Large Language Models for Wearable Data Analysis and Interpretation

Conference Paper

Author(s):
Böhi, Simon; Gashi, Shkurta

Publication date:
2024

Permanent link:
https://doi.org/10.3929/ethz-b-000661839

Large Language Models for Wearable Data Analysis and Interpretation

Simon Böhi
Department of Computer Science
ETH Zürich, Switzerland
boehis@student.ethz.ch

Shkurta Gashi
ETH AI Center
ETH Zürich, Switzerland
shkurta.gashi@ai.ethz.ch

Abstract

In this paper, we investigate the application of large language models (LLMs) to zero-shot and few-shot prediction and classification of multimodal wearable sensor data. Using data from the large-scale HomeKit2020 dataset, we explore health tasks including cardiac activity monitoring, metabolic health prediction, and sleep detection. We demonstrate that LLMs perform feature extraction, prediction, and classification with comparable or higher performance than classical machine learning approaches even in the zero-shot scenario. Our findings show promising results in using LLMs for wearable data analysis and interpretation.

1 Introduction

Foundation large language models (LLMs) have recently demonstrated strong capabilities to solve a diverse range of tasks, including, diagnostic reasoning Wu et al. (2023); Moor et al. (2023), visual referring expression comprehension Sui et al. (2023), time series modeling Sun et al. (2023), tabular Hegselmann et al. (2023), and health data analysis Belyaeva et al. (2023). In this paper, we propose using LLMs to analyze and interpret wearable sensor data. The main advantages of LLMs in comparison to state-of-the-art techniques would be to ingest complex multimodal data to gain an understanding of individual health risks and to be able to do that with little amount of data, which is a known challenge in the wearable health community. Only a few researchers investigated the use of LLMs for wearable sensor data Liu et al. (2023); Spathis & Kawsar (2023); Sooriya Patabandige et al. (2023). These approaches either employ self-curated toy datasets Liu et al. (2023), investigate zero-shot learning only, do not report baseline model accuracy making it difficult to interpret the results Sooriya Patabandige et al. (2023) or focus on identifying challenges related to the application of LLM to wearable data Spathis & Kawsar (2023). We build upon this work and perform a systematic evaluation of the LLMs for feature extraction, prediction, and classification tasks on wearable data. We investigate for the first time the use of LLMs for sleep detection and the impact of varying window lengths of wearable sensor data collected in natural settings.

The objectives of the paper are: 1) investigate whether LLMs can be used out-of-the-box to predict health conditions from wearable sensor data, and 2) evaluate the number of samples needed by the LLMs to learn the task with sufficient accuracy. We formulate eight tasks related to cardiac activity, metabolic health, and sleep monitoring using wearable devices. We then use an off-the-shelf pretrained LLM, LLaMa 2 Touvron et al. (2023), and in context learning to show the feasibility of our approach. We compare the LLM score with supervised machine learning methods such as gradient-boosted decision trees. To this end, we use a large-scale, real-world dataset called HomeKit2020 Merrill et al. (2023), which was collected from $5,034$ participants over four months in real-world settings. We evaluate different ways of prompting and passing data to the LLM.

2 Large Language Models for Wearable Sensor Data Analysis

Model. We use LlaMa 2 Touvron et al. (2023) as a pretrained LLM. In a recent study, it has been demonstrated that LLaMa 2 surpasses GPT-4 in arithmetic tasks by breaking down each digit into a separate token, ensuring a uniform tokenization of numbers Liu & Low (2023). We perform *zero-shot* – by using the LLM out-of-the-box to predict health outcomes from provided wearable sensor

Table 1: Comparison of performance between LLMs and baselines trained using few-shot learning across the health tasks explored in this work. The metric for sleep detection is accuracy and for the other tasks is mean absolute error (MAE). HR refers to heart rate. S refers to steps. C refers to context. Sle refers to sleep. Cal refers to calories.

Task	**XGB**		**LR/SVM**		**LlaMa 2**		
	3-shot	*10-shot*	*3-shot*	*10-shot*	*0-shot*	*3-shot*	*10-shot*
HR_{Avg}	9.38 (.4)	5.40 (.1)	2.62 (.3)	**0.08 (.0)**	0.77 (.1)	**0.75 (.1)**	0.93 (.1)
HR_{Max}	11.3 (.7)	7.18 (.5)	**6.08 (.9)**	12.0 (1.8)	0.82 (.1)	**0.43 (.0)**	0.73 (.1)
HR_{Min}	8.86 (.4)	5.19 (.1)	**4.76 (.4)**	9.23 (.6)	**0.38 (.0)**	0.56 (.1)	1.00 (.2)
HR_{Std}	2.68 (.3)	**2.48 (.2)**	2.90 (.3)	6.47 (.9)	11.6 (.6)	**1.11 (.1)**	2.17 (.1)
Cal_{HR}	875 (42)	830 (35)	**828 (44)**	958 (63)	**572 (35)**	1167 (116)	1047 (72)
Sle_{HR}	0.50 (.0)	**0.69 (.0)**	0.62 (.0)	0.64 (.0)	0.49 (.0)	**0.53 (.2)**	0.51 (.1)
Sle_{HR+S}	0.50 (.0)	0.75 (.0)	0.70 (.0)	**0.79 (.0)**	0.50 (.0)	0.59 (.0)	**0.60 (.0)**
Sle_{HR+S+C}	0.50 (.0)	0.75 (.0)	0.70 (.0)	**0.79 (.0)**	0.58 (.0)	**0.70 (.0)**	0.58 (.0)

data – and *few-shot learning* – by providing tuples of wearable sensor data and health conditions to the network similar to a supervised learning setting. Few-shot learning, also known as *in context learning* Brown et al. (2020), removes the need to further train the LLM, which might be challenging in the wearable data domain without suitable samples. We test the pipeline on unseen, test samples for *feature extraction*, *classification*, and *regression* tasks. Feature extraction refers to the prediction of the mean, maximum, minimum, and standard deviation of heart rate (HR). Regression task predicts the amount of calories burned during an activity from HR. Classification refers to inferring whether the user is asleep or awake from HR and the number of steps. To the best of our knowledge, we are the first to explore LLMs for sleep detection, which is a common wearable-based task.

Dataset. To evaluate our approach, we use HomeKit2020 dataset Merrill et al. (2023), which is a large-scale, high-resolution, and publicly available dataset. It consists of minute-level HR, number of steps, and sleep measurements collected using Fitbit devices from $5,034$ participants over 4 months in real-world settings. We randomly sampled 3 and 10 examples for the few-shot experiments and 200 samples as the test sets. We repeat these experiments five times and each time pick a different set of samples to evaluate the performance. We report the average results (and standard deviation) over all the runs of an experiment. Table 2 shows the tasks and prompts used in this work.

Baselines. We compare LLM's performance to gradient boosting, (XGB), linear regression (LR), and support vector machine (SVM) for regression and classification tasks respectively.

Results. Table 1 shows a summary of the results we achieved. Our findings show that LLMs extract HR features with lower or comparable errors to baseline methods for both the zero-shot and few-shot learning scenarios. The LLM predicts calories burned during a day with a significantly lower error than classical machine learning algorithms. The performance of LLM for sleep detection is overall lower than that of baseline classifiers. In particular, with only 3 samples the LLM achieves an accuracy of 70% for sleep detection, which is the same as SVM in the 3-shot learning scenario. We further find that including more information, such as the number of steps and offering hints concerning HR changes during sleep and wake stages, as shown in the last query in Table 2, contributes to further improving the LLM's results. These findings demonstrate LLMs' potential beyond text processing and show their capability for analyzing and interpreting wearable sensor data.

3 Conclusion

This paper contributes to the application of LLMs for wearable health monitoring. We demonstrate LLM's potential to perform tasks related to cardiac activity, metabolic health, and sleep activity with comparable or higher performance to traditional methods. Our results warrant further investigation with a diverse set of LLMs and health tasks to enable personalized health recommendations and enhance user experience. Investigation of privacy and security risks as well as energy consumption are needed to overcome the challenges of LLMs application to wearable health technology.

Acknowledgements

S.G. is partially funded by PHRT (grant number: 2021-801), SDSC (grant number: C21-18P) and an ETH AI Center postdoctoral fellowship.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of the ICLR 2024 Tiny Papers Track.

References

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pp. 86–102. Springer, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 2020.

Shkurta Gashi, Chulhong Min, Alessandro Montanari, Silvia Santini, and Fahim Kawsar. A multi-device and multimodal dataset for human energy expenditure estimation using wearable devices. *Scientific Data*, 9(1):537, 2022.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Vanessa Ibáñez, Josep Silva, and Omar Cauli. A survey on sleep assessment methods. *PeerJ*, 6: e4849, 2018.

Tiedong Liu and Bryan Kian Hsiang Low. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *arXiv preprint arXiv:2305.14201*, 2023.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023.

Mike A Merrill, Esteban Safranchik, Arinbjörn Kolbeinsson, Piyusha Gade, Ernesto Ramirez, Ludwig Schmidt, Luca Foshchini, and Tim Althoff. Homekit2020: A benchmark for time series classification on a large mobile sensing dataset with laboratory tested ground truth of influenza infections. In *Conference on Health, Inference, and Learning*. Proceedings of Machine Learning Research, 2023.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023.

Pramuka Medaranga Sooriya Patabandige, Steven Antya Orvala Waskito, Kunjun Li, Kai Jie Leow, Shantanu Chakrabarty, and Ambuj Varshney. Poster: Rethinking embedded sensor data processing and analysis with large language models. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pp. 561–562, 2023.

Dimitris Spathis and Fahim Kawsar. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *arXiv preprint arXiv:2309.06236*, 2023.

Xiuchao Sui, Shaohua Li, Hong Yang, Hongyuan Zhu, and Yan Wu. Language models can do zero-shot visual referring expression comprehension. 2023.

Chenxi Sun, Yaliang Li, Hongyan Li, and Shenda Hong. Test: Text prototype aligned embedding to activate llm's ability for time series. *arXiv preprint arXiv:2308.08241*, 2023.

Table 2: Example of the tasks and prompts used to query the LLM. HR refers to heart rate, S refers to steps, C refers to context, Avg to average, Min to minimum, Max to maximum, Std to standard deviation, Sle refers to sleep, Cal to calories.

Type	Task	Prompt	Output
Feature Extraction	HR$_{Avg}$	Answer only with the numerical average of the provided heart rate data. Exclude all other information or calculations. [HR]	`mean([HR])`
	HR$_{Min}$	Answer only with the numerical minimum of the provided heart rate data. Exclude all other information or calculations. [HR]	`min([HR])`
	HR$_{Max}$	Answer only with the numerical maximum of the provided heart rate data. Exclude all other information or calculations. [HR]	`max([HR])`
	HR$_{Std}$	Answer only with the standard deviation of the provided heart rate data. Exclude all other information or calculations. [HR]	`std([HR])`
Regression	Cal$_{HR}$	Your task is to provide a direct prediction of the total calories burned during a single day for an adult, based on heart rate collected by a Fitbit device. Use the given data for this calculation and respond with only the predicted calorie count. [HR]	`[Calories]`
Classification	Sle$_{HR}$	Predict sleep status (awake or asleep) based on these consecutive heart rates over a 30-minute interval. [HR]	`[Sleep]`
	Sle$_{HR+S}$	Predict sleep status (awake or asleep) based on these consecutive heart rates over a 30-minute interval and the total steps. [HR] and steps.	`[Sleep]`
	Sle$_{HR+S+C}$	You are a helpful assistant that analyzes heart rate data and total steps to predict sleep status. Consider that higher and more abruptly changing heart rates typically indicate the user is awake, while steadier and lower rates suggest the user is asleep. Very low step count indicates that the user is sleeping. Use this intuition to analyze the provided minute-level heart rate data from an adult, measured by a Fitbit, and predict whether the user is awake or asleep. [HR] and steps.	`[Sleep]`

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. Large language models perform diagnostic reasoning. *arXiv preprint arXiv:2307.08922*, 2023.

A Dataset

This paper uses the Homekit2020 dataset (Merrill et al., 2023) collected as part of the Home Testing of Respiratory Illness Study by Evidation Health and described in Synapse. The study was conducted in partnership with the Biomedical Advanced Research and Development Authority (BARDA), an existing office of the U.S. Department of Health and Human Services, and Audere, a non-profit digital health technology corporation.

Homekit2020 consists of 14 million hours of minute-level Fitbit data, such as HR, number of steps, and sleep stage. The dataset is accessible through a request process to Synapse. The study that collected this dataset was approved by the Western Institutional Review Board (WIRB, Puyallup, WA, USA) and the University of Washington IRB (Study #1271380).

A.1 Health Tasks

Heart rate metrics. This task derives the average, minimum, maximum, and standard deviation of heart rate (HR) from a series of instantaneous HR measurements. These tasks are framed as regression tasks to be able to compare the performance with conventional machine learning techniques.

Calorie prediction. Energy expenditure prediction refers to the estimation of calories burned during an activity. It is an essential aspect of health and well-being monitoring. The traditional techniques to estimate calories include direct calorimetry, indirect calorimetry, doubly labeled water, and the more recent wearable devices Gashi et al. (2022). In this work, we investigate the challenging problem of estimating calories based on HR only and we frame the problem as a regression task.

Sleep detection. Sleep detection refers to the process of identifying whether an individual is asleep or not, which allows deriving metrics like sleep duration. It is crucial for understanding the overall health and well-being of an individual. Sleep is commonly measured through various methods such as polysomnography, actigraphy, and wearable devices Ibáñez et al. (2018). In this work, we use HR and the number of steps measured with wearable devices and predict whether the user is asleep or awake. We frame the problem of sleep detection as a binary classification task.

B Results

In this section, we show graphical representations of the results presented in Table 1.

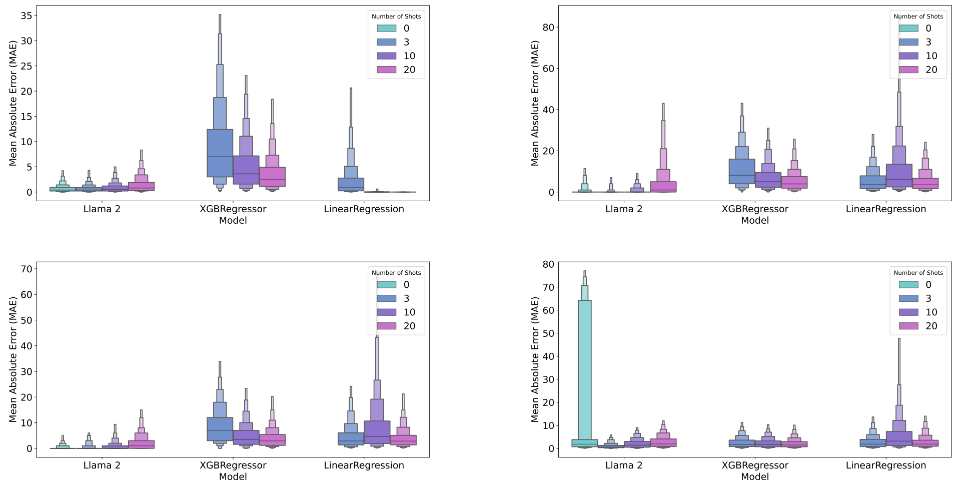

Figure 1: Mean absolute error (MAE) for the LLM and supervised baselines to predict HR average (top-left), maximum (top-right), minimum (bottom-left), and standard deviation (bottom-right).

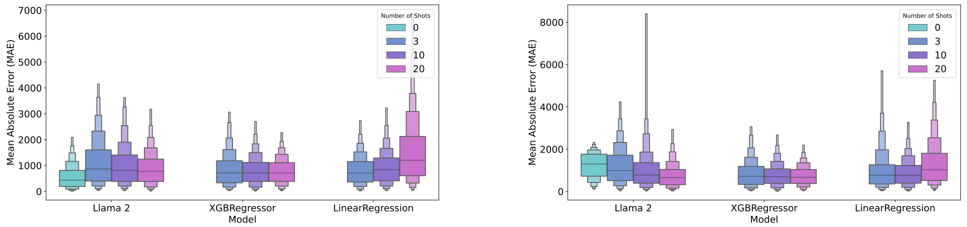

Figure 2: Mean absolute error (MAE) for the LLM and supervised baselines to predict calories burned during the day using HR (left) and HR combined with steps (right).

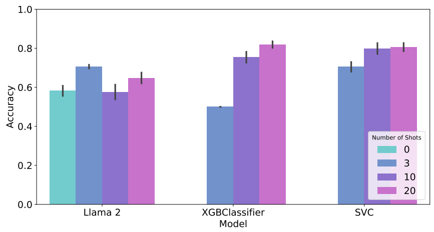

Figure 3: Accuracy for the LLM (Llama 2), gradient boosting (XGBClassifier), and support vector machine (SVC) classifiers to recognize sleep from HR data.

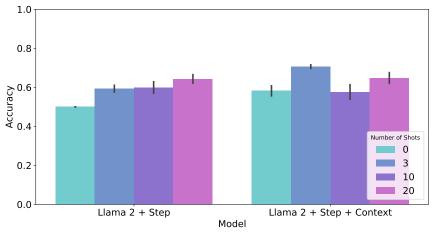

Figure 4: Accuracy for the LLM to recognize sleep from HR and steps (left) as well as HR and steps with more contextual information about the pattern of HR and steps during sleep.