Tilburg University

USING UNEMPLOYMENT AND EDUCATION TO PREDICT MENTAL HEALTH OUTCOMES

EVIDENCE FROM UNDERSTANDING SOCIETY AND CFPS

WANGBO TAO

thesis proposal
data science & society

student number

2137948

committee

Dr. Federico Zamberlan

location

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

date

September 3rd, 2025

word count

1815

USING UNEMPLOYMENT AND EDUCATION TO PREDICT MENTAL HEALTH OUTCOMES

EVIDENCE FROM UNDERSTANDING SOCIETY AND CFPS

wangbo tao

1 project definition, motivation & relevance

Since the Covid-19 pandemic and the recent wave of artificial intelligence, unemployment has become a major concern in many societies, especially for young graduates and middle-aged workers. In the European Union and the Eurozone, youth unemployment rates remain high at more than 13–14%. In the United Kingdom, youth unemployment is around 14%, while in China, the youth unemployment rate reached a record 21.3% in 2023, forcing the National Bureau of Statistics to suspend its routine publication of youth unemployment figures.

While many previous studies have shown that unemployment and education are related to mental health, fewer have looked at the predictive value of more detailed unemployment features, beyond a simple unemployment indicator. Even fewer studies have compared these patterns across different labor markets and cultural contexts.

This Thesis will investigate to what extent unemployment status and educational attainment—together with detailed features of unemployment histories (such as duration, frequency, and age at first unemployment)—predict individual mental health outcomes in a cross-sectional setting. The focus will be on two large and nationally representative household panel datasets: Understanding Society: the UK Household Longitudinal Study (UKHLS) and the China Family Panel Studies (CFPS) (ukhls; cfps).

From a societal perspective, predicting mental health risk can help identify vulnerable groups and support early intervention for people exposed to unstable work situations. From a scientific perspective, testing the added value of unemployment history features helps us understand the mechanisms that connect labor market experiences to mental health, and whether education acts as a buffer or moderator (clark2001unemployment; paul2009does). The study uses a cross-sectional design (one wave from each study) to focus on prediction and comparability. Possible longitudinal extensions will be discussed as future work.

2 literature review

Previous research discovered strong links between unemployment and mental health. For example, Yang et al. (2024) analyzed data from 201 countries between 1970 and 2020 and found that unemployment is significantly associated with higher risks of mental disorders, especially anxiety and depression. These findings indicate that unemployment is not only an economic challenge but also a public health issue.

More detailed studies have investigated how both current unemployment and past unemployment history matter. Using large-scale Finnish register data, Junna et al. (2022) found that current unemployment remains linked to poor mental health even after controlling for stable personal characteristics. They also reported that longer unemployment histories increase risks, especially for men in younger age groups. This highlights the importance of considering both present and past labor market experiences when predicting mental health.

A broader evidence base is provided by systematic reviews and meta-analyses. Sterud et al. (2025) reviewed longitudinal studies and concluded that unemployment increases the risk of mental health problems, while re-employment tends to decrease the risks. Although evidence certainty is sometimes limited, the general pattern supports the idea that transitions in and out of employment shape mental health trends.

In addition, new methodological approaches from data science are increasingly used. Machine learning has been applied to detect depression through social media traces and survey data (Gadzama et al., 2024; Zhang et al., 2025). These studies demonstrate the potential of advanced algorithms to identify risk factors and predict mental health outcomes, though many focus on social media data rather than labor market dynamics.

In summary, the literature confirms that unemployment is harmful for mental health, that detailed histories matter, and that re-employment can help reduce risks. However, few studies explicitly test the added predictive value of detailed unemployment features compared to a simple unemployment indicator, and even fewer do so across different cultural and labor market contexts. This project aims to fill that gap.

3 research strategy & research questions

This Thesis addresses the following research question:

To what extent do unemployment features and education level predict individual mental health outcomes in a cross-sectional setting?

The following sub-questions and hypotheses guide the empirical strategy of this thesis. Specific sub-questions and strategy are refined as follows:

1. Do detailed unemployment features (e.g., duration, frequency, age at first unemployment) improve prediction performance compared to simple unemployment indicators?

   Strategy: Develop two sets of models: a baseline model using a simple binary indicator of unemployment (with education), and an extended model including detailed unemployment-history features. Compare predictive performance (accuracy, precision, recall, F1-score, ROC-AUC).

   Expected Contribution: Tests whether richer labor-market information provides incremental predictive value beyond the simplest measures.

2. Does incorporating demographic characteristics (e.g., gender, age, urban/rural residence), either as features or through subgroup analysis, improve model prediction?

   Strategy: Compare models with and without demographic variables as predictors. Additionally, conduct subgroup analyses (e.g., separate models for men and women, urban vs rural). Assess differences in predictive performance.

   Expected Contribution: Clarifies whether demographic context strengthens prediction and reveals subgroup-specific vulnerabilities.

3. To what extent do predictive models trained in one country generalize to another, and what does this reveal about cross-national differences in the unemployment–education–mental health link?

   Strategy: Train and validate models on one dataset (e.g., UKHLS), then test predictive performance on the other dataset (e.g., CFPS), and vice versa. Ensure comparable features are extracted from both datasets.

   Expected Contribution: Provides evidence on whether predictive relationships are country-specific or generalizable, offering insights into the role of cultural and institutional labor-market differences.

Hypotheses:

H1: Models including detailed unemployment features (duration, frequency, age at first unemployment) will achieve higher predictive performance for mental health outcomes than models using only a simple unemployment indicator.

H2: Adding demographic characteristics (e.g., gender, age, urban/rural residence) as features or through subgroup analysis will improve predictive performance compared to models without these characteristics.

H3: Predictive models trained on one national dataset (e.g., UKHLS) will show limited generalizability when tested on the other dataset (e.g., CFPS), reflecting cross-national differences in the link between unemployment, education, and mental health.

4 methodology and evaluation

4.1 Dataset Description

This Thesis will use data from two major panel datasets: the China Family Panel Studies (CFPS 2022) and Understanding Society (UKHLS).

CFPS. The CFPS dataset (2022) includes 27,001 individual instances. It provides a rich set of variables relevant for this study. for education, we will focus on: the highest degree and the total term of education. For employment, we will include variables such as: current employment status, family economic status (related to employment/income), whether had a job in the past 12 months (proxy respondent), job security: whether unemployment insurance is provided, reason for not having a job, whether the first job was as an employee or self-employed, and occupation code of the first job (based on KGD4 text, coded using ISCO standards). Based on these employment variables and demographic variables, we will also construct derived indicators such as early-career unemployment and mid-career unemployment. For health outcomes, we will use disease variables related to mental disorders and neurological diseases.

4.2 Algorithms and Software

The analysis will be mainly conducted using Python.

The target variables on mental health are primarily categorical. While some predictors (e.g., income, education years) are continuous, most unemployment- and education-related predictors are categorical, and other demographic features can also be encoded as categorical variables. Therefore, predictive models will focus on classification algorithms such as logistic regression, random forest classifiers, and gradient boosting classifiers.

Feature engineering techniques will be applied to encode categorical variables, construct derived unemployment indicators, and assess the relative contribution of detailed unemployment features versus broad indicators.

The use of multiple algorithms ensures robustness and allows systematic comparison across models.

Table 1: Overview of Key Variables Used in the Study

Category	Variables
Education	Highest degree; Total term of education; Background and attainments; Expectations; Recent education and training
Employment	Current employment status; Family economic status (employment/income); Job in past 12 months (proxy); Job security (unemployment insurance); Reason for not having a job; First job type (employee vs. self-employed); Occupation code of first job (ISCO); Employment history; Hours worked; Self-employment; Wages and deductions; Derived indicators: early-career unemployment, mid-career unemployment
Diseases (Mental Health)	Mental disorders; Neurological diseases; Self-reported health problems (anxiety, depression); Treatment for mental health conditions

4.3 Evaluation Method

Model performance will be evaluated using standard predictive accuracy metrics such as accuracy, precision, recall, F1-score, and ROC-AUC for classification tasks. Cross-validation techniques will be employed to avoid overfitting. Comparisons will be made between models with simple unemployment indicators and those including detailed unemployment features, testing whether additional variables significantly improve predictive performance. Grouped analysis (e.g., by gender, age cohort) will also be explored to assess effects across countries with different economies and cultures.

5 milestones and plan

- Weeks 1–4: Literature review and dataset familiarization.
- Weeks 4–5: Data preprocessing and feature engineering.
- Weeks 6–9: Model development and evaluation.
- Weeks 10–11: Analysis and validation.
- Weeks 12–14: Writing and finalization of the Thesis.

references

Gadzama, W. A., Gabi, D., Argungu, M. S., & Suru, H. U. (2024). The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review. Personalized Medicine in Psychiatry, 45-46, 100125. https://doi.org/10.1016/j.pmip.2024.100125

Junna, L., Moustgaard, H., & Martikainen, P. (2022). Current unemployment, unemployment history, and mental health: A fixed-effects model approach. American Journal of Epidemiology, 191(8), 1459–1469. https://doi.org/10.1093/aje/kwac077

Sterud, T., Lunde, L., Berg, R., et al. (2025). Mental health effects of unemployment and re-employment: A systematic review and meta-analysis of longitudinal studies. Occupational and Environmental Medicine. https://doi.org/10.1136/oemed-2025-110194

Yang, Y., Niu, L., Amin, S., & Yasin, I. (2024). Unemployment and mental health: A global study of unemployments influence on diverse mental disorders. Frontiers in Public Health, 12, 1440403. https://doi.org/10.3389/fpubh.2024.1440403

Zhang, Y., Wang, Z., Ding, Z., Tian, Y., Dai, J., Shen, X., Liu, Y., & Cao, Y. (2025). Employing machine learning and deep learning models for mental illness detection. Computation, 13(8), 186. https://doi.org/10.3390/computation13080186