  $\mathbf{F}$  is obtained by concatenating the multi-head output with original features, followed by layer normalization:  $\mathbf{F} = \text{LayerNorm}(\text{Concat}[\mathbf{F}_{\text{multi}}, \mathbf{H}_t, \mathbf{H}_f])$ .

*Multi-Head Self-Attention Block.* This block enhances the fused representations by capturing long-range dependencies within the integrated features. Unlike standard transformer blocks, our implementation incorporates specialized components to improve performance for RRI time series classification. Given the fused representation  $\mathbf{F}$  from the Cross-Attention Fusion Block, the multi-head self-attention operation is applied as  $\mathbf{A} = \text{SelfAttn}(\mathbf{F}, \mathbf{F}, \mathbf{F})$ , where the multi-head self-attention is computed similarly to the cross-attention mechanism, but with the same sequence serving as queries, keys, and values. To enhance the expressiveness of the model, we incorporate an attention-gating mechanism  $\mathbf{G} = \sigma(\mathbf{W}_g \mathbf{A} + \mathbf{b}_g)$ , where  $\mathbf{W}_g$  and  $\mathbf{b}_g$  are learnable parameters, and  $\sigma$  is the sigmoid activation function. The gated attention output is then computed as  $\mathbf{A}_g = \mathbf{A} \odot \mathbf{G}$  where  $\odot$  denotes element-wise multiplication. To maintain training stability, we apply channel-wise scaling with a learnable

parameter  $\alpha$  (initialized to a small value) to the gated attention output. This is followed by a position-wise feed-forward network with expansion factor to increase model capacity:

$$\mathbf{F}' = \mathbf{F} + \alpha \mathbf{A}_g, \quad \mathbf{F}'' = \text{LayerNorm}(\mathbf{F}' + \text{Dropout}(\text{FFN}(\mathbf{F}'))), \quad (5)$$

where  $\mathbf{F}''$  represents the final output of the self-attention block. This multi-head self-attention mechanism effectively captures complex temporal-frequency dependencies in the fused representation, allowing the model to learn complex patterns in RRI time series.

**Classifier Head Block.** This block transforms the representation learned by the Multi-Head Self-Attention Block into the final classification decision. This block employs multiple techniques to maximize information extraction and classification performance. First, we apply both global average pooling and global max pooling operations to capture different aspects of the sequence representation. These pooled representations are concatenated to form a comprehensive feature vector. To ensure stable training, we apply batch normalization to the concatenated features as:

$$\mathbf{h}_{\text{avg}} = \frac{1}{T} \sum_{t=1}^T \mathbf{F}''_t, \quad \mathbf{h}_{\text{max}} = \max_{t \in \{1, \dots, T\}} \mathbf{F}''_t, \quad \mathbf{h} = \text{Concat}[\mathbf{h}_{\text{avg}}, \mathbf{h}_{\text{max}}], \quad \mathbf{h}_{\text{norm}} = \text{BatchNorm}(\mathbf{h}). \quad (6)$$

The normalized features pass through dense layers with attention-weighted residual connections, enhancing feature extraction while maintaining gradient flow, where  $\sigma_g$  is GELU, and  $\odot$  denotes element-wise multiplication:

$$\mathbf{h}_1 = \sigma_g(\mathbf{W}_1 \mathbf{h}_{\text{norm}} + \mathbf{b}_1), \quad \mathbf{a} = \sigma(\mathbf{W}_a \mathbf{h}_1 + \mathbf{b}_a); \quad \mathbf{h}_{\text{out}} = \text{GrpNorm}((\mathbf{h}_1 \odot \mathbf{a}) + (\mathbf{W}_r \mathbf{h}_{\text{norm}} + \mathbf{b}_r)). \quad (7)$$

Finally, the output layer produces the probability of the input RRI time series belonging to the positive class:

$$\hat{y} = \sigma(\mathbf{w}_o^T \mathbf{h}_{\text{out}} + b_o). \quad (8)$$

**4.2.3 Built-in Safeguard: Self-Adversarial Explanations (SAE).** Research on explanation methods has demonstrated that different techniques can produce varying attributions, with some consistently highlighting clinically meaningful features across cases while others show inconsistent patterns. Furthermore, studies have found that explanation attributions often differ systematically between correct and incorrect predictions [32, 47, 54]. Hence, as the core built-in safeguard of CAI framework, SAE is designed to detect the discrepancies between different explanation maps, then identify potential inconsistencies in the MSTFT model's decision-making process. As shown in Fig. 7, SAE identifies discrepancies by comparing two complementary explanation methods: *attention-based explanation* reveals where the model focuses during forward inference, while *gradient-based explanation* uses backpropagation to identify features that most strongly influence the final prediction. Contrasting these forward and backward computational paths provides valuable insights into model faithfulness. When the methods agree, the model's forward attention aligns with features that genuinely drive the prediction, indicating decisions based on clinically relevant patterns. Conversely, when they diverge significantly, the model during inference may be attending to regions that do not actually influence the final classification when analyzed through gradients, signaling potentially unreliable or unfaithful reasoning that warrants further clinical review. The implementation details for both explanation methods are presented as pseudocode in Appendix D.

**Attention-based Explanation** leverages the attention weights from transformer layers (i.e., cross-attention and self-attention) to identify regions in the input that contribute most to the model's predictions. Given a sequence  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ , we extract attention weights from target layers defined as  $\mathcal{L}$ . For each layer  $l \in \mathcal{L}$ , we compute the attention map  $\mathbf{A}^{(l)}$  by averaging across all attention heads:  $\mathbf{A}^{(l)} = \frac{1}{H} \sum_{h=1}^H \mathbf{A}_h^{(l)}$ , where  $H$  is the number of attention heads and  $\mathbf{A}_h^{(l)}$  is the attention weight matrix from the  $h$ -th head in layer  $l$ . The combined attention-based explanation map  $\mathbf{E}_{\text{attn}}$  is generated by averaging across all target layers:  $\mathbf{E}_{\text{attn}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathbf{A}^{(l)}$ . To adapt the attention map to the original signal length, we employ an expansion function  $\mathcal{E}$  that projects the sequence-level explanations to the time-domain:  $\mathbf{E}_{\text{attn}}^T = \mathcal{E}(\mathbf{E}_{\text{attn}}, T)$ , where  $T$  is the original signal length, and  $\mathcal{E}$

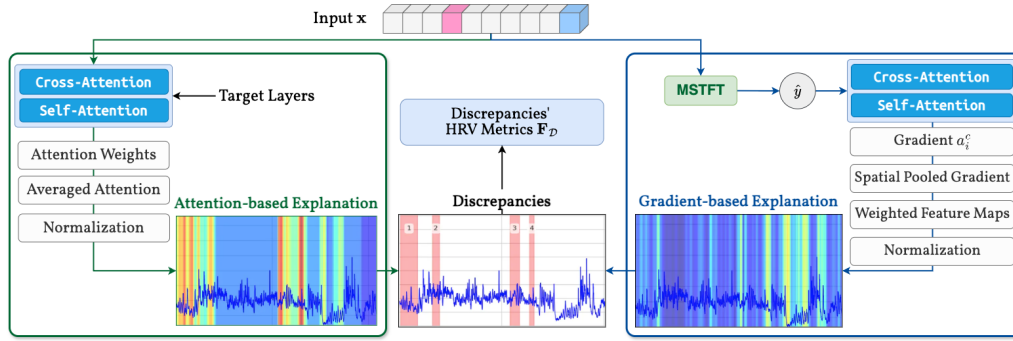


Fig. 7. Self-Adversarial Explanations (SAE) Mechanism as the Built-in Safeguard (See Appendix D for pseudocode).

performs a weighted distribution of attention values across overlapping segments. The final expanded map is normalized using z-score standardization followed by min-max scaling to ensure values lie within  $[0, 1]$ .

**Gradient-based Explanation** employs class activation mapping to identify important regions for model prediction. For each layer  $l \in \mathcal{L}$ , we compute the gradient  $\mathbf{G}^{(l)} = \partial \mathbf{y}_c / \partial \mathbf{F}^{(l)}$ , where  $\mathbf{y}_c$  is the prediction for target class  $c$  and  $\mathbf{F}^{(l)}$  is layer  $l$ 's activation output. We average gradients to obtain importance weights  $\alpha_k^{(l)} = \frac{1}{Z} \sum_{i=1}^Z \partial \mathbf{y}_c / \partial \mathbf{F}_{i,k}^{(l)}$  for each feature map  $k$ , where  $Z$  is the feature maps' spatial dimension. The gradient-weighted activation map is  $\mathbf{L}_{\text{grad}}^{(l)} = \text{ReLU}(\sum_k \alpha_k^{(l)} \cdot \mathbf{F}_k^{(l)})$ . The combined explanation map  $\mathbf{E}_{\text{grad}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathbf{L}_{\text{grad}}^{(l)}$  averages across all target layers. Similar to the attention-based explanation, we expand the gradient-based explanation to match the original signal length and normalize it to ensure values lie within  $[0, 1]$ :  $\mathbf{E}_{\text{grad}}^T = \mathcal{E}(\mathbf{E}_{\text{grad}}, T)$ .

**Discrepancies Detection.** The self-adversarial nature of SAE emerges from comparing the two different explanation methods. To enable a temporally consistent comparison, the attention-based explanation is aligned to the gradient-based explanation using Dynamic Time Warping (DTW) as  $\mathbf{E}_{\text{attn}}^T = \text{DTW}(\mathbf{E}_{\text{attn}}^T, \mathbf{E}_{\text{grad}}^T)$  ensuring that both explanation sequences are matched in temporal structure before discrepancy analysis. The discrepancy map is calculated as the absolute difference between the aligned attention-based and the gradient-based explanation. High discrepancy regions indicate areas where the two explanation methods diverge, potentially revealing unreliable patterns that warrant further human inspection. The discrepancy regions are defined as:

$$\mathcal{D} = \mathcal{M}_\delta \left[ \mathbb{I}(|\mathbf{E}_{\text{attn}}^T - \mathbf{E}_{\text{grad}}^T| > \rho) \right], \quad (9)$$

where  $\mathbb{I}(\cdot)$  is the indicator function that marks time points exceeding the discrepancy threshold  $\rho$ . The operator  $\mathcal{M}_\delta(\cdot)$  denotes a  $\delta$ -merge function that joins adjacent positive indices separated by gaps smaller than  $\delta$ , producing a set of contiguous intervals  $\mathcal{D} = (s_i, e_i)_{i=1}^n$ . These intervals represent temporally coherent regions of explanation disagreement, guiding clinician review and interpretability assessment.

**4.2.4 Contestable LLM.** Contestable LLM transforms MSTFT predictions and SAE discrepancies into interpretable clinical reasoning that enables human-AI collaboration. By processing structured physiological evidence, including baseline HRV metrics and regional metrics from discrepancy regions, LLM grounds its explanations in established cardiac biomarkers of autonomic function. This design allows clinicians to engage in natural language dialogue to validate correct predictions or challenge potentially unfaithful decisions flagged by SAE.

**HRV Metrics Calculation.** To enable clinically meaningful contestability, we extract established HRV metrics from both complete RRI time series and SAE discrepancies. For each discrepancy region  $r_i = (s_i, e_i) \in \mathcal{D}$ , we



extract time-domain and frequency-domain HRV features, namely  $\mathbf{F}_{\mathcal{D}}$ . Time-domain features include mean RR (MRR), Root Mean Square of Successive RRI Differences (RMSSD), Standard Deviation of Normal-to-Normal intervals (SDNN), Percentage of successive RRI that differ by more than 50 ms (pNN50):

$$\begin{aligned} \text{MRR}[r_i] &= \frac{1}{e_i - s_i + 1} \sum_{t=s_i}^{e_i} X(t); \quad \text{RMSSD}[r_i] = \sqrt{\frac{1}{e_i - s_i} \sum_{t=s_i}^{e_i-1} (X(t+1) - X(t))^2} \\ \text{SDNN}[r_i] &= \sqrt{\frac{1}{e_i - s_i + 1} \sum_{t=s_i}^{e_i} (X(t) - \text{MRR}[r_i])^2}; \quad \text{pNN50}[r_i] = \frac{100}{e_i - s_i} \sum_{t=s_i}^{e_i-1} \mathbf{1}(|X(t+1) - X(t)| > 50). \end{aligned}$$

Frequency-domain features include LF power (0.04-0.15 Hz) and HF power (0.15-0.40 Hz) estimated using Welch's method, where  $P_{xx}(f)$  is the power spectral density at frequency  $f$ ,  $w(t)$  is a window function,  $L$  is the window length,  $R$  is the window shift, and  $K$  is the number of windows:

$$\begin{aligned} \text{Power Spectral Density: } P_{xx}(f) &= \frac{1}{K} \sum_{k=0}^{K-1} \left| \sum_{t=0}^{L-1} w(t) X_{r_i}(t + kR) e^{-j2\pi f t} \right|^2 \\ \text{LF Power}[r_i] &= \int_{0.04}^{0.15} P_{xx}(f) df, \quad \text{HF Power}[r_i] = \int_{0.15}^{0.40} P_{xx}(f) df. \end{aligned}$$

*Human-CAI Collaboration Chatbot.* We employ recent open-source LLMs (llama-4-maverick(17B) [52], phi-4-mini(3.8B) [1], gemma-3(27B) [83]) as the human-CAI collaboration chatbot, as they demonstrate competitive performance on medical domain benchmarks [46]. Each LLM receives identical inputs through a structured prompt template (see Appendix E) consisting of: patient profile information (when available), initial MSTFT prediction  $\hat{y}$ , baseline HRV metrics from the complete RRI series  $\mathbf{F}_{\mathcal{R}}$ , and regional HRV metrics from discrepancy regions  $\mathbf{F}_{\mathcal{D}}$  (when present). The generated explanations provide an intuitive summary of psychiatric disorder prediction, allowing clinicians to understand the model's rationale behind its decisions, contest the faithfulness of the model's decisions, and collaborate to finalize decisions.

## 5 Experiments and Results

We evaluate Heart2Mind in 3 stages: (1) assessing MSTFT's psychiatric disorder prediction accuracy, (2) examining SAE's ability to assess MSTFT decision reliability, and (3) measuring how effectively domain experts collaborate with contestable LLM to finalize decisions through the proposed Human-CAI Consensus Rate (HCCR) metric.

### 5.1 Dataset

We employed the raw RRI time series of 60 participants of the HRV-ACC dataset [41]: 30 diagnosed with schizophrenia/bipolar disorder (labeled as "treatment"/positive) and 30 controls (labeled as "control"/negative) with data samples shown in Appendix C. This dataset is considered balanced for training the model. Each participant contributed 1.5–2 hours (minimum 70 minutes) of ECG recordings using a wearable Polar H10 sensor, during which they followed light free-living protocols that included short corridor walks interleaved with seated rest periods. For model training, we transformed each RRI series into a set of overlapping input sequences  $\mathbf{x}_i = (\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_T})$ ,  $T = 300$ , producing  $N - T + 1$  sequences from an original length  $N$ . Consecutive sequences share  $T - 1$  RRIs ( $\mathbf{x}_i \cap \mathbf{x}_{i+1} = \{\mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_T}\}$ ). Thus every inner interval  $\mathbf{x}_k$  appears in exactly  $T$  contexts  $\{\mathbf{x}_{k-T+1}, \dots, \mathbf{x}_k\}$ . We selected  $T = 300$  to align with standard 5-minute short-term HRV analysis windows, which provide sufficient duration to capture meaningful autonomic dynamics across both time-domain and frequency-domain features while remaining practical for continuous wearable monitoring [23]. This sliding-window strategy exposes the network to diverse local temporal neighborhoods, enhancing its ability to capture short- and mid-range autonomic dynamics. Before windowing, we rescaled each participant's RRI signal to zero mean and unit

Table 2. Comparison of model performance. The best results are in **bold**, the second-best are underlined, and (–) indicates that the metric was not reported in the original study. Models marked with <sup>†</sup> correspond to the ablation variants of MSTFT.

Model	#Params	5-fold cross-validation					Leave-one-out cross-validation				
		Acc	Pre	Rec	F1	AUC	Acc	Pre	Rec	F1	AUC
Misgar et al. [56]	8.1M	0.766	<u>0.833</u>	0.667	0.741	<u>0.928</u>	0.833	0.884	0.767	0.836	0.906
Buza et al. [13]	–	–	–	–	–	–	0.833	–	–	–	<u>0.910</u>
Książek et al. [40]	–	<u>0.830</u>	–	–	–	–	0.800	–	–	–	–
MTC/Temporal-only <sup>†</sup>	4.5M	0.787	0.747	0.826	0.784	0.895	0.800	0.781	0.833	0.806	0.826
WT/Frequency-only <sup>†</sup>	4.9M	0.776	0.745	0.740	0.742	0.853	0.800	0.765	<u>0.867</u>	0.813	0.813
MTC+WT+Concat <sup>†</sup>	5.6M	<u>0.830</u>	0.813	<u>0.899</u>	<u>0.854</u>	0.805	<u>0.850</u>	<u>0.920</u>	0.767	0.836	0.850
<b>MSTFT (Ours)</b>	6.3M	<b>0.891</b>	<b>0.877</b>	<b>0.903</b>	<b>0.890</b>	<b>0.961</b>	<b>0.917</b>	<b>0.963</b>	<b>0.867</b>	<b>0.913</b>	<b>0.940</b>

variance to remove inter-subject scale differences. The resulting sample distribution remains balanced (30 positive vs 30 negative) and fully anonymized.

## 5.2 MSTFT – Base Prediction Model Evaluation: Psychiatric Disorder Prediction

This section evaluates the performance of the proposed MSTFT model for predicting psychiatric disorders from RRI time series and compares it with several established baselines using two validation protocols.

**5.2.1 Experimental Setup.** To evaluate MSTFT, we compared it with three strong baselines: the attention-guided method of Misgar et al. [56], the convolutional nearest-neighbor model of Buza et al. [13], and the GRU–SVM ensemble of Książek et al. [40]. For the implemented models (including our re-implementation of Misgar et al. [56]), hyperparameters and architectures (Appendix B) were optimized using the Keras Random Search Tuner [65], reserving 20% of the training data in each split as an inner validation set. For Buza et al. [13] and Książek et al. [40], we report results from the original studies for comparability. To quantify the contribution of each component, we performed an ablation study with three MSTFT variants: (1) multi-scale temporal convolutions (MTC; temporal-only), (2) wavelet transforms (WT; frequency-only), and (3) MTC+WT+Concat (direct concatenation of temporal and frequency features). We report Accuracy, Precision, Recall, F1, and ROC–AUC under two validation schemes: (i) 5-fold cross-validation (48 participants train, 12 test per fold; averaged across folds) and (ii) leave-one-out cross-validation (train on 59 participants, test on 1; repeated 60 times and averaged).

**5.2.2 Results.** Across both validation schemes, MSTFT consistently demonstrated high performance and strong generalization capability. As summarized in Table 2, MSTFT achieved the best results in every metric. Under 5-fold cross-validation, the model reached an accuracy of 0.891 and recall of 0.903, outperforming the strongest baseline [40] by 6.1% in accuracy. Its F1 score of 0.890 and AUC of 0.961 indicated clear separability between control and treatment cases. The ablation variants revealed distinct performance gaps: MTC (temporal-only) and WT (frequency-only) paths achieved moderate results, with accuracy of 0.787 and 0.776, respectively, showing that single-domain features captured only partial dynamics of the RRI signal. The MTC+WT+Concat variant improved accuracy to 0.830, yet the direct concatenation of temporal and frequency representations still lacked adaptive weighting, limiting the interaction between the two modalities. Under LOOCV, which provides a more rigorous, subject-independent evaluation, MSTFT maintained robustness with an accuracy of 0.917 and an AUC of 0.940, outperforming all baselines. The consistent improvement across both 5-fold and LOOCV schemes indicates that MSTFT generalizes well to unseen individuals, an important property for clinical deployment. Notably, this level of accuracy was achieved with only 6.3M parameters, fewer than the 8.1M of Misgar et al. [56], showing

that MSTFT attains an effective balance between performance and model size through its cross-attention and stochastic-skip design. In contrast, MTC (4.5M), WT (4.9M), and MTC+WT+Concat (5.6M) demonstrated that smaller models without adaptive multi-scale fusion could not achieve comparable results.

These results confirm that accurate prediction of psychiatric disorders depends on jointly modeling temporal rhythm variability and frequency-domain information. The MSTFT architecture, through its cross-attention fusion and stochastic depth, effectively captures discriminative time-frequency relationships, yielding both higher accuracy and more stable generalization compared to single-path or non-adaptive fusion methods.

### 5.3 SAE – Built-in Safeguard Evaluation: Base Model Decision Reliability Assessment

This section presents the evaluation of the SAE mechanism, which serves as the built-in safeguard of the base MSTFT model, assessing the faithfulness and internal consistency of the model’s decision-making process. SAE evaluates discrepancies between the attention-based and gradient-based explanation maps, identifying when the model’s attention during inference diverges from the features that most strongly influence its predictions. Such divergences may signal potential unfaithful reasoning, thereby serving as early indicators of unreliable predictions that warrant further human review. For this evaluation, we employed the best-performing MSTFT checkpoint from the 5-fold cross-validation, which yielded balanced outcomes: 27 true positives (TP), 27 true negatives (TN), 3 false positives (FP), and 3 false negatives (FN). To identify meaningful discrepancy regions, we set the gap tolerance  $\delta = 300$ , matching the input window size  $T$ , to merge nearby discrepancies that likely reflect the same underlying autonomic event rather than treating them as separate regions. Using this configuration, Fig. 8 showed the systematic analysis of the discrepancy statistics across prediction categories and multiple threshold values.

**5.3.1 Discrepancy Patterns.** Fig. 8a illustrates that true predictions (TP:  $\mu = 0.046$ ,  $\sigma = 0.075$ , IQR = (0.004, 0.058); TN:  $\mu = 0.072$ ,  $\sigma = 0.085$ , IQR = (0.011, 0.090)) exhibit much lower discrepancy values compared to false predictions (FP:  $\mu = 0.216$ ,  $\sigma = 0.200$ , IQR = (0.025, 0.365); FN:  $\mu = 0.275$ ,  $\sigma = 0.192$ , IQR = (0.109, 0.421)). The consistent alignment between the attention-based and gradient-based explanations in true prediction cases suggests that the model reasons faithfully where its attention mechanisms accurately capture the physiologically meaningful RRI segments that most significantly influence classification. Conversely, the wider spread and

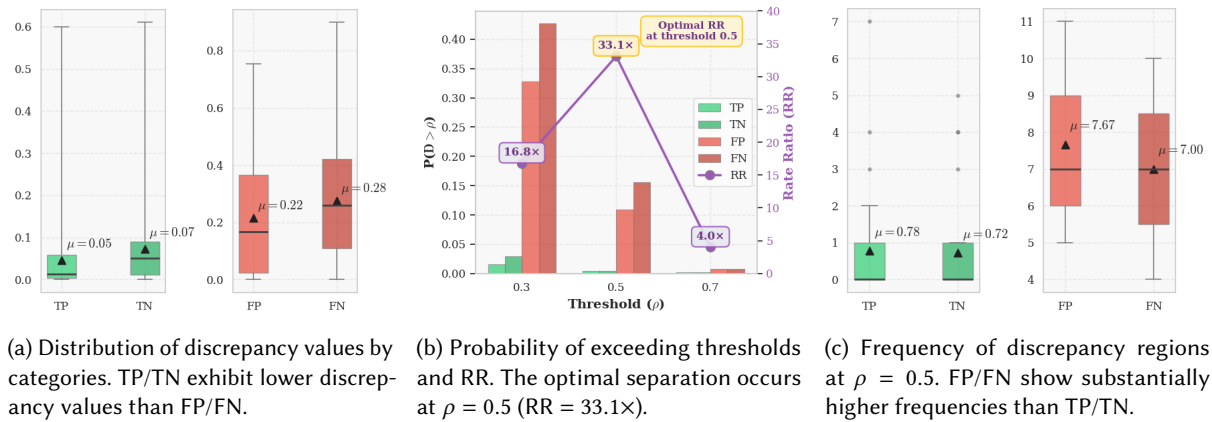


Fig. 8. Analysis of SAE as the built-in safeguard. (a) Discrepancy values of true predictions are lower than those of false predictions, (b) Threshold analysis identifies  $\rho = 0.5$  as the optimal discrimination point with maximum rate ratio (RR), (c) At  $\rho = 0.5$ , false predictions exceed the threshold substantially more frequently than true predictions.

elevated mean discrepancies observed in false predictions reflect instability in the model's focus during inference, where attention may shift toward irrelevant temporal regions or noisy fluctuations. This pattern supports the interpretation that faithful model reasoning corresponds to low-discrepancy regions, where the model's attention-based map is spatially coherent with the gradient-based one, and thus aligns with the features that drive its outputs.

**5.3.2 Threshold ( $\rho$ ) Ablation Study.** To further strengthen the safeguard performance, we conducted an ablation study to determine the optimal discrepancy threshold  $\rho$  that maximizes the separation between faithful and unfaithful model behaviors. To quantify this separation, we defined the rate ratio (RR) as  $RR(\rho) = \frac{P(\mathcal{D} > \rho | \text{error})}{P(\mathcal{D} > \rho | \text{correct})}$ . A higher RR indicates stronger discriminative power at a given threshold  $\rho$ . We selected  $\rho \in \{0.3, 0.5, 0.7\}$  for the ablation study to represent low, moderate, and high discrepancy sensitivities, enabling evaluation of the safeguard's behavior across a broad range of flagging thresholds.

As illustrated in Fig. 8b, the RR peaks at  $\rho = 0.5$ , yielding a  $33.1\times$  enrichment of errors over correct predictions where only  $\approx 0.4\%$  of TP/TN cases are flagged, whereas  $\approx 13.3\%$  of FP/FN cases exceed the limit (FP 10.9%, FN 15.6%). Lowering  $\rho = 0.3$  results in over-flagging (37.8% of false cases, 2.3% of true cases,  $RR = 16.8\times$ ), while raising  $\rho = 0.7$  under-flags (0.8% of false cases, 0.1% of true cases,  $RR = 4.0\times$ ). Therefore,  $\rho = 0.5$  offers the most balanced trade-off, capturing the majority of unfaithful predictions without producing excessive false alarms. Combined with the higher medians and interquartile ranges of  $\mathcal{D}$  in FP/FN relative to TP/TN (Section 5.3.1), this cutoff maximizes discrimination while maintaining interpretability and manageable flagging volume.

Under this optimal  $\rho = 0.5$ , the discrepancy frequency patterns in Fig. 8c further support the safeguard's effectiveness. Correct predictions displayed strong alignment between the two explanation modalities: TN cases had a mean discrepancy count of 0.78 and TP cases 0.72, indicating that disagreements between attention and gradient maps were minimal and localized. A few outliers reached up to 5 (TP) and 7 (TN) discrepancies but still yielded correct outcomes, demonstrating robustness to moderate inconsistencies. In contrast, false predictions exhibited a different profile, with FP cases showing an average of 7.67 discrepancies, ranging from a maximum of 11, while FN cases averaged 7.0, with a maximum of 10.

The results also reveal a clear threshold effect, where prediction reliability begins to decline noticeably once the number of discrepancy regions exceeds approximately 5 to 6. This observation carries important clinical implications, as monitoring the frequency of such high-discrepancy regions allows the system to automatically flag predictions that may be unreliable. In doing so, the SAE provides an inherent mechanism for quantifying uncertainty that operates directly during inference, supporting proactive human oversight.

## 5.4 Contestable LLM Evaluation: Base Model Decision Justification and Enhancement

Building on the SAE safeguard analysis in Sec. 5.3, this section tests whether a contestable LLM can operationalize those safeguards by interpreting SAE discrepancies to justify or revise MSTFT predictions. We evaluate each model's reliability as a human-CAI collaboration chatbot, focusing on its ability to confirm correct outputs and challenge incorrect ones using only physiological evidence. To isolate reasoning over SAE discrepancies and HRV metrics, we excluded domain-specific clinical knowledge from the prompts. We evaluated three open-source LLMs at different scales: phi-4-mini(3.8B), llama-4-maverick(17B), and gemma-3(27B). Each model received the same structured inputs, including patient profile information, the baseline MSTFT prediction, global HRV features, and regional HRV metrics from discrepancy segments. For fair comparison, we fixed generation settings across models: maximum tokens 2048, temperature 0.8, and top- $P$  0.1.

As shown in Table 3, all three contestable LLMs successfully [retained](#) every correct MSTFT prediction (27 TP + 27 TN), confirming their ability to interpret coherent HRV patterns when SAE discrepancies are minimal. This alignment reflects that in faithful reasoning scenarios, characterized by low disagreement regions ( $\rho = 0.5$ ), LLMs can reproduce clinically consistent explanations. More critically, contestable LLM's effectiveness was tested on

Table 3. Performance of contestable LLM in justifying or overturning MSTFT predictions with SAE’s discrepancies. Arrows (↑/↓) denote the desired direction of improvement.

Model	Retain (TP)↑	Retain (TN)↑	Overturn (TP/TN)↓	Overturn (FN)↑	Overturn (FP)↑	Retain (FN/FP)↓
llama-4-maverick(17B)	27	27	0	0	1	5
phi-4-mini(3.8B)	27	27	0	1	0	5
gemma-3(27B)	27	27	0	2	1	3

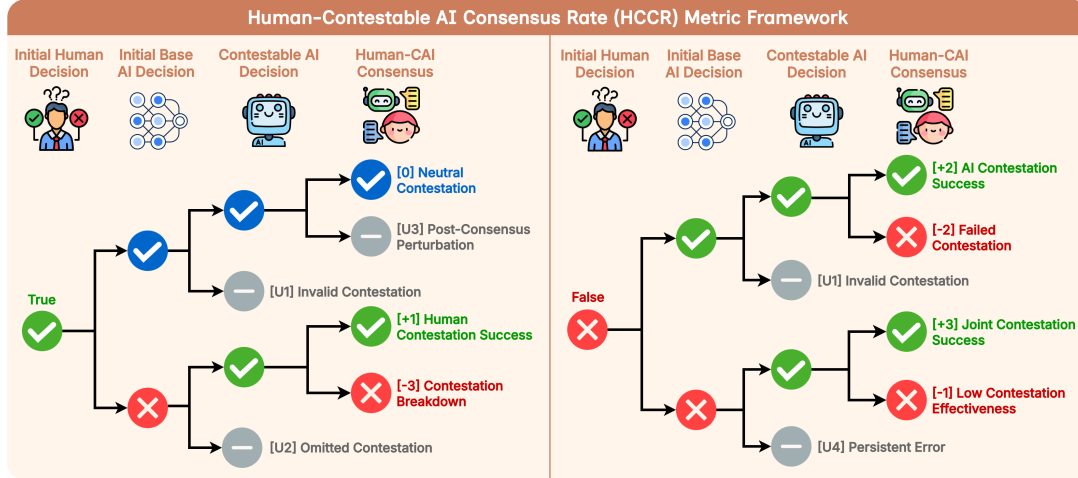


Fig. 9. Human-Contestable AI Consensus Rate (HCCR) Metric Framework.

the 6 erroneous MSTFT predictions (3 FN, 3 FP). All LLMs **overturned** at least one incorrect case, demonstrating varying degrees of autonomous contestation. The largest model, gemma-3, achieved the highest correction rate, **overturning** 3 of 6 errors (2 FN, 1 FP), followed by llama-4-maverick with one FP correction, and phi-4-mini with one FN correction. These results suggest that larger models exhibit stronger interpretive reasoning for identifying unfaithful MSTFT decisions, although smaller models display complementary strengths across error types. Overall, this evaluation shows that contestable LLM can effectively leverage SAE discrepancies to both affirm reliable predictions and contest unfaithful ones, advancing from static explainability to active contestability.

### 5.5 Human-centered Contestability Evaluation: Human-CAI Consensus Rate (HCCR)

To evaluate human-centered contestability within Heart2Mind, we propose the **Human-CAI Consensus Rate (HCCR)** metric, designed to measure how effectively humans collaborate with CAI to reach the correct final decision. This metric captures both the outcomes and processes of contestation, emphasizing not only the accuracy achieved after collaboration but also how contestation improves or undermines the reliability of decisions.

**5.5.1 Metric Definition.** As shown in Fig. 9, HCCR traces the full trajectory of decision-making through four stages: (1) **Initial human decision**; (2) **Initial base AI decision**, (3) **CAI decision** informed by explanations, and (4) **Human-CAI consensus** after collaboration. Each case is categorized as either a **true** or **false** prediction, depending on its consistency with the ground truth. HCCR defines every possible decision path connecting these four stages and assigns an ordinal level that represents how contestation influenced the final outcome. Defined levels, ranging from **-3** to **+3**, quantify the effectiveness of contestation, while undefined states (U1–U4) correspond



Table 4. Summary of HCCR outcome levels (defined and undefined states). Each row represents a trajectory of the decision process across four stages: initial human decision  $\rightarrow$  initial base AI output  $\rightarrow$  CAI decision  $\rightarrow$  Human–CAI consensus, each evaluated against the ground truth ( $\checkmark$  = true,  $\times$  = false). Defined levels  $[0, \pm 1, \pm 2, \pm 3]$  quantify contestation effectiveness, while U1–U4 are non-scorable states, reflecting system or interaction irregularities rather than genuine contestation outcomes.

Level	Formal name	Path	Description
[-3]	Contestation Break-down	$\checkmark \rightarrow \times \rightarrow \checkmark \rightarrow \times$	Human begins correctly, the base AI issues an incorrect output, the CAI provides the correct recommendation, yet the final consensus turns wrong. This reflects miscalibrated trust that overrides both the initial human judgment and CAI.
[-2]	Failed Contestation	$\times \rightarrow \checkmark \rightarrow \checkmark \rightarrow \times$	Human begins incorrectly, base AI and CAI identify the correct label with justification, but fail to persuade, so the final consensus remains wrong.
[-1]	Low Contestation Effectiveness	$\times \rightarrow \times \rightarrow \checkmark \rightarrow \times$	Human and base AI begin incorrectly, the CAI is correct, but its evidence or interaction is insufficient to reverse the error, leaving the final consensus wrong.
[0]	Neutral Contestation	$\checkmark \rightarrow \checkmark \rightarrow \checkmark \rightarrow \checkmark$	Human, base AI, and CAI are all correct, and the dialogue simply confirms the correct decision.
[+1]	Human Contestation Success	$\checkmark \rightarrow \times \rightarrow \checkmark \rightarrow \checkmark$	The human is correct while the base AI is wrong; the CAI aligns with the human and helps retain the correct decision, yielding a correct consensus.
[+2]	AI Contestation Success	$\times \rightarrow \checkmark \rightarrow \checkmark \rightarrow \checkmark$	The human is wrong while both the base AI and the CAI are correct; dialogue overturns the human decision and reaches a correct consensus.
[+3]	Joint Contestation Success	$\times \rightarrow \times \rightarrow \checkmark \rightarrow \checkmark$	Human and base AI are wrong; the CAI supplies the correct alternative and successfully steers the discussion to a correct consensus.
[U1]	Invalid Contestation	$\checkmark/\times \rightarrow \checkmark \rightarrow \times \rightarrow \emptyset$	The base AI is correct, but the CAI outputs an incorrect challenge. By design, the CAI must not flip a correct base decision, so this path is blocked.
[U2]	Omitted Contestation	$\checkmark \rightarrow \times \rightarrow \times \rightarrow \emptyset$	The human is correct and the base AI is wrong, yet the CAI fails to intervene or to challenge the error; there is no contestation episode to score, so no consensus state is recorded.
[U3]	Post-Consensus Perturbation	$\checkmark \rightarrow \checkmark \rightarrow \checkmark \rightarrow \times$	After all stages are correct, a consensus flip occurs without new evidence, logged as decision instability and not a valid contestation outcome.
[U4]	Persistent Error	$\times \rightarrow \times \rightarrow \times \rightarrow \emptyset$	Human, base AI, and CAI are all wrong, and no contestation progresses to a consensus state; this logs a persistent error rather than a scorable outcome.

Table 5. Summary of human-centered contestability evaluation in HCCR, readability, and interactability. Time to Decision (TTD) and Dialogue Length (DL) are reported as median [min-max].

	Human-CAI Consensus Rate (HCCR)							Readability	Interactability	
	[-3]	[-2]	[-1]	[0]	[+1]	[+2]	[+3]	Mean FKGL	TTD (min:s)	DL (turns)
TP		0		7		5		14.6	7:24 [3:45–14:41]	3 [1:4]
FN	0		1		8		3	15.4	8:30 [3:12–17:34]	4 [1:6]

to irregular or disallowed interaction patterns that do not constitute valid contestation. As detailed in Table 4, lower scores indicate breakdowns or failed contestation, while higher scores reflect successful collaboration between human and AI agents. HCCR can also generalize beyond one-to-one human–AI settings to more complex interactions such as many-to-one (e.g., multiple healthcare specialists collaborating with a single CAI) or many-to-many (e.g., multiple collaborations between different specialists and CAI agents). This flexibility allows HCCR to serve as a generalizable measure of collaborative decision reliability across human-centered AI systems.

**5.5.2 Experimental Setup.** Within the psychiatric disorder diagnosis setting, the four stages are defined as follows: (1) Initial human decision: the clinician analyzes the RRI time series and issues a preliminary judgment without AI support; (2) Initial base AI decision: base MSTFT model predicts the class label; 3) CAI decision: contestable LLM (gemma-3 was chosen due to its stable and superior performance evaluated in Section 5.4), supported by

SAE discrepancies, provides its own classification and rationale; and (4) Human–CAI consensus: the user, after interacting with the CAI through CDI and contestable LLM chatbot, finalizes the decision. For evaluation<sup>2</sup>, 12 participants with diverse expertise were recruited: 1 in cognitive psychology, 2 in biomedical engineering, 7 in AI, and 2 in human–computer interaction (HCI). Each participant completed two diagnostic sessions using the CDI interface: (1) Correct prediction case: base MSTFT model correctly classified the sample, and contestable LLM was expected to validate and **retain** this outcome. This case evaluated the CAI’s ability to confirm consistent reasoning and support correct decisions; (2) Incorrect prediction case: base MSTFT model incorrectly classified a sample, and contestable LLM was expected to **overturn** the misclassification. This study tested whether human–CAI collaboration could identify inconsistencies and guide decisions toward the correct outcome. To complement HCCR, additional human-centered measures were included to provide a more comprehensive evaluation of system usability and interpretability: (1) **Flesch-Kincaid Grade Level (FKGL) [Readability]** estimates the U.S. school grade needed to comprehend a text [16]. FKGL values around 16–17 are typical for specialist medical literature, which supports using a similar grade-level ceiling when writing for clinicians and domain experts [7]; (2) **Time to Decision (TTD) [Interactability]** records elapsed time from case presentation to final decision; (3) **Dialogue Length (DL) [Interactability]** records the number of conversational turns to reach the final decision.

**5.5.3 True Positive Case Analysis.** In the correction prediction (TP) case (*treatment\_40*), the base MSTFT model correctly classified the case as “treatment,” and contestable LLM was expected to **retain** this decision. This case contained 5 SAE discrepancy regions (Fig. 10), which represented the highest number of inconsistencies observed among correct predictions. Despite this, all three contestable LLMs converged on the same “treatment” classification, showing consistent physiological reasoning and high interpretability. llama-4-maverick focused on “significant variability across different regions” and “lower parasympathetic activity indicators”, emphasizing differences in local HRV patterns that aligned with reduced vagal tone. phi-4-mini identified “regional discrepancies in HRV metrics” and “variability in LF/HF ratios across regions”, showing similar regional reasoning but with slightly less depth in autonomic interpretation. gemma-3 described “regional discrepancies indicating autonomic instability” and “low vagal tone indicators (RMSSD, pNN50)”, offering a more comprehensive and clinically precise explanation of autonomic imbalance. Regarding response time, llama-4-maverick produced its output in just 10.04 seconds, significantly faster than the smaller phi-4-mini (24.40 seconds), which highlights its potential suitability for real-time clinical applications where rapid validation of correct decisions is crucial.

Human-centered evaluation supported these observations. As shown in Table 5, out of 12 sessions, 7 were [0] **Neutral Contestation** and 5 were [+2] **AI Contestation Success**, showing that when experts initially disagreed, CAI successfully guided them back to the correct decision. The responses were concise and readable (FKGL 14.6), with efficient interactions (median time to decision: 7:24 minutes, range: 3:45 to 14:41; median dialogue length: 3 turns, range: 1 to 4).

**5.5.4 False Negative Case Analysis.** In the incorrect prediction (FN) case (*treatment\_1*), the base MSTFT model incorrectly classified a “treatment” sample as “control,” and contestable LLM was expected to **overturn** this error. This case contained 7 SAE discrepancy regions (Fig. 12), the highest observed among all misclassified samples, illustrating substantial disagreement in HRV patterns across regions. The case was therefore particularly suitable for evaluating how the CAI reasoned through uncertainty and complexity.

Both gemma-3 and phi-4-mini correctly identified autonomic dysregulation patterns characteristic of psychiatric conditions and successfully **overturned** the baseline “control” prediction to “treatment.” phi-4-mini highlighted “reduced parasympathetic activity (low RMSSD and pNN50)” and “reduced overall HRV (low SDNN in some regions)”, showing an accurate focus on diminished HRV and parasympathetic withdrawal. gemma-3 described a “pattern of autonomic instability” and “fluctuations in LF/HF balance with periods of reduced HRV,”

<sup>2</sup>Participant session videos are available at <https://www.youtube.com/playlist?list=PLWTLzhO-RmiwdSjadQCfp8hpAF8RducVL>

capturing both temporal and spatial irregularities in HRV that correspond to clinical features of psychiatric dysregulation. In contrast, llama-4-maverick **retained** the incorrect “control” classification, justifying that “*the fluctuations are within the realm of normal variability*”. This reflected a more conservative interpretation of the HRV variance, suggesting narrower sensitivity in recognizing autonomic pathology.

Human-centered evaluation confirmed the CAI’s effectiveness in corrective collaboration. Out of 12 sessions, 8 were **[+1] Human Contestation Success**, 3 were **[+3] Joint Contestation Success**. Notably, there was 1 case of **[-1] Low Contestation Effectiveness**, where the participant interpreted the irregular RRI pattern as a signal artifact rather than a treatment-related feature, underscoring a current CAI limitation discussed in Section 6.2 toward future multimodal integration. Most of sessions required slightly more deliberation (median time to decision: 8:30 minutes, range 3:12 to 17:34; dialogue length: 4 turns, range: 1 to 6) and used more technical language (FKGL 15.4), consistent with the higher reasoning demands of error correction. The larger gemma-3 model produced the most accurate and detailed rationales but required longer generation time (36-38 seconds), whereas the smaller models were faster but less reliable when confronted with ambiguous patterns.

Together, TP and FN analyses demonstrate that human-CAI collaboration was effective in both validating correct outcomes and recovering from model errors. In the TP case, the CAI strengthened clinician confidence by

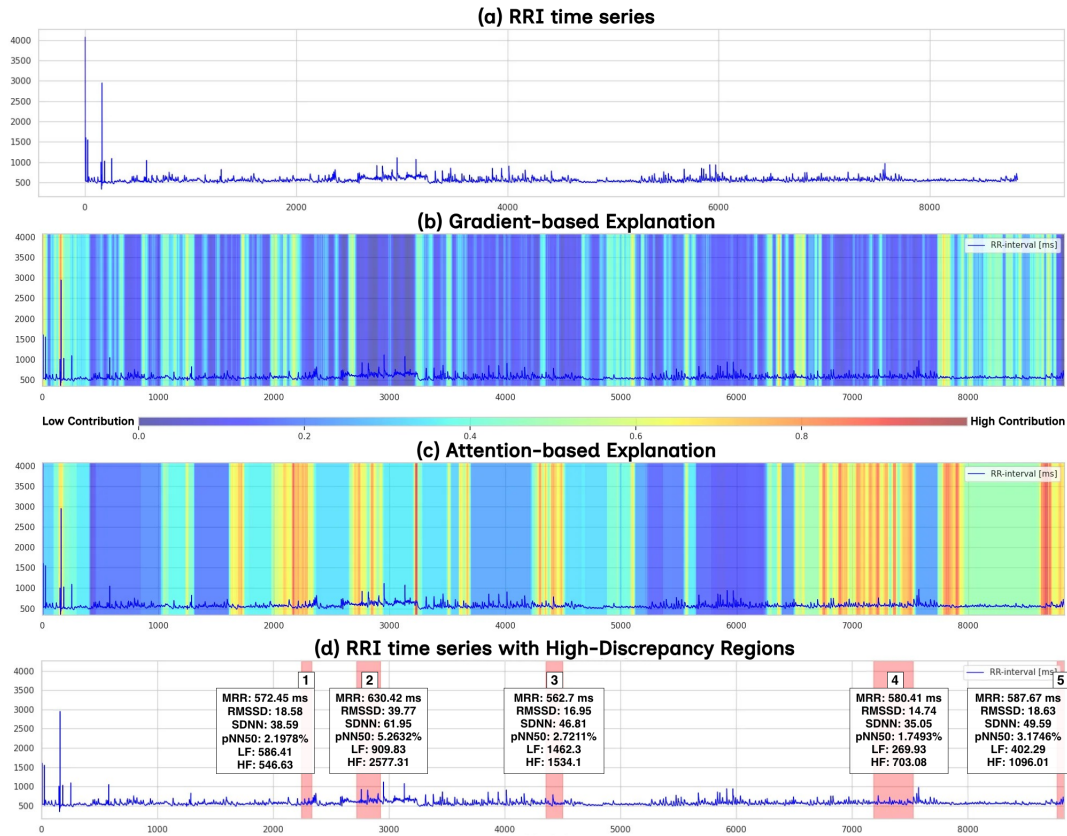


Fig. 10. SAE discrepancies of the TP case (*treatment\_40*): (a) Raw input RRI time series, (b) Gradient-based Explanation, (c) Attention-based Explanation, and (d) 5 discrepancies detected by SAEs.

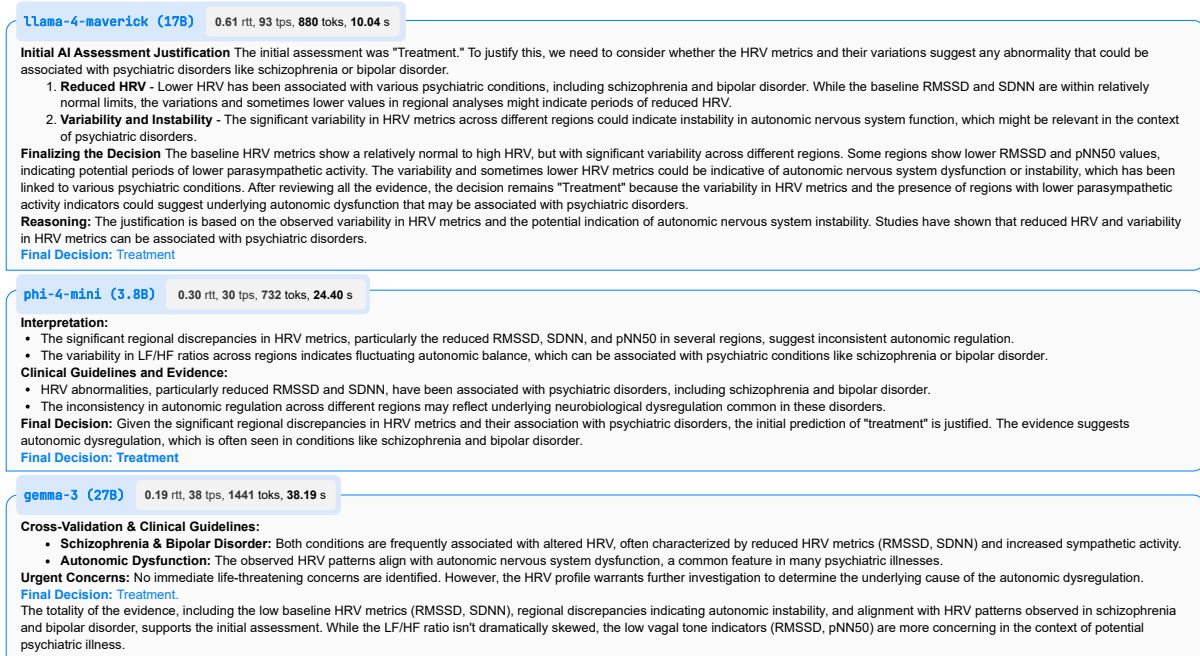


Fig. 11. Response excerpts of contestable LLMs in TP case, where they **retained** initial correct prediction of baseline MSTFT. Output metrics: time to first token (rtt), token per second (tps), number of output tokens (toks), output time (s).

providing confirmatory reasoning, whereas in the FN case, it actively supported corrective reasoning, resulting in accurate consensus in 11 of 12 sessions. HCCR distributions reflected this adaptability: stable confirmation under low uncertainty and productive contestation under high discrepancy. Larger models such as gemma-3 consistently exhibited stronger interpretive capacity, whereas llama-4-maverick offered superior responsiveness, highlighting a tradeoff between reasoning depth and latency. Overall, these findings emphasize the promise of contestable LLMs as transparent, collaborative diagnostic partners capable of maintaining accuracy, interpretability, and clinical efficiency across both stable and uncertain decision contexts.

## 6 Discussion

### 6.1 Potential of Wearable ECG Devices in Psychiatric Disorder Prediction

Wearable ECG devices are emerging as powerful psychiatric care tools, measuring heart activity and HRV to assess ANS function, which shows dysregulation in schizophrenia and bipolar disorder. Our explanation methods consistently highlight rapid RRI changes, aligning with research showing lower HRV in these disorders compared to healthy controls [9, 39, 40]. Both conditions exhibit ANS dysregulation with reduced parasympathetic tone and elevated sympathetic activity, linked to symptom severity and illness progression, making HRV a promising objective biomarker. Continuous ECG monitoring enables real-time detection of subtle cardiac changes, opening avenues for early detection and monitoring. The ease of ambulatory measurement has expanded HRV interest in mental health through smartphone apps paired with wearables [4]. This data stream complements traditional assessments: for instance, sustained vagal tone decline in bipolar patients could alert clinicians to depressive or manic shifts, while heightened autonomic arousal in schizophrenia might indicate stress or early

psychotic symptoms. Wearable ECGs extend psychiatric monitoring from clinics to daily environments, capturing physiological changes that would otherwise go unnoticed. Although wearables are often cost-effective compared to conventional diagnostics [20], their utility depends on disease context and healthcare access [41]. Because they measure general physiological signals, they cannot serve as stand-alone diagnostic tools. Our future work should integrate additional modalities (e.g., skin conductance, temperature, activity) and expand datasets to include more diverse populations and disorders. A multimodal approach combining various data streams would provide a holistic understanding of physiology, thereby improving psychiatric disorder prediction accuracy.

## 6.2 Toward Human-centered CAI Systems in Healthcare: Opportunities, Challenges and Visions

**6.2.1 Opportunities and Challenges.** Our study highlights the difficulty of designing comprehensive clinical evaluations for a human-centered CAI system. This difficulty stems from the multidisciplinary expertise required for a thorough assessment, which brings together specialists who can jointly reason about psychiatric disorders, cardiac signals, and human-AI interaction, a combination that remains relatively rare in clinical practice. Our current human-centered evaluation, based on the HCCR and dialogue readability, primarily captures the accuracy dimension (i.e., the frequency with which humans and CAI collaborate to reach the correct final decision).

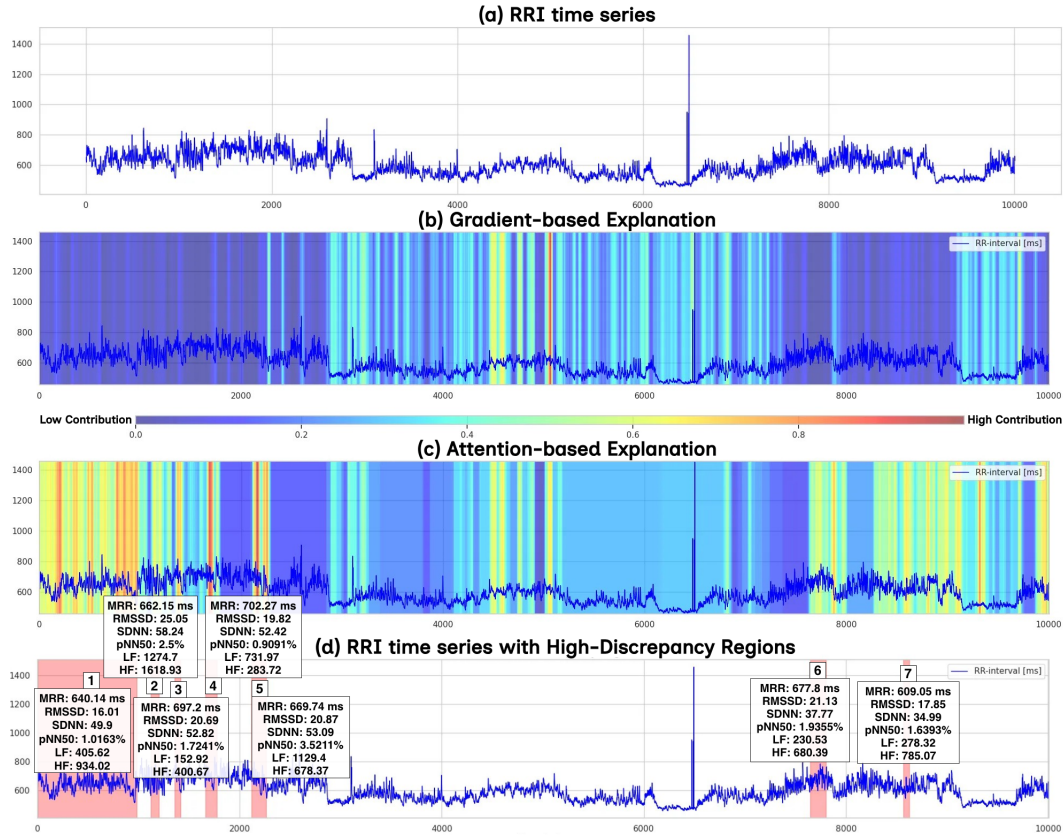


Fig. 12. SAE discrepancies of the FN case (*treatment\_1*): (a) Raw RRI time series, (b) Gradient-based Explanation, (c) Attention-based Explanation, and (d) 7 discrepancies detected by SAEs.



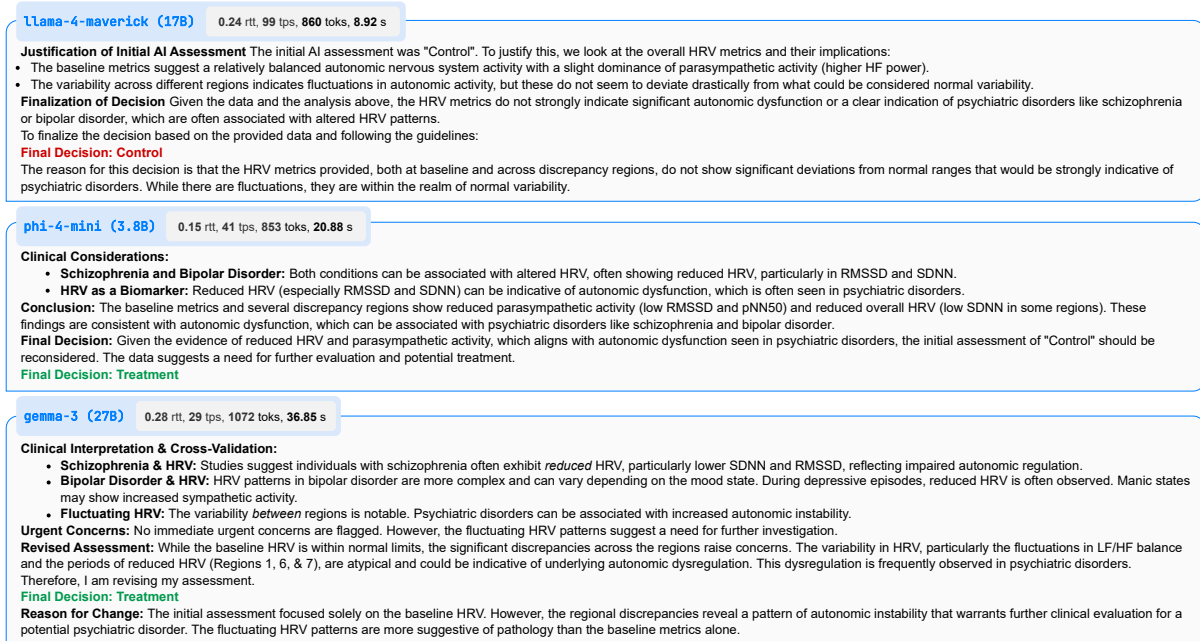


Fig. 13. Response excerpts of contestable LLM in FN case. The final decision (in red) indicates the model retains the incorrect initial prediction of base MSTFT model. The final decision (in green) indicates that the model overturns the incorrect initial prediction of base MSTFT model to the correct prediction. Output metrics: time to first token (rtt), token per second (tps), number of output tokens (toks), output time (s).

Our work demonstrates that contestable LLM systems can provide initial diagnostic reasoning and refine their conclusions when clinicians raise valid counter-evidence, with evaluation results showing that clinicians and CAI reach correct consensus through concise and readable dialogues. This reflects a future in which machines can evaluate human challenges and adjust their decisions accordingly [43], aligning with calls in medical AI ethics for context-dependent explainability rather than a single fixed standard. Future work should extend this framework to cover additional dimensions such as user experience, perceived workload, trust calibration, user-adaptive readability of explanations, and the impact on clinician learning and confidence over time. Furthermore, meaningful clinical validation must consider broader factors, including clinical utility, workflow integration, and long-term effects on treatment decisions and outcomes. While our preliminary technical and human-centered validation on the HRV-ACC dataset is encouraging, full clinical implementation will require large-scale real-world studies across diverse patient populations, regions, clinical environments, and healthcare providers, in line with emerging best practices that emphasize contextual performance rather than technical metrics alone [25, 86].

Our evaluation demonstrates opportunities for human-centered CAI in psychiatric disorder prediction. SAE discrepancy analysis establishes a practical safeguard mechanism, where a threshold of  $\rho = 0.5$  effectively enriches flagged errors while maintaining low false alarm rates. The finding that predictions with more than approximately five discrepancy regions are especially likely to require clinician review provides actionable guidance for clinical deployment. Building on this foundation, contestable LLMs can leverage SAE signals combined with HRV summaries to operationalize human-AI collaboration, either confirming correct outputs or facilitating the overturning of misclassifications. However, current LLMs still exhibit inconsistent medical knowledge. The

variation in contestation effectiveness across architectures highlights limitations in their pretrained knowledge about psychiatric disorders and HRV interpretation. Studies on psychiatric note analysis have found that LLMs *lack a robust understanding of meaning and nuances* in mental health codes [11], emphasizing the need for domain-specific optimization and the incorporation of clinical expertise [78].

Although limited by the size and homogeneity of the available dataset and the unimodal physiological input [41], our work represents a first attempt at building a CAI system for continuous healthcare assessment using short recordings. We are currently developing partnerships with clinical institutions to support the next phase of evaluation, which will incorporate both quantitative performance measures and qualitative feedback from healthcare providers and patients. Clinical deployment will allow us to examine how clinicians use discrepancy regions flagged by SAE, how the contestable LLM supports revision of initial decisions, and how CDI integrates into existing diagnostic routines.

**6.2.2 Visions for Future CAI Systems.** The vision of CAI extends beyond technical progress on a single task to include broader aspects of agency, ethics, and human–AI interaction. The transferability of our CAI framework (detailed in Appendix F) demonstrates that each component (i.e., base prediction model, built-in safeguard mechanism, and human–CAI collaboration interface) can be implemented using different technical approaches suited to specific clinical domains and data modalities, while preserving the core contestability principles that ensure human oversight and collaborative decision-making.

A promising direction is adaptive explainability, where explanation style, complexity, and detail automatically adjust to match the specific context and user needs, moving beyond heuristic prompt engineering approaches. This could mean simplifying explanations for a junior clinician or providing more technical, detailed justifications for an expert specialist. It could also mean highlighting different content, such as emphasizing pathophysiology for a physician versus explaining in lay terms for patient-facing contexts. Future research may explore methods for AI to gauge clinicians’ needs, which could be achieved through user profiles or real-time interactions, and then tailor its explanations accordingly. Moreover, learning to explain frameworks might be developed, where models are trained not just to maximize accuracy but also to optimize explanation human-centered utility metrics, such as usefulness, plausibility, faithfulness, and fairness [38, 59]. Incorporating clinician feedback on the effectiveness of explanations enables LLMs to refine their contestability strategies over time, thereby improving both efficiency and understanding in clinical practice. Future CAI should employ multi-model agreement, where cross-checking between models or algorithmic rules determines consensus. For patient vital signs or sensor data (e.g., continuous ECG monitoring), signal processing algorithms like MSTFT and contestable LLM can cross-validate each other’s interpretations. Research is needed to integrate these verification steps into clinical workflows without introducing delays. We also expect new governance mechanisms, such as hospital AI oversight committees, continuous monitoring of AI models in deployment to detect drifts or biases that could affect contestability (e.g., AI whose explanations degrade over time due to data shifts), and comprehensive audit trails that log all AI recommendations and clinician overrides to support accountability. Feedback loops in governance are equally important, where clinicians and patients should have avenues to report AI errors or problematic decisions, prompting model improvements or regulatory action [51].

### 6.3 Limitations and Future Works

Despite the promising results, several limitations should be acknowledged to guide future research directions.

**6.3.1 Scope of Clinical Validation.** Our empirical evaluation is currently limited to schizophrenia and bipolar disorder. While these conditions are well-motivated targets due to their documented ANS dysregulation and established HRV abnormalities [36, 37, 57, 76, 90], psychiatric disorders encompass a much broader spectrum with heterogeneous physiological signatures. Consequently, the generalizability of the proposed framework to

other conditions (e.g., depression, anxiety disorders) remains unexplored. Future work should extend validation to a wider range of ECG-based psychiatric diagnoses and symptom profiles [81, 85], enabling a more comprehensive assessment of robustness and clinical applicability across mental health contexts.

**6.3.2 Model Complexity.** The size of our base MSTFT model ( $\approx 6.3\text{M}$  trainable parameters; see Appendix B) is relatively large compared to the available dataset, as this study prioritizes predictive performance. While regularization and cross-validation were used to reduce overfitting, the model complexity may still limit generalization. In addition, the absence of an independent external dataset restricts evaluation under distribution shifts commonly seen in real-world clinical settings. Our future work should collect larger and more diverse datasets with external validation, and explore more parameter-efficient models or techniques, such as pruning, quantization, or knowledge distillation [61], to support deployment on a wider range of clinical and edge devices.

**6.3.3 Subject-Specific Prompt Template.** The contestable LLM component relies on subject-specific inputs (see Appendix E) to support clinical reasoning and contestation. While these inputs enhance interpretability and contextual relevance, they also introduce a potential limitation. Specifically, the system’s performance and explanations may depend on the completeness, accuracy, and availability of such individualized information, which may vary across clinical environments. Moreover, this design choice raises questions about how well the system would perform when certain patient attributes are missing, noisy, or intentionally withheld. Future work should investigate prompt robustness under partial or uncertain inputs, such as with retrieval-augmented generation (RAG), argumentative computation [24] or iterative prompt refinement [50], on historical medical records, and explore standardized prompt configurations that balance personalization with broader applicability.

## 7 Conclusion

Heart2Mind integrates wearable ECG monitoring with a human-centered CAI system to reduce diagnostic subjectivity, address monitoring gaps, and preserve clinician oversight. The base MSTFT model achieves 91.7% accuracy, demonstrating the effectiveness of a multi-scale cardiac biomarker approach. Built-in safeguard SAE discrepancies, calibrated at  $\rho = 0.5$ , sharply enrich error detection and direct clinician attention to unreliable predictions. Contestable LLM uses these discrepancies to validate correct outputs or overturn misclassifications, and our human-centered evaluation shows that experts and CAI achieve high agreement through concise, readable dialogues. Future work should incorporate additional physiological modalities, explore parameter-efficient models, standardize prompt configurations, diversify datasets, and conduct large-scale clinical validation across care settings to strengthen reliability and ensure systems like Heart2Mind function as trustworthy, collaborative tools for earlier detection and more personalized psychiatric care, while upholding clinician agency.

## Acknowledgments

This study is supported by the UNB-FCS Startup Fund (22–23 START UP/H CAO) and NSERC Discovery Grant No RGPIN 2025-00129. It is also funded in part by the National Research Council of Canada’s Aging in Place Challenge program.

## References

- [1] Abdelrahman Abouelenin et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743* (2025).
- [2] ARSS Izzatunnisa Ainunhusna et al. 2020. Bipolar disorder classification based on electrocardiogram signal using support vector machine. *INTERNATIONAL JOURNAL OF SCIENTIFIC and TECHNOLOGY RESEARCH* 9, 1 (2020), 4.
- [3] Kars Alfrink et al. 2023. Contestable AI by design: Towards a framework. *Minds and Machines* 33, 4 (2023), 613–639.
- [4] Navya Alugubelli et al. 2022. Wearable devices for remote monitoring of heart rate and heart rate variability—what we know and what is coming. *Sensors* 22, 22 (2022), 8903.

- [5] Nancy C Andreasen et al. 1995. Correlational studies of the Scale for the Assessment of Negative Symptoms and the Scale for the Assessment of Positive Symptoms: an overview and update. *Psychopathology* 28, 1 (1995), 7–17.
- [6] Javiera T Arias and César A Astudillo. 2023. Enhancing Schizophrenia Prediction Using Class Balancing and SHAP Explainability Techniques on EEG Data. In *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*. IEEE, 1–5.
- [7] Nadir Noureldin Abdella Bahta et al. 2024. Assessing the readability in medical research papers. *Medical Research Archives* 12, 7 (2024).
- [8] İsmail Baydili et al. 2025. Artificial Intelligence in Psychiatry: A Review of Biological and Behavioral Data Analyses. *Diagnostics* 15, 4 (2025), 434.
- [9] Beatrice R Benjamin et al. 2021. Heart rate variability is associated with disease severity in psychosis spectrum disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 111 (2021), 110108.
- [10] Gary G Berntson et al. 1997. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology* 34, 6 (1997), 623–648.
- [11] Lokesh Boggavarapu et al. 2024. Evaluating Enhanced LLMs for Precise Mental Health Diagnosis from Clinical Notes. *medRxiv* (2024), 2024–12.
- [12] Louise Brådvik. 2018. Suicide risk and mental disorders. 2028 pages.
- [13] Krisztian Buza et al. 2023. A Simple and Effective Classifier for the Detection of Psychotic Disorders based on Heart Rate Variability Time Series. In *Workshop on Bioinformatics and Computational Biology WBCB 2023*.
- [14] Health Canada. 2025. Pre-market guidance for machine learning-enabled medical devices. <https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/pre-market-guidance-machine-learning-enabled-medical-devices.html>
- [15] Matteo Cella et al. 2018. Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. *Schizophrenia research* 195 (2018), 537–542.
- [16] Douglas W Challener et al. 2025. Flesch-Kincaid Grade Level Readability Scores to Evaluate Readability of Clinical Documentation During an Electronic Health Record Transition. *Advances in Health Information Science and Practice* 1, 1 (2025), VBWY7913.
- [17] Chun-Fu Richard Chen et al. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 357–366.
- [18] Hao-Fei Cheng et al. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [19] Filippo Corponi et al. 2024. A Bayesian analysis of heart rate variability changes over acute episodes of bipolar disorder. *npj Mental Health Research* 3, 1 (2024), 44.
- [20] Julia Cullen et al. 2021. Experience of a smartphone ambulatory ECG clinic for emergency department patients with palpitation: a single-centre cohort study. *European Journal of Emergency Medicine* 28, 6 (2021), 463–468.
- [21] Richard Dazeley et al. 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence* 299 (2021), 103525.
- [22] U.S. Food and Drug Administration. 2025. Clinical Decision Support Software - Guidance. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>
- [23] Task Force. 1996. Standards of measurement, physiological interpretation and clinical use. Task force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation* 93, 5 (1996), 1043–1065.
- [24] Gabriel Freedman et al. 2025. Argumentative Large Language Models for Explainable and Contestable Claim Verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 14930–14939. Issue: 14.
- [25] Nils Freyer et al. 2024. The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. *BMC Medical Ethics* 25, 1 (2024), 104.
- [26] Maurizio Garbarino et al. 2014. Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th international conference on wireless mobile communication and healthcare-transforming healthcare through innovations in mobile and wireless technologies (MOBIHEALTH)*. IEEE, 39–42.
- [27] MD Hill et al. 2024. Precision medicine for psychotic disorders: objective assessment, risk prediction, and pharmacogenomics. *Molecular Psychiatry* (2024), 1–22.
- [28] Katrina Hinde et al. 2021. Wearable devices suitable for monitoring twenty four hour heart rate variability in military populations. *Sensors* 21, 4 (2021), 1061.
- [29] Tad Hirsch et al. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 95–99.
- [30] Gao Huang et al. 2016. Deep networks with stochastic depth. In *European conference on computer vision*. Springer, 646–661.
- [31] Tomoko Inoue et al. 2022. The development and clinical application of a novel schizophrenia screening system using yoga-induced autonomic nervous system responses. *Frontiers in Physiology* 13 (2022), 902979.
- [32] Sarthak Jain and Byron C Wallace. 2019. Attention is not Explanation. In *Proceedings of NAACL-HLT*. 3543–3556.
- [33] Carmen Jimenez-Mesa et al. 2024. Machine learning in small sample neuroimaging studies: Novel measures for schizophrenia analysis. *Human Brain Mapping* 45, 5 (2024), e26555.

- [34] Amir-Hossein Karimi et al. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics*. PMLR, 895–905.
- [35] Smith K Khare et al. 2023. ECGPsychNet: An optimized hybrid ensemble model for automatic detection of psychiatric disorders using ECG signals. *Physiological Measurement* 44, 11 (2023), 115004.
- [36] Hye-Geum Kim et al. 2018. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation* 15, 3 (2018), 235.
- [37] Joel EW Koh et al. 2022. Automated classification of attention deficit hyperactivity disorder and conduct disorder using entropy features with ECG signals. *Computers in biology and medicine* 140 (2022), 105120.
- [38] Xiangwei Kong et al. 2024. Toward Human-centered XAI in Practice: A survey. *Machine Intelligence Research* 21, 4 (2024), 740–770.
- [39] Kamil Michał Książek et al. 2023. The analysis of heart rate variability and accelerometer mobility data in the assessment of symptom severity in psychotic disorder patients using a wearable Polar H10 sensor. *medRxiv* (2023), 2023–08.
- [40] Kamil Michał Książek et al. 2025. Deep learning approach for automatic assessment of schizophrenia and bipolar disorder in patients using RR intervals. *medRxiv* (2025), 2025–03.
- [41] K Książek et al. [n. d.]. HRV-ACC: a dataset with RR intervals and accelerometer data for the diagnosis of psychotic disorders using a Polar H10 wearable sensor, 2023b. URL <https://doi.org/10.5281/zenodo.8171266> ([n. d.]).
- [42] Emily Lam et al. 2020. Measuring heart rate variability in free-living conditions using consumer-grade photoplethysmography: Validation study. *JMIR Biomedical Engineering* 5, 1 (2020), e17355.
- [43] Francesco Leofante et al. 2024. Contestable AI needs computational argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*. 888–896.
- [44] Hui Li and Xiao-Jun Wu. 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion* 103 (2024), 102147.
- [45] Xiangchen Li et al. 2023. Evaluation of measurement accuracy of wearable devices for heart rate variability. *Iscience* 26, 11 (2023).
- [46] Percy Liang et al. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). Featured Certification, Expert Certification.
- [47] Yibing Liu et al. 2022. Rethinking attention-model explainability through faithfulness violation test. In *International conference on machine learning*. PMLR, 13807–13824.
- [48] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [49] Henrietta Lyons et al. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.
- [50] Aman Madaan et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2023), 46534–46594.
- [51] Gennie Mansi et al. 2025. Legally-Informed Explainable AI. *arXiv preprint arXiv:2504.10708* (2025).
- [52] AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, April 2025.
- [53] Jacqueline Michelle Metsch et al. 2024. CLARUS: an interactive explainable AI platform for manual counterfactuals in graph neural networks. *Journal of Biomedical Informatics* 150 (2024), 104600.
- [54] Jacqueline Michelle Metsch et al. 2025. Ensemble Post-hoc Explainable AI in Multivariate Time Series: Identifying Medical Features Driving Disease Prediction. *bioRxiv* (2025), 2025–02.
- [55] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [56] Muzafar Mehraj Misgar et al. 2024. Unveiling psychotic disorder patterns: A deep learning model analysing motor activity time-series data with explainable AI. *Biomedical Signal Processing and Control* 91 (2024), 106000.
- [57] Nicola Montano et al. 2009. Heart rate variability explored in the frequency domain: a tool to investigate the link between heart and behavior. *Neuroscience & Biobehavioral Reviews* 33, 2 (2009), 71–80.
- [58] RJ Neuwirth. 2022. The EU Artificial Intelligence Act. *The EU Artificial Intelligence Act* 106 (2022).
- [59] Hung Nguyen et al. 2023. Towards trust of explainable ai in thyroid nodule diagnosis. In *International Workshop on Health Intelligence*. Springer, 11–26.
- [60] Hung Nguyen et al. 2024. LangXAI: Integrating Large Vision Models for Generating Textual Explanations to Enhance Explainability in Visual Perception Tasks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. 8754–8758.
- [61] Hung Nguyen et al. 2024. XEdgeAI: A human-centered industrial inspection framework with data-centric Explainable Edge AI approach. *Information Fusion* (2024), 102782.
- [62] Hung Nguyen et al. 2025. Human-centered Explainable Psychiatric Disorder Diagnosis System using Wearable ECG Monitors. In *The 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2025)*.
- [63] Loc Phuc Truong Nguyen, Hung Thanh Do, Hung Truong Thanh Nguyen, and Hung Cao. 2025. Motion2Meaning: A Clinician-Centered Framework for Contestable LLM in Parkinson’s Disease Gait Interpretation. In *Proceedings of 9th International Symposium on Chatbots and Human-centred AI (CONVERSATIONS) 2025*.



- [64] Government of Canada. 2019. Directive on Automated Decision-Making- Canada.ca. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>
- [65] Tom O'Malley et al. 2019. KerasTuner. <https://github.com/keras-team/keras-tuner>.
- [66] Montréal Declaration on Responsible AI. 2023. Montréal Declaration on Responsible AI. <https://montrealdeclaration-responsibleai.com/>
- [67] Michael J Owen et al. 2023. Genomic findings in schizophrenia and their implications. *Molecular psychiatry* 28, 9 (2023), 3638–3647.
- [68] Cristiano Patrício et al. 2025. CBVLM: Training-free Explainable Concept-based Large Vision Language Models for Medical Image Classification. *arXiv preprint arXiv:2501.12266* (2025).
- [69] Carsten Bøcker Pedersen et al. 2014. A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA psychiatry* 71, 5 (2014), 573–581.
- [70] Thomas Ploug and Søren Holm. 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial intelligence in medicine* 107 (2020), 101901.
- [71] Medicines & Healthcare products Regulatory Agency. 2024. Transparency for machine learning-enabled medical devices: guiding principles. <https://www.gov.uk/government/publications/machine-learning-medical-devices-transparency-principles/transparency-for-machine-learning-enabled-medical-devices-guiding-principles>
- [72] Kavita Rawat and Trapti Sharma. 2025. PsyneuroNet architecture for multi-class prediction of neurological disorders. *Biomedical Signal Processing and Control* 100 (2025), 107080.
- [73] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* 679 (2016), 2016.
- [74] Marco Tulio Ribeiro et al. 2016. “Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [75] Ramprasaath R Selvaraju et al. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision* 128 (2020), 336–359.
- [76] Andrea Sgoifo et al. 2009. The inevitable link between heart and behavior. New insights from biomedical research and implications for clinical practice. (2009).
- [77] Alessandro Silvani et al. 2016. Brain–heart interactions: physiology and clinical implications. *Philosophical transactions of the royal society A: Mathematical, physical and engineering sciences* 374, 2067 (2016), 20150181.
- [78] Karan Singhal et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.
- [79] Fabrizio Smeraldi. 2025. *BleakHeart*. [doi:10.5281/zenodo.10992864](https://doi.org/10.5281/zenodo.10992864)
- [80] Alessandro Taddei et al. 1992. The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. *European heart journal* 13, 9 (1992), 1164–1172.
- [81] Burak Tasci et al. 2024. A novel ternary pattern-based automatic psychiatric disorders classification using ECG signals. *Cognitive Neurodynamics* 18, 1 (2024), 95–108.
- [82] BURAK TAŞCI. 2024. *Psychiatry-ECG*. <https://www.kaggle.com/datasets/buraktaci/psychiatry-ecg>
- [83] Gemma Team et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).
- [84] Hardik Telangore et al. 2025. A novel ECG-based approach for classifying psychiatric disorders: Leveraging wavelet scattering networks. *Medical Engineering & Physics* 135 (2025), 104275.
- [85] Erhan Tiriyaki et al. 2021. Real-time CNN based ST depression episode detection using single-lead ECG. In *2021 22nd International Symposium on Quality Electronic Design (ISQED)*. IEEE, 566–570.
- [86] Naoko Tomitani et al. 2021. Regional differences in office and self-measured home heart rates in Asian hypertensive patients: AsiaBP@ Home study. 23, 3 (2021), 606–613.
- [87] Gaetano Valenza et al. 2016. Predicting mood changes in bipolar disorder through heartbeat nonlinear dynamics. *IEEE journal of biomedical and health informatics* 20, 4 (2016), 1034–1043.
- [88] Chang Wang et al. 2024. Schizophrenia classification and abnormalities reveal of brain region functional connection by deep-learning multiple sparsely connected network. *Biomedical Signal Processing and Control* 96 (2024), 106580.
- [89] Danding Wang et al. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [90] Zuxing Wang et al. 2025. Heart rate variability in mental disorders: an umbrella review of meta-analyses. *Translational Psychiatry* 15, 1 (2025), 104.
- [91] I WHO. 2007. International classification of diseases (ICD).
- [92] Xiaohan Zang et al. 2022. End-to-end depression recognition based on a one-dimensional convolution neural network model using two-lead ECG signal. *Journal of Medical and Biological Engineering* 42, 2 (2022), 225–233.
- [93] Tian Hong Zhang et al. 2022. Imbalance model of heart rate variability and pulse wave velocity in psychotic and nonpsychotic disorders. *Schizophrenia Bulletin* 48, 1 (2022), 154–165.
- [94] Wencan Zhang and Brian Y Lim. 2022. Towards relatable explainable AI with the perceptual process. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–24.

### A Cardiac Monitoring Interface (CMI) Workflow

Fig. 14 illustrates the detailed flow of a recording session, showcasing our modular architecture that separates data acquisition, state management, and user interface (UI) components.

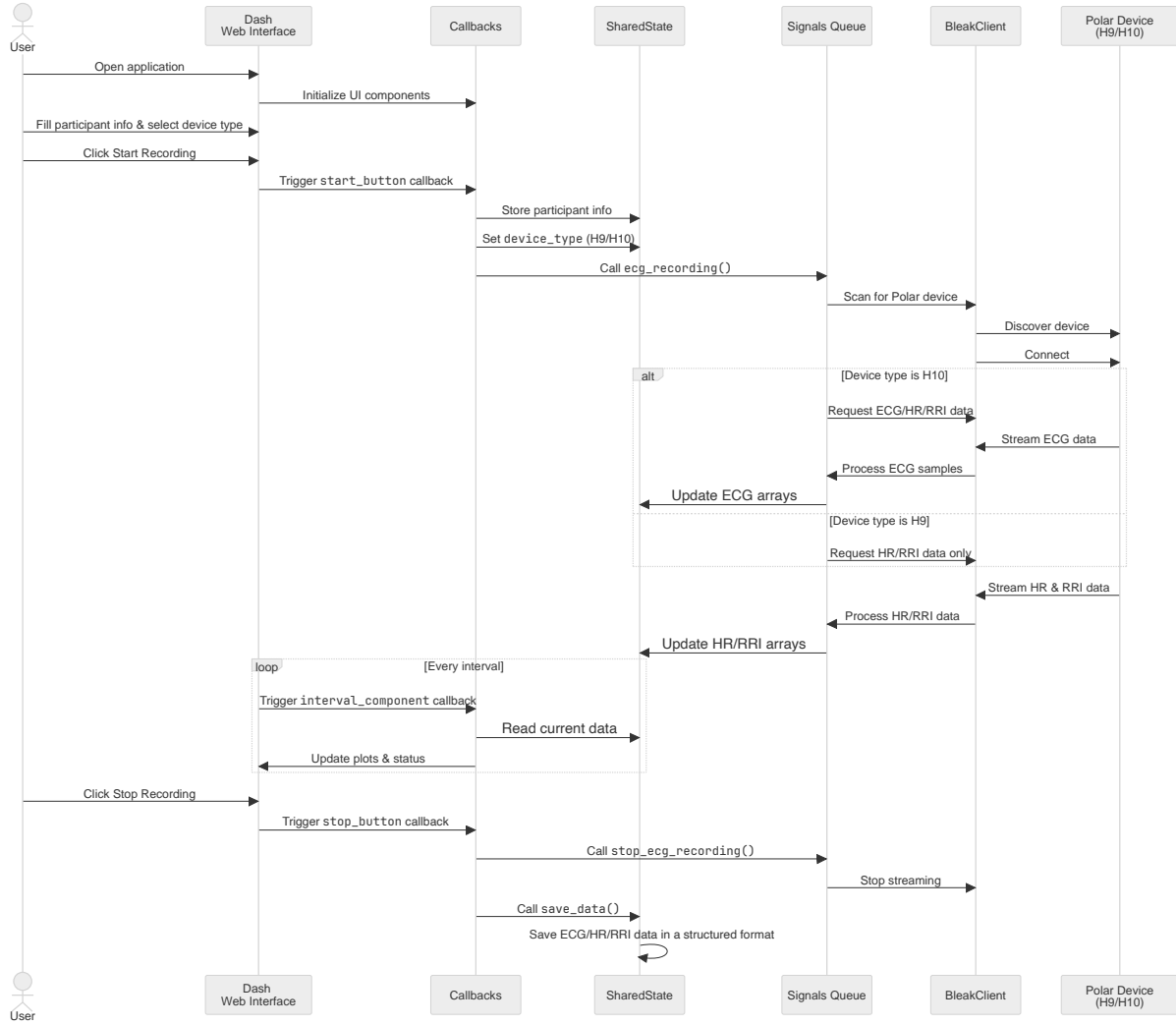


Fig. 14. Sequence diagram illustrating the cardiac signal recording workflow in the CMI. The diagram shows the interaction flow between user actions, web interface components, and Polar device communication, including device-specific branching for H9 (HR/RRR only) and H10 (ECG/HR/RRR) configurations.

## B Model Architecture and Hyperparameters

This appendix provides detailed architectural specifications and optimized hyperparameters for all evaluated models, including the Misgar et al. baseline [56] (Table 6), the proposed MSTFT model (Table 7), its temporal-only variant MTC (Table 8), frequency-only variant WT (Table 9), and the direct-fusion variant MTC+WT+Concat (Table 10), to support full transparency and reproducibility.

Table 6. Misgar et al. [56] Model Architecture with Optimized Hyperparameters.

Layer	Shape	#Parameters	Hyperparameters
Input (sequence)	$(T, 1)$	0	$T$ = sequence length
Lambda (x branch)	$(T, 1)$	0	stride = 1
Lambda (y branch)	$(T, 1)$	0	stride = 1
<i>x branch</i>			
Conv1D	$(T, 512)$	7,168	filters = 512, padding=same
Conv1D	$(T, 512)$	3,932,672	filters = 512, padding=same
MaxPooling1D	$(T/2, 512)$	–	pool = 2
MaxPooling1D	$(T/4, 512)$	–	pool = 2
<i>y branch</i>			
Conv1D	$(T, 256)$	3,840	filters = 256, padding=same
Conv1D	$(T, 256)$	1,048,832	filters = 256, padding=same
MaxPooling1D	$(T/2, 256)$	–	pool = 2
MaxPooling1D	$(T/4, 256)$	–	pool = 2
Concatenate (x, y)	$(T/4, 768)$	–	–
Conv1D	$(T/4, 128)$	1,671,296	filters = 128, padding=same
InstanceNormalization	$(T/4, 128)$	256	$\gamma, \beta$ per channel
PReLU	$(T/4, 128)$	128	channel-wise $\alpha$
Dropout	$(T/4, 128)$	–	rate = 0.5
MaxPooling1D	$(T/8, 128)$	–	pool = 2
MultiHeadAttention	$(T/8, 128)$	263,808	heads = 4, key_dim = 128
Add (z, z)	$(T/8, 128)$	–	–
LayerNormalization	$(T/8, 128)$	256	–
Conv1D	$(T/8, 512)$	1,180,160	filters = 512, padding=same
InstanceNormalization	$(T/8, 512)$	1,024	$\gamma, \beta$ per channel
PReLU	$(T/8, 512)$	512	channel-wise $\alpha$
Dropout	$(T/8, 512)$	–	rate = 0.6
MaxPooling1D	$(T/16, 512)$	–	pool = 2
Dense (ReLU)	$(T/16, 32)$	16,416	–
Dense (ReLU)	$(T/16, 16)$	528	–
Dense (ReLU)	$(T/16, 8)$	136	–
Dense (ReLU)	$(T/16, 4)$	36	–
InstanceNormalization	$(T/16, 4)$	8	$\gamma, \beta$ per channel
Flatten	$(T/4)$	–	–
Dense (Sigmoid)	(1)	$T/4 + 1$	Binary output
<b>Total Trainable Parameters</b>		<b>8,130,126</b>	

Table 7. MSTFT Model Architecture with Optimized Hyperparameters.

Layer	Shape	#Parameters	Hyperparameters
Input (sequence)	$(T, 1)$	0	Sequence length = $T$
GaussianNoise	$(T, 1)$	–	$\lambda^2 = 0.1$
Conv1D (positional projection)	$(T, 64)$	64	$d = 64, \tau = 10^4$
<i>Multi-Scale Temporal Convolution Branch (<math>x_{\text{temporal}}</math>)</i>			
SpectralNorm Conv1D (Block 1)	$(T, 1024)$	$\approx 3.1 \times 10^5$	$n_t = 2, \text{filters}=1024, \text{kernel}=3, L_2 = 10^{-4}$
GroupNorm + GELU + StochasticSkip	$(T, 1024)$	–	$p_s = 0.8$
AvgPooling + SpatialDropout(0.2)	$(T/2, 1024)$	–	dropout=0.2
SpectralNorm Conv1D (Block 2)	$(T/2, 512)$	$\approx 1.6 \times 10^5$	filters=512
GroupNorm + GELU + StochasticSkip	$(T/2, 512)$	–	–
AvgPooling + SpatialDropout(0.2)	$(T/4, 512)$	–	–
<i>Learnable Wavelet Frequency Branch (<math>x_{\text{freq}}</math>)</i>			
Conv1D (initial)	$(T, 64)$	704	filters=64, kernel=11
SeparableConv1D (Block 1)	$(T, 512)$	$\approx 1.3 \times 10^5$	$n_f = 2, \text{filters}=512, \text{kernel}=5, \text{depth\_mult}=2$
GroupNorm + GELU + AdaptiveAvgPool	$(T/4, 512)$	–	adaptive pooling
SeparableConv1D (Block 2)	$(T/4, 1024)$	$\approx 5.2 \times 10^5$	filters=1024
GroupNorm + GELU	$(T/4, 1024)$	–	–
<i>Cross-Attention Fusion</i>			
Dense projection ( $x_{\text{temporal}} / x_{\text{freq}}$ )	$(T/4, 1024)$	$\approx 1.05 \times 10^6$	$d_p = 1024$
MultiHeadAttention	$(T/4, 512)$	$8.4 \times 10^4$	$d_k = 512, h = 16, \text{attention dropout}=0.1$
Concatenate + LayerNorm	$(T/4, 3072)$	–	–
<i>Transformer Encoder Block</i>			
MultiHeadAttention + Gating	$(T/4, 3072)$	$\approx 2.6 \times 10^5$	$h = 16, d_k = 512, \text{gating sigmoid}$
Dense (FFN, GELU) + Dropout(0.2) + Dense	$(T/4, 3072)$	$\approx 3.8 \times 10^5$	$f_{\text{drop}} = 0.2$
LayerNormalization	$(T/4, 3072)$	6,144	–
<i>Classifier Head</i>			
GlobalAvg + GlobalMaxPooling	(6144)	–	–
BatchNormalization	(6144)	12,288	–
Dense (GELU)	(512)	3,145,728	$d_p/2 = 512$
Dropout (0.4) + Gated Attention	(512)	–	dropout=0.4
Dense (Residual) + Add	(512)	262,656	–
GroupNormalization(8)	(512)	1,024	–
Output Dense (Sigmoid)	(1)	513	Binary output
<b>Total trainable parameters</b>		<b>6,323,121</b>	

Table 8. MTC (MSTFT with Temporal-only Path) Model Architecture with Optimized Hyperparameters.

Layer	Shape	#Parameters	Hyperparameters
Input (sequence)	$(T, 1)$	0	$T$ = sequence length
GaussianNoise	$(T, 1)$	–	std = 0.1
Conv1D (positional projection)	$(T, 64)$	128	$d = 64, \tau = 10^4$
<i>Temporal Branch (only), <math>n_t = 2</math></i>			
SpectralNorm Conv1D (Block 1)	$(T, 512)$	98,816	filters=512, kernel=3, dilation= $2^0$ , causal, $L_2 = 10^{-4}$
GroupNormalization(8) + GELU	$(T, 512)$	1,024	$\gamma, \beta$
Residual 1×1 (64→512)	$(T, 512)$	33,280	with SpectralNorm
StochasticSkip + SpatialDropout(0.2)	$(T, 512)$	–	$p_s = 0.8$ , dropout=0.2
SpectralNorm Conv1D (Block 2)	$(T, 256)$	393,472	filters=256, kernel=3, dilation= $2^1$ , causal
GroupNormalization(8) + GELU	$(T, 256)$	512	$\gamma, \beta$
Residual 1×1 (512→256)	$(T, 256)$	131,328	with SpectralNorm
<i>Projection + Transformer Encoder</i>			
Dense (projection)	$(T, 672)$	172,704	$d_p = 672$
LayerNormalization	$(T, 672)$	1,344	–
MultiHeadAttention	$(T, 672)$	861,792	$h = 10$ , total $d_k = 320$ (per-head 32)
Dense (gate)	$(T, 1)$	673	sigmoid gate
Trainable scale (residual)	$(T, 672)$	1	scalar
FFN: Dropout(0.2) + Dense	$(T, 672)$	1,808,352	$2\times$ expansion
LayerNormalization	$(T, 672)$	1,344	–
<i>Classifier Head</i>			
GlobalAvg + GlobalMaxPooling	(1344)	–	–
BatchNormalization	(1344)	2,688	–
Dense (GELU)	(336)	451,920	$1344 \rightarrow 336$
Dropout (0.4) + Gated Attention (Dense 336→336)	(336)	113,232	dropout=0.4
Dense (Residual, 1344→336) + Add	(336)	451,920	–
GroupNormalization(8)	(336)	672	–
Output Dense (Sigmoid)	(1)	337	Binary output
<b>Total trainable parameters</b>		<b>4,525,539</b>	



Table 9. WT (MSTFT with Frequency-only Path) Model Architecture with Optimized Hyperparameters.

Layer	Shape	#Parameters	Hyperparameters
Input (sequence)	$(T, 1)$	0	$T$ = sequence length
GaussianNoise	$(T, 1)$	–	std = 0.1
Conv1D (positional projection)	$(T, 64)$	128	$d = 64, \tau = 10^4$
<i>Frequency Branch (only)</i>			
Conv1D (initial)	$(T, 64)$	45,120	filters=64, kernel=11, padding=same
SeparableConv1D (Block 1)	$(T, 256)$	16,960	filters=256, kernel=5, depth_mult=1
GroupNormalization(8) + GELU	$(T, 256)$	512	$\gamma, \beta$
SeparableConv1D (Block 2)	$(T, 512)$	132,864	filters=512, kernel=5, depth_mult=1
GroupNormalization(8) + GELU	$(T, 512)$	1,024	$\gamma, \beta$
<i>Projection + Transformer Encoder</i>			
Dense (projection)	$(T, 736)$	377,568	$d_p = 736$
LayerNormalization	$(T, 736)$	1,472	–
MultiHeadAttention	$(T, 736)$	1,038,080	$h = 11$ , total $d_k = 352$ (per-head 32)
Dense (gate)	$(T, 1)$	737	sigmoid gate
Trainable scale (residual)	$(T, 736)$	1	scalar
FFN: Dense(736→1472) + Dropout(0.2) + Dense(1472→736)	$(T, 736)$	2,168,992	2× expansion
LayerNormalization	$(T, 736)$	1,472	–
<i>Classifier Head</i>			
GlobalAvg + GlobalMaxPooling	(1472)	–	–
BatchNormalization	(1472)	2,944	–
Dense (GELU)	(336)	494,928	$1472 \rightarrow 336$
Dropout (0.4) + Gated Attention (Dense 336→336)	(336)	113,232	dropout=0.4
Dense (Residual, 1472→336) + Add	(336)	494,928	–
GroupNormalization(8)	(336)	672	–
Output Dense (Sigmoid)	(1)	337	Binary output
<b>Total trainable parameters</b>		<b>4,891,971</b>	

Table 10. MTC+WT+Concat (MSTFT Temporal and Frequency with Direct Concatenation) Model Architecture with Optimized Hyperparameters.

Layer	Shape	#Parameters	Hyperparameters
Input (sequence)	$(T, 1)$	0	$T$ = sequence length
GaussianNoise	$(T, 1)$	–	std = 0.1
Conv1D (positional projection)	$(T, 64)$	128	$d = 64, \tau = 10^4$
<i>Temporal Branch (only), <math>n_t = 2</math></i>			
SpectralNorm Conv1D (Block 1)	$(T, 512)$	98,816	filters=512, kernel=3, causal, $L_2 = 10^{-4}$
GroupNormalization(8) + GELU	$(T, 512)$	1,024	$\gamma, \beta$
Residual $1 \times 1$ ( $64 \rightarrow 512$ )	$(T, 512)$	33,280	
StochasticSkip + SpatialDropout(0.2)	$(T, 512)$	–	$p_s = 0.8$ , dropout=0.2
SpectralNorm Conv1D (Block 2)	$(T, 256)$	393,472	filters=256, kernel=3, causal
GroupNormalization(8) + GELU	$(T, 256)$	512	$\gamma, \beta$
Residual $1 \times 1$ ( $512 \rightarrow 256$ )	$(T, 256)$	131,328	
<i>Frequency Branch (only), <math>n_f = 2</math></i>			
Conv1D (initial)	$(T, 64)$	45,120	filters=64, kernel=11, padding=same
SeparableConv1D (Block 1)	$(T, 256)$	16,960	filters=256, kernel=5, depth_mult=1
GroupNormalization(8) + GELU	$(T, 256)$	512	$\gamma, \beta$
SeparableConv1D (Block 2)	$(T, 512)$	132,864	filters=512, kernel=5, depth_mult=1
GroupNormalization(8) + GELU	$(T, 512)$	1,024	$\gamma, \beta$
AdaptiveAvgPool (align length)	$(T, 512)$	–	
<i>Concatenation + Projection</i>			
Dense projection (temporal)	$(T, 640)$	164,480	$256 \rightarrow 640$
Dense projection (frequency)	$(T, 640)$	328,320	$512 \rightarrow 640$
Concatenate	$(T, 1280)$	–	–
Dense ( $1280 \rightarrow 640$ )	$(T, 640)$	819,840	project to target width
LayerNormalization	$(T, 640)$	1,280	–
<i>Transformer Encoder</i>			
MultiHeadAttention	$(T, 640)$	820,800	$h = 10$ , total $d_k = 320$ (per-head 32)
Dense (gate)	$(T, 1)$	641	sigmoid gate
Trainable scale (residual)	$(T, 640)$	1	scalar
FFN: Dropout(0.2) + Dense( $1280 \rightarrow 640$ )	$(T, 640)$	1,640,320	$2 \times$ expansion
LayerNormalization	$(T, 640)$	1,280	–
<i>Classifier Head</i>			
GlobalAvg + GlobalMaxPooling	(1280)	–	–
BatchNormalization	(1280)	2,560	–
Dense (GELU)	(320)	409,920	$1280 \rightarrow 320$
Dropout (0.4) + Gated Attention (Dense $320 \rightarrow 320$ )	(320)	102,720	dropout=0.4
Dense (Residual, $1280 \rightarrow 320$ ) + Add	(320)	409,920	–
GroupNormalization(8)	(320)	640	–
Output Dense (Sigmoid)	(1)	321	Binary output
<b>Total trainable parameters</b>		<b>5,558,083</b>	

## C Dataset

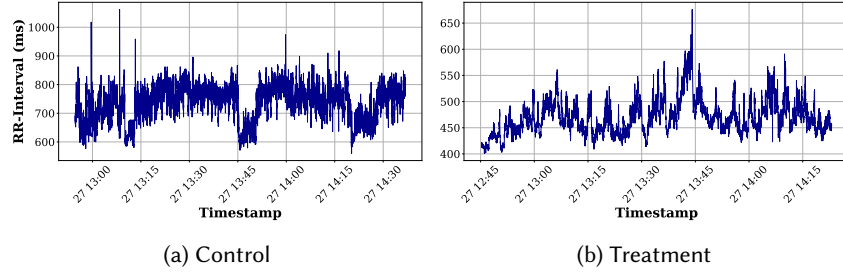


Fig. 15. Data samples of (a) “control” and (b) “treatment” (schizophrenia/bipolar disorder) groups in the HRV-ACC dataset.

## D Pseudocode of Attention-based and Gradient-based Explanation Methods

In this section, we present the pseudocode for the attention-based explanation method (Algorithm 1) and the gradient-based explanation method (Algorithm 2).

---

### Algorithm 1: Attention-based Explanation

---

**Input:** Model  $M$ , Input  $\mathbf{X} \in \mathbb{R}^{T \times 1}$ , Target layers  $\mathcal{L}$   
**Output:** Expanded attention map  $\mathbf{E}_{\text{attn}}^T$

```

// Extract multi-head attention weights
1 for each layer  $l \in \mathcal{L}$  do
2   | Extract attention weights  $\mathbf{A}^{(l)}$ ;
3   | // Average across attention heads
4   |  $\mathbf{A}^{(l)} \leftarrow \frac{1}{H} \sum_{h=1}^H \mathbf{A}_h^{(l)}$ ;
5 end
// Combine attention maps across layers
6  $\mathbf{E}_{\text{attn}} \leftarrow \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathbf{A}^{(l)}$ ;
// Expand to original signal length
7  $\mathbf{E}_{\text{attn}}^T \leftarrow \text{Expand}(\mathbf{E}_{\text{attn}}, T)$ ;
// Normalize using z-score then min-max
8  $\mathbf{E}_{\text{attn}}^T \leftarrow \frac{\mathbf{E}_{\text{attn}}^T - \mu(\mathbf{E}_{\text{attn}}^T)}{\sigma(\mathbf{E}_{\text{attn}}^T)}$ ;
9  $\mathbf{E}_{\text{attn}}^T \leftarrow \frac{\mathbf{E}_{\text{attn}}^T - \min(\mathbf{E}_{\text{attn}}^T)}{\max(\mathbf{E}_{\text{attn}}^T) - \min(\mathbf{E}_{\text{attn}}^T)}$ ;
10 return  $\mathbf{E}_{\text{attn}}^T$ 

```

---



---

### Algorithm 2: Gradient-based Explanation

---

**Input:** Model  $M$ , Input  $\mathbf{X} \in \mathbb{R}^{T \times 1}$ , Target layers  $\mathcal{L}$ , Predicted class  $c$   
**Output:** Expanded gradient map  $\mathbf{E}_{\text{grad}}^T$

```

// Process each layer
1 for each layer  $l \in \mathcal{L}$  do
2   | Forward pass:  $\mathbf{F}^{(l)} \leftarrow M_l(\mathbf{X})$ ;
3   | Compute prediction:  $y^c \leftarrow M(\mathbf{X})_c$ ;
4   | // Compute importance weights
5   | for  $k \leftarrow 1$  to  $K$  (channels) do
6   | |  $\alpha_k^{(l)} \leftarrow \frac{1}{Z} \sum_{i=1}^Z \frac{\partial y^c}{\partial \mathbf{F}_{i,k}^{(l)}}$ ;
7   | end
8   | // Weighted combination with ReLU
9   |  $\mathbf{L}^{(l)} \leftarrow \text{ReLU}(\sum_k \alpha_k^{(l)} \cdot \mathbf{F}_k^{(l)})$ ;
10 end
// Average gradient maps across layers
11  $\mathbf{E}_{\text{grad}} \leftarrow \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathbf{L}^{(l)}$ ;
// Expand and normalize
12  $\mathbf{E}_{\text{grad}}^T \leftarrow \text{Expand}(\mathbf{E}_{\text{grad}}, T)$ ;
13  $\mathbf{E}_{\text{grad}}^T \leftarrow \frac{\mathbf{E}_{\text{grad}}^T - \min(\mathbf{E}_{\text{grad}}^T)}{\max(\mathbf{E}_{\text{grad}}^T) - \min(\mathbf{E}_{\text{grad}}^T)}$ ;
14 return  $\mathbf{E}_{\text{grad}}^T$ 

```

---

## E Contestable LLM Prompt Template

This appendix presents the structured prompt template used by the contestable LLM system, outlining the system message configuration.

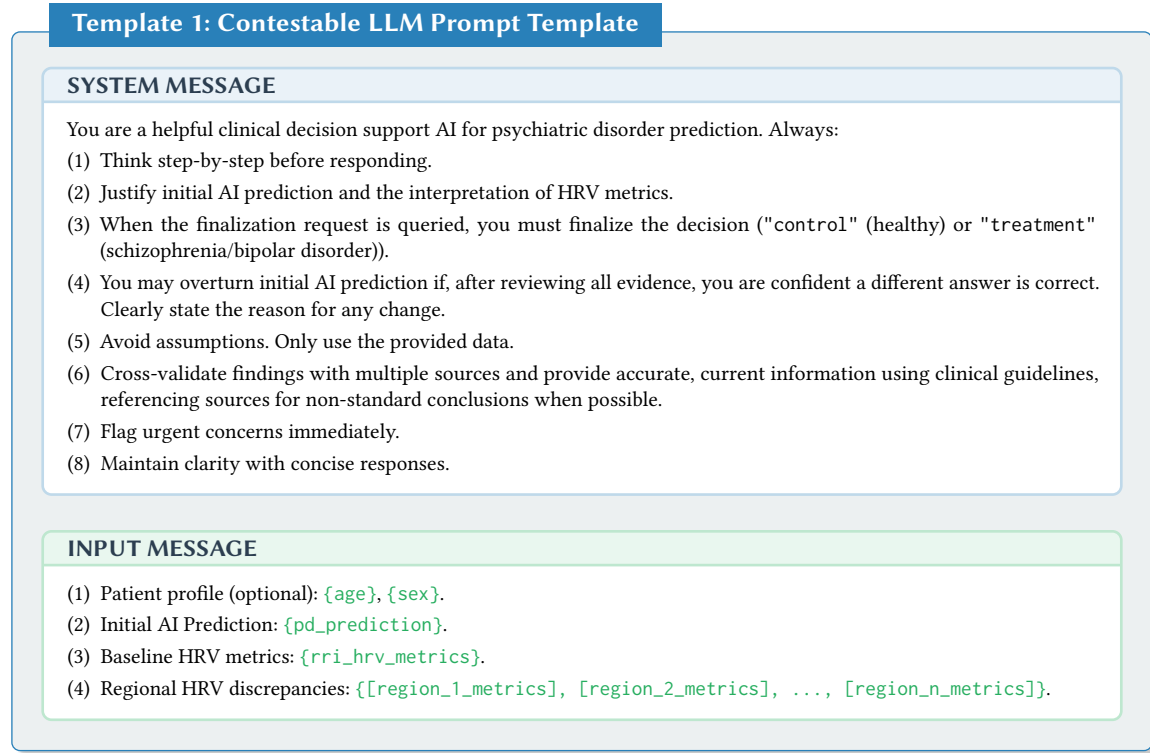


Fig. 16. Contestable LLM Prompt Template structure showing system instructions and required input format.

## F Transferability

Our proposed CAI framework is demonstrated through the Heart2Mind system for psychiatric disorder prediction, which is composed of three components: a base model for initial prediction, a built-in safeguard that verifies prediction reliability, and a contestable LLM that facilitates human review and challenge. As each component is designed to be modality and domain-agnostic, rather than tied specifically to ECG or psychiatric diagnosis, the same design can be applied to different data modalities, prediction targets, and interaction workflows while preserving the core contestability logic.

Here, we describe how Heart2Mind is transferred to another system, Motion2Meaning, for Parkinson's disease gait prediction. In Motion2Meaning (its architecture shown in Fig. 17), the base model is a one-dimensional convolutional neural network (1D-CNN) to perform multiclass classification of Hoehn and Yahr severity stages from wearable vertical ground reaction force (vGRF) time series. The built-in safeguard, SAE, compares class activation mapping (CAM)-based and backpropagation-based maps and flags samples with high discrepancies. As demonstrated in Fig. 18, the incorrect prediction case shows about a fivefold higher discrepancy rate than

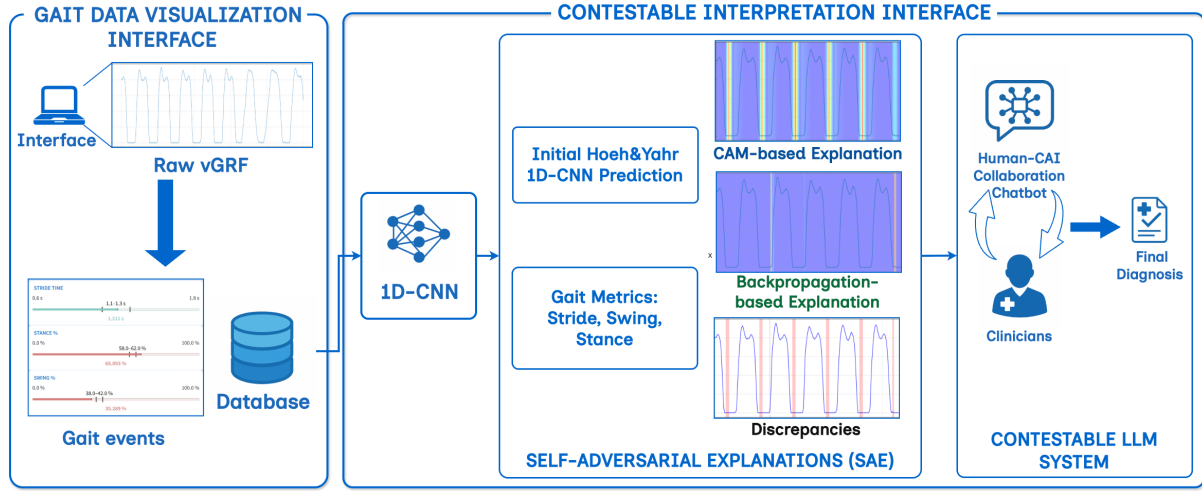


Fig. 17. Architecture of Motion2Meaning: Contestable Parkinson's Disease Gait Prediction System [63].

the correct one, indicating that attributional inconsistency is a stable signal of prediction unreliability. The human-CAI collaboration chatbot is implemented as a contestable LLM interaction layer that uses discrepancy regions, gait metrics, and the initial prediction to drive a structured collaboration workflow in which LLMs validate the most accurate predictions and overturn a subset of incorrect ones.

This reflects the transferability of our proposed CAI framework. The base model can be any supervised model that outputs a clear clinical prediction, including binary or multiclass classifiers, multilabel predictors,

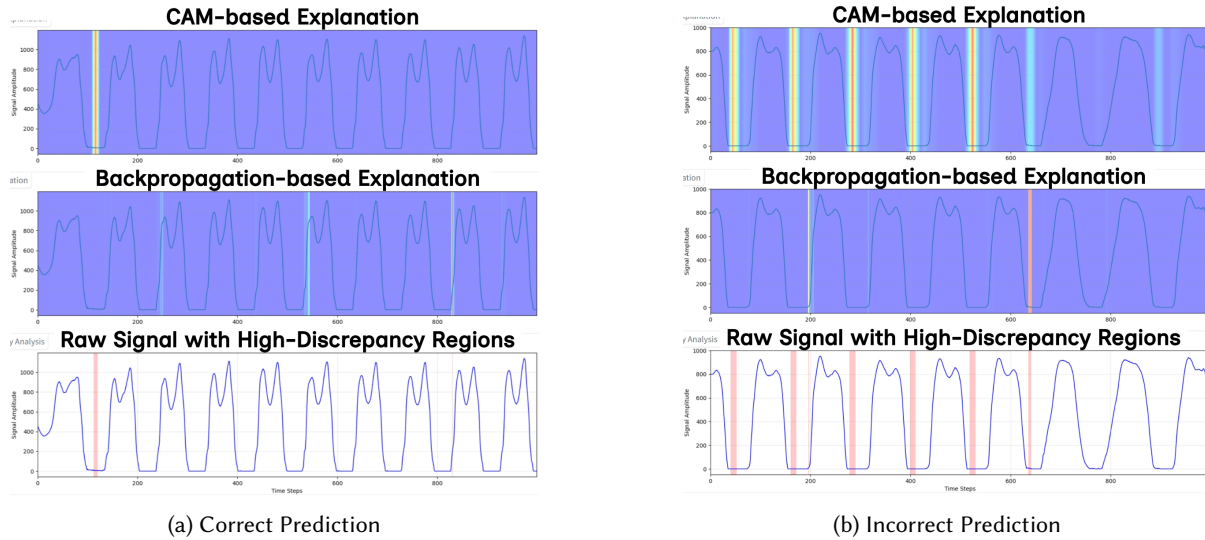


Fig. 18. Visualization for (a) correct and (b) incorrect prediction cases.





sequence models, or regression models. The built-in safeguard can use different XAI and reliability techniques, such as computational argumentation, uncertainty scores, out-of-distribution detectors, and domain-specific rule checks. The human-CAI collaboration chatbot can be implemented with different LLM families or with hybrid rule-based and retrieval-augmented models, configured to prioritize readability, factual grounding, or conservative behavior according to clinical needs. The contestation path remains the same, where the system presents a clinically meaningful prediction with its supporting evidence, automatically highlights cases that warrant attention, supports structured human challenge, and records final decisions.

## G Human-CAI Collaboration Dialogues


This appendix presents notable dialogues from human-CAI interaction sessions, highlighting how participants reviewed initial model decisions, SAE discrepancies, and collaborated with the contestable LLM to finalize decisions.

### G.1 True Positive Case Dialogues (*treatment\_40*):

 **[+2] AI Contestation Success**

 **AI Expert #6:**

Are you sure that the low vagal tone indicators (RMSSD, pNN50) can affect the result?


 **Contestable LLM (gemma-3):**

Yes, I am certain that the low vagal tone indicators (RMSSD and pNN50) significantly affect the result and strongly support the “Treatment” decision.

Here is the justification based on clinical evidence:

- (1) **Vagal Tone as a Biomarker:** RMSSD (Root Mean Square of Successive Differences) and pNN50 (Percentage of successive differences greater than 50 ms) are the primary time-domain metrics reflecting parasympathetic nervous system (PNS) activity, often referred to as vagal tone. The PNS is crucial for regulating emotional responses, stress recovery, and overall autonomic flexibility.
- (2) **Clinical Significance in Psychiatry:** Reduced vagal tone (low RMSSD and pNN50) is one of the most consistent findings in major psychiatric disorders, including Major Depressive Disorder, Anxiety Disorders, Schizophrenia, and Bipolar Disorder. This reduction indicates autonomic dysfunction, characterized by a shift toward sympathetic dominance or overall reduced cardiac control.
- (3) **Interpretation of the Data:**
  - **pNN50 (2.3208%):** This value is critically low. A healthy pNN50 typically ranges much higher. A value this low strongly suggests severely impaired vagal modulation and significant autonomic dysregulation.
  - **RMSSD (58.23):** While not the lowest possible value, when combined with the critically low pNN50 and the fast heart rate (Mean RR 568.34 ms, approximately 105 bpm), it reinforces the conclusion of reduced parasympathetic influence.
- (4) **Conclusion:** The profound reduction in pNN50, specifically, is a robust indicator of the autonomic dysregulation frequently observed in individuals requiring psychiatric intervention (“Treatment”). Therefore, these metrics are central to the final assessment.

✓ [0] Neutral Contestation

 Cognitive Psychology Expert #5:

Are these patterns specific to psychiatric disorders?

 Contestable LLM (gemma-3):

No, these patterns are **not specific** to psychiatric disorders.

The observed HRV profile, characterized by low overall variability (SDNN), severely reduced vagal tone (pNN50), and a high resting heart rate (low Mean RR), is indicative of **general autonomic nervous system (ANS) dysfunction** and chronic sympathetic dominance.

**Conditions that also cause these HRV patterns**

Autonomic dysregulation is a common feature across numerous medical conditions. The patterns observed here can also be strongly associated with:

- (1) **Cardiovascular Disease (CVD)** Low HRV is a powerful, independent predictor of mortality and morbidity, particularly following myocardial infarction (MI) or in patients with heart failure.
- (2) **Metabolic Disorders** Diabetic autonomic neuropathy causes significant reduction in all HRV metrics, especially vagal indicators (RMSSD, pNN50).
- (3) **Inflammatory Conditions** Chronic systemic inflammation (for example autoimmune disorders) is linked to reduced HRV.
- (4) **Endocrine Issues** Conditions such as hyperthyroidism can cause persistent tachycardia and reduced overall variability.
- (5) **Sleep Disorders** Severe sleep apnea can lead to significant regional instability and overall reduced HRV.

**Role of HRV in Psychiatric Diagnosis**

In psychiatry, low HRV is considered a **non specific biomarker of risk and severity**, reflecting:

- **Impaired Stress Response** The inability of the ANS to flexibly adapt to internal or external demands.
- **Allostatic Load** The cumulative wear and tear on the body due to chronic stress.

Therefore, while these HRV findings strongly support the presence of autonomic dysregulation that is frequently seen in psychiatric illness, they must be interpreted within a comprehensive clinical context that rules out primary cardiovascular, metabolic, or other systemic medical causes.

## G.2 False Negative Case Dialogues (*treatment\_1*):

✗ [-1] Low Contestation Effectiveness

 AI Expert #2:

Are you sure the fluctuation is enough for this decision?

 Contestable LLM (gemma-3):

Yes, I am confident in the decision to flag this case as **Treatment** based on the observed autonomic instability.

Here is the justification:

- (1) **Significant Regional Reduction in HRV:** While the overall average HRV metrics are borderline, the regional analysis reveals periods of clinically significant reduction in autonomic function. Specifically:
  - **Region 7:** SDNN is 34.99 ms, which is well below the threshold often associated with poor health outcomes and severe autonomic dysregulation ( $SDNN \leq 50$  ms is often considered low).

- **Region 1:** RMSSD is 16.01 ms, indicating a marked reduction in beat to beat variability (parasympathetic tone) during this extended period.
- (2) **Autonomic Instability as a Marker of Pathology:** The rapid and significant fluctuation between regions (for example, SDNN 34.99 ms in R7 versus 58.24 ms in R2) is a key indicator of autonomic dysregulation. Healthy individuals typically maintain a more stable, high level of HRV. This instability, meaning the inability of the system to maintain robust homeostatic control, is frequently observed across various psychiatric disorders (for example, anxiety, depression, schizophrenia) and warrants clinical investigation.
- (3) **Masking Effect of Averaging:** Relying solely on the overall average (SDNN 73.07 ms) masks the critical periods of low parasympathetic activity. The regional data provides a more accurate, dynamic picture of a system under stress or lacking flexibility.

Therefore, the presence of sustained, significantly low HRV metrics in specific regions, coupled with high overall instability, strongly supports the classification of **Treatment**.

### ✓ [+3] Joint Contestation Success

#### Biomedical Engineering Expert #3:

Are you sure that this should be a treatment case? This goes against both the AI and my predictions. Please revise carefully.

#### Contestable LLM (gemma-3):

This is a critical point, and I appreciate the request for careful revision. While the overall SDNN (73.07 ms) might not appear critically low, a deeper clinical interpretation of the full dataset, especially the regional discrepancies, strongly supports the “Treatment” classification due to evidence of significant autonomic dysregulation.

Here is the step-by-step justification for maintaining the **Treatment** decision:

##### 1. Assessment of Overall Vagal Tone

- **RMSSD (23.42 ms):** This metric reflects short-term, beat-to-beat variability and is the primary indicator of parasympathetic (vagal) activity. A value of 23.42 ms is considered low for a healthy population, suggesting chronically reduced vagal modulation.
- **Clinical Relevance:** Reduced vagal tone (low RMSSD) is a highly consistent finding across numerous psychiatric disorders, including Major Depressive Disorder, Generalized Anxiety Disorder, and Schizophrenia, reflecting impaired emotional and physiological regulation.

##### 2. Assessment of Autonomic Instability (Regional Discrepancies)

The most compelling evidence for pathology lies in the significant fluctuations between the recorded regions, indicating poor autonomic flexibility and stability.

Region	RMSSD (Vagal Tone)	LF/HF Balance Implication	Clinical Interpretation
Overall	23.42 ms (Low)	HF > LF (Slight Vagal Bias)	Reduced overall capacity
Region 1	16.01 ms (Very Low)	HF 934.02 (High Vagal)	Low variability despite high power; potential instability
Region 4	19.82 ms (Low)	LF 731.97, HF 283.72	<b>Strong Sympathetic Dominance</b> (High LF/HF ratio)
Region 7	17.85 ms (Very Low)	HF 785.07, LF 278.32	Low variability, high vagal power

- **Pathological Finding:** The shift from periods of high sympathetic dominance (Region 4) to periods of high vagal power (Regions 1 and 7) with consistently low RMSSD values (indicating poor beat-to-beat variability) demonstrates **autonomic inflexibility and instability**.

- **Clinical Relevance:** Autonomic instability—the inability of the system to maintain a stable regulatory state—is a key indicator of allostatic load and is frequently observed in individuals requiring psychiatric intervention, as the system struggles to adapt to internal or external stressors.

### 3. Conclusion and Final Justification

The combination of **chronically reduced overall vagal tone (low RMSSD)** and **significant regional autonomic instability** provides strong clinical evidence of dysregulation. In the context of psychiatric diagnosis support, these findings warrant a recommendation for further clinical evaluation and potential intervention.

#### Final Decision: Treatment

**Justification:** The data reveals pathological autonomic dysregulation characterized by low overall RMSSD and marked instability across regions, which is highly correlated with underlying psychiatric conditions.

## ✓ [+1] Human Contestation Success

### Human-Computer Interaction (HCI) Expert #9:

It would be good to have colour coded and verbal explanation before the initial assessment step.

### Contestable LLM (gemma-3):

#### HRV Data Analysis and Interpretation

**1. Color-Coded HRV Metrics and Verbal Explanation** The following metrics reflect the overall autonomic nervous system (ANS) balance, where time-domain metrics (RMSSD, SDNN) primarily assess variability and parasympathetic tone, and frequency-domain metrics (LF, HF) assess sympathetic and parasympathetic contributions.

Metric	Value	Color Code	Verbal Explanation
Mean RR	597.43 ms	Red	Corresponds to a high average heart rate (about 100 BPM). Suggests potential resting tachycardia or high sympathetic drive.
RMSSD	23.42	Red	Root Mean Square of Successive Differences. A key indicator of short-term beat to beat variability (vagal tone). This value is low, indicating significantly reduced parasympathetic activity.
SDNN	73.07	Orange	Standard Deviation of NN intervals. Reflects overall variability from both sympathetic and parasympathetic influence. This value is borderline or low, suggesting reduced overall autonomic flexibility.
pNN50	1.1692%	Red	Percentage of successive differences greater than 50 ms. A highly specific measure of vagal tone. This value is very low, confirming poor parasympathetic function.
LF Power	358.97	Green	Low Frequency Power (0.04 to 0.15 Hz). Reflects a mix of sympathetic and vagal activity and is often associated with baroreflex function.
HF Power	507.87	Green	High Frequency Power (0.15 to 0.4 Hz). A pure measure of parasympathetic (vagal) activity. The absolute value is moderate.
LF/HF Ratio	0.71	Green	Ratio of sympathetic to parasympathetic influence. A ratio below 1.0 suggests relative parasympathetic dominance in the frequency domain despite poor time-domain metrics.

#### 2. Step by Step Thinking

- (1) **Analyze Time Domain Metrics (RMSSD, SDNN, pNN50):** The time-domain metrics (RMSSD 23.42 and pNN50 1.1692%) are significantly low. Low RMSSD and pNN50 are strong indicators of poor vagal tone and reduced parasympathetic activity, which is a common finding in various psychiatric disorders such as anxiety, depression, and schizophrenia.

- (2) **Analyze Mean RR:** The Mean RR (597.43 ms) corresponds to a high resting heart rate (about 100 BPM), suggesting chronic sympathetic activation or reduced vagal braking.
- (3) **Analyze Frequency Domain Metrics (LF, HF, LF/HF):** The LF/HF ratio (0.71) suggests relative parasympathetic dominance in the frequency domain, which conflicts with the severe reduction seen in the time domain (RMSSD, pNN50). This discrepancy suggests complex or unstable autonomic regulation.
- (4) **Evaluate Regional Discrepancies:** Review the regional data for stability. Significant fluctuations in Mean RR, SDNN, and power metrics across regions, for example SDNN ranging from 34.99 to 73.07, indicate autonomic instability, a key feature associated with affective disorders and schizophrenia.
- (5) **Cross Validate Findings:** Low overall HRV, especially RMSSD and pNN50, combined with high resting heart rate, is a robust biomarker for autonomic dysregulation that is frequently observed in clinical populations requiring psychiatric treatment.
- (6) **Formulate Initial Assessment:** Based on the severely reduced time-domain HRV metrics and elevated heart rate, the evidence strongly points toward significant autonomic dysregulation that is consistent with a potential psychiatric disorder (Treatment group).

### 3. Justification for Initial Assessment

The overall HRV profile demonstrates significant autonomic dysregulation characterized by:

- (1) **Reduced Vagal Tone:** RMSSD (23.42) and pNN50 (1.1692%) are critically low, indicating impaired parasympathetic function.
- (2) **Elevated Heart Rate:** Mean RR (597.43 ms) suggests resting tachycardia, often linked to chronic stress or sympathetic hyperactivity.
- (3) **Clinical Relevance:** Reduced HRV, particularly low RMSSD and SDNN, is a well established biomarker for autonomic dysfunction that is associated with major psychiatric conditions, including Major Depressive Disorder, Generalized Anxiety Disorder, and Schizophrenia.

**Initial Assessment:** The data strongly supports the presence of autonomic dysregulation consistent with a **Treatment** requirement.

### 4. Interpretation of Regional HRV Metrics

Region	Mean RR (ms)	RMSSD	SDNN	pNN50 (%)	LF Power	HF Power	Interpretation
Overall	597.43	23.42	73.07	1.17	358.97	507.87	Severely reduced vagal tone (RMSSD, pNN50).
Region 1	640.14	16.01	49.97	1.02	405.62	934.02	Extremely low SDNN and RMSSD. Severe vagal withdrawal.
Region 7	609.05	17.85	34.99	1.64	278.32	785.07	Critically low SDNN (34.99). Indicates a period of severe autonomic rigidity.

**Discrepancies Detected:** The regional analysis confirms significant autonomic instability. Periods of extremely low SDNN (Region 7 with 34.99) and RMSSD (Region 1 with 16.01) highlight moments of severe autonomic rigidity and vagal withdrawal, which are highly atypical for a healthy control subject.

### 5. Clinical Interpretation and Cross Validation

- **Schizophrenia and HRV:** Reduced HRV, with low SDNN and RMSSD, is a consistent finding in schizophrenia and reflects impaired autonomic regulation and increased cardiovascular risk.
- **Affective Disorders and HRV:** Low HRV is also strongly associated with depression and anxiety and often reflects chronic stress and allostatic load.



- **Cross Validation:** The combination of low time-domain metrics (RMSSD, pNN50) and high resting heart rate, coupled with significant regional instability with periods of SDNN below 50, provides robust evidence for autonomic dysfunction that warrants clinical intervention or monitoring for a psychiatric disorder.

**Urgent Concerns:** No immediate urgent medical concerns are flagged, but the persistent pattern of low vagal tone and high heart rate requires clinical follow up.

#### 6. Finalization Request

Based on the comprehensive analysis of the overall and regional HRV metrics, please finalize the decision as either “control” or “treatment”.

#### Final Decision: Treatment

**Reason for Decision:** The decision is consistently defended as **Treatment** because the data exhibits severe autonomic dysregulation. Specifically, the critically low RMSSD (23.42) and pNN50 (1.1692%), combined with an elevated resting heart rate (Mean RR 597.43 ms), are strong physiological markers of impaired vagal tone and chronic sympathetic activation. Furthermore, the regional analysis reveals periods of extreme autonomic rigidity, for example SDNN 34.99 in Region 7, reinforcing the conclusion of significant autonomic instability that is commonly observed in psychiatric patient populations.

## H Table of Acronyms

The detailed acronyms and definitions used in the paper are listed in Table 11.

Table 11. Acronyms and Definitions

Term	Definition	Term	Definition
AI	Artificial Intelligence	IChi2	Interactive Chi-square
ANN	Artificial Neural Network	IMV	Iterative Majority Voting
ANS	Autonomic Nervous System	LF	Low-Frequency
BP	Blood Pressure	LIME	Local Interpretable Model-agnostic Explanations
BD	Bipolar Disorder	LLM	Large Language Models
BF	Blood Flow	LOOCV	Leave-One-Out Cross-Validation
BLE	Bluetooth Low Energy	MDWT	Multilevel Discrete Wavelet Transform
BPM	Beats Per Minute	MRI	Magnetic Resonance Imaging
CAI	Contestable AI	MRR	Mean R-R intervals
CAM	Class Activation Mapping	MSTFT	Multi-scale Temporal-Frequency Transformer
CDI	Contestable Diagnosis Interface	NN	Neural Network
CMI	Cardiac Monitoring Interface	non-EUT	Non-euthymic
CNN	Convolutional Neural Network	PANSS	Positive and Negative Syndrome Scale
DL	Deep Learning	PD	Psychiatric Disorders
DP	Depression	PPG	Photoplethysmography
DTW	Dynamic Time Warping	RPA	Recurrence Plot Analysis
EEG	Electroencephalography	RRI	R-R Interval
EMD	Empirical Mode Decomposition	RMSSD	Root Mean Square of Successive R-R Interval Differences
EUT	Euthymic	SAE	Self-Adversarial Explanation
fMRI	functional Magnetic Resonance Imaging	SHAP	SHapley Additive exPlanations
FKNN	Fine K-Nearest Neighbor	SZ	Schizophrenia
FN	False Negative	TN	True Negative
FP	False Positive	TP	True Positive
GDPR	General Data Protection Regulation	TQWT	Tunable-Q Wavelet Transform
HC	Healthy Control	VLF	Very Low Frequency
HF	High-Frequency	VMD	Variational Mode Decomposition
HR	Heart Rate	WSN	Wavelet Scattering Network
HRV	Heart Rate Variability	XAI	Explainable AI