Leveraging Periodicity for Robustness with Multi-modal Mood Pattern Models

Jaya Narain[*]
Apple
jnarain@apple.com

Jenny Qinhua Sun*[†]
University of California, Irvine
qinhuas@uci.com

Oussama Elachqar
elachqaroussama@gmail.com

Haraldur Hallgrimsson
Apple
hth@apple.com

Feng Zhu
Apple
feng_zhu2@apple.com

Shirley Ren
Apple
shirleyr@apple.com

Abstract

Data from wearable sensors (e.g., heart rate, step count) can be used to model mood patterns. We characterize feature representations and modeling strategies with multi-modal discrete time series data for mood pattern classification with a large dataset with naturalistic missingness (n=116,819 participants) using 12 wearable data streams, with a focus on capturing periodic trends in data. Considering both performance and robustness, periodicity-based aggregate feature representations with gradient boosting models outperformed other representations and architectures studied. The use of periodic features improved the model performance compared to temporal statistics, and gradient boosting models were more robust to dataset size and shifts in missingness distributions than a deep learning time series model.

1 Introduction and Related Work

Data collected from mobile and wearable devices can provide insights into mood patterns. Prior approaches to mood pattern modeling with data from wearables often use handcrafted features with temporal aggregations including multi-week averages, daily averages, and multi-hour epochs Sano et al. [2018], Xu et al. [2019, 2023], Wahle et al. [2016], Chikersal et al. [2021], Sefidgar et al. [2019]. In prior work, models were trained using data from tens Wang et al. [2018], Wahle et al. [2016], or hundreds Xu et al. [2023], Sano et al. [2018], Chikersal et al. [2021], Sefidgar et al. [2019] of participants.

We train and evaluate mood pattern models with a large (n=413,749 samples from 116,819 participants) multi-modal time series dataset with self-reported mood pattern labels from surveys around depression and anxiety symptoms. We focus on feature representations and modeling strategies that can capture periodic trends in the data, which have been shown previously to have signal for mood related models (e.g., sleep routine consistency features in Sano et al. [2018]) and for time series models more generally including models for COVID-19 using data from wearables Klein et al. [2023], Merrill and Althoff [2023] and models of motor development with motion data Li et al. [2021]. We investigate the impact of missingness and dataset size – factors that can significantly influence model performance Barda et al. [2020], Jungo et al. [2024], Che et al. [2018], Yıldız et al. [2022], Hashimoto [2021], Kaplan et al. [2020] – to better understand how data representation strategies intersect with real-world training considerations.

[*] Equal Contribution
[†] Work done during Apple internship

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

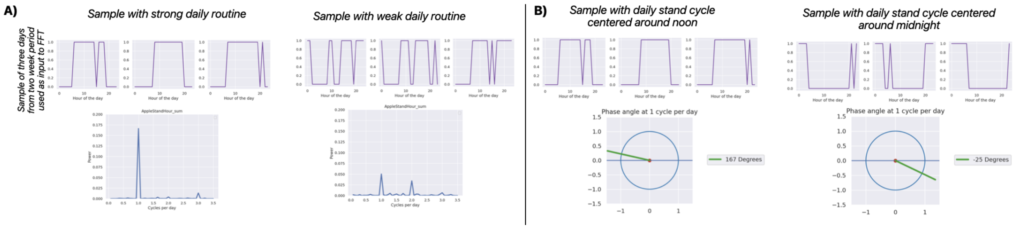

Figure 1: A) The power component is related to routine regularity. B) The phase component is related to routine timing.

2 Dataset

We used an ambulatory longitudinal dataset collected from an IRB-approved study. Data collection included collecting physiological and activity-related data streams from the iPhone and Apple Watch. Participants volunteered for the study and provided informed consent before enrollment. Participants control what data is shared with the study, and can un-enroll at any time. Participants included in this analysis responded to four survey questions on mood patterns and had data available for the utilized streams.

The dataset used in this analyses included 12 multi-modal data streams: heart rate (HR), resting HR, walking HR, HR variability (HRV), walking speed, flights climbed, exercise time, step count, stand time, standing hours, sleep time, and active energy burned (documentation at Apple Developer Documentation and details in A.1. Data (n=413,749 samples from 116,819 participants) was split into train, validation, and test splits with a 70/15/15 split, stratified by user.

Study participants take a survey four times a year that includes four questions related to mood patterns – two questions regarding selected anxiety mood patterns from the General Anxiety Disorder (GAD) questionnaire Spitzer et al. [2006] and two questions regarding selected depression mood patterns from the Patient Health Questionnaire (PHQ) Kroenke et al. [2009]. Labels were converted to binary high/low labels for depression mood patterns and anxiety mood patterns (see A.1). Note that participants' responses to the questions capture information about specific patterns related to mood but do not capture clinical diagnoses.

3 Data representations

Each data stream was summarized across hourly and daily time windows, covering up to two weeks of data preceding a survey. Resting HR, walking HR, HR, HRV, and walking speed were summarized by the average in each time window (e.g., average daily or hourly HR) and the remaining data streams were summarized by the sum in each time window (e.g., total daily or hourly steps). Time series models used hourly and daily summary values directly.

Additional statistics were then calculated from daily and hourly values for each data stream (A.1. Temporal statistics (average, median, minimum, maximum, variance, skew, and kurtosis) were calculated from the daily feature summaries for both 7- and 14-day periods preceding each survey. To capture weekly periodic structures, weekend-weekday averaged statistics were also calculated across daily summary values. To capture periodic trends, 24-hour per hour features (e.g., average step count from 9-10 AM in the two weeks preceding the survey) and FFT-based features (Figure 1) were calculated from the hourly summary values. The FFT yields information about power and phase, related to routine regularity and timing. Figure 1 shows dataset samples for the stand hour feature, illustrating the relationship between routine strength and power and between routine timing and phase. The periodic combined feature set included 24-hour averages along with specific frequency and phase features that were selected for their interpretability (e.g., at frequencies representing daily and weekly routines). Demographic features (age, gender, and BMI) were analyzed as a baseline.

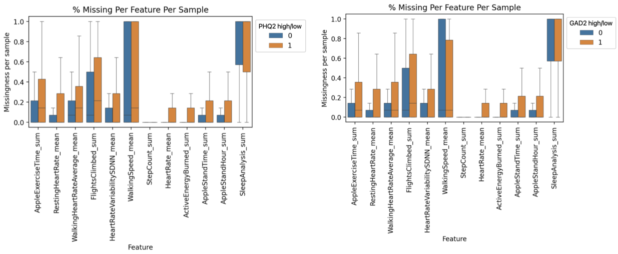

Figure 2: Daily missingness for each recorded data type, stratified by binary mood pattern labels

4 Methods

Model Architectures Binary classifiers were trained independently for PHQ2 mood patterns and GAD2 mood patterns – gradient boosting trees for aggregate feature representations and 1D convolutional networks for the time series representations. 1D convolutional networks have shown strong performance with other time series tasks Abbaspourazad et al. [2023], Thapa et al. [2024], Gopal et al. [2021] and have the capacity to learn kernels that capture periodic trends. The 1D convolutional network with daily data had 6,105 trainable parameters and the network used with hourly data had 1,610,329 trainable parameters. Network structure and size was tuned on the validation data, and larger networks were also explored. Architectural details are provided in A.2

Missingness and Dataset Size The dataset missingness is representative of user habits for the enrolled population of volunteer participants in the longitudinal study (Figure 2) To evaluate the impact of data missingness and model performance, two missingness filters were applied to the datasets (A.2): a lax filter (removing data with missingness at or above 75th percentile) and a stringent filter (removing at or above 25th percentile missingness). To explore the impact of dataset size on model performance, the dataset was randomly downsampled at each missingness level from 10% to 90% of the original dataset, in 10% increments. Downsampling was stratified by label.

Downsampling and missingness filters were applied to four representations: the demographics baseline, 14-day averaged temporal statistics, periodic combined features, and hourly time series. These representations were chosen to span representation and model types, and because of their performance with the full dataset. Each combination of missingness level and downsampling level was evaluated on three test datasets: a test dataset sampled using the same missingness and downsampling parameters as the training dataset, the full test dataset at the same missingness level as the training dataset (the "full in-distribution" dataset), and the full test dataset with no missingness filters or downsampling (the "full dataset").

Table 1: Best performing representation for each model architecture highlighted in bold. The number of statistics describing each data stream (e.g., walking speed) is shown for each representation

Data representation	Num. features/ data stream	Model	AUC	
			PHQ2	GAD2
Demographics	3 total	XGBoost	0.639	0.696
14-day statistics	7	XGBoost	0.695	0.696
7-day statistics	7	XGBoost	0.692	0.691
Weekend-weekday average	14	XGBoost	0.677	0.682
Daily tabular	14	XGBoost	0.687	0.689
Hourly tabular	336	XGBoost	0.697	0.691
Frequency + phase	335	XGBoost	0.703	0.693
24 hour per-hour averages	24	XGBoost	0.703	0.700
Periodic combined	33	XGBoost	**0.708**	**0.705**
Daily time series	14	1DConv	0.690	0.694
Hourly time series	336	1DConv	**0.708**	**0.711**

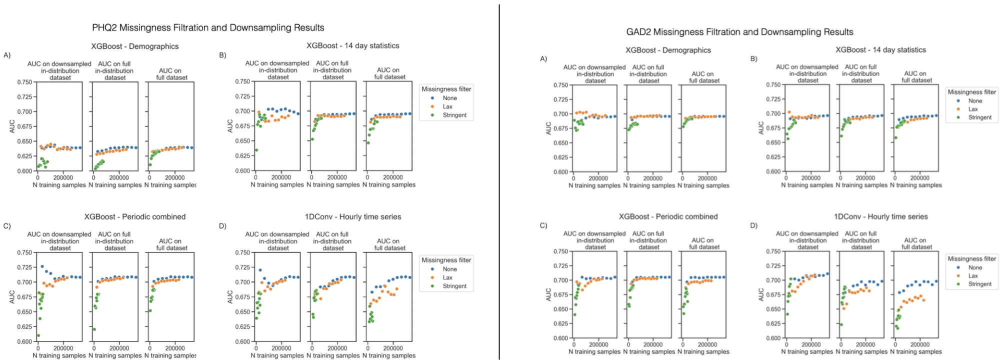

Figure 3: AUC with varying missingness and downsampled training data for selected representations

5 Results and Discussion

5.1 Full dataset results

Table 1 reports the model performance on the full dataset for each evaluated representation. Demographics were highly predictive, particularly for anxiety. The best performing aggregate representation was the periodic combined representation (Table 1), outperforming traditional temporal statistics and higher dimensional feature sets. Adding demographics to the best performing feature set improved model performance by 1 point for PHQ2 (AUC of 0.718) and by 3.3 points for GAD (AUC of 0.738), confirming that the sensor data contains some signal beyond underlying demographics.

5.2 Missingness and Downsampling Experiments

Periodic combined features had consistent, robust performance. A strength of the periodic combined representation is its strong performance on data across missingness filters – it generally had a higher AUC than 14-day statistics across dataset size and missingness level, except for with the stringent dataset filter for GAD2 (Figures 3B & C). Training with no missingness filter generally resulted in the best performance, especially for the hourly time series dataset. The higher performance of the models trained on data with no missingness filter may be due to the correlations between missingness and binary label value (Figure 2)

Small evaluation sets overestimated AUC. In-distribution downsampled data performances (leftmost plot for each representation) were anomalous for the smallest dataset sizes, often over-estimating AUC compared to the same model when evaluated on the full in-distribution dataset (middle plot for each representation). Note that the smallest evaluation dataset used in this analysis (the stringent missingness filter downsampled to 10%) had 456 samples from 396 participants, which is larger than many datasets often used for evaluating similar models.

Impact of non-representative missingness in training data. Models trained on missingness filtered data and evaluated on the full dataset had lower performance than when evaluated on in-distribution (missingness filtered data) for the 1D convolutional time series model. Such a scenario might arise if a model is trained on study data with low missingness due to participation incentives (e.g., payment for high compliance with protocols during a study) and then deployed on real-world data.

Demographics and missingness. The demographics dataset had lower performance at a given dataset size with the stringent missingness filter (Figure 3), suggesting a non-random relationship between demographics and missingness - further analyses is a topic for future work.

5.3 Limitations

Limitations of this study include the use of non-clinical self-reported binarized labels, and the limited number of architectures explored. The studied population may not represent the overall population – participants in the study had at least one qualifying device and chose to self-enroll in the study. Confounding factors that might impact data and label patterns were not explored. Future work will explore the impact of data representations for regression tasks with higher richness labels and additional modeling strategies.

5.4 Acknowledgements

Thank you to to Guillermo Sapiro for his support of this work. Thanks to Vikram Mitra and Vincent Chan for their feedback on this manuscript.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

Apple Developer Documentation. Healthkit data types. URL https://developer.apple.com/documentation/healthkit/data_types.

Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, et al. Developing a covid-19 mortality risk prediction model when individual-level data are not available. *Nature communications*, 11(1):4439, 2020.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G Creswell, Jennifer Mankoff, J David Creswell, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(1):1–41, 2021.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pages 156–167. PMLR, 2021.

Tatsunori Hashimoto. Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, pages 4107–4116. PMLR, 2021.

Janosch Jungo, Yutong Xiang, Shkurta Gashi, and Christian Holz. Representation learning for wearable-based applications in the case of missing data. *arXiv preprint arXiv:2401.05437*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Amit Klein, Varun Kumar Viswanath, Benjamin Smarr, and Edward Jay Wang. Detecting periodic biases in wearable-based illness detection models. In *ICLR 2023 Workshop on Time Series Representation Learning for Health*, 2023.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

Xinyue Li, Michael Kane, Yunting Zhang, Wanqi Sun, Yuanjin Song, Shumei Dong, Qingmin Lin, Qi Zhu, Fan Jiang, and Hongyu Zhao. Circadian rhythm analysis using wearable device data: Novel penalized machine learning approach. *Journal of Medical Internet Research*, 23(10):e18403, 2021.

Mika A Merrill and Tim Althoff. Self-supervised pretraining and transfer learning enable\titlebreak flu and covid-19 predictions in small mobile sensing datasets. In *Conference on Health, Inference, and Learning*, pages 191–206. PMLR, 2023.

Akane Sano, Sara Taylor, Andrew W McHill, Andrew JK Phillips, Laura K Barger, Elizabeth Klerman, and Rosalind Picard. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. *Journal of medical Internet research*, 20(6):e210, 2018.

Yasaman S Sefidgar, Woosuk Seo, Kevin S Kuehn, Tim Althoff, Anne Browning, Eve Riskin, Paula S Nurius, Anind K Dey, and Jennifer Mankoff. Passively-sensed behavioral correlates of discrimination events in college students. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–29, 2019.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097, 2006.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Y Zou. Sleepfm: Multi-modal representation learning for sleep across ecg, eeg and respiratory signals. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, Steffi Weidt, et al. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth*, 4(3): e5960, 2016.

Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

xgboost developers. Xgboost python package. URL https://xgboost.readthedocs.io/en/stable/python/python_api.html.

Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–33, 2019.

Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34, 2023.

A Yarkın Yıldız, Emirhan Koç, and Aykut Koç. Multivariate time series imputation with transformers. *IEEE Signal Processing Letters*, 29:2517–2521, 2022.

A Appendix

A.1 Dataset details

Data streams The 12 data streams used in the analyses were:

- Heart rate, HR (*heartRate*): heart rate
- Resting heart rate (*restingHeartRate*): heart rate while at rest
- Walking heart rate (*walkingHeartRate*): heart rate while walking
- Heart rate variability, HRV (*heartRateVariabilitySDNN*): standard deviation of NN intervals
- Walking speed (*walkingSpeed*): average speed walking steadily on flat ground
- Flights climbed (*flightsClimbed*): number of flights of stairs climbed
- Exercise time (*appleExerciseTime*): time spent exercising; any activity beyond a "brisk walk" is considered exercise
- Step count (*stepCount*): number of steps

- Stand time (*appleStandTime*): amount of time spent standing
- Stand hours (*appleStandHour*): Number of hours during the day where the wearer had some stand time; this is converted to a binary value at the hourly level
- Sleep time (*HKCategoryValueSleepAnalysis*); this is converted to a binary value at the hourly level
- Active energy burned (*appleExerciseTime*); active energy burned

Additional documentation on each feature is available at Apple Developer Documentation. The field name for each data stream as accessible during development is provided in italics.

Data representations The data representations included in the analysis were:

- **Demographic baseline**: age, gender, and BMI (self-reported by participants)
- **14-day statistics**: Temporal aggregate statistics calculated across daily values for each feature for the 14 days preceding the completed survey - average, median, minimum, maximum, variance, skew, and kurtosis; N = 7 per data stream and 84 total
- **7-day statistics**: Temporal aggregate statistics calculated across daily values for each feature for the 7 days preceding the completed survey - average, median, minimum, maximum, variance, skew, and kurtosis; N = 7 per data stream and 84 total
- **Weekend-weekday average**: Average of daily values over 7-day period beginning on the Saturday preceding the survey, split into weekend and weekday time periods; N = 14 per data stream and 168 total
- **24-hour per hour average**: Average of hourly values over 14-day period (e.g., average step count from 9-10 AM)
- **Phase and power**: Phase and frequency representation from FFT on hourly data
- **Periodic combined**: Includes 24-hour per hour average features along with phase and power features (power at 0, 0.14, 1, 2, and 3 cycles per day and phase at 0, 1, 2, and 3 cycles per day) selected for their semantic meaning.
- **Daily time series**: Time series of daily averages and sums for each feature. In XGBoost models, these were formatted as tabular data (e.g., day 1 average step count, day 2 average step count...)
- **Hourly time series**: Time series of hourly averages and sums for each feature. In XGBoost models, these were formatted as tabular data (e.g., hour 1 average step count, hour 2 average step count...)

Self-reported mood pattern labels The questions from the GAD asked users to rate (1) whether they felt nervous/anxious and (2) whether they felt unable to control or stop worrying. The questions from the PHQ asked users to rate (1) whether they had little interest or pleasure in doing things and (2) whether they felt down, depressed, or hopeless. Participants provided Likert-scale ratings for each question, selecting from the options: "Not at all", "Several days", "More than half the days", and "Nearly every day". Participants' responses were converted to numeric values, ranging from 0 for "Not at all" to 3 for "Nearly every day". For classification, a positive label for self-reported depression mood patterns was a total score ≥ 4 across both questions from the PHQ and a positive label for self-reported anxiety mood patterns was a total score ≥ 4 across both questions from the GAD.

The depression-related mood pattern dataset (PHQ2 dataset) had 331,008 samples from 93,974 participants in the training split; 41,728 samples from 11,170 participants in the validation split; and 41,013 samples from 11,675 participants in the test split. The anxiety-related mood pattern dataset (GAD2 dataset) had 331,328 samples from 94,049 participants in the training split; 41,728 samples from 11,737 participants in the validation split; and 41,052 samples from 11,676 participants in the test split. Participants were in the same split for the GAD2 and PHQ2 datasets (i.e., if a participant was in the training split in the PHQ2 dataset, then they were also in the training split in the GAD2 dataset). Small differences in numbers are due to survey completion differences for the included questions.

A.2 Methods

Feature extraction FFT-based features were extracted by taking the power of the Fourier transform of the hourly features, normalized by length.

Model architectures. The boosting models xgboost developers had a maximum depth of 3, a learning rate of 0.1, and an L2 regularization constanst of 0.1. The 1D convolutional network with daily data had 6,105 trainable parameters. The network included 4 convolutional layers of kernel size 3 followed by an MLP head. From the first layer to the final layer, the number of output channels per convolutional layer were [16, 16, 32, 32]. Each convolutional layer was followed by max pooling (kernel size 3 with stride 2), and a ReLu operation. The 1D convolutional network with hourly data had 1,610,329 trainable parameters. The network included 4 convolutional layers of kernel size 48 followed by a multilayer perceptron head (MLP). From the first layer to the final layer, the number of output channels were [64, 64, 128, 128]. Each convolutional layer was followed by max pooling (kernel size 6 with stride 3), and a ReLu operation. Early stopping was used to prevent overfitting.

Models were tuned on the full dataset with realistic missingness. A limited number of experiments were run with smaller models tuned for the smaller datasets. The experiments suggested that small (0.005 AUC improvements) improvements could be achieved with models tuned to the smaller datasets for 1DConv models. No performance differences were seen with tuning to smaller datasets for XGBoost models. Future work will explore performance and robustness trends with model hyperparameters selected systemically for each dataset size. Still, the observation of performance over-estimation 5 using a smaller dataset would still hold, and perhaps be magnified, with a model tuned with the smaller dataset.

Missingness Filters The filters were applied based on the missingness distributions in the training set for three features: heart rate, step count, and sleep. These features were chosen because they encompass device usage habits across devices and wake and sleep periods.

The filters were created and applied separately for the PHQ2 and GAD2 datasets. The PHQ2 dataset with the lax missingness filter applied had 255,040 training samples from 77,320 subjects; 32,128 validation samples from 9,745 subjects; and 31,581 test samples from 9,604 subjects. The PHQ2 dataset with the stringent missingness filter applied had 76,736 training samples from 28,915 subjects; 9,472 validation samples from 3,607 subjects; and 9,248 test samples from 3,547 subjects. The GAD2 dataset with the lax missingness filter applied had 255,296 training samples from 77,369 subjects; 32,128 validation samples from 9,735 subjects; and 31,606 test samples from 9,609 subjects. The GAD2 dataset with the stringent missingness filter applied had 76,736 training samples from 28,822 subjects; 9,472 validation samples from 3,602 subjects; and 9,254 test samples from 3,551 subjects.