Interpretable-MTLNet: A Kolmogorov–Arnold Network for Multitask Mental Health Prediction

Mai Ali*, Zhenghao Ni*, Deepa Kundur
Department of Electrical and Computer Engineering
University of Toronto
Toronto, ON, Canada
{mai.ali, zhenghao.ni}@mail.utoronto.ca, dkundur@ece.utoronto.ca

Abstract

Depression and anxiety are among the most prevalent mental health disorders worldwide, yet often underdiagnosed due to reliance on self-reporting and infrequent clinical assessments. Wearable devices offer a scalable path toward continuous mental health monitoring, but existing models often sacrifice interpretability for accuracy. We present Interpretable-MTLNet, a multitask neural architecture that jointly detects depression and anxiety from daily wearable time series while preserving mathematical transparency. The model couples multi-scale temporal convolutions with Kolmogorov–Arnold Network (KAN) layers to learn scale-aware embeddings and task-specific spline-based heads. This design provides consistent global and local explanations via activation-weighted univariate curves and symbolic surrogates. Evaluated on 40,000 participants with Fitbit data from the All of Us Research Program, Interpretable-MTLNet achieves a macro-AUROC of 0.731 across tasks, outperforming strong baselines under subject-level splits and remaining robust in imbalanced settings. These findings suggest that KAN-based architectures can deliver both accuracy and interpretability for detecting mental health problems from wearable data, advancing trustworthy digital phenotyping.

1 Introduction

Depression and anxiety are frequently comorbid, with approximately 45.7% of individuals diagnosed with depression also meeting the criteria for an anxiety disorder [9]. Consequently, predictive models that explicitly capture the correlation between these two disorders and produce joint predictions offer greater value for both diagnostic precision and treatment planning [3]. Concurrently, consumer wearables provide continuous, multimodal physiological data streams (e.g., heart-rate variability, sleep architecture, physical activity) that enable scalable and objective population-level screening.

Current machine learning approaches face two barriers: most treat depression and anxiety as separate tasks [14], without shared physiological signatures and joint representation learning; and state-of-the-art deep models remain black boxes with limited mathematical interpretability. The convergence of large-scale wearable datasets and interpretable neural architectures now enables principled multi-task learning that leverages cross-disorder structure *and* preserves mathematical transparency, achieving superior detection performance while remaining suitable for clinical adoption [1].

We present a hard-parameter-sharing multi-task model that couples a multi-scale temporal convolutional front-end with KAN layers and per-task logistic heads; we develop an interpretation toolkit that reveals **activation-weighted B-spline edge curves**, organizes signals into a clinically grounded system (activity level, variability, circadian, behavioral patterns, physiological, temporal trends, social

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body
* Mai Ali, and Zhenghao Ni are co-first authors.

rhythms), and provides task-level summaries. This approach outperforms strong baselines under identical data splits and preprocessing, while offering transparent global and local explanations.

2 Related Work

2.1 KANs in Time-Series and Healthcare Applications

Recent work has begun to explore the application of KAN in time-series analysis, often validating a hybrid approach where KAN is integrated with established sequential architectures to enhance non-linear feature mapping. One study explicitly note that KANs perform particularly well on key features extracted from time [16], a strategy conceptually similar to our model's use of pooled temporal representations. This has spurred applications across the clinical domain, where KANs have shown significant promise. For instance, CoxKAN [10]offers an interpretable, high-performance model for survival analysis , MedKAN [17]adapts KANs for robust medical image classification , and other work has validated their utility on physiological time-series data such as ECG signals. More specifically within mental health, KANs have been applied to classification from social media text [8], where they outperformed traditional baselines.Our work extends this growing body of research by addressing the distinct challenges of continuous, multimodal wearable data and leveraging the intrinsic, mathematical transparency of KANs for multitask prediction.

2.2 Interpretable Machine Learning Approaches

Given the black-box nature of deep models, interpretability has become crucial for clinical adoption of wearable-based mental health systems. Recent work integrates explainable AI techniques to understand model decisions. For instance, Shah et al. [15] used Shapley additive explanations (SHAP) in a personalized mood prediction model to identify important features for each individual—revealing diverse drivers of depressed mood ranging from sleep patterns and physical activity to comorbid anxiety levels. Ko et al. [11] similarly employed SHAP and LIME in a hybrid deep learning framework for depression detection, highlighting key predictors such as nightly sleep duration and resting heart rate. Other approaches incorporate attention mechanisms within neural networks to highlight salient time periods or physiological signals linked to emotional changes [7]. These interpretability efforts help clinicians and researchers trust the models and glean insights (e.g., which behavioral patterns signal worsening anxiety), although achieving both high accuracy and clear explanations remains an ongoing trade-off in current wearable mental health models.

3 Methods

3.1 Model Overview

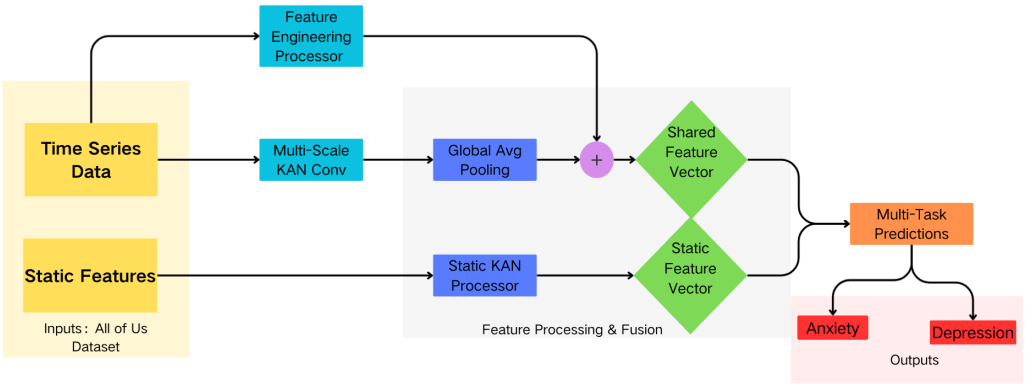

Figure 1: Architecture of *Interpretable-MTLNet*.

As shown in Figure 1, *Interpretable-MTLNet* jointly predicts depression and anxiety from daily wearable time series (e.g., steps, heart rate, sleep) and static covariates (e.g., demographics). The architecture has three components:

(1) Multi-Scale KAN Convolution. Spline-based filters at multiple temporal scales capture short-, medium-, and long-range dynamics, yielding scale-aware feature maps for downstream reasoning over the sequence [4].

(2) Temporal Aggregation & KAN Mixing. For each scale, we summarize the feature map by global average pooling over time and then fuse all scales with one or more KAN layers to form a compact shared embedding. Temporal dependencies are modeled within the multi-scale convolutions, while the KAN layers provide flexible nonlinear mixing of aggregated descriptors.

(3) KAN-based Multi-Task Heads (hard sharing). The shared embedding is concatenated with static covariates and fed to task-specific KAN heads that output sigmoid probabilities. All layers before the heads are *hard-parameter-shared* across tasks. All nonlinear transformations are implemented with Kolmogorov–Arnold Network (KAN) layers, enabling consistent visualization of feature–response curves across modules.

3.2 Kolmogorov–Arnold Networks Foundation

KAN layers [13] replace fixed activations with learned *univariate spline* functions. For input $\mathbf{x} \in \mathbb{R}^{n_{\text{in}}}$ and output $\mathbf{y} \in \mathbb{R}^{n_{\text{out}}}$, each output dimension is

$$y_j \;=\; \sum_{i=1}^{n_{\text{in}}} \phi_{j,i}(x_i) + b_j, \tag{1}$$

where each scalar map $\phi_{j,i}$ is parameterized as a B-spline expansion with a linear skip:

$$\phi_{j,i}(x) \;=\; \sum_{k=0}^{G+p} c_{j,i,k}\, B_{k,p}(x) \;+\; w_{j,i}\, \sigma(x), \tag{2}$$

with spline coefficients $c_{j,i,k}$, basis $B_{k,p}$ defined on knots $\{t_k\}$, and base activation σ=SiLU. See Fig. 2 for a schematic of a KAN layer.

3.3 Multi-Scale KAN Convolution

To capture temporal patterns at multiple resolutions, we apply 1D KAN convolution filters [12] of different lengths to the input time series. For each kernel size $k \in \{3, 5, 7\}$ (short-, medium-, long-range contexts), a KAN-based convolutional layer [6] slides a spline filter of width k over each sequence, producing feature maps $\mathbf{H}^{(k)}$:

$$H^{(k)}_{b,c,t} \;=\; \sum_{i=1}^{d} \sum_{j=0}^{k-1} \phi^{(k)}_{c,i,j}\big(X_{b,(t-j),i}\big), \tag{3}$$

where $X_{b,t,i}$ is the value of feature i on day t for participant b, and $\phi^{(k)}_{c,i,j}$ is the spline filter applied to offset j of feature i to produce channel c.

3.4 Temporal Aggregation and KAN Mixing

For each scale k, we apply global average pooling over time to obtain a per-scale vector

$$\mathbf{h}^{(k)} \;=\; \mathrm{GAP}_t\big(\mathbf{H}^{(k)}\big) \in \mathbb{R}^{C_k}. \tag{4}$$

We then concatenate all scales to form $\mathbf{z} = [\mathbf{h}^{(3)}; \mathbf{h}^{(5)}; \mathbf{h}^{(7)}] \in \mathbb{R}^{d_h}$ and pass it through shared KAN layer(s) to obtain the shared embedding.

3.5 KAN-based Multi-Task Learning (MTL)

Let $\mathbf{u} \in \mathbb{R}^{d_u}$ be the shared embedding and $\mathbf{s} \in \mathbb{R}^{d_s}$ the static covariates. For each task $t \in \{\text{dep}, \text{anx}\}$, a task-specific KAN head $g_t(\cdot)$ maps the concatenation to a logit $z^{(t)}$, followed by a sigmoid:

$$z^{(t)} \;=\; g_t\big([\mathbf{u}; \mathbf{s}]\big), \qquad p^{(t)} = \mathrm{sigmoid}(z^{(t)})). \tag{5}$$

We adopt *hard parameter sharing* for all layers before $\{g_t\}$ to reduce overfitting and encourage representation reuse, while keeping the task heads fully interpretable via their spline-based transformations.

Training objective (focal loss). Given labels $y_i^{(t)} \in \{0,1\}$ and predictions $p_i^{(t)}$ for sample i in task t, we minimize a weighted sum of task-wise focal losses:

$$\mathcal{L} \;=\; \sum_{t\in\{\text{dep, anx}\}} w_t \cdot \frac{1}{N_t} \sum_{i=1}^{N_t} \mathrm{FL}\Big(y_i^{(t)},\, p_i^{(t)};\, \alpha_t,\, \gamma_t\Big)\,, \tag{6}$$

where N_t is the number of labeled samples for task t, w_t is an optional task weight, and

$$\mathrm{FL}(y,p;\alpha,\gamma) \;=\; -\,\alpha\, y\,(1-p)^{\gamma}\log p \;-\; (1-\alpha)\,(1-y)\,p^{\gamma}\log(1-p). \tag{7}$$

Here $\alpha_t \in (0,1)$ mitigates class imbalance and $\gamma_t > 0$ down-weights majority class examples, while the task weight w_t controls the relative contribution of each task to the overall loss.

4 Experiment

4.1 Datasets and Evaluation Metrics

All of Us provides longitudinal data from ∼40,000 participants with 60-day Fitbit streams (daily steps, calories, HR/HRV, and sleep stages) [2]. Targets are clinically validated depression (5.87%) and anxiety (8.03%) from standardized interviews and medical records. All data were de-identified to protect personal privacy.

Our feature engineering generates five groups: (1) basic activity metrics (steps, calories, active minutes), (2) cardiovascular signals (average/resting heart rate, heart rate variability), (3) sleep patterns (duration, efficiency, REM/deep sleep percentages), (4) engineered variability indicators (coefficient of variation across 3-, 7-, and 14-day windows), and (5) behavioral regularity metrics (activity consistency, circadian rhythm proxies, weekend vs. weekday patterns). Missing data patterns are explicitly modeled through dedicated indicator features. Evaluation Metrics Details are in Appendix A.

4.2 Experimental Setup and Performance Comparison

We use subject-level data splits (70%/10%/20% train/val/test) to prevent leakage. The multi-scale KAN model processes temporal patterns using kernels of sizes $\{3,5,7\}$ with 64 channels, followed by shared mixing and task-specific heads. Training uses AdamW optimizer with focal loss to handle class imbalance. Implementation details are in Appendix B. Table 1 compares our approach with baseline methods.

5 Results

To benchmark our proposed architecture, we compared Interpretable-MTLNet against several strong multitask baselines, including CNN-LSTM, Transformer models, gradient-boosted trees, and the recently introduced Wearnet. As shown in Table 1, Interpretable-MTLNet consistently achieved the highest performance across both depression and anxiety prediction tasks, with macro-average AUROC and AUC-PR values of 0.731 and 0.332, respectively.

Table 1: Performance comparison on All of Us dataset. Values are mean±std over five seeds.

Model	Depression		Anxiety		Macro-Average	
	AUROC	AUC-PR	AUROC	AUC-PR	AUROC	AUC-PR
CNN-LSTM + FC	0.684±0.023	0.312±0.018	0.701±0.019	0.326±0.022	0.693±0.021	0.319±0.020
Transformer	0.662±0.019	0.321±0.016	0.678±0.021	0.341±0.019	0.670±0.020	0.331±0.018
Gradient-boosted Trees	0.615±0.027	0.308±0.021	0.632±0.024	0.319±0.017	0.624±0.026	0.314±0.019
Wearnet [5]	0.717±0.009	0.487±0.008	–	–	–	–
Interpretable-MTLNet	**0.728±0.015**	**0.330±0.012**	**0.734±0.018**	**0.334±0.014**	**0.731±0.012**	**0.332±0.013**

5.1 Interpretability

The interpretability of our model stems directly from its use of KAN layers, which replace conventional fixed activation functions with learnable, univariate spline functions. This architecture makes

the relationship between inputs and outputs mathematically transparent and directly inspectable. As shown in Figure 2, the model identifies clinically relevant signals, such as `activity_rhythm`, `weekly_steps_cv` (weekly steps coefficient of variation), and `weekday_hr_mean` (weekday heart rate mean), as top predictors for both depression and anxiety. This visualization not only confirms that the model learns meaningful patterns but also provides a clear view of the shared physiological and behavioral signatures between the two comorbid conditions, reinforcing the value of the multi-task learning approach.

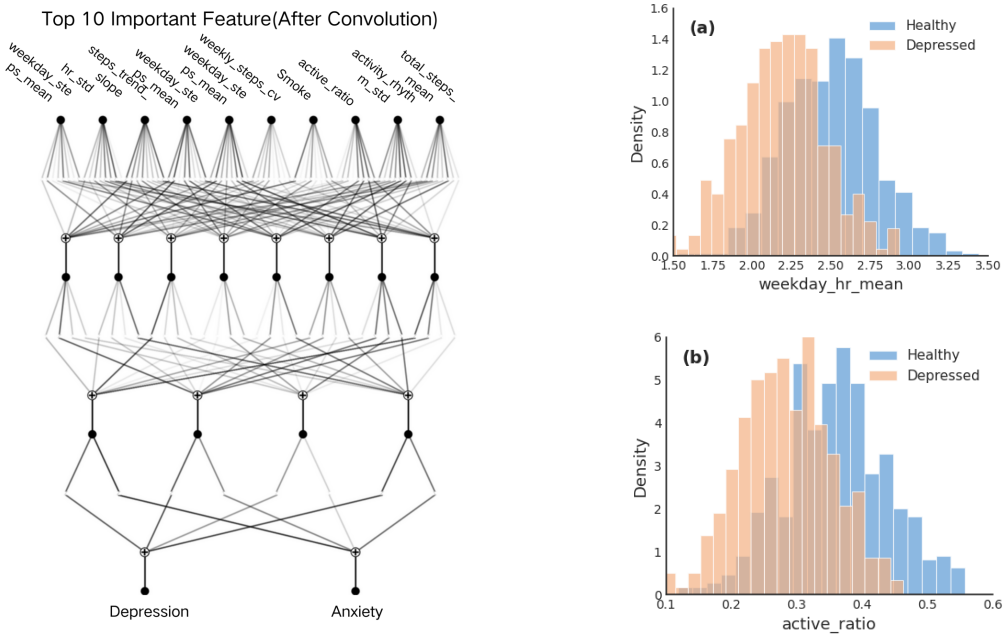

Figure 2: The network structure of **Interpretable-MTLNet** and distribution of key features. (a) Distribution of weekday heart rate mean. (b) Distribution of active ratio.

Figure 2 (a)–(b) presents two of the most informative actigraphy-derived features identified by our *Interpretable-MTLNet* model, developed for interpretable depression prediction from wearable sensor data. The KAN framework revealed *weekday_hr_mean* and *active_ratio* as the top-ranked predictors, providing physiologically and behaviorally meaningful differentiation between healthy and depressed participants. As shown in panel (a), healthy individuals exhibit higher mean weekday activity levels (approximately 2.8–3.0) relative to their depressed counterparts (approximately 2.0–2.4), reflecting attenuated diurnal rhythmicity and reduced daytime engagement among the latter. Panel (b) illustrates the *active_ratio*, representing the proportion of time spent in active movement states, which is similarly lower in the depressed cohort (approximately 0.25–0.35) compared to healthy controls (approximately 0.35–0.45). These patterns align with established clinical observations of psychomotor slowing, fatigue, and diminished behavioral activation in depressive disorders. By decomposing nonlinear relationships into interpretable univariate functional mappings, the KAN model highlights how specific behavioral rhythms contribute to prediction outcomes, offering a transparent link between model inference and clinically recognized symptoms. This interpretability advances the potential of actigraphy-based systems to provide clinically actionable insights for adolescent depression monitoring and relapse risk assessment.

5.2 Ablation Study

To validate our model's design, we conducted an ablation study that confirmed the importance of its key components. As shown in Table 2, the results demonstrated that leveraging temporal data from wearables is critical, as removing these features led to the most significant performance drop of over 20% in AUROC. The multi-scale convolutional approach also proved vital; using only a single scale resulted in a notable performance decrease. Furthermore, the MTL framework was shown to be

beneficial, as a single-task version of the model performed less effectively than the joint prediction model. These findings collectively underscore the effectiveness of our architecture's design choices. After multiple trials, we set the task weights to $w_{\text{dep}} = 1.2$ and $w_{\text{anx}} = 0.8$, which yielded the best validation performance and were kept fixed for testing.

Table 2: Ablation Study of **Interpretable-MTLNet** Components on the *All of Us* Test Set.

Model Variant	**AUROC**	**AUC-PR**	**Δ vs. Full Model (AUROC)**
Interpretable-MTLNet (Full Model)	0.731 ± 0.008	0.332 ± 0.003	–
(A) Single-Task (ST) Learning	0.715 ± 0.010	0.330 ± 0.015	-2.2%
(B) MLP-based	0.720 ± 0.009	0.329 ± 0.007	-1.5%
(C) Single-Scale Conv ($k = 5$)	0.697 ± 0.007	0.305 ± 0.008	-4.7%
(D) No Temporal Features (static only)	0.578 ± 0.025	0.009 ± 0.018	-20.9%

6 Conclusion

Our work demonstrates that interpretable neural architectures can achieve state-of-the-art performance without compromising transparency in mental health applications. The proposed *Interpretable-MTLNet* attains an AUROC of 0.731 while enabling clinician-inspectable spline-based activations, providing direct visibility into the contribution of behavioral and physiological predictors. Its multi-scale, multi-task design captures cross-disorder dependencies by leveraging shared circadian and activity-related dynamics to improve joint prediction of depression and anxiety. Beyond performance gains, this approach establishes a critical bridge between modern deep learning and clinical interpretability, demonstrating that transparent architectures can yield meaningful insights into symptom expression and comorbidity structure. The spline-based representation allows practitioners to visualize individualized feature–outcome relationships, facilitating model auditing, trust calibration, and hypothesis generation for personalized interventions.

Future research will extend this framework along several axes: (*i*) integrating adaptive personalization layers to capture inter-individual variability in behavioral rhythms; (*ii*) incorporating fairness-aware debiasing to ensure equitable performance across demographic subgroups; and (*iii*) generalizing the model to additional psychiatric and somatic conditions through multimodal fusion of actigraphy, physiological, and ecological momentary assessment data. Collectively, these directions advance interpretable machine learning toward clinically deployable, human-centered decision support tools for mental health care.

7 Acknowledgment

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.

References

[1] Mai Ali, Christopher Lucasius, Tanmay Pranav Patel, Madison Aitken, Jacob Vorstman, Peter Szatmari, Marco Battaglia, and Deepa Kundur. A multi-task LLM framework for multimodal speech-based mental health prediction. In *IEEE-EMBS International Conference on Body Sensor Networks 2025*, 2025.

[2] All of Us Research Program Investigators, Joshua C Denny, Joni L Rutter, David B Goldstein, Anthony Philippakis, Jordan W Smoller, Gwynne Jenkins, and Eric Dishman. The "All of Us" research program. *N Engl J Med*, 381(7):668–676, August 2019.

[3] C. Beard, A. J. Millner, M. J. C. Forgeard, E. I. Fried, K. J. Hsu, M. T. Treadway, C. V. Leonard, S. J. Kertz, and T. Björgvinsson. Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine*, 46:3359–3369, 2016.

[4] Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, and Santiago Pourteau. Convolutional kolmogorov-arnold networks, 2025.

[5] Ruixuan Dai, Thomas Kannampallil, Seunghwan Kim, Vera Thornton, Laura Bierut, and Chenyang Lu. Detecting mental disorders with wearables: A large cohort study. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, IoTDI '23, page 39–51, New York, NY, USA, 2023. Association for Computing Machinery.

[6] Ivan Drokin. Torch conv kan. `https://github.com/IvanDrokin/torch-conv-kan`, 2024. Accessed: 2025-08-05.

[7] Zhentao Huang, Yahong Ma, Rongrong Wang, Weisu Li, and Yongsheng Dai. A model for eeg-based emotion recognition: Cnn-bi-lstm with attention mechanism. *Electronics*, 12(14), 2023.

[8] Ajay Surya Jampana, Mohitha Velagapudi, Neethu Mohan, and Sachin Kumar S. Exploring kolmogorov arnold networks for interpretable mental health detection and classification from social media text. In Sobha Lalitha Devi and Karunesh Arora, editors, *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 206–214, AU-KBC Research Centre, Chennai, India, December 2024. NLP Association of India (NLPAI).

[9] Ned H. Kalin. The critical relationship between anxiety and depression. *American Journal of Psychiatry*, 177:365–367, 2020.

[10] William Knottenbelt, Zeyu Gao, Rebecca Wray, Woody Zhidong Zhang, Jiashuai Liu, and Mireia Crispin-Ortuzar. Coxkan: Kolmogorov-arnold networks for interpretable, high-performance survival analysis, 2024.

[11] Jaehoon Ko, Somin Oh, Doljinsuren Enkhbayar, Jin-kyung Lee, Moo-Kwon Chung, Taeksoo Shin, Min-Hyuk Kim, Hyo-Sang Lim, Erdenebayar Urtnasan, and Jaehong Key. Interpretable Feature Selection and Hybrid Deep Learning Models for Major Depressive Disorder Prediction from Wearable Device Data. *Research Square*, jun 2025. PREPRINT (Version 1).

[12] Yuanhang Li, Shuo Liu, Jie Wu, Weichao Sun, Qingke Wen, Yibiao Wu, Xiujuan Qin, and Yanyou Qiao. Multi-scale kolmogorov-arnold network (kan)-based linear attention network: Multi-scale feature fusion with kan and deformable convolution for urban scene image semantic segmentation. *Remote Sensing*, 17(5), 2025.

[13] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Christopher Lucasius, Mai Ali, Tanmay Patel, Deepa Kundur, Peter Szatmari, John Strauss, and Marco Battaglia. A procedural overview of why, when and how to use machine learning for psychiatry. *Nature Mental Health*, 3(1):8–18, 2025.

[15] Rutvik V. Shah, Gillian Grennan, Mariam Zafar-Khan, Fahad Alim, Sujit Dey, Dhakshin Ramanathan, and Jyoti Mishra. Personalized machine learning of depressed mood using wearables. *Translational Psychiatry*, 11(1):338, 2021. Published online 2021-06-09.

[16] Cristian J. Vaca-Rubio, Luis Blanco, Roberto Pereira, and Màrius Caus. Kolmogorov-arnold networks (kans) for time series analysis, 2024.

[17] Zhuoqin Yang, Jiansong Zhang, Xiaoling Luo, Zheng Lu, and Linlin Shen. Medkan: An advanced kolmogorov-arnold network for medical image classification, 2025.

Appendix A Evaluation Metrics Details

For task $t \in \{\texttt{dep}, \texttt{anx}\}$ with N_t samples, labels $\{y_i^{(t)}\}_{i=1}^{N_t} \in \{0,1\}$ and prediction scores $\{s_i^{(t)}\}_{i=1}^{N_t} \in [0,1]$, we compute per-task AUROC (A_t) and AUC-PR (P_t) in the standard way. The aggregated metrics are:

$$\text{Macro-AUROC} = \frac{1}{K}\sum_{t=1}^{K} A_t, \qquad \text{Macro-AUC-PR} = \frac{1}{K}\sum_{t=1}^{K} P_t, \quad (K{=}2) \tag{8}$$

$$\text{Micro-AUROC} = \text{AUROC}\big(\{(y_i^{(t)}, s_i^{(t)})\}_{t,i}\big), \tag{9}$$

$$\text{Micro-AUC-PR} = \text{AUC-PR}\big(\{(y_i^{(t)}, s_i^{(t)})\}_{t,i}\big). \tag{10}$$

Macro gives each task equal weight regardless of sample size, while **Micro** pools all predictions across tasks before computing metrics, effectively weighting by N_t. Due to class imbalance, the no-skill AUC-PR baseline equals the positive prevalence $\pi_t{=}\frac{1}{N_t}\sum_i y_i^{(t)}$ (5.87% for depression, 8.03% for anxiety); the macro baseline is $\frac{1}{2}(\pi_{\texttt{dep}}{+}\pi_{\texttt{anx}})$. All reported values represent mean $\pm$ standard deviation over five training runs with different random seeds, evaluated on the same held-out test participants. Macro values in Table 1 are computed as the arithmetic mean of per-task results, e.g., $\frac{1}{2}(0.728{+}0.734) = 0.731$ for AUROC.

Appendix B Implementation Details

Data Preprocessing and Splits. We perform subject-level splits with 70%/10%/20% allocation for train/validation/test sets to avoid data leakage across multiple days from the same participant. Continuous features are standardized using z-score normalization with training-set statistics (mean and standard deviation). Missing values are imputed using the training-set median, and paired binary missingness indicators are retained as additional features to preserve information about data availability patterns.

Model Architecture. The multi-scale KAN convolution module employs three parallel branches with kernel sizes $\{3, 5, 7\}$ to capture temporal patterns at different scales. After concatenation and temporal pooling, this yields $d_h{=}64$ total channels. The shared KAN mixing layer consists of one hidden layer with width 64, enabling cross-feature interactions. Task-specific heads are implemented that produce the final predictions for depression and anxiety classification.

Training Configuration. We use the AdamW optimizer with learning rate $1{\times}10^{-3}$ and weight decay $1{\times}10^{-4}$ for regularization. Training is conducted with batch size 128 and employs early stopping based on validation Macro-AUROC with a patience of 10 epochs to prevent overfitting. The focal loss function is configured with hyperparameters $(\alpha, \gamma){=}(0.25, 2.0)$ for both tasks to address class imbalance. All reported metrics are computed on the frozen held-out test set after model selection.