







Multitask Learning with Kolmogorov–Arnold Networks for Mental Health Comorbidity Prediction

Interpretable Mental Health

Farah Alshiha
farah.shiha@kaust.edu.sa

Mentor: Deepa Kundur
dkundur@ece.utoronto.ca

Mentor: Mai Ali
maia.ali@mail.utoronto.ca

August 27, 2025

Abstract

The World Health Organization estimates that depression affects one in seven adolescents [1] and is often comorbid with anxiety and sleep disturbance [2], underscoring the need for predictive tools that capture these complex relationships. While multilayer perceptrons (MLPs) have been effective in such prediction tasks, their black-box nature limits clinical trust. To address this, we evaluated a multitask Kolmogorov–Arnold Network (KAN) trained on Fitbit activity and sleep data from approximately 700 students in the NetHealth Project, comparing its performance to that of MLPs. Results showed that the multitask MLP achieved higher balanced accuracy, ROC AUC, and PR AUC (e.g., 68.30, 75.40, 58.20) than the multitask KAN (53.20, 60.78, 41.69). Although the KAN underperformed in predictive accuracy, it provided interpretable symbolic formulas linking wearable-derived features to mental health outcomes, offering potential value for transparent clinical decision-support.

Table of Contents

Abstract **1**

1 Introduction **3**
1.1 Problem Motivation 3
1.2 Key Assumptions 3
1.3 Objective 3

2 System Overview **3**

3 Methodology **4**
3.1 Tools, Libraries, or Frameworks Used 4
3.2 Dataset 4
3.3 Data Preprocessing 4
3.4 Loss Function for Multitask Learning 6
3.5 Kolmogorov–Arnold Network 6

4 Results Discussion **7**
4.1 Performance Metrics 7
4.2 Symbolic Formula 8

5 Reflections on Challenges **8**

6 Conclusion and Future Work **9**

1 Introduction

1.1 Problem Motivation

Comorbid mental health conditions such as depression, anxiety, and sleep disturbance often occur together within individuals, making comorbidity the rule rather than the exception in clinical practice. Traditional predictive models, such as multilayer perceptrons (MLPs), can achieve strong diagnostic performance, particularly when applied in multitask frameworks that exploit shared representations across related conditions [3]. However, their black-box nature limits interpretability, meaning clinicians cannot easily discern how predictions are made. This lack of transparency compromises trust and accountability, especially within a clinical setting, thus underscoring the need for models that combine predictive accuracy with interpretability.

1.2 Key Assumptions

There remains a gap between the prevalence of depression in students and formal diagnosis. Many students experiencing depression are unlikely to actively seek professional help, which highlights the importance of passive and accessible diagnostic tools. This study follows approximately 700 university students, collecting longitudinal data on physical activity and sleep patterns through Fitbit wearables. These devices provide a non-intrusive way to capture behavioral and physiological measures that may serve as predictors of potential mental health problems. While this approach may seem promising, it carries limitations, namely the noise contained in wearable data, as well as inconsistent self-reported labels. This may introduce bias into the dataset, and accurately modeling the correlation among comorbidities will pose as a challenge.

1.3 Objective

The objective of this project is to design a multitask predictive system that enhances interpretability without sacrificing performance. By using Kolmogorov–Arnold Networks (KANs) [4], we aim to extract symbolic formulas that reveal the relationships between wearable-derived features and mental health outcomes. This reframes the problem as: How can we design a multitask learning model for comorbidity prediction that addresses the black-box limitations of standard neural networks while maintaining predictive accuracy? To this end, we propose a multitask KAN framework trained on wearable and survey data to jointly predict depression, anxiety, and sleep disturbance. The system's performance is evaluated by balanced accuracy, ROC AUC, and PR AUC.

2 System Overview

The KAN model was built on the basis of a multitask learning framework, predicting three comorbid conditions simultaneously. The dataset contains 26 features. The interpretability of the model lies in the extraction of its symbolic formula, one derived for each task, which shows the functional contribution of the input features. Figure 1 illustrates the main workflow of the system, from data acquisition to model prediction.

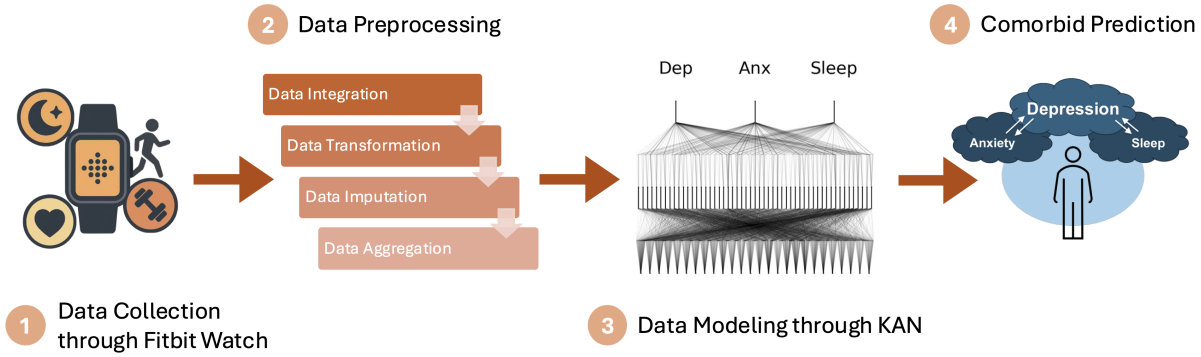


Figure 1: Overview of the multitask KAN system.

3 Methodology

3.1 Tools, Libraries, or Frameworks Used

The implementation of this study was written in Python using the official pykan library, alongside PyTorch, Scikit-learn, Seaborn, and Matplotlib. Data preprocessing was handled with Pandas, NumPy, and Imbalanced-learn.

3.2 Dataset

The dataset was derived from the NetHealth Project [5], combining Fitbit activity and sleep logs with longitudinal survey-based labels for mental health problems. This study focuses on depression, anxiety, and sleep disturbance. Participant data spans from August 2015 (orientation) through spring 2019, though most analyses focus on data up to May 2017 due to wear compliance and academic term consistency. The dataset is imbalanced, with depression positives comprising 28.7%. There are roughly 357,827 records in total. Fitbit features include:

- Steps and floors climbed
- Minutes spent in activity zones: sedentary, light, fairly, and very active
- Minutes in heart-rate zones (low, fat-burn, cardio, peak) and corresponding zone-wise calories burned
- Derived metrics: mean daily heart rate, standard deviation, and compliance (% of day wearable worn)

3.3 Data Preprocessing

The dataset integrates two primary modalities from a wearable Fitbit device: daily activity data (e.g., step counts, calories, heart rate) and sleep data (e.g., total sleep time, bedtime, wake-up time). To link these continuous streams of wearable data with mental health outcomes, survey responses were used as ground-truth labels. Each individual completed periodic surveys, known as survey waves, conducted no more than six months apart. For labeling, each Fitbit record was aligned with the closest following survey date, ensuring that the input features corresponded to the most clinically relevant mental health

outcome. This temporal alignment strategy reflects how wearable signals precede and potentially predict survey-reported symptoms. Figure 2 presents the labeling procedure, where each bubble represents a survey wave, and the vertical arrow represents a row of Fitbit measurements aggregated over a day for one individual.

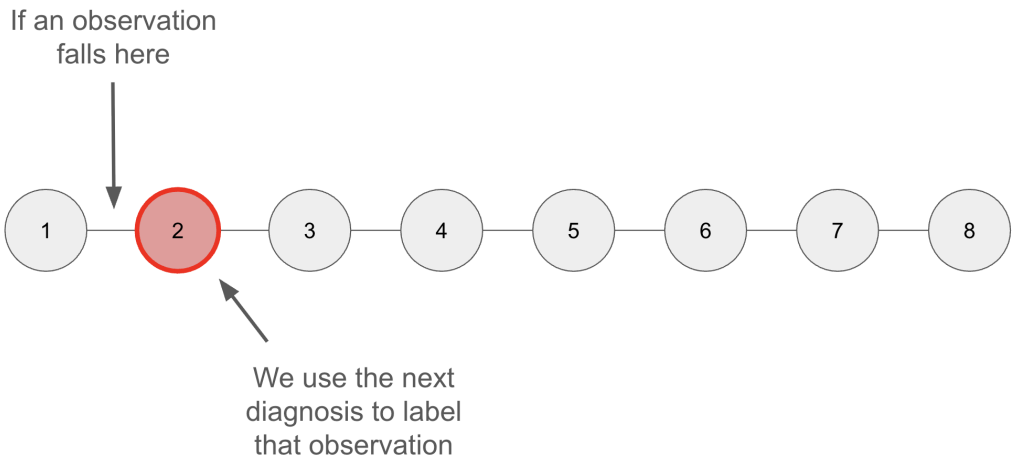


Figure 2: Labeling Procedure

Like most longitudinal wearable datasets, many entries were missing due to compliance issues. To address this, missing values were imputed using Bayesian Ridge regression, which estimates missing entries as conditional distributions informed by other features. Bayesian Ridge was benchmarked against other imputation strategies and demonstrated more consistent results in metrics like RMSE, MAE, and R^2. Evaluation of the different imputation strategies can be seen in Figure 3.

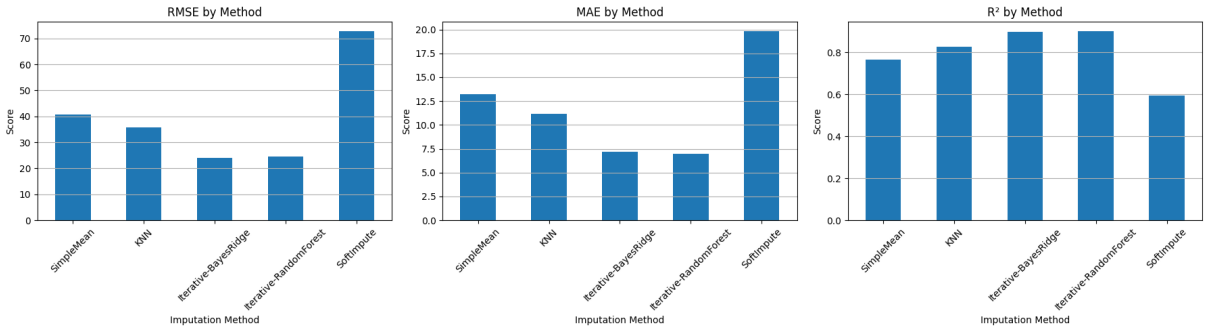


Figure 3: Labeling Procedure

Class imbalance posed a significant challenge, particularly for depression, where the positive class accounted for just over one quarter of the samples (Figure 4). To counter this, multiple balancing strategies were explored. Variants of SMOTE (Synthetic Minority Oversampling Technique) were applied to generate synthetic samples of the minority class. Downsampling the majority class was also tested, though at the cost of discarding data and severely hurting model performance. Finally, pos_weight adjustments were applied directly in the loss function (BCEWithLogitsLoss), scaling errors from the minority class more heavily. Among these, SMOTE achieved the best performance gains.

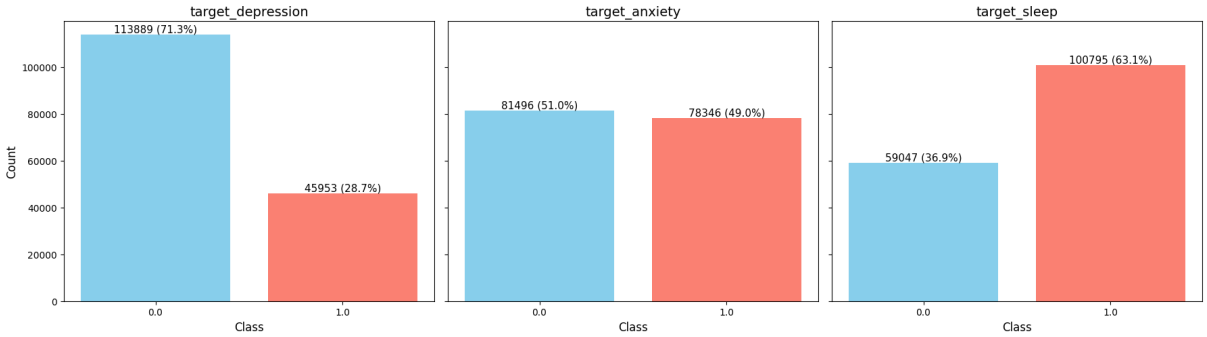


Figure 4: Class distribution for Tasks

Because each individual contributed multiple days of Fitbit records, the raw dataset contained tens of thousands of rows. To align these with survey outcomes, rows were aggregated by unique participant IDs and survey waves. This ensured that all wearable features within a given time window leading up to a survey were summarized into one fixed representation. Features were aggregated using the mean statistic, which reduced the dataset from 333,289 rows to just 1,261 rows. This also accelerated KAN training, which is roughly ten times slower than MLP training.

3.4 Loss Function for Multitask Learning

The design of the loss function is central to the multitask framework adopted in this paper. Standard approaches often employ an unweighted sum of binary cross-entropy (BCE) losses across tasks, but this can lead to suboptimal results when tasks are imbalanced. To address this, we implemented a weighted multitask loss inspired by the paper *Speech as a Multimodal Digital Phenotype for Multi-Task LLM-based Mental Health Prediction* [6]:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{depression}} + \lambda\big(\mathcal{L}_{\text{anxiety}} + \mathcal{L}_{\text{sleep}}\big). \tag{1}$$

This formulation allows the model to place various amounts of emphasis on depression prediction while still giving weight to auxiliary signals from related comorbidities. By varying λ, we explored trade-offs between tasks to better understand the balance between predictive performance and shared feature learning. λ=1 consistently outperformed models with λ=0.25, 5, and 0.75 across various tests, with both the KAN and MLP models.

3.5 Kolmogorov–Arnold Network

KANs are a recently proposed neural network architecture inspired by the Kolmogorov–Arnold representation theorem, which states that any multivariate continuous function can be expressed as a composition of univariate functions and addition [4]. Unlike multilayer perceptrons (MLPs), which rely on fixed nonlinear activation functions and learn weights on connections between neurons, KANs parameterize the activation functions themselves. This is typically achieved using splines or other basis functions that adapt during training.

The advantage of this approach is twofold. Firstly, it supports interpretability by extracting symbolic expressions from the learned univariate functions, providing human-readable formulas that highlight feature contributions. Secondly, it is flexible by learning activation

functions rather than fixing them, which allows KANs to approximate complex nonlinearities with fewer hidden units. Table 1 outlines the hyperparameter settings of both the KAN and MLP models.

Table 1: Comparison of hyperparameters for KAN and MLP models

Setting	**KAN (Proposed)**	**MLP (Baseline)**
Input dimension	26 features	26 features
Output dimension	3 tasks (Dep/Anx/Sleep)	3 tasks (Dep/Anx/Sleep)
Hidden layers	1	3
Width per layer	53	128, 64, 32
Basis functions / Splines	Grid = 5, k = 3	N/A
Optimizer	LBFGS	Adam
Learning rate	1	1e-3
Batch size	Full	128
Epochs / Steps	20	40
Regularization	λ, λ_{l1}, $\lambda_{entropy}$ = 0.01, 1.0, 2.0	Dropout = 0.2, L2 = 1e-4

4 Results Discussion

4.1 Performance Metrics

The results highlight a clear performance gap between the two model architectures. The multilayer perceptron (MLP) outperformed the Kolmogorov–Arnold Network (KAN) across both single-task and multitask settings. As shown in Table 2, the single-task MLP achieved a balanced accuracy of 0.660, compared to 0.503 for KAN. In the multitask setting, both models improved slightly, but the MLP again led with 0.683, surpassing 0.532 for KAN.

Table 2: Balanced Accuracy Results for Single-Task and Multitask Models

Model	**Single-Task**	**Multitask**
MLP	0.660	0.683
KAN	0.503	0.532

This suggests that while multitask learning provided a modest performance gain for both models—likely due to the shared representation of depression, anxiety, and sleep disturbance—the KAN struggled to match the predictive strength of the MLP. The underperformance of KANs may stem from their relatively recent development, sensitivity to hyperparameter tuning, and the complexity of adapting spline-based activations to the noisy and imbalanced wearable dataset.

Table 3: Comparison of Multitask MLP and KAN Performance

Task	Model	Balanced Acc	ROC AUC	PR AUC
Depression	MLP	0.683	0.754	0.582
	KAN	0.532	0.608	0.417
Anxiety	MLP	0.652	0.715	0.708
	KAN	0.621	0.627	0.580
Sleep Dist.	MLP	0.633	0.694	0.806
	KAN	0.559	0.618	0.680

Across all three tasks, the multitask MLP model consistently outperforms the KAN in balanced accuracy, ROC AUC, and PR AUC. While the KAN provides interpretability through its symbolic formulas, its predictive power remains weaker, possibly due to sensitivity to hyperparameter tuning. MLPs demonstrate stronger discriminative ability, particularly in ROC AUC and PR AUC, which are crucial under conditions of imbalance. This confirms a key trade-off, where MLPs excel in accuracy, while KANs are better suited for interpretability. For clinical settings, a hybrid approach (e.g., using KANs for explainability while relying on MLPs for prediction) could be a valuable future direction.

4.2 Symbolic Formula

$$
\begin{aligned}
f(x) = {} & 0.00807\,x_{10} + 0.00240\,x_{11} - 0.00115\,x_{12} + 0.00310\,x_{13} + 0.01678\,x_{14} - 0.00108\,x_{16} \\
& + 0.01173\,x_{17} + 0.00601\,x_{19} - 0.00452\,x_{2} - 0.00049\,x_{22} - 0.00013\,x_{23} \\
& + 0.00456\,x_{25} - 0.00523\,x_{26} + 0.00044\,x_{3} + 0.00601\,x_{4} + 0.01249\,x_{5} \\
& - 0.00105\,x_{6} + 0.00777\,x_{7} + 0.00137\,x_{8} - 0.00433\,x_{9} \\
& - 3.6577\,e^{0.01081\,x_{12} - 0.01453\,x_{23}} \\
& + 2.1396\,e^{-0.02504\,x_{23} + 0.03202\,x_{8}} \\
& - 1.0183\,e^{0.00172\,x_{13} - 0.00083\,x_{23} - 0.00229\,x_{8}} \\
& + 2.1096 \times 10^{-6}\,e^{-0.01352\,x_{12} + 0.02090\,x_{13} + 0.02139\,x_{22} - 0.01775\,x_{23} + 0.00624\,x_{8}} \\
& + 0.1375\,e^{0.00316\,x_{12} + 0.00738\,x_{13} - 0.00726\,x_{23} - 0.00749\,x_{6} - 0.00835\,x_{8}} \\
& - 0.00025\,e^{0.00700\,x_{13} + 0.00705\,x_{16} + 0.00905\,x_{22} - 0.00780\,x_{23} - 0.00613\,x_{8}} \\
& - 0.7105 - 4.8213 \times 10^{-356}\,e^{-0.11278\,x_{23}}
\end{aligned}
$$

The expression above is one of the three extracted symbolic formulas showing how specific Fitbit-derived features combine to predict depression. The coefficients indicate the strength and direction of each feature's influence, while the exponential terms capture nonlinear effects. Unlike black-box models, this symbolic formula makes the decision process transparent and clinically interpretable.

5 Reflections on Challenges

The main challenges of this project stemmed from both the dataset and the model architecture. The dataset was highly imbalanced, particularly for the depression label where positive cases accounted for only a minority of the population. This imbalance created difficulties in training and evaluation, as models tended to overpredict the majority class

and required additional techniques to compensate. On the modeling side, KANs presented their own challenges. As a relatively new architecture, KANs have limited documentation, scarce practical implementations, and evolving best practices. This lack of footing made experimentation and tuning more difficult compared to the more mature multilayer perceptron framework. Together, these factors accounted for the major challenge of working with complex, imperfect real-world data while also adapting a novel, less-documented model paradigm.

6 Conclusion and Future Work

This project demonstrates that multitask KANs have the potential to predict comorbid mental health conditions while overcoming the black-box problem through its symbolic formula extraction. The extracted formula highlights key risk factors, enhancing the clinical applicability of machine learning in mental health diagnostics. However, limitations remain in the training instability of KANs through their long training time and high sensitivity to hyperparameter-tuning, as well as acquiring balanced datasets with activity and sleep measurements obtained from a wearable device. Future work will explore hybrid architectures between MLPs and KANs, larger datasets across populations, and user-centered evaluations with clinicians to assess interpretability in practice.

References

[1] Institute for Health Metrics and Evaluation. *Global Health Data Exchange (GHDx)*. https://vizhub.healthdata.org/gbd-results/. Accessed: 9 July 2025. 2025.

[2] J. M. Valderas et al. "Defining Comorbidity: Implications for Understanding Health and Health Services". In: *The Annals of Family Medicine* 7.4 (July 2009), pp. 357–363. DOI: 10.1370/afm.983. URL: https://doi.org/10.1370/afm.983.

[3] Rich Caruana. "Multitask Learning". In: *Learning to Learn*. Ed. by Sebastian Thrun and Lorien Pratt. Boston, MA: Springer, 1998, pp. 95–133. DOI: 10.1007/978-1-4615-5529-2_5. URL: https://doi.org/10.1007/978-1-4615-5529-2_5.

[4] Ziming Liu et al. "KAN: Kolmogorov–Arnold Networks". In: *arXiv preprint arXiv:2404.19756* (2024).

[5] University of Notre Dame. *NetHealth*. https://sites.nd.edu/nethealth/. Accessed: 9 July 2025. 2015.

[6] Mai Ali, Christopher Lucasius, and Patel Tanmay P. "Speech as a Multimodal Digital Phenotype for Multi-Task LLM-based Mental Health Prediction". In: (2025).