Large Language Models-based Feature Selection for Mental Health Prediction via Wearable Sensor Data

Pakin Siwathammarat
Vidyasirimedhi Institute of Science and Technology
SIPGA 2024

A REPORT SUBMITTED TO
The National Science and Technology Development Agency (NSTDA)

Supervisor	Dr. Song Yuting
Attachment	Institute of High Performance Computing (IHPC), A*STAR, Singapore
Date	1 September 2024 - 31 January 2025

Large Language Models-based Feature Selection for Mental Health Prediction via Wearable Sensor Data

Abstract

Pakin Siwathammarat

The integration of wearable sensor data with advanced machine learning techniques presents a transformative opportunity for mental health prediction, enabling early diagnosis and personalized interventions. However, existing models face challenges due to high-dimensional, noisy, and complex data, which limits scalability and interpretability. Traditional feature selection methods, while effective, often require access to extensive user data, raising privacy concerns and regulatory constraints. This work introduces an agentic framework leveraging large language models (LLMs) for feature selection in mental health prediction tasks. By synthesizing and reasoning about wearable sensor data, our approach identifies a compact yet highly informative subset of features, enhancing model performance while preserving interpretability. Our findings demonstrate that LLM-driven feature selection improves predictive accuracy, reduces computational overhead, and enhances privacy, offering a scalable and practical solution for clinical applications.

Keywords: Wearable sensors, mental health prediction, machine learning, large language models (LLMs), feature selection, predictive modeling, healthcare AI, time series analysis.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

The rapid advancement of wearable sensor technology and machine learning has created new opportunities for improving mental health prediction. Wearable devices continuously generate rich time-series data streams, capturing behavioral signals such as physical activity, sleep patterns, and physiological responses. When effectively analyzed, this data can enable early diagnosis and personalized interventions, ultimately improving patient outcomes. However, leveraging this wealth of information for mental health prediction requires sophisticated machine learning techniques to handle complex, high-dimensional data.

1.2 Machine Learning in Mental Health Prediction

Recent research has employed classical machine learning and deep learning approaches using wearable sensor data to predict mental health conditions. Studies have demonstrated correlations between behavioral signals and mental health metrics, such as depression and anxiety scores. While these models have shown promise, they often rely on domain-specific algorithms and feature engineering, which may limit generalizability. There has been growing interest in using large language models (LLMs) to synthesize and reason about sensor data. For example, studies utilizing the GLOBEM dataset and LLM-driven healthcare applications suggest that these models can provide valuable insights into complex health patterns.

1.3 Challenges in Wearable Sensor Data Analysis

Despite promising advancements, several challenges persist in applying machine learning to wearable sensor data:

- High Dimensionality: Wearable devices collect vast amounts of features, many of which may be redundant or non-informative.
- Inherent Noise: Sensor data is prone to inconsistencies, missing values, and variations due to environmental or user-specific factors.
- Non-Linearity and Complexity: Mental health conditions are influenced by intricate, multi-factorial relationships that are difficult to model accurately.

To address these issues, feature selection is crucial in refining the input data. By selecting the most relevant features, models can achieve higher predictive accuracy, improved interpretability, and reduced computational overhead. However, traditional feature selection methods often require access to large-scale user datasets, subject to

strict privacy regulations and data-sharing restrictions, limiting their real-world applicability.

1.4 Machine Learning in Mental Health Prediction

Given the limitations of traditional feature selection techniques**,** LLMs present a novel opportunity to enhance feature selection in wearable-based mental health prediction. Unlike conventional methods, LLMs can analyze, interpret, and prioritize features based on contextual significance, reducing dependence on large user datasets while preserving privacy. Despite their potential, LLM-driven feature selection remains an underexplored area in the field of mental health prediction.

1.5 Proposed Approach and Contributions

We propose an agentic framework that leverages LLMs for feature selection in mental health prediction tasks to address these challenges. Our approach differs from conventional methods by harnessing LLMs' generative and reasoning capabilities to identify a compact yet highly informative subset of features from wearable sensor data. This framework:

- Improves predictive accuracy by selecting the most meaningful features.
- Enhances interpretability, making models more transparent and clinically relevant.
- Reduces computational complexity, enabling scalability for real-world applications.
- Promotes privacy preservation, minimizing the need for extensive user data.

Our findings demonstrate that LLM-driven feature selection significantly enhances model performance while reducing computational burden, offering a scalable, efficient, and privacy-conscious solution for clinical applications in mental health prediction.

Chapter 2

Literature Reviews

2.1 Mental health prediction

The prediction of mental health states using machine learning has gained significant traction with the rise of wearable and mobile sensing technologies. Tasks such as depression detection have been explored using data collected from sensors that track activity, sleep, and phone usage patterns. The GLOBEM dataset has been instrumental in this domain, offering a multi-year longitudinal dataset with rich behavioral signals and validated mental health assessments such as the Patient Health Questionnaire-4 (PHQ-4).

Classical machine learning approaches have employed algorithms such as Random Forest, Support Vector Machines, and Logistic Regression for mental health prediction. These methods typically rely on handcrafted feature engineering and domain-specific knowledge. However, they often struggle with high-dimensional, noisy, and multi-modal data, leading to challenges in generalization and scalability.

The use of Large Language Models (LLMs) in mental health prediction is a more recent development. LLMs, such as GPT-4 and PaLM, can synthesize multi-modal inputs and reason about temporal data, making them well-suited for tasks like depression classification and PHQ-4 score prediction. Initial show that LLMs outperform classical models in capturing complex patterns, offering a promising avenue for future research.

2.2 Feature selection

Feature selection is the process of identifying and retaining the most relevant features from a dataset that contribute significantly to the predictive power of a model. By reducing the number of input variables, it simplifies the model, enhances interpretability, and improves computational efficiency.

Feature selection is particularly crucial in domains like healthcare, where datasets are often high-dimensional and noisy. It mitigates overfitting, ensures the model focuses on

meaningful patterns, and reduces the reliance on sensitive or costly-to-collect data. In healthcare applications, selecting fewer but highly informative features also aligns with privacy regulations, as it minimizes exposure of patient information while maintaining strong predictive performance.

2.2.1 Traditional feature selection

Traditional feature selection techniques have been widely used to identify informative features, enhancing model interpretability and computational efficiency. Commonly applied methods include:

- Lasso Regression: Uses L1 regularization to shrink irrelevant features to zero, resulting in a sparse feature set.
- Mutual Information (MI): Evaluates the dependency between features and target variables, selecting the most informative ones.
- Recursive Feature Elimination (RFE): Iteratively removes the least important features based on model weights, optimizing feature subsets.

While effective in structured datasets, these methods are often limited in multi-modal and high-dimensional contexts, particularly in domains like healthcare, where privacy concerns restrict data availability.

2.2.2 LLM for feature selection

The application of LLMs for feature selection represents a transformative shift. LLM-Select demonstrates that LLMs can identify predictive features using only feature names and task descriptions without accessing the downstream data.

Key approaches include:

- LLM-Score: Assigns numerical importance scores to features based on their relevance to the task.
- LLM-Rank: Produces a ranked list of features by their conceptual importance.
- LLM-Seq: Sequentially selects features in a dialogue-based manner, refining the feature set dynamically.

These methods rival classical data-driven techniques like Lasso and Mutual Information in predictive performance, highlighting the potential of LLMs in feature selection for domains such as healthcare.

Chapter 3

Methodology

3.1 Agentic workflow

The Agentic Workflow integrates LLMs into iterative processes for feature selection and model optimization. In this workflow, LLMs act as agents that utilize domain-specific prompts, interpretability-driven criteria, and feedback loops to refine feature sets. This approach enables:

1. Dynamic Adaptation: Adjusting feature selection strategies based on task-specific needs.
2. Contextual Integration: Incorporating domain knowledge, patient demographics, and temporal patterns for enhanced relevance.
3. Scalable Design: Extending workflows to multi-modal, longitudinal data such as those in the GLOBEM dataset.

The workflow bridges the gap between classical methods and LLM-driven insights, offering a robust framework for developing clinically interpretable and high-performing

3.2 Dataset

This study uses the GLOBEM dataset, a multi-year longitudinal dataset collected from wearable sensors and smartphone applications. The dataset inclues heart rate, step counts, sleep duration, phouse usage, and other behavioral signals paired with weekly mental health assessments. These assessments provide ground-truth labels for evaluating mental health status using the Patient Health Questionnaire-4 (PHQ-4). Classification task is addressed in this study:

Predicting the presence or absence of depression based on weekly survey responses. The evaluation metrics for this task include:

- **Accuracy**: The proportion of correctly classified instances.
- **Precision**: The proportion of predicted positive cases that are actually positive, measuring the model's reliability in identifying depression.
- **Recall**: The proportion of actual positive cases correctly identified, reflecting the model’s sensitivity.
- **F1-score**: The harmonic mean of precision and recall, balancing false positives and false negatives.

The GLOBEM dataset is particularly suitable for this research due to its richness in multi-data and longitudinal design, which captures dynamic changes in behavioral and physiological patterns. This diversity allows for a robust analysis of how wearable data correlates with mental health. Using PHQ-4 scores ensures that the proposed agentic framework is adaptables and clinically relevant.

3.3 Data preparation

GLOBEM data has 4 waves, which are conducted through 4 different years for 10 weeks each. More data information can be found here. I combined waves 2, 3, and 4 for the overall data for this project. Wave 1 was not included because weekly PHQ scores were not taken. PHQ score is the primary variable of interest for prediction.

I categorized the data into none, mild, moderate, and severe depression levels based on the participant's score on the BDI-II, taken during the pre-survey period. From each category, I randomly selected 10% of the participants to use as testing data. After taking out the testing data participants, I randomly selected 100 participants for the training data to match the amount of participants in the testing data. For both groups, I excluded participants who had insufficient features data (more than 5 different features that had more than 20% missing values).

3.4 Baselines

To benchmark performance, I employ both classical feature selection techniques and feature selection using Large Lanuge Models (LLMs). For traditional feature selection, I evaluate Lasso Regression, Mutual Information (MI), and Recursive Feature Elimination (RFE)**,** which have been widely used for improving model efficiency and interpretability. However, these methods often struggle with high-dimensional, multi-modal data and require direct access to training datasets, raising privacy concerns.

For LLM-based feature selection, I consider LLM-Score, LLM-Rank, and LLM-Seq**,** which leverage LLMs to identify relevant features using only feature names and task descriptions. These approaches enable feature selection without direct access to raw data, offering a more privacy-conscious alternative.

By comparing these baselines, I assess the effectiveness of LLM-driven feature selection in improving model performance while addressing scalability and privacy constraints in mental health prediction.

Chapter 4

Results and Discussion

4.1 Classification Results

For the binary classification task, the proposed agentic framework leveraging LLM-based feature selection demonstrated superior performance compared to traditional feature selection methods. I evaluated two feature sets: Human's selected features and sleep and step features, using accuracy, true positive rate (TPR), and false positive rate (FPR) as evaluation metrics.

4.1.1 Table

	Metrics	Accuracy	Precision	Recall	F1
All		0.61	0.50	0.67	0.57
Trditional	lasso	0.49	0.49	0.41	0.45
	mi	0.55	0.51	0.57	0.54
	rfe	0.57	0.56	0.6	0.58
LLM	score	0.57	0.57	0.58	0.57
	rank	0.51	0.53	0.72	0.61
	seq	0.59	0.59	0.59	0.59
Our	without data	0.55	0.56	0.48	0.52
	data	0.60	0.60	0.59	0.59

4.1.2 Result Analysis

The results highlight the robustness of the LLM-based agentic framework, particularly in optimizing predictive performance with fewer and more interpretable features. The LLM-Seq method consistently outperformed other approaches, demonstrating the advantage of iterative, context-aware feature selection in improving classification accuracy while reducing false positives.

Chapter 5

Conclusion

5.1 Conclusion

This study proposed an agentic framework leveraging Large Language Models (LLMs) for feature selection in mental health prediction tasks using wearable sensor data. By incorporating LLM-driven methods such as LLM-Score, LLM-Rank, and LLM-Seq, our approach enhanced predictive accuracy, reduced computational overhead, and improved interpretability compared to traditional feature selection techniques. The results demonstrated that LLM-based feature selection effectively identifies compact, high-quality feature subsets, leading to superior performance in both classification and regression tasks. These findings highlight the potential of LLM-driven methodologies in optimizing data-driven mental health prediction, offering a scalable, privacy-preserving, and interpretable solution for real-world clinical applications.

5.2 Future work

While this study demonstrates the effectiveness of LLM-driven feature selection for mental health prediction using wearable sensor data, several avenues remain for future exploration:

- Enhancing Generalizability: Future research should explore the adaptability of the proposed framework across diverse datasets and populations to ensure robustness in real-world applications.
- Privacy-Preserving Techniques: Investigating federated learning and differential privacy approaches could further enhance data security while maintaining model performance.
- Fine-Tuning LLMs for Feature Selection: Optimizing LLMs for domain-specific feature selection could lead to more efficient and interpretable models tailored for healthcare applications.

By addressing these challenges, future work can further refine LLM-based feature selection techniques, paving the way for more effective, ethical, and scalable mental health prediction solutions.

References

1. Canzian, L., & Musolesi, M. (2015). Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 1293–1304). Association for Computing Machinery.
2. Saeb, S., Zhang, M., Karr, C., Schueller, S., Corden, M., Kording, K., & Mohr, D. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. J. Med. Internet Res., 17(7), e175.
3. Chikersal, P., Doryab, A., Tumminia, M., Villalba, D., Dutcher, J., Liu, X., Cohen, S., Creswell, K., Mankoff, J., Creswell, J., Goel, M., & Dey, A. (2021). Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection. ACM Trans. Comput.-Hum. Interact., 28(1).
4. Xu, X., Liu, X., Zhang, H., Wang, W., Nepal, S., Sefidgar, Y., Seo, W., Kuehn, K., Huckins, J., Morris, M., Nurius, P., Riskin, E., Patel, S., Althoff, T., Campbell, A., Dey, A., & Mankoff, J. (2023). GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 6(4).
5. Kim, Y., Xu, X., McDuff, D., Breazeal, C., & Park, H. (2024). Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. In Proceedings of the fifth Conference on Health, Inference, and Learning (pp. 522–539). PMLR.
6. Englhardt, Z., Ma, C., Morris, M., Chang, C.C., Xu, X., Qin, L., McDuff, D., Liu, X., Patel, S., & Iyer, V. (2024). From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 8(2).

7. Robert Tibshirani (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.
8. Lewis, D. (1992). Feature Selection and Feature Extraction for Text Categorization. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.
9. Isabelle M Guyon, Jason Weston, Stephen D. Barnhill, & Vladimir Naumovich Vapnik (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46, 389-422.