Explain the Why: A Dual-Layer Interpretable Framework for Mental Health Detection

Bo Yuan, Yulin Chen*

Zhejiang University, Hangzhou, China

{byuan,yulinchen}@zju.edu.cn

Abstract

The growing prevalence of mental health disorders has emerged as a critical societal issue, drawing increasing attention from both academia and industry. While recent deep learning methods have achieved impressive progress in mental health monitoring, their clinical deployment remains hindered by limited interpretability and a lack of insight into underlying psychological mechanisms. To address these challenges, we propose Dual-level Rule Learning (DRL) — a novel framework for trustworthy and explainable mental health monitoring. DRL introduces a dual-level explainability paradigm, encompassing both recognition-level and decision-level reasoning through psychologically grounded symbolic learning. The framework integrates established psychological constructs (e.g., PANAS scales) with the contextual reasoning capabilities of large language models (LLMs) to induce human-interpretable logical predicates that connect observable symptom patterns with latent causal factors. Specifically, the cognition module leverages LLMs to identify potential psychopathological indicators during representation learning, while the decision module aggregates these insights into generalizable logical rules that govern the model's predictions. This design enhances both transparency and etiological interpretability, bridging the gap between data-driven inference and psychological theory. Extensive experiments across multiple benchmarks demonstrate that DRL achieves strong predictive performance while offering consistent and interpretable reasoning traces.

Introduction

The global burden of mental disorders has reached unprecedented levels, with depressive and anxiety disorders alone affecting over 450 million people worldwide and constituting 14% of the global disease burden. In response, artificial intelligence (AI)—particularly deep learning—has emerged as a pivotal tool for enhancing diagnostic precision and treatment personalization in mental healthcare (Kasanneni et al. 2025; ZHA 2025).

Despite demonstrating superior predictive accuracy in detecting conditions like depression and generalized anxiety, these neural network-based approaches face critical adoption barriers due to their black-box nature. Current AI diagnostic systems suffer from two interrelated limitations: (1) Inherent opacity of black-box models obscures clinical reasoning pathways, undermining clinician trust and auditability (Li et al. 2024); and (2) limited ability to unveil the cause for mental illnesses ($e.g.$, socioeconomic stressors or neurobiological markers) essential for personalized interventions (Gorrepati et al. 2025). Moreover, the increasing integration with LLMs is prone to uncontrollable risks due to hallucinations and sociocultural biases. Thus, a growing awareness emphasizes the trustworthiness of these systems (Li et al. 2024). Unfortunately, the characteristics of a trustworthy mental health monitor remain an open question. Hence, based on recent surveys of explainable AI (Martino and Delmastro 2023) and mental health monitoring (Cao et al. 2024), we identify two crucial aspects that go beyond accuracy performance for trustworthy mental health AI: decision-level explainability and recognition-level explainability.

Decision-level explainability requires a transparent decision-making process for predictions, such as symptom-to-diagnosis mapping describing how the depressive symptoms inform classifications, thereby enabling clinicians to verify mental monitoring mechanisms and identify model limitations — a fundamental prerequisite for clinical adoption (Martino and Delmastro 2023). Complementing this, recognition-level explainability addresses the crucial etiological dimension by uncovering underlying biopsychosocial determinants ($e.g.$, unemployment precipitating anxiety or genetic predispositions increasing depression susceptibility), drawing upon the biopsychosocial model to provide clinicians with actionable insights for identifying modifiable risk factors, personalizing treatment plans, and anticipating comorbidity patterns (Garg et al. 2022; Azmee et al. 2023). This dual explanatory capacity effectively transforms mental health monitoring systems from mere monitoring tools into comprehensive clinical decision-support systems capable of generating both diagnostic conclusions and their causative evidence base. While existing approaches may separately address decision-level or recognition-level explainability, they frequently fail to integrate both dimensions systematically, limiting their clinical utility.

To bridge this gap, we propose a Dual-level logical Rule Learning (DRL) framework for explainable mental health monitoring, drawing inspiration from the dual-system theory

*Equal contribution, Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

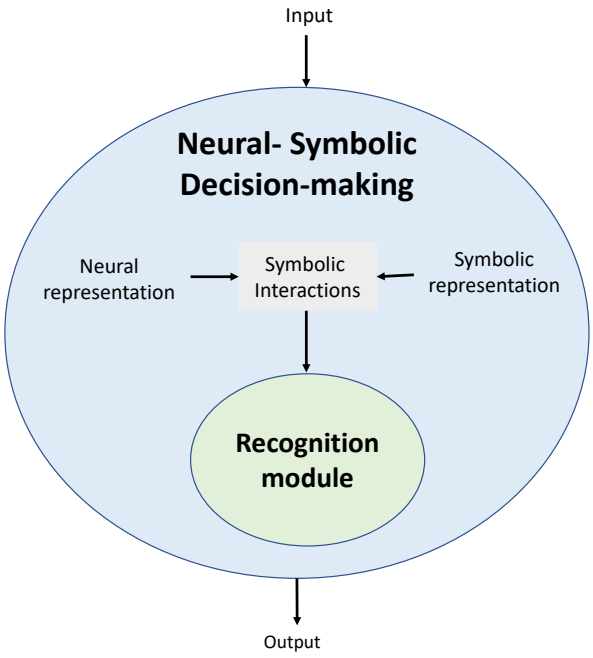

Figure 1: The diagram of our proposed framework, where we incorporate the recognition module into the neural-symbolic decision process.

(Liu et al. 2024). This framework systematically integrates decision-level monitoring transparency with recognition-level etiological insight into the existing neural-symbolic pipeline. As shown in Figure 1, neural representations of input samples are transformed into higher-order first-order logic predicates through symbolic interactions with established psychological constructs from PANAS scales, then the cognition module is embedded with the explainable decision framework, thereby achieving dual-level FOL rule learning. Specifically, our methodology leverages large language models (LLMs) to generate clinically validated annotations of potential psychopathological determinants ($e.g.$, educational stressors, bias, physical illnesses) which serve as recognition-level ground truth labels, while maintaining conventional depression classification as the decision-level output. This creates a hierarchical explanatory architecture that provides clinicians with two clinically actionable outputs: a standardized mental illness monitoring (decision-level) accompanied by the potential etiological profiles (recognition-level), effectively satisfying the critical need for etiological understanding – the recognition-level explainability that clinicians require for personalized intervention planning.

Our framework ensures human-in-the-loop explainability by incorporating domain expert knowledge and a transparent decision-making process based on logic rules. This explainability further enables the flexibility to adjust rules through pre-defining the domain concepts for symbolic interactions. Additionally, recognition-level explainability is optimized with the estimated latent factors of depressive manifestations and LLM-annotated ground truth. In this way, the output of the LLMs is used as optimization references rather than direct explanations, which can enhance the model's robustness against false LLM predictions and guarantee controllability. Moreover, our model exhibits generalizability, attributed to the generalizable performance of LLMs combined with reliable domain knowledge as guidance and the utilization of the neural-symbolic model, which can learn domain-generalizable rules across datasets in mental health monitoring tasks.

To summarize, the contributions of this work include: (1) The first unified framework simultaneously achieving decision-level and recognition-level explainability in mental health AI through dual-level logical rule learning; (2) A novel application of LLMs as clinical knowledge distillators for generating psychologically valid etiological annotations for learning recognition-level rules; (3) Empirical results demonstrate that joint optimization of recognition-level rule learning and decision-level rule learning can achieve state-of-the-art performance while maintaining interpretability.

Related Work

Mental disorders, characterized by their epidemiological pervasiveness and debilitating societal consequences, constitute a critical public health burden through multidimensional quality-of-life impairments (Angadi et al. 2024). Recent mental illness detection research has witnessed a notable paradigm shift from prioritizing accuracy to considering trustworthiness (Ding et al. 2025; Joyce et al. 2023). In line with our work, our investigation focuses on two pivotal dimensions of explainability: decision-level explainability and recognition-level explainability.

Regarding decision-level explainability, (Yin et al. 2025; Dalal et al. 2025; Wiegreffe and Pinter 2019) suggested obtaining key evidence for interpretation based on feature importance or attention weights, while (Ahmed et al. 2025; Malhotra and Jindal 2024) provides post-hoc explanations for the decisions of deep models (Ahmed et al. 2025; Malhotra and Jindal 2024) . However, the reliability of these methods is constrained by inherent stochasticity in deep neural architectures and the "explanation gap" between model internals and human-comprehensible justifications. Furthermore, another group of works (Spitale, Axelsson, and Gunes 2025), explored large generative language models (Cai et al. 2025) and regarded the self-generated explanation as an explanation after decision-making. Nevertheless, these explanations may not be reliable due to the hallucination phenomenon or inconsistency (Heo, Son, and Park 2025). Moving on to recognition-level explainability, most methods, such as (Garg et al. 2022; Gorrepati et al. 2025), unveil the underlying cause for mental health issues through deep learning methods, which contravene the explicability requirements of clinical decision-support systems due to the inherent opaque decision-making process of deep models. To address these limitations, we propose a Dual-level Rule Learning (DRL) framework that can simultaneously achieve decision-level explainability and recognition-level explainability in mental health monitoring tasks.

Prelimilary

Problem formulation

This paper proposes a novel reformulation of explainable mental health monitoring as a dual-level logical rule learning framework, addressing both decision-level and recognition-level interpretability during the monitoring processes. Building upon established concept-based eXplainable AI (XAI) methodologies (Doumanoglou, Asteriadis, and Zarpalas 2024), we present a principled approach to the two-level rule learning challenge in mental health monitoring by integrating domain-specific expert knowledge with cognitive-level insights, as formally defined in Definition .

Let D be the input dataset and M be the neural representations from a small language model ; the dual-level logical rule learning is formally defined as $D_{RL} := D \times K_1 \times K_2 \times M \rightarrow \{\mathcal{R}_r, \mathcal{R}_d\}$. Among them, K_1 denotes the recognition-level knowledge data from LLMs, K_2 denotes the domain expert knowledge data derived from standardized PANAS psychological scales (Crawford and Henry 2004); $\mathcal{R}_r$ is the set of recognition-level rules over the symbolic representations, and $\mathcal{R}_d$ is the set of decision rules derived based on $\mathcal{R}_r$.

Given observational data $D = X, Y$, we formalize mental health monitoring as a dual-level rule learning paradigm. Each symbolic rule $r : a \Rightarrow b$ consists of: (a) An antecedent a expressed in conjunctive normal form (CNF): $\bigwedge_{i \in N_p} \phi_i$, where ϕ_i represents first order logic (FOL) predicates ($e.g.$, $Depressed(X)$) and N_p indicates the number of predicates. (b) A consequent b representing the rule's predictive outcome, quantifying the causal effect for covered instances. The framework's expressiveness is enhanced through the incorporation of existential ($\exists$) and universal ($\forall$) quantifiers, enabling comprehensive representation of mental health monitoring patterns.

Knowledge representation based on first order logic

As mentioned above, the semantic knowledge during the decision-making process can be represented by first order logic predicates. Let X_i be an arbitrary sequence and M_j be the concept from domain expert knowledge; then we employ $P_1(X_i,M_j)$ predicates to denote that X_i is conceptually relevant to M_j, and $\neg P_1(X_i,M_j)$ predicates to denote that X_i is conceptually irrelevant to M_j. Additionally, to model the recognition-level explainability of sequences, we employ $P_2(X_i,C_k)$ to underlying attributes for the predictions, which means that X_i is associate with C_k, and $\neg P_2(X_i,C_k)$ predicates denote that X_i is not associated with C_k. Additionally, we also employ the predicate $P_d(X_i,D_y)$ to denote that the prediction of X_i is the label D_y. Based on these FOL predicates, we can model the neural-symbolic interactions in the feature space and derive high-level rule-based explanations.

Method

As previously discussed, our task is to learn dual-level rules to provide comprehensive explanations for mental health monitoring. Let us denote B as a base model ($e.g.$, small langugae model) that estimates the probability distribution $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{x}$ is an input data sample and $\mathbf{y}$ is a candidate class. Decision-level explainability means that we upgrade model B to a self-explaining version incorporating domain expert knowledge $\boldsymbol{k}_e$ into the input, then we can reformulate the probability distribution as

$$p(\mathbf{y}|\mathbf{x}) \rightsquigarrow p(\mathbf{y}|\mathbf{x}, \boldsymbol{k}_e) \tag{1}$$

To further achieve recognition-level explainability during the decision-making process, we integrate the inner cause into the explainable decision-making system by adding the recognition variables $\boldsymbol{q}$ into the decision process, then we can reformulate the probability distribution as

$$\begin{aligned} p(y|x, \boldsymbol{k}_e) &= \sum_{\boldsymbol{q}} p(y|x, \boldsymbol{k}_e, \boldsymbol{q}) p(\boldsymbol{q}|x, \boldsymbol{k}_e) \\ &= \sum_{\boldsymbol{q}} p(y|\boldsymbol{q}) p(\boldsymbol{q}|x, \boldsymbol{k}_e) \end{aligned} \tag{2}$$

As shown in Figure 2, the derived two terms correspond to the two main modules of the proposed framework, DRL:

- **domain expert knowledge $\boldsymbol{k_e}$** is used to help comprehend the decision process of black-box models by forming the symbolic representation with symbolic interactions, which is the key to surface-level explainability.
- **recognition variable $\boldsymbol{q}$** is used to unveil the underlying cause for the decision process, which is the key to core-level explainability.

To import decision-level explainability into the deep learning paradigm, one simple but effective way is to perform symbolic interactions during the training process. In this way, the decision-making process is inherently interpretable and can be represented by symbol-grounded rules. Inspired by the recent research that demonstrates the effectiveness of the concept-based models in deriving interpretable decisions (Doumanoglou, Asteriadis, and Zarpalas 2024), we propose to align the domain concepts and the embeddings from small language models, thereby forming symbol-grounded predicates for neural-symbolic interactions.

Domain expert knowledge from PANAS scales To establish clinically interpretable mappings between deep neural representations and validated psychological constructs, we operationalize the Positive and Negative Affect Schedule (PANAS) as domain-expert knowledge (as shown in Table 1), a psychometrically validated instrument that quantifies two orthogonal affective dimensions through self-reported emotional descriptors. The positive affect (PA) subscale measures adaptive emotional tendencies, including enthusiasm and alertness, while the negative affect (NA) subscale assesses maladaptive states such as distress and irritability, with this bipolar structure providing a clinically grounded framework for interpreting model outputs in mental health monitoring systems. Leveraging the PANAS scales to bridge the data-driven features and established psychopathological constructs, DRL enables explainable mental health monitoring for generating interpretations aligned with diagnostic

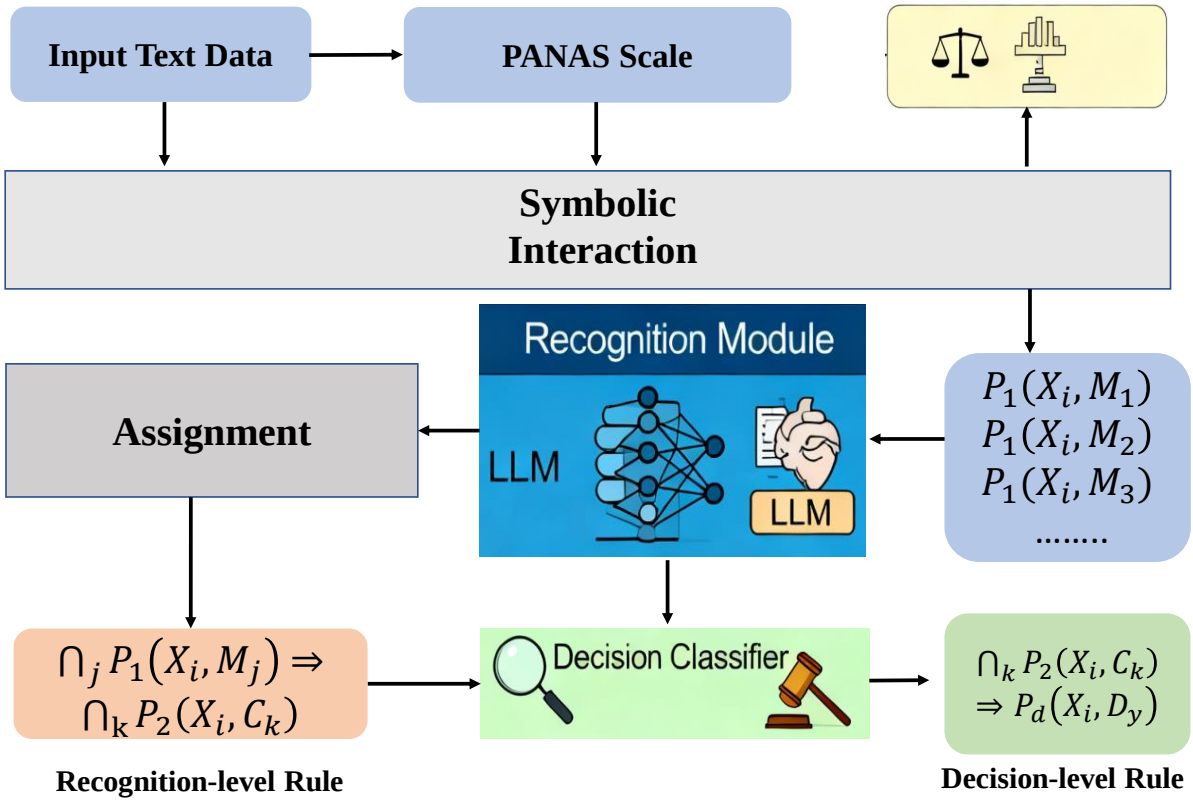

Figure 2: The architecture of our proposed Dual-level logical Rule Learning (DRL) framework.

Table 1: Notions and definitions of domain concepts with PANAS scales from (Crawford and Henry 2004)

Category	Domain concepts				
Positive	M_1:Attentive	M_2:Active	M_3:Alert	M_4:Excited	M_5:Enthusiastic
	M_6:Determined	M_7:Inspired	M_8:Proud	M_9:Interested	M_{10}:Strong
Negative	M_{11}:Hostile	M_{12}:Irritable	M_{13}:Ashamed	M_{14}:Guilty	M_{15}:Distressed
	M_{16}:Upset	M_{17}:Scared	M_{18}:Afraid	M_{19}:Jittery	M_{20}:Nervous

Table 2: Notions and descriptions of potential cause for mental disorders from (Garg et al. 2022)

Notions	Descriptions
C_0	no relevance to the specific determinants below (default category, regardless of mental health presence
C_1	bias, discrimination, or abuse (e.g., harassment, prejudice)
C_2	jobs or career including the education, employment and other financial aspects
C_3	physical/emotional illness or medication use
C_4	interpersonal relationship such as friendship or kinship
C_5	the feeling of life such as meaninglessness, purposefulness, or cultural estrangement/integration

mechanisms while maintaining the validated relationships between affective states and personality dimensions.

Semantic Interaction Semantic interaction operation is used for detecting different semantic relations, such as conceptual relevance. Given the dataset $D = \{X, Y\}$, where X denotes the input samples and Y denotes the corresponding labels. For each sample $X_i \in X$, we employ a small language model (SLM) to get feature vectors (hidden states) of each sample as $\mathbf{X} = [x_1, x_2, x_3, \ldots, x_N]$, where x_1 denotes the sentence embedding of X_i through the SLM and and N denotes the whole number of samples in the training data. In this work, we employ Bert as the SLM. With the obtained embedding vectors from the SLM, we then perform sym-

bolic interactions with the domain concepts. Considering the embedding vectors of domain concepts that M=[m_1, m_2, $\ldots$, m_{N_m}] denotes the feature vectors of domain concepts with a PLM Bert, where N_m denotes the number of domain concepts, then the truth value the predicates $P_1(X_i,M_j)$ are formulated as follows:

$$\alpha_j^{(i)} = \sigma(\mathbf{w}^\top[\mathbf{x}_i \odot \mathbf{m}_j]), \quad \hat{\alpha}_j = \mathbb{I}(\alpha_j^{(i)} \geq 0.5) \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function, ($\odot$ denotes dot product, and $\mathbf{w}$ is the trainable parameter matrix. $\mathbb{I}(\cdot)$ is the indicator function for symbolic truth valuation of predicates $P_1(X_i,M_j)$ based on α_j during the decision-making process. Intuitively, we can directly estimate the logical consequences based on the conjunction of these obtained logical predicates, achieving the decision-level explainability. However, to explore the root cause of the decisions, we integrate a recognition module to infer the recognition-related rules, and then derive the final prediction based on the recognition-related rules.

Recognition module

To estimate the recognition-related rules based on predicates $P_1(X_i,M_j)$, we firstly construct the knowledge-augmented representation for the input samples in the neural space as $\mathbf{v}_i = \frac{\sum_j \alpha_j^{(i)}\mathbf{m}_j}{|\sum_j \alpha_j^{(i)}\mathbf{m}_j|_2}$. Then the recognition rule is estimated through the recognition loss as:

$$\mathcal{L}_c = -\sum_{k=1}^{|\mathcal{C}|} y_c^{(k)} \log P(c^{(k)}|\mathbf{v}_i) \tag{4}$$

where $y_c^{(k)} \in \mathcal{C}$ = { $C_0, C_1, \ldots, C_5$} (as shown in Table 2), which are generated via querying the LLM ($i.e.$, Deepseek-V3) templates with the prompt in Appendix **??**. $P(c^{(k)}|\mathbf{v}_i)$ is calculated by softmax$(\mathbf{w}_c\mathbf{v}_i + \mathbf{b}_c))$ with two tainable parameters $\mathbf{w}_c$ and $\mathbf{b}_c$. The loss function ensures alignment with LLM-generated causal labels (y_c). Then the truth valuation of the consequence of the recognition rule $\mathcal{R}_{r(X_i)}$ is given by:

$$\forall X_i, \quad P_2(M_i, C_k) = \begin{cases} \text{True} & \text{if } P(c_k|\mathbf{v}_i) \geq 0.5 \\ \text{False} & \text{if } P(c_k|\mathbf{v}_i) < 0.5 \end{cases} \tag{5}$$

Based on the above assignment, the recognition-level rule $\mathcal{R}_r$ is obtained as:

$$\forall X_i, \quad \left(\bigwedge_j P_1(X_i, M_j)\right) \Rightarrow \left(\bigwedge_k P_2(X_i, C_k)\right) \tag{6}$$

where ($\wedge$) denotes trainable AND operations. The recognition-level rule of all training samples is optimized with LLM-generated content during the training process.

Decision-Level rule learning

The results of the recognition-level rule optimization serve as an interpretable feature, which is incorporated into the decision-making process to derive the final decisions as below:

$$P(y|\mathbf{q}) = \text{softmax}(\mathbf{w}_d^\top \mathbf{q} + b_d) \tag{7}$$

where $\mathbf{q} = [P(c^{(1)}|\mathbf{v}), ..., P(c^{(k)}|\mathbf{v})])$. Then the total loss combines both decision-level loss and recognition loss is defined as:

$$\mathcal{L} = \lambda\mathcal{L}_c + (1-\lambda)\mathcal{L}_d \tag{8}$$

where $\mathcal{L}_d$ is the cross-entropy for $P(y|\mathbf{q})$ and $\lambda \in [0,1]$ controlling decision-recognition. Since the ground truth of the prediction for recognition-level rule learning is provided by LLMs, we set $\lambda = 0.2$ to achieve a good trade-off between performance and interoperability. To simplify the rule-based explanation for users, we employ the predicates with the highest confidence as the final decisions. Specifically, let D_y denotes the prediction of X_i with maximum probability based on the estimated $P(y|\mathbf{q})$, then the final decision-level rule $\mathcal{R}_d$ is obtained as:

$$\forall X_i, \quad \left(\bigwedge_k P_2(X_i, C_k)\right) \Rightarrow P_d(X_i, D_y) \tag{9}$$

where the rule antecedent $\bigwedge_k P_2(X_i, C_k)$ denotes the conjunction of concequences from the recognition rule $\mathcal{R}_r$. By incorporating the $\mathcal{R}_r$ into the whole decision-level rule learning process, the hierarchical structure implements a causal chain ($i.e.$, $P(y|\mathbf{q}) = \sum_{c\in\mathcal{C}} \underbrace{P(y|c)}_{\text{Decision-level}} \cdot \underbrace{P(c|\mathbf{v})}_{\text{Recognition-level}}$), which satisfies the explainability of gradient-based attribution with $\frac{\partial P(y)}{\partial P(c)}$) quantifying each root-cause's contribution. In addition, DRL also ensures counterfactual validity by manipulating $P(c|\mathbf{v})$, such as setting $P(\text{Interpersonal relationship}|\mathbf{v}) = 1$.

Experiment setup

Dataset

To establish rigorous benchmarks for explainable mental health assessment, our experiments are conducted on three real-world mental health monitoring datasets as follows:

- **MHR** [1] (Mental Health Recognition) corpus constitutes a diverse collection of professionally annotated textual expressions from individuals exhibiting mental disorders. The dataset is partitioned into 8,000 training instances and 2,000 test instances, ensuring model training and evaluation.
- **RER** [2] (Reddit Emotional Repository) (RER) is a diverse corpus of mental health narratives extracted through ethically compliant web scraping protocols, subsequently processed using clinical NLP pipelines. The dataset is partitioned into 19989 training instances and 3527 test instances, ensuring model training and evaluation.

[1] https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus/discussion

[2] https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned

- **DAR** [3] (Depression and Anxiety Record) captures mental health issues through social media discourse, employing psycholinguistic filtering to isolate depression/anxiety indicators from Facebook comments, posts, etc. The dataset is partitioned into 5574 training instances and 1395 test instances, ensuring model training and evaluation.

Baselines

Our experimental design incorporated both black-box baselines and white-box baselines. As for black-box baselines, we benchmark cutting-edge large language models including **GPT-4** (Moëll 2024) and **Deepseek-V3** (Xian, Ni, and Wang 2025), selected for their state-of-the-art performance in clinical NLP tasks despite their inherent interpretability limitations. While white-box models have proliferated theoretically, their clinical adoption remains constrained by domain-specific implementation challenges; our study therefore, focuses on common white-box models within the context of medical and health tasks. Following the interpretable mental health monitoring work, we conduct the following baselines: **Linear Regression** (Wongkoblap, Vadillo, and Curcin 2017), **K Nearest Neighbors** (Pamungkas, Isnanto, and Nugraheni 2024), **Decision Tree** (Huang et al. 2022) and **Explainable Boosting Machine** (Hegselmann et al. 2022).

Evaluation metric

To evaluate model performance in clinical mental health applications, we employ macro-averaged metrics including precision (P), recall (R), and F1 scores to ensure equitable evaluation across different mental health categories, since DAR is an imbalanced dataset (Segura-Bedmar and Raez 2019). The macro-averaging approach provides population-level insights by weighting all classes equally, crucial for identifying model biases in minority groups - a critical consideration for explainable AI systems in mental healthcare. These metrics collectively assess the model's capacity to maintain both statistical accuracy and clinical interpretability when processing psycholinguistic markers of mental health conditions.

Implementation details

In our primary experiments, we uniformly set the batch size to 128 and set the maximum number of epochs to 200, while utilize early stopping to manage the training process, with $min_delta = 1e-4$ and $patience = 6$. For the small language model, we incorporated a dense layer with a hidden size of 128 for all evaluation datasets using Bert. The parameters were trained using the default learning rate of the Adam optimizer in Keras ($i.e.$, $learning_rate = 0.001$) and the "sparse_categorical_crossentropy" loss function and the Adam Optimizer (Kingma and Ba 2015). White-box baselines are implemented following the original hyperparameter specifications of the referenced literature. All experiments were conducted on NVIDIA GTX 2080 Ti/RTX 3090 GPUs running Ubuntu systems, with Python 3 implementations. To ensure statistical reliability, we performed five independent training runs for each configuration, reporting mean performance metrics with variance indicated via subscript notation. For large language model evaluations, we employed the GPT-4-Turbo ($e.g.$, GPT-4) and Deepseek-V3 APIs (temperature=0.5) through carefully designed prompts. Due to computational resource constraints, we performed single inference passes on each test set instance. Complete prompt engineering details are provided below:

Prompt of LLMs for mental health monitoring: `You are an NLP classifier that can identify potential mental health issues based on input sentences. Label 1 indicates that the input sentence may contain mental health issues such as depression and anxiety. Label 0 indicates that no mental health issues can be detected from the input sentence. Please classify the input sentence into 0 or 1, and return only the category type (0 or 1) without explanation or reason.`

Prompt of LLMs for detecting the potential cause for mental health issues: `You are an NLP classifier for mental health determinant identification. Categorize the input sentence strictly into one of six integer labels based on its primary association with potential causes of mental health issues: Label 0 indicates no relevance to the specific determinants below (default category, regardless of mental health presence). Label 1 indicates that the input sentence is related to bias, discrimination, or abuse (e.g., harassment, prejudice). Label 2 indicates that the input sentence is related to the jobs or career including the education, employment and other financial aspects. Label 3 indicates that the input sentence is related to physical/emotional illness or medication use. Label 4 indicates that the input sentence is related to interpersonal relationship such as friendship or kinship. Label 5 indicates that the input sentence is related to the feeling of life such as meaninglessness, purposefulness, or cultural estrangement/integration. Please classify the input sentence into 0 or 1 or 2 or 3 or 4 or 5, and return only the category type (0 or 1 or 2 or 3 or 4 or 5) without`

[3] https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset/data

Table 3: Evaluation results of baselines and our methods, with weighted average metrics on three datasets. **Bold** denotes the optimal performance on each dataset; the mark $^\diamondsuit$ denotes the explainabel baselines.

Datasets	Models	Precision	Recall	F_1 score
MHR	Linear Regression$^\diamondsuit$	$53.65\pm_{1e-3}$	$53.57\pm_{1e-3}$	$53.14\pm_{1e-3}$
	Decision Tree$^\diamondsuit$	$35.44\pm_{3e-4}$	$49.70\pm_{1e-7}$	$33.37\pm_{2e-7}$
	K Nearest Neighbors$^\diamondsuit$	$48.21\pm_{0e-0}$	$48.31\pm_{0e-0}$	$47.59\pm_{0e-0}$
	Explainable Boosting Machine $^\diamondsuit$	$71.84\pm_{1e-5}$	$71.83\pm_{1e-5}$	$71.83\pm_{1e-5}$
	GPT-4	80.19	77.52	76.88
	Deepseek-V3	87.36	86.93	86.84
	DRL(ours)	**89.27**$\pm_{4e-5}$	**89.22**$\pm_{1e-4}$	**89.22**$\pm_{1e-4}$
RER	Linear Regression$^\diamondsuit$	$70.04\pm_{1e-3}$	$69.08\pm_{1e-3}$	$68.62\pm_{1e-3}$
	Decision Tree$^\diamondsuit$	$75.32\pm_{1e-5}$	$50.24\pm_{1e-7}$	$34.08\pm_{4e-7}$
	K Nearest Neighbors$^\diamondsuit$	$62.56\pm_{0e-0}$	$62.30\pm_{0e-0}$	$62.05\pm_{0e-0}$
	Explainable Boosting Machine $^\diamondsuit$	$88.15\pm_{1e-3}$	$83.7\pm_{1e-2}$ 6	$83.13\pm_{1e-2}$
	GPT-4	81.95	78.82	78.27
	Deepseek-V3	86.04	84.91	84.78
	DRL(ours)	**95.54**$\pm_{1e-5}$	**95.35**$\pm_{1e-5}$	**95.38**$\pm_{1e-5}$
DAR	Linear Regression$^\diamondsuit$	$49.50\pm_{1e-4}$	$48.92\pm_{3e-4}$	$47.16\pm_{1e-3}$
	Decision Tree$^\diamondsuit$	$44.37\pm_{4e-10}$	$49.98\pm_{4e-8}$	$47.01\pm_{1e-8}$
	K Nearest Neighbors$^\diamondsuit$	$50.56\pm_{0e-0}$	$50.42\pm_{0e-0}$	$50.34\pm_{0e-0}$
	Explainable Boosting Machine $^\diamondsuit$	$76.06\pm_{1e-4}$	$55.06\pm_{2e-5}$	$56.42\pm_{1e-4}$
	GPT-4	68.78	87.56	71.83
	Deepseek-V3	75.82	88.27	80.09
	DRL(ours)	**96.69**$\pm_{3e-5}$	**96.01**$\pm_{1e-4}$	**96.35**$\pm_{4e-5}$

Table 4: Descriptions of dual-level logical rule generation in Example 1.

	Example 1
Case	X_{116} =" *i have nothing to offer anybody that they can t and don t get from a dozen other better people i m ashamed to even exist in front of others let alone pretend that i m a real interesting in any way attractive person i m tired of pretending though i m tired of feeling like knowing that nobody care and for good reason other people are just unconcerned with me all of them always i m going to be alone until i kill myself nothing help nothing will help i m just a reject which make so much of life off limit to me and it s part that feel extremely important"*
Label set	{Mental_Disorders(y^D), Mental_Wellness(y^W)}
$\mathcal{R}_r(X_{116})$	$\neg P_1(X_{116}, M_1) \wedge \neg P_1(X_{116}, M_2) \wedge \neg P_1(X_{116}, M_3) \wedge \neg P_1(X_{116}, M_4) \wedge \neg P_1(X_{116}, M_5) \wedge \neg P_1(X_{116}, M_6) \wedge \neg P_1(X_{116}, M_7) \wedge \neg P_1(X_{116}, M_8) \wedge \neg P_1(X_{116}, M_9) \wedge \neg P_1(X_{116}, M_{10}) \wedge \neg P_1(X_{116}, M_{11}) \wedge \neg P_1(X_{116}, M_{12}) \wedge P_1(X_{116}, M_{13}) \wedge P_1(X_{116}, M_{14}) \wedge \neg P_1(X_{116}, M_{15}) \wedge P_1(X_{116}, M_{16}) \wedge P_1(X_{116}, M_{17}) \wedge \neg P_1(X_{116}, M_{18}) \wedge P_1(X_{116}, M_{19}) \wedge \neg P_1(X_{116}, M_{20}) \Rightarrow \neg P_2(X_{116}, C_1) \wedge \neg P_2(X_{116}, C_2) \wedge \neg P_2(X_{116}, C_3) \wedge \neg P_2(X_{116}, C_4) \wedge P_2(X_{116}, C_5)$
$\mathcal{R}_d(X_{116})$	$\neg P_2(X_{116}, C_1) \wedge \neg P_2(X_{116}, C_2) \wedge \neg P_2(X_{116}, C_3) \wedge \neg P_2(X_{116}, C_4) \wedge P_2(X_{116}, C_5) \Rightarrow P_d(X_{116}, y^D)$

```
explanation or reason.
```

Results and Discussions

Model performance

The experimental results demonstrate that our proposed DRL framework achieves state-of-the-art performance while inherently integrating dual-level explainability into its decision-making architecture, addressing a critical limitation of both traditional black-box models and white-box models. As shown in Table 3, DRL consistently outperforms all baseline methods across all evaluation datasets in terms of three evaluation metrics, including precision, recall and F_1 score. Particularly, DRL achieves statistically significant improvements of 2.38%–61.3% in F_1-scores across all datasets compared to large language models ($e.g.$, GPT-4 and Deepseek-V3) and white-box approaches ($e.g.$, Explainable Boosting Machine and Decision Trees). This performance advantage is particularly meaningful when considering the explainability-performance trade-off that has historically plagued explainable AI systems. While white-box models like Explainable Boosting Machine demonstrate higher performance than other white-box models, their substantial performance gap (12.25%-39.93% lower F1-scores than our method) highlights the inherent trade-off between transparency and efficacy in existing methods—a trade-off our framework successfully achieves. These findings position DRL as a paradigm-shifting solution in XAI, achieving dual superiority in both accuracy and explainability that meets the rigorous demands of high-stakes applications requiring transparent and reliable decisions.

Model explainability

We present a case study (Example 1 from RAR test data) to demonstrate DRL's explainability capabilities through the learned logical rules, as illustrated in Table 4. The framework first extracts semantically meaningful predicates ($e.g.$, $\neg P_1(X_{116}, M_1)$ and $\neg P_1(X_{116}, M_2)$) via symbolic interactions, then the recognition-level rule $\mathcal{R}_r(X_{116})$ and decision-level rule $\mathcal{R}_d(X_{116})$ are learned based on these established predicates. From the learned rule $\mathcal{R}_r(X_{116})$ and $\mathcal{R}_d(X_{116})$, users can readily understand the model-decision-making process, $i.e.$, the prediction $P_d(X_{116}, y^D)$ is derived through the logical conjunction of conceptually relevant predicates ($i.e.$, $P_1(X_{116}, M_{13})$ and $P_1(X_{116}, M_{14})$ and $P_1(X_{116}, M_{16})$ and $P_1(X_{116}, M_{17})$ and $P_1(X_{116}, M_{19})$), combined with recognition-level rule consequences ($e.g.$, $P_2(X_{116}, C_5)$). Considering the truth value of predicates during the rule learning process, we can infer that the input case X_{116} is predicted as "has mental issues" because the semantic information of input instance `H` is relevant with the negative concepts (M_{13}, M_{14},M_{16},M_{17},M_{19}) and potential psychological causes C_5 (meaning of life). Furthermore, each predicate in the antecedents corresponds to the relationship with a specific psychological concept (as shown in Tables 1 and 2), which can be used to automatically generate rule-based explanations and provide guidance for users to design personalized clinician interventions.

Conclusion

In this work, we tackle the persistent challenges of reliability and user trust in AI-driven mental health monitoring. We argue that building a trustworthy system requires two complementary forms of explainability: recognition-level and decision-level interpretability. Guided by this principle, we introduce DRL, a dual-level explainable framework that integrates a recognition module into a neural–symbolic reasoning pipeline. DRL not only identifies potential etiological factors underlying mental health states (recognition-level explainability) but also provides transparent and traceable decision pathways (decision-level explainability). Extensive experiments across multiple datasets verify the robustness, interpretability, and clinical relevance of our approach.

Despite these advances, a notable limitation lies in the validation of recognition-level rules, as current ground-truth supervision relies on LLM-generated content. Future work will focus on (1) developing alternative supervisory signals to mitigate dependency on LLM outputs, and (2) designing clinically grounded validation protocols for learned symbolic predicates. While LLM-generated content currently serves only as optimization guidance—thus preserving decision-level transparency—the potential inaccuracies in cognitive-level rules highlight an important direction for ensuring clinical reliability.

Impact Statement

This work establishes a novel neural-symbolic framework that formalizes mental health monitoring through a dual-level rule-learning architecture, integrating cognitive modules to simultaneously identify etiological factors and deliver interpretable diagnostic decisions. Our approach represents a significant advancement in explainable AI (XAI) for mental health by bridging the critical gap between deep learning's predictive power and clinical psychology's need for explainability. The research has profound implications for both clinical practice and public mental health initiatives, particularly in developing scalable yet trustworthy monitoring systems tools.

References

2025. The burden of mental disorders, substance use disorders and self-harm among young people in Asia, 20192021: Findings from the global burden of disease study 2021. *Psychiatry Research*, 345: 116370.

Ahmed, A.; Saleem, M.; Alzeen, M.; Birur, B.; Fargason, R. E.; Burk, B. G.; Harkins, H. R.; Alhassan, A.; and Al-Garadi, M. A. 2025. Leveraging Large Language Models to Enhance Machine Learning Interpretability and Predictive Performance: A Case Study on Emergency Department Returns for Mental Health Patients. arXiv:2502.00025.

Angadi, N. A.; Eedula, N.; Gopali, K. P.; and Jayashree, R. 2024. Supervised learning models to predict mental illness and its severity from Reddit posts. *Int. J. Comput. Sci. Eng.*, 27(3): 352–363.

Azmee, A. A.; Murikipudi, M.; Khan, M. A. A. H.; and Pei, Y. 2023. Sentence Level Analysis for Detecting Mental Health Causes Using Social Media Posts. In Shahriar,

H.; Teranishi, Y.; Cuzzocrea, A.; Sharmin, M.; Towey, D.; Majumder, A. K. M. J. A.; Kashiwazaki, H.; Yang, J.; Takemoto, M.; Sakib, N.; Banno, R.; and Ahamed, S. I., eds., *47th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2023, Torino, Italy, June 26-30, 2023*, 1388–1393. IEEE.

Cai, Z.; Fang, H.; Liu, J.; Xu, G.; Long, Y.; Guan, Y.; and Ke, T. 2025. Improving unified information extraction in Chinese mental health domain with instruction-tuned LLMs and type-verification component. *Artif. Intell. Medicine*, 162: 103087.

Cao, Y.; Dai, J.; Wang, Z.; Zhang, Y.; Shen, X.; Liu, Y.; and Tian, Y. 2024. Systematic Review: Text Processing Algorithms in Machine Learning and Deep Learning for Mental Health Detection on Social Media. *CoRR*, abs/2410.16204.

Crawford, J. R.; and Henry, J. D. 2004. The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3): 245–265.

Dalal, S.; Tilwani, D.; Gaur, M.; Jain, S.; Shalin, V. L.; and Sheth, A. P. 2025. A Cross Attention Approach to Diagnostic Explainability Using Clinical Practice Guidelines for Depression. *IEEE J. Biomed. Health Informatics*, 29(2): 1333–1342.

Ding, Z.; Wang, Z.; Zhang, Y.; Cao, Y.; Liu, Y.; Shen, X.; Tian, Y.; and Dai, J. 2025. Efficient or Powerful? Trade-offs Between Machine Learning and Deep Learning for Mental Illness Detection on Social Media. *CoRR*, abs/2503.01082.

Doumanoglou, A.; Asteriadis, S.; and Zarpalas, D. 2024. Unsupervised Interpretable Basis Extraction for Concept-Based Visual Explanations. *IEEE Trans. Artif. Intell.*, 5(4): 1496–1510.

Garg, M.; Saxena, C.; Saha, S.; Krishnan, V.; Joshi, R.; and Mago, V. 2022. CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, 6387–6396. European Language Resources Association.

Gorrepati, L. P.; Sonani, R.; Velugubantla, V.; Potla, R. T.; and Sathvik, M. 2025. Mental Health and Relations: Detection of Mental Health Disorders Related to Relationship Issues Through Reddit Posts. In Long, G.; Blumestein, M.; Chang, Y.; Lewin-Eytan, L.; Huang, Z. H.; and Yom-Tov, E., eds., *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, 1885–1889. ACM.

Hegselmann, S.; Ertmer, C.; Volkert, T.; Gottschalk, A.; Dugas, M.; and Varghese, J. 2022. Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines. *Frontiers in Medicine*, 9: 960296.

Heo, S.; Son, S.; and Park, H. 2025. HaluCheck: Explainable and verifiable automation for detecting hallucinations in LLM responses. *Expert Syst. Appl.*, 272: 126712.

Huang, Y.; Li, S.; Lin, B.; Ma, S.; Guo, J.; and Wang, C. 2022. Early detection of college students' psychological problems based on decision tree model. *Frontiers in Psychology*, 13: 946998.

Joyce, D. W.; Kormilitzin, A.; Smith, K. A.; and Cipriani, A. 2023. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, 6(1): 6.

Kasanneni, Y.; Duggal, A.; Ragupathi, S.; and Raja, S. P. 2025. Effective Analysis of Machine and Deep Learning Methods for Diagnosing Mental Health Using Social Media Conversations. *IEEE Trans. Comput. Soc. Syst.*, 12(1): 274–294.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR*.

Li, W.; Zhu, Y.; Lin, X.; Li, M.; Jiang, Z.; and Zeng, Z. 2024. Zero-shot Explainable Mental Health Analysis on Social Media by Incorporating Mental Scales. In Chua, T.; Ngo, C.; Lee, R. K.; Kumar, R.; and Lauw, H. W., eds., *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, 959–962. ACM.

Liu, H.; Wang, W.; Li, H.; and Li, H. 2024. TELLER: A Trustworthy Framework for Explainable, Generalizable and Controllable Fake News Detection. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 15556–15583. Association for Computational Linguistics.

Malhotra, A.; and Jindal, R. 2024. XAI Transformer based Approach for Interpreting Depressed and Suicidal User Behavior on Online Social Networks. *Cogn. Syst. Res.*, 84: 101186.

Martino, F. D.; and Delmastro, F. 2023. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artif. Intell. Rev.*, 56(6): 5261–5315.

Moëll, B. 2024. Comparing the Efficacy of GPT-4 and Chat-GPT in Mental Health Care: A Blind Assessment of Large Language Models for Psychological Support. *CoRR*, abs/2405.09300.

Pamungkas, A.; Isnanto, R. R.; and Nugraheni, D. M. K. 2024. Implementation of K-Nearest Neighbor in Case-Based Reasoning for Mental Health Diagnosis Systems. *Scientific Journal of Informatics*, 11(4): 1109–1120.

Segura-Bedmar, I.; and Raez, P. 2019. Cohort selection for clinical trials using deep learning models. *Journal of the American Medical Informatics Association*, 26(11): 1181–1188.

Spitale, M.; Axelsson, M.; and Gunes, H. 2025. VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching. *ACM Trans. Hum. Robot Interact.*, 14(2): 38:1–38:28.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 11–20. Association for Computational Linguistics.

Wongkoblap, A.; Vadillo, M. A.; and Curcin, V. 2017. Researching mental health disorders in the era of social media: systematic review. *Journal of medical Internet research*, 19(6): e228.

Xian, L.; Ni, J.; and Wang, M. 2025. Leveraging Large Language Models for Cost-Effective, Multilingual Depression Detection and Severity Assessment. *arXiv preprint arXiv:2504.04891*.

Yin, Q.; Yang, L.; Kuang, J.; and Yan, Z. 2025. A two-stage multi-task learning model for predicting and interpreting received answers in online mental health communities. *Expert Syst. Appl.*, 270: 126583.