Depression prediction using behavioral indicators and survey datasets

Di Wu, Tianchen Huang
Supervisor: James Spall

Johns Hopkins University | Whiting School of Engineering | Baltimore, MD

Study design and datasets

- Objective: Predict depression using two complementary public datasets
- Datasets:
  - **Actigraphy** (pilot): Small dataset with only **55 samples**; Minute-level activity data; used for exploratory behavioral analysis
  - **National Health and Nutrition Examination Survey** (NHANES, benchmark): Large dataset with **4836 samples**; Demographic and questionnaire data; used as the main benchmark because of larger sample size and a more stable structure
- Methods: Regression models and neural nets
- Rationale: Compare a small, behaviorally rich pilot dataset with a larger, statistically stronger benchmark dataset

Data preprocessing

Actigraphy:

- Cleaned minute-level records, checked timestamp continuity, and summarized valid monitoring days
- Group differences were mainly **time-structured**, with the strongest separation during **daytime / early morning**

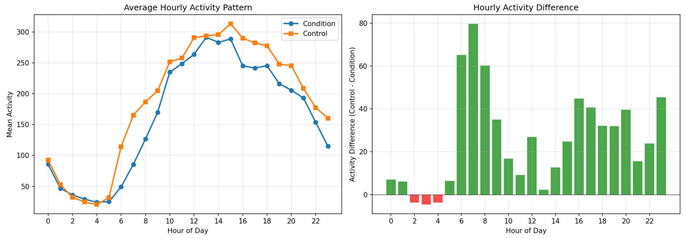

Figure 1. Average hourly activity patterns and between-group differences in the actigraphy dataset

NHANES:

- Defined depression as **Patient Health Questionnaire-9** (PHQ-9) ≥10 after cleaning and merging survey modules
- PHQ-9 scores were strongly right-skewed, indicating a clear class imbalance
- Most participants were in the none or mild severity range; relatively few were moderate or severe

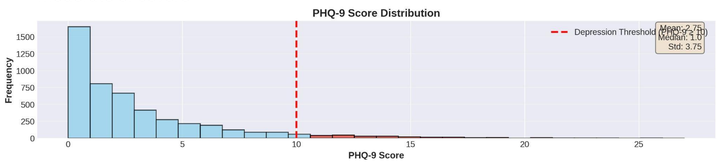

Figure 2. Distribution of PHQ-9 scores in the NHANES dataset

Figure 3. PHQ-9 score distributions by depression status in the NHANES dataset

Feature engineering and feature selection

Actigraphy:

- Transformed **minute-level records → daily features → participant-level summaries**
- Pruned redundant features
- Reduced the feature space from 108 to 42 features
- **Feature types:**
  - Daily activity summaries (e.g., daytime mean activity)
  - Variability features (e.g., coefficient of variation)
  - Circadian rhythm features (e.g., relative amplitude, acrophase)
  - Sleep-proxy features (e.g., nighttime inactivity)

NHANES:

- Pruned highly correlated, duplicate, and redundant variables
- Reduced the feature space from **70 to 54** features before downstream modeling
- **Feature types:**
  - Numerical variables (e.g., age, household size)
  - Binary variables (e.g., gender)
  - Ordinal variables (e.g., education)
  - One-hot encoded categorical variables (e.g., marital status, income brackets)

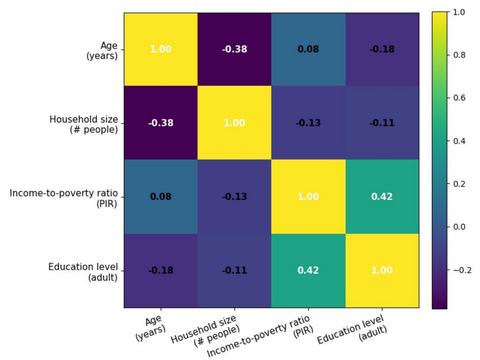

Figure 4. Representative 4×4 feature correlation matrix for the final NHANES feature set

Model Training and Results

Actigraph:

- Train regression models to predict depression scores based on all 33 features.
- Remove the features that has coefficient close to zero and train the models on these features.
- Test the significance of madrs1 feature by comparing performance of models trained with / without madrs1. (madrs1: depression score at the beginning of the measurement)

Feature Kept	Interpretation
MADRS1	Depression Score before Test
RA_std	How consistent the day/night contrast in activity is across measurement days.
ar1_day_normalized_rmse	Unpredictability of daytime activity from minute to minute, normalised by daytime standard deviation.
night_cv_trend	Whether sleep irregularity is worsening or improving over time.
cv_trend	How activity consistency is changing across days.

Table 1. The five features that aren't zeroed out

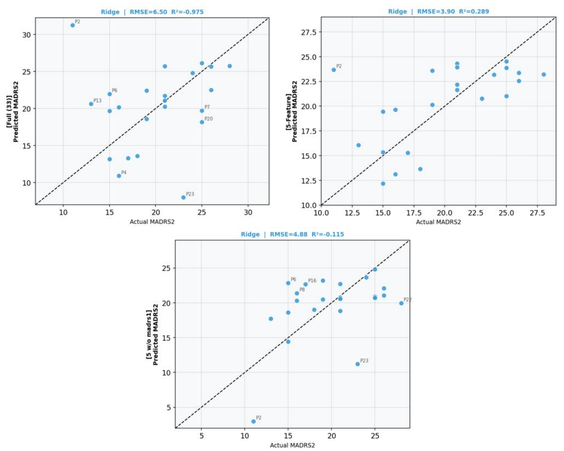

Figure 5. Results using different feature sets

NHANES:

- Highly imbalanced data for this dataset where 93.7% of the data is tagged as healthy.
- Two approaches for model training.
  - Train the model on raw data
  - Use **Synthetic Minority Over-Sampling Technique** (SMOTE) and class weights to make up for imbalanced data.
- Metrics:
  - **Accuracy**: Percentage of correct prediction
  - **F1 Score**: 2TP / (2TP + FP + FN) (TP: True Positive; FP: False Positive; FN: Negative Positive)

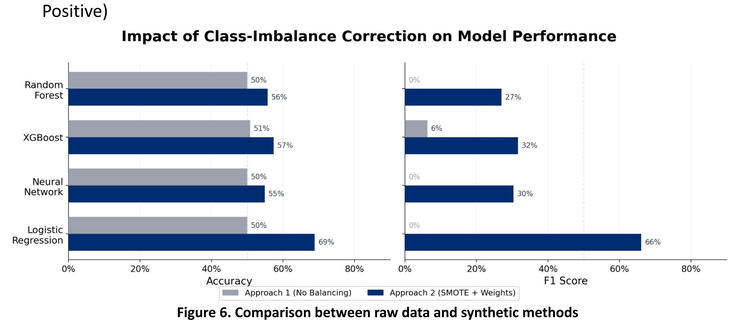

Figure 6. Comparison between raw data and synthetic methods

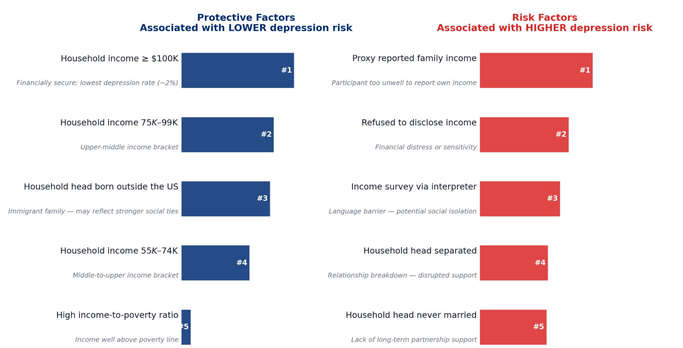

Figure 7. Most influential features (ranked by coefficient value)

- Features ⇨ lower risk: **high income, strong social ties**
- Features ⇨ higher risk: **low income, family issues**

Conclusion

- **Actigraph data (small dataset):** ElasticNet explained 51.5% of variance in follow-up depression severity — but sample size limits generalizability.
- **NHANES data (large dataset):**
  - without imbalance correction all models fail
  - with SMOTE and class weighting, Logistic Regression achieves best result.
- Across both phases, regularized linear models consistently match or outperform complex alternatives.
- Future work should integrate wearable physiological data with demographic context at adequate scale.