A Machine Learning Analysis of Risk Factors for Somatic Symptoms in the PHQ-15 Among Young Adults

Miao Yu, Jessica Tanchone

Data Science in Human Behavior, Department of Psychology, University of Wisconsin–Madison

Abstract: Somatic symptoms differ widely in their psychological and physiological origins, yet they are often summarized with a single PHQ-15 total score. This study used multiple machine-learning models to predict 13 individual symptoms in young adults. Fatigue and sleep problems showed the strongest discriminability (high F1 and ROC–AUC), while many other symptoms were difficult to predict. SHAP analyses revealed distinct psychological predictors, underscoring the value of item-level modeling for more precise assessment.

Background:

The PHQ-15 measures somatic symptom burden.
It's usually scored by summing all 15 items, assuming one common cause.
But symptoms differ: fatigue, dizziness, and chest pain may have distinct biological, psychological, or behavioral drivers.
A single total score can hide these differences and oversimplify diagnosis.
Prior work (Löwe et al., 2022; Tomenson et al., 2013) links somatic symptoms with depression and anxiety, but few studies test symptom-specific risk factors.
Finding these unique patterns could improve diagnostic accuracy and support more personalized care.

Problem:

The **DSM-5** acknowledges that psychological distress and symptom interpretation are central to **Somatic Symptom Disorder (SSD)**, yet diagnosis still depends on **subjective clinician judgment** of what counts as "excessive" concern—leading to inconsistent outcomes.
Using only PHQ-15 total scores prevents researchers from discerning which symptoms reflect **medical conditions** versus **psychological distress**.
Clarifying these distinctions is crucial for improving **clinical assessment and treatment**.

Objectives:

1. Predict the 13 PHQ-15 symptoms using psychological and demographic variables.
2. Compare the performance of Logistic regression(GLM net), K-nearest neighbors, XGboost, Catboost, Light Gradient Boosting, Random forest, Extratree, Voting Ensemble, Neural network (Tabnet-deep) models.
3. Interpret and visualize the most influential predictors using SHAP (Shapley Additive Explanations) values.

Method:

Dataset:

- **Source:** EAMMi2 open dataset (Grahe et al., 2018).
- **Sample:** 4000+ participants
  - Participants's age range from 18 to 29. 72.8% of respondents were women (*Nwomen* = 2280; *Nmen* = 771) (*Mage* = 21.10, *SD* = 4.83)
  - Racial and ethnicity: White/European American (63.5%),Black/African-American (7.6%), Hispanic/Latino/Latina (8.7%), Asian/Pacific Islander (6.5%), Native American (0.4%),"Other" Race (2.2%)

Predictors:

Psychological Predictors	Demographic and Attitudinal Predictors
Individual Differences & Personality • *IDEA-8*: Identity exploration and development • *NPI-13*: Narcissism • *Interpersonal Exploitativeness* (3 items) • *Self-Efficacy / Competence* • *Mindfulness Scale* **Social & Interpersonal Factors** • *Perceived Social Support* (12 items) • *Need to Belong* (10 items) • *Interpersonal Transgressions* (4 items) • *Social Media Use*: Maintaining / Making connections, Seeking information **Stress & Well-Being** • *Perceived Stress Scale* • *Subjective Well-Being* (6 items) • *Disability Identity and Status* (22 items combined from Q10–Q14) • *Belief in the "American Dream"* (2 items) **Markers of Adulthood** • *Achievement* and *Importance* subscales (10 items each)	• Sex: male, female, other • Education level • Race / Ethnicity • Household income • School attended • Parental marriage status: recoded categorical • Siblings: recoded binary (no siblings vs. ≥1 sibling) • Importance of marriage: single-item rating

Outcomes:

Item No.	Symptom	Item No.	Symptom
1	Stomach pain	8	Palpitations
2	Back pain	9	Shortness of breath
3	Limb/joint pain	10	Constipation
4	Headache	11	Dyspepsia/bloating
5	Chest pain	12	Fatigue
6	Dizziness	13	Trouble sleeping
7	Fainting spells		

Given severe class imbalance (most participants reported no symptoms on most PHQ-15 items),we evaluated models with balanced accuracy and AUC, which better reflect performance under unequal classes.

Analysis:

Step	Description
Preprocessing	Cleaned data, dummy-coded categorical variables, standardized predictors.
Modeling	Implemented Logistic regression(GLM net), K-nearest neighbors, XGboost, Catboost, Light Gradient Boosting, Random forest, Extratree, Voting Ensemble, Neural network (Tabnet-deep)
Evaluation	10-fold cross-validation, metrics: Accuracy, Balanced Accuracy, ROC-AUC.
Interpretation	Feature importance SHAP values for explainability

Table 2. Machine Learning Workflow Summary

Result:

Model Performance

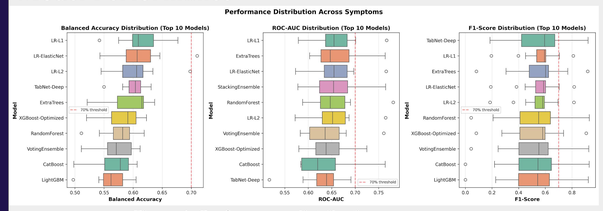


Figure 1. Performance distribution across symptoms

Across symptoms, median ROC–AUC values cluster around .65–.69 for most models, with ExtraTrees and Random Forest showing the highest upper ranges, while balanced accuracy medians remain below the .70 reference line for all models. Linear baselines (LR-ElasticNet/LASSO/Ridge) are competitively placed with relatively tight interquartile ranges, suggesting stable performance across symptoms. F1-scores are lower and more variable, especially for TabNet-Deep and boosting/ensemble methods, consistent with class imbalance and uneven positive class detection.

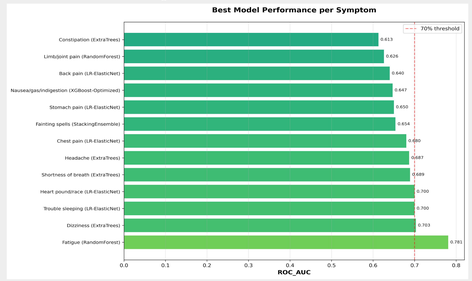


Figure 2. ROC–AUC of best-performing models for all somatic symptoms.

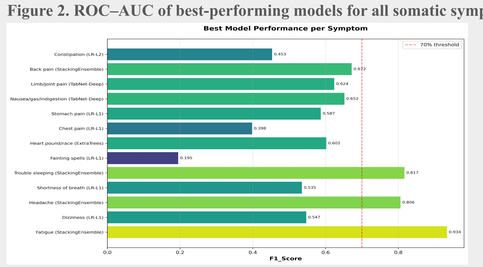
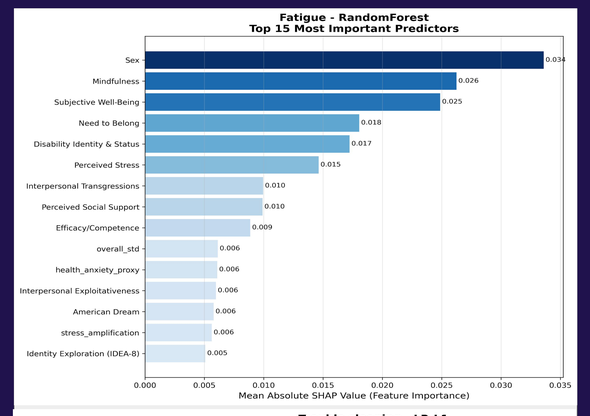


Figure 3. F1 score of best-performing models for all somatic symptoms.

According to figure 3 and figure 4, only a small subset of symptoms reached acceptable performance on both metrics. Using thresholds of F1 ≥ .70 and ROC–AUC ≥ .70, **fatigue**, **trouble sleeping**, **headache**, and **dizziness** emerged as the only symptoms with consistently strong prediction. Fatigue showed the highest scores overall (F1 = .934; AUC = .781). Trouble sleeping and headache achieved strong F1 scores (≥ .80) with moderate AUC (≈ .70).Dizziness met the AUC threshold (.703) with moderate F1. All other symptoms fell below at least one threshold, particularly on F1, indicating weak or inconsistent predictive signal. These results suggest that psychological and demographic factors are most informative for stress-linked symptoms, whereas rarer or physiologically driven symptoms (e.g., chest pain, constipation, fainting) were not reliably predictable from this dataset.

Feature importance

Feature importance indicates how much each feature contributes to the model prediction. Here we only here examined SHAP plots for **fatigue** and **sleep problems**, which had both high F1 adn AUC ROC score. Because ROC–AUC captures performance across all thresholds, we use it to select the best model for each symptom and then apply SHAP to those models to understand their predictions.



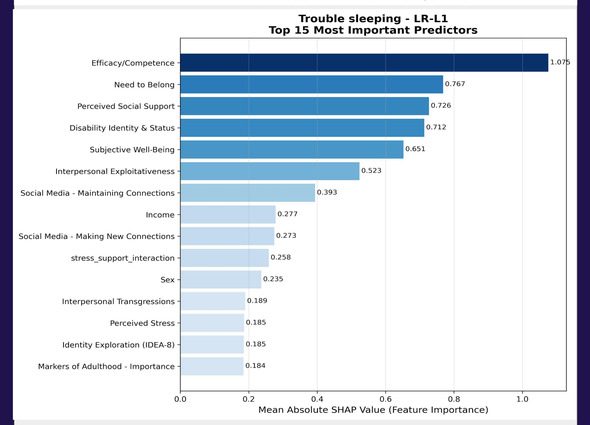


Figure 3. Shap plots of fatigue and sleep problem

We summarized global importance using mean absolute SHAP values from the best tree ensemble for each outcome. For fatigue, the most influential predictors were sex, mindfulness, subjective well-being, need to belong, disability identity and status, and perceived stress. For sleep problem The strongest predictor is efficacy or competence, this indicates that people's sense of capability and control over their lives is strongly associated with sleep problems. Need to belong, perceived social support, and disability identity also appear high on the list.

Discussion & Conclusion:

Finding:

- Different symptoms have different predictors.
- Psychosocial factors matter more for some symptoms (fatigue, sleep problem, headache, dizziness).
- Physiological factors likely play a larger role for others (chest pain, shortness of breath).

Limitation:

- The Dataset only includes young adults, Generalizability
- Limited symptom set in dataset

Future Direction:

Expand to additional somatic symptoms
(include more PHQ items and symptom-specific models)
Incorporate predictors related to psychopathology
(e.g., anxiety, depression, trauma-related symptoms)
Validate our models using a national dataset (Add Health)
– Test model performance in a large, representative U.S. sample
– Examine whether relationships replicate across populations
Build a more generalized prediction framework