Leveraging Large Language Models for Generating Mobile Sensing Strategies in Human Behavior Modeling

Nan Gao
Tsinghua University
Beijing, China
University of New South Wales (UNSW)
Sydney, Australia
nangao@tsinghua.edu.cn

Zhuolei Yu
Tsinghua University
Beijing, China
yuzl21@mails.tsinghua.edu.cn

Yue Xu
Tsinghua University
Beijing, China
yue-xu22@mails.tsinghua.edu.cn

Chun Yu*
Tsinghua University
Beijing, China
chunyu@mail.tsinghua.edu.cn

Yuntao Wang
Tsinghua University
Beijing, China
yuntaowang@tsinghua.edu.cn

Flora D. Salim
University of New South Wales (UNSW)
Sydney, Australia
flora.salim@unsw.edu.au

Yuanchun Shi
Tsinghua University
Beijing, China
shiyc@tsinghua.edu.cn

ABSTRACT

Mobile sensing plays a crucial role in generating digital traces to understand human daily lives. However, studying behaviours like mood or sleep quality in smartphone users requires carefully designed mobile sensing strategies such as sensor selection and feature construction. This process is time-consuming, burdensome, and requires expertise in multiple domains. Furthermore, the resulting sensing framework lacks generalizability, making it difficult to apply to different scenarios. In the research, we propose an automated mobile sensing strategy for human behaviour understanding. First, we establish a knowledge base and consolidate rules for data collection and effective feature construction. Then, we introduce the multi-granular human behaviour representation and design procedures for leveraging large language models to generate strategies. Our approach is validated through blind comparative studies and usability evaluation. Ultimately, our approach holds the potential to revolutionise the field of mobile sensing and its applications.

CCS CONCEPTS

• **Applied computing**;

KEYWORDS

Self-report survey, Mobile sensing, Human behavioural modelling, Large language models, Human-computer collaboration

*Corresponding author.

UbiComp Companion '24, October 5–9, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1058-2/24/10.
https://doi.org/10.1145/3675094.3678423

ACM Reference Format:
Nan Gao, Zhuolei Yu, Yue Xu, Chun Yu, Yuntao Wang, Flora D. Salim, and Yuanchun Shi. 2024. Leveraging Large Language Models for Generating Mobile Sensing Strategies in Human Behavior Modeling. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '24), October 5–9, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3675094.3678423

1 INTRODUCTION

The development of the Internet of Things (IoT) has transformed how we capture and analyze digital traces of daily life. Mobile sensing, a form of passive sensing using smartphone sensor data, plays a crucial role in this transformation. By leveraging data from software and hardware sensors, mobile sensing provides a comprehensive understanding of human behaviours [10, 36, 40]. Compared to wearable and environmental sensing, mobile sensing offers unobtrusive, long-term data collection in real-world settings, reducing user burden and providing convenience without additional devices [23]. Additionally, multiple sensors in mobile devices yield rich, diverse data, facilitating a contextual understanding of the surroundings.

Recently, mobile sensing has become popular for understanding human behaviours, such as affective states [36], academic performance [38], life satisfaction [42], and personality [10]. It serves as an effective *Quantified-Self* tool [24] to enhance self-awareness and well-being, with applications in health monitoring and personalized services. For example, Gao et al. [10] predicted Big-5 personality traits using call logs, message logs, and accelerometer data. Wampfler et al. [36] predicted affective states using touch and IMU data. Wang et al. [38] used activity, conversational interaction, and mobility data to predict college students' GPA.

However, understanding human behaviours through mobile sensing presents significant challenges, especially for complex behaviours like well-being and personality traits. One one hand, researchers need a deep understanding of relevant domain knowledge (e.g., psychology [10, 15], medicine [29], education [9, 38]) to extract pertinent features effectively. Expertise in sensor combinations, battery optimization, device settings, and data-driven modeling is essential for accurate models. For instance, a study on social functioning in individuals with schizophrenia [40] used various mobile sensing data types and extracted features related to social functioning, highlighting the need for domain knowledge and machine learning skills.

On the other hand, traditional mobile sensing studies often focus on specific research objectives (e.g., measuring depression during COVID-19 [28], identifying time-killing moments on smartphones [3], predicting weekend nightlife drinking behaviour [26]), resulting in frameworks that lack generalizability. This makes it challenging to apply them to different scenarios and participants, especially with minor variations in sensor usage.

Therefore, we aim to explore the automation of mobile sensing strategies for dynamic research objectives. While automation has been implemented in traditional modeling tasks (e.g., AutoML [17] and Auto-Sklearn [6]), most focus on traditional tabular data rather than mobile sensing settings and do not effectively utilize semantic information. Our research questions are: *1. What specific types of data should be collected to achieve different research objectives using mobile sensing technologies? 2. How can the collected data be effectively utilized to generate meaningful features that align with the research objective? 3. Which models can be utilized, and what is the estimated performance based on the research objective?*

To address these questions, we propose an automated mobile sensing strategy generation system. We reviewed mobile sensing studies from top venues, building a knowledge base. From this, we consolidated rules for feature construction, sensor selection, and model suggestions. We also developed a multi-granular human behavior decomposition mechanism to understand behaviors at varying levels. Large Language Models (LLMs) were utilized in five steps of strategy generation. The system outputs automated mobile sensing strategies that dynamically respond to user inquiries. Our contributions are as follows:

- We establish a mobile sensing knowledge base from 55 studies in reputable venues such as CHI and IMWUT, identifying rules for effective feature construction and sensor selection.
- We develop a multi-granular human behaviour representation mechanism for understanding behaviours in mobile sensing settings, aiding in effective feature construction.
- We propose an automated mobile sensing strategy that provides suggestions for data selection, feature construction, model building and performance estimation.

2 RELATED WORKS

2.1 Modelling Human Behaviours using Mobile Sensing Technologies

Mobile sensing technologies revolutionize understanding human behavior, enabling predictions of personality traits [10], depression [41], stress-resilience [1], social anxiety [30], and schizophrenia [39]. They also explore links with alcohol consumption [26], behavior post-promotion [27], time-killing on smartphones [3], and notification response time [13]. Capturing real-time data in natural settings, mobile sensing offers unprecedented insights into human life.

While mobile sensing infers various aspects of human behavior, each requires comprehensive study design, data collection, and feature construction. Researchers typically invest significant time in these areas. Traditionally, data is collected via background apps (e.g., SensingKit [20], AWARE [5], AWARE-Light [35], CARP [2]), but excessive data collection poses challenges like unused data, battery drain, and privacy concerns, reducing participant willingness and requiring extensive post-processing [35]. Limited data collection, however, restricts understanding due to budget and ethical constraints. We propose **RQ1** to optimize data collection.

Feature engineering, creating new features from raw data [22], is time-consuming and requires multi-domain expertise. Effective features enhance model performance, while poor features yield poor results. Traditional features, like statistical measures [4], have limited effectiveness due to human behavior's complexity. We propose **RQ2** to enable automated feature construction, reducing reliance on human expertise and streamlining data collection by rationalizing sensor data selection.

Accurate prediction models are essential for understanding behavior, approached through regression [9, 10, 40] or classification [3, 25]. Traditional models include *Random Forest* (RF) [33], *Gradient Boosting* (GB) [7], and *Naive Bayes* (NB) [31]. Neural networks and deep learning are limited by small participant samples. Researchers need to estimate model performance before studies. We propose **RQ3** to help identify suitable models, understand expected performance, aid informed decisions, and recognize prediction limitations.

2.2 AutoML and Large Language Models

Auto Machine Learning (AutoML) [17] offers automated solutions for identifying efficient machine learning pipelines, with notable successes including AutoSklearn [6], Auto-WEKA [21], and Auto-Pytorch [18]. However, these focus on traditional features, overlooking semantic information. Large Language Models (LLMs) excel in natural language processing [34], encapsulating a wealth of domain knowledge. Hollmann [14] proposed CAAFE, leveraging LLMs for semantically meaningful feature engineering from dataset descriptions, but focused on simple tabular data. Applying such methods to complex sensing data and human behavior research remains underexplored. The prevalence of LLMs has opened up new possibilities for understanding human needs by exploring behavior-related variable correlations. Their embedded domain knowledge can automate data science tasks involving intricate contextual information. This intersection of AutoML and LLMs presents a promising direction for future research.

3 METHODOLOGY

3.1 Construction of Knowledge Base

We focused on papers using only mobile sensing, excluding other sources like wearables, to construct our knowledge base. We selected articles from top venues in mobile sensing and ubiquitous

Table 1: An overview of features components summarised from the knowledge base

Component	Category	Descriptions	Example values
Time span	Duration	Daily epoches Past to present	Morning, afternoon, night In the last 30 minutes
	Periodicity	Recurrence	Daily, weekly, monthly
Metrics	Statistical	Central tendency Dispersion Shape Direct	Mean, median, mode Standard deviation, variance, range Skewness, kurtosis Temperature, screen on state, location
		Others	Count, magnitude, sum, slope, max, min frequency, ratio, proportion
	Regularity	Regularity	Mean Squares Successive Difference (MSSD), regularity index, consistency score
		Circadian rhythms	Same as above
	Relation	Correlation Ranking	Pearson, Spearman, Kendall tau correlation The most frequent place visited
	Diversity	Diversity of values	Shannon entropy
	Similarity	Similarity	Cosine, Jaccard, Hamming distance
	Spatial	Spatial	Distance, density, location
	Temporal	Temporal	Duration, frequency, trend
	Other	Other measures	Fast Fourier Transform (FFT), Mel Frequency Cepstral Coefficient (MFCC)

computing: CHI (Conference on Human Factors in Computing Systems) and IMWUT (Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies). The selection process included: 1. Accessing the ACM advanced search website [1]. 2. Search Within: 'Title' = (mobile OR smartphone) AND (sensing OR sensors OR sensor OR sense) NOT (wearable OR wristband OR desktop OR wrist-worn OR environmental OR environment OR laptop) 3. Apply the filters successively to ensure the inclusion of relevant research articles: a) Select 'UbiComp: Ubiquitous Computing' AND 'Research Article'. b) Select 'CHI: Conference On Human Factors In Computing Systems' AND 'Research Article'. c) Select 'Proceedings Of The ACM On Interactive, Mobile, Wearable And Ubiquitous Technologies' AND 'Research Article' In total, we collected 121 papers: 42 from IMWUT, 22 from CHI, and 57 from Ubicomp. After meticulous review, we retained 55 papers that exclusively used mobile sensing and focused on human behaviour.

3.2 Overview of Data Sources

From the reviewed papers, we identified commonly used sensors for mobile sensing studies, excluding those used in fewer than two papers due to potential data collection difficulties. We also standardized the names of data sources for consistency. The most commonly used sensors are:

- **Hardware Sensors**. Integral components in mobile devices that monitor physical activities. Common sensors include: *[Accelerometer, Gyroscope, Light, Magnetometer, Gravity, Temperature, Humidity, Orientation, Barometer, Proximity, Microphone, Bluetooth, WiFi]*.
- **Software Senors**. Derived from hardware sensors combined with software models to deduce new variables. Common sensors include: *[Application, Calls, Message, GPS/Location, Notification, Keyboard]*.
- **Contextual Information**. Provides insight into the surrounding environment or circumstances of device usage, essential for understanding user behaviour and preferences. Common data includes: *[Screen, Time, Date, Battery]*.

[1] https://dl.acm.org/search/advanced

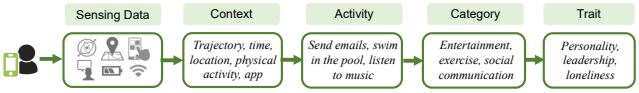


Figure 1: Multi-granular human behaviour representation

3.3 Overview of Features

Our review of the 55 studies revealed inconsistencies in feature construction. Some studies relied solely on statistical features, while others incorporated meaningful features but lacked organized granularity of human behaviour. Time spans were often inconsistently applied or not mentioned at all. This area lacks a clear principle for designing effective features.

After analyzing current mobile sensing studies, we found that effective features for human-centered mobile sensing typically consist of three components: the time span of the sensing data, the metrics used for measurements, and the specific human behaviours being studied. For instance, the feature *"Duration of screen time per weeknight"* includes the metric *"Duration of time"*, the atomic behaviour *"screen"* and the time span *"weeknight"*. Table 1 summarizes commonly used time spans and metrics.

3.4 Model and Performance

Analysis of 55 mobile sensing studies shows that most research uses similar machine learning models: *Random Forest*, *Gradient Boosting Machine*, *Linear Regression*, *Gaussian Mixture Model*, *Support Vector Machine*, *Naive Bayes*, *K-nearest Neighbour*, and *Logistic Regression*. Some models, like *Random Forest* and *Gradient Boosting*, handle complex relationships well, making them robust for high-dimensional data. The primary goal of using these models is to evaluate the effectiveness of features in predicting research objectives. However, providing recommendations on model choice is useful. Knowing the approximate performance level for the research objective helps guide researchers' expectations and decisions.

4 MULTI-GRANULAR HUMAN BEHAVIOUR REPRESENTATION

Understanding human behaviours deeply is key to effective feature construction and successful mobile sensing studies. Translating sensing signals, smartphone usage, and context descriptors into specific human behaviours (e.g., emotion, alcohol consumption) remains challenging due to the complex nature of human behaviours [8, 12]. Many studies overlook this aspect, extracting data like statistical features or trajectory data without considering the broader context. This not only wastes time but also fails to cover all facets of human behaviour comprehensively.

Human behaviour refers to the potential and expressed capacity for physical, mental, and social activity in response to internal

and external stimuli throughout life [19]. It has been explored by various fields such as psychology, sociology, ethology, and human-centered design. While there are many facets of human behaviour, no single definition or field of study can encapsulate its entirety. For example, behaviour can be decomposed by temporal phases (pre-natal life, infancy, childhood, adolescence, adulthood, and old age) [19], reactive modes (reactive and deliberative behaviours) [32], or dimensions (actions, cognition, and emotion). To capture human behaviours through smartphones, we propose a multi-granular human behaviour representation mechanism. This mechanism serves as a foundation for constructing meaningful features in mobile sensing research. It comprises four dimensions that encompass human behaviours at varying levels of granularity: contexts, activities, categories, and traits (see Figure 1).

- *Context:* This level includes information directly inferred or easily calculated from smartphone sensors, such as location/trajectory (GPS), physical activity (Android Activity Recognition API [2]), time, and screen usage.
- *Activity:* This level identifies specific activities or behaviours exhibited by individuals, such as sending emails, swimming, or listening to music.
- *Category:* At this level, similar behaviours are grouped based on shared characteristics or attributes, allowing the identification of commonalities and patterns. Categories may include entertainment, exercise, communication, and social activities.
- *Trait:* This level considers enduring characteristics or traits intrinsic to individuals, reflecting their behaviour patterns, such as personality traits, social abilities, leadership, and loneliness.

For example, to investigate someone's mood instability (a trait), we can identify relevant categories like stress, happiness, and sadness. Within these categories, activities contributing to mood fluctuations might include work-related tasks, spending time with loved ones, or engaging in hobbies. At the context level, we can examine specific atomic activities like using the smartphone, opening social media apps, texting, and going to bed. By considering these different levels of granularity—from traits to categories to activities and contexts—we can construct a comprehensive representation of human behaviour, enabling a deeper understanding of complex phenomena like mood instability.

5 AUTOMATED MOBILE SENSING FRAMEWORK

5.1 Design Rules for Mobile Sensing Strategies

To achieve effective mobile sensing strategies, we follow five steps (see Figure 5). The system outputs the strategy based on the user's inquiry after these steps.

5.1.1 Information Extraction (Step 1). The user initiates an inquiry, such as *"I wish to understand the mood instability of this user during the night."* The system extracts the research objective, which in this case is *"mood instability during the night"*. Next, the system defines the level of human behaviour (trait, category, activity, or context) based on the multi-granular human behaviour mechanism described in Section 4. Since *"mood instability during the night"* is an intrinsic trait affecting behaviour patterns, it is considered at the *trait level*.

[2]Android Activity Recognition API: https://developers.google.com/location-context/activity-recognition

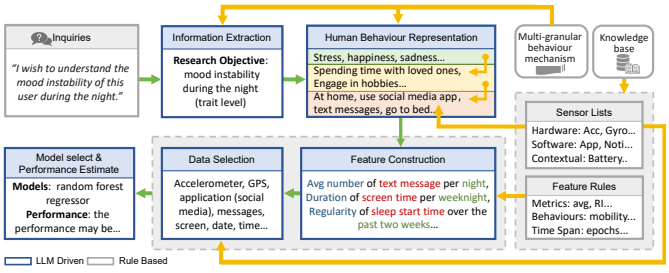


Figure 2: The generation process of mobile sensing strategies involves two main data flows: the user's inquiry in natural language (green arrows) and the designed rules (yellow arrows). These flows merge to produce the final mobile sensing strategies.

5.1.2 Human Behavior Representation (Step 2). The system extracts multi-granular behaviours based on the objective, moving hierarchically from category to activity and context levels. Context-level behaviours are inferred from smartphone sensing data, using sensor lists from the knowledge base to generate relevant behaviours.

5.1.3 Feature Construction (Step 3). The system constructs comprehensive features for modeling the research objective. Effective features consist of the time span of the data, measurement metrics, and specific behaviours. Using context-level behaviours from Step 2, the system selects appropriate metrics and time spans. For instance, a feature for *"mood instability during the night"* could be the *"regularity of sleep start time over the past two weeks"*.

5.1.4 Data Selection (Step 4). The system determines the data to compute the features from Step 3. The sensing data source is selected from the identified sensors in Section 3.2, including hardware, software sensor data, and contextual information. For example, the feature *"regularity of sleep start time"* requires time and sleep data, identified using sensors like the accelerometer and gyroscope.

5.1.5 Model and its Estimated Performance (Step 5). The system suggests a machine learning model based on the research objective, selected data sources, and constructed features. Good features enhance performance, while limited data sources may hinder accuracy. The system estimates performance using natural language and provides reasoning. For example, modeling psychological traits may result in lower performance due to their complexity and variability.

5.2 Prompt Structure

Our prompt structure, inspired by [37], consists of three main components: (1) **Prefix**. A clear and concise introduction outlining the prompt's purpose and design rules, as detailed in Section 5.1, provding a high-level overview and sets the context for the examples that follow. (2) **Examples**. Each example is divided into three parts:

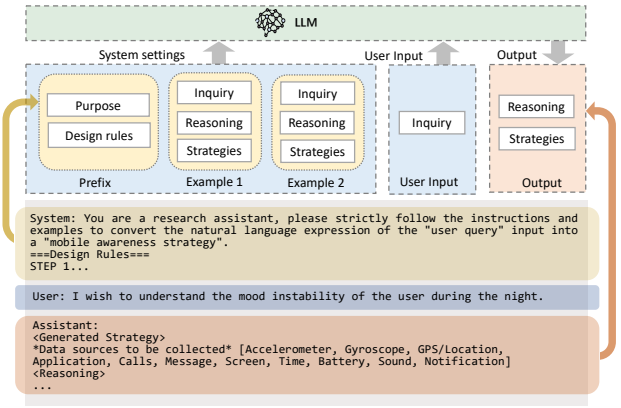


Figure 3: An example illustrating the proposed prompt structure

a) *Inquiry*. Presents a natural language inquiry to enhance understanding, such as, *"I wish to understand the mood instability of this user during the night."* b) *Reasoning*. Explains the reasoning behind each design decision with step-by-step justifications, following the design rules. c) *Mobile Sensing Strategy*. Outlines the chosen strategies, specifying data to be collected and features to be constructed. (3) **User Input**. Users provide their own inquiry related to their objective, expressed in natural language.

This approach provides a comprehensive framework that guides LLMs in reasoning and formulating mobile sensing strategies based on the given inquiries. The number of examples can be adjusted according to user requirements. Figure 3 shows an example of our prompt structure.

6 EVALUATION

In our experiment, we used GPT-3.5-turbo for its robust capabilities in generating coherent, context-rich responses to complex prompts. Users initiated interactions with the model using a prefixed indicator "INPUT". For example, a researcher wishing to model mood instability could type: *INPUT: I wish to understand the mood instability of the user during the night.*

6.1 Expert Evaluators

Unlike typical user studies that rely on easily recruited ordinary participants, our research targets experts in the field of mobile sensing to provide valuable insights into human behaviours. We enlisted 8 experts with significant experience in modeling human behaviours using mobile sensing technologies, averaging 4.25 years of research experience. Although the number of evaluators is limited, their profound understanding of mobile sensing techniques provides substantial insights into the system.

6.2 Procedure

We conducted two evaluation studies: a comparative study and a usability study. The comparative study evaluated the effectiveness of the automated mobile sensing strategy against existing strategies, using the *Blind Comparison* method [11]. In the usability study, experts typed any inquiry they wanted and then completed a survey and participated in an interview.

6.2.1 Comparative Study. For the comparative study, we selected two highly cited mobile sensing tasks (over 100 citations). We extracted research objectives, selected data sources, and constructed features based on existing descriptions, then applied our sensing strategy generation system. To ensure fairness, we did not include selected models and performance metrics. We maintained consistent data and feature descriptions, excluding sensor details. The primary difference between the existing and automated strategies was the sensing data and features used.

Experts were presented with both existing and auto-generated strategies in a randomized order to avoid *Order Effects* [11]. They compared the strategies, assessing effectiveness, interpretability, relevance, and completeness on a 5-point Likert scale from 1 (very negative) to 5 (very positive). This assessment was conducted through semi-structured interviews, repeated for each expert until both tasks were completed.

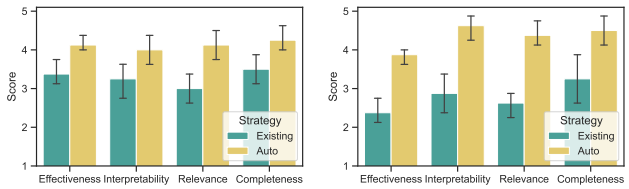
6.2.2 Usability Study. In this usability study, experts independently used the automatic sensing strategies system. They typed inquiries into the system to understand various human behaviours through smartphones. The system generated strategies with a step-by-step reasoning process, including data sources to be collected and features to be constructed. We used an adapted NASA-TLX [16] evaluation method to assess the generated strategies, excluding questions on temporal demand or effort as the system required minimal waiting time. Participants rated the following on a 5-point Likert scale, with 1 being the most negative and 5 the most positive: (1) Mental demand: *How mentally demanding was the task?* (2) Physical demand: *How physically demanding was the task?* (3) Performance: *How successful were you in accomplishing what you planned to do?*

Participants also evaluated their overall experience on a 5-point Likert scale, with 1 indicating 'not at all' and 5 indicating 'very much': (1) Satisfaction: *How satisfied are you with the automated generated strategy?* (2) Enhanced Understanding: *Does the automated strategy enhance your understanding of the research objective?* (3) Ease of use: *How easy was the system to use?* (4) Willingness to reuse: *How likely are you to use this assistant again in the future?*

A concluding interview was conducted to gain deeper insights into the experts' thoughts on the automated mobile sensing strategies, including the system's effectiveness, its impact on their research process, and suggestions for improvements.

6.3 Result and Discussion

6.3.1 Comparative Performance Analysis. We compared automated and existing strategies in two studies: predicting *Brain Functional Connectivity* (Study A) and understanding *Compound Emotion* (Study B). Figure 4 shows the results for effectiveness, interpretability, relevance, and completeness based on expert opinions. The automated strategy ('Auto') consistently outperformed the existing strategy ('Existing') in all dimensions: Effectiveness: Auto 4.0 (STD = 0.37) vs. Existing 2.88 (STD = 0.72). Interpretability: Auto 4.31 (STD = 0.60) vs. Existing 3.06 (STD = 0.77). Relevance: Auto 4.25 (STD = 0.58) vs. Existing 2.81 (STD = 0.54). Completeness: Auto 4.375 (STD = 0.5) vs. Existing 3.375 (STD = 0.81).



(a) Task: Study A (b) Task: Study B

Figure 4: The evaluation results for both studies from experts

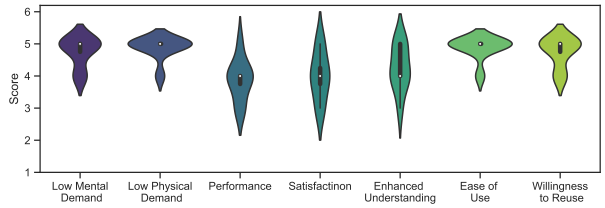


Figure 5: Ratings for the automated mobile sensing strategy from experts

Overall, the automated strategies outperformed the existing strategies in all dimensions. However, Performance varied between studies due to different research objectives. Study B's reliance on basic statistical features (e.g., "Longitude, Altitude, Latitude of GPS") was less relevant compared to our proposed features (e.g., "Distance travelled per day/weeknight").

Two experts raised concerns about the feasibility of computing the proposed features, as they are more intricate than low-level statistical features. However, most features can still be computed using mature algorithms. Three experts found the proposed features insightful and beneficial for understanding user behaviour. For example, Expert 2 noted, *"I was pleasantly surprised to find that application data were used in the automated strategy. Obviously, it would be useful for understanding user brain function connectivity"*.

6.3.2 Usability Analysis. Experts tested the system independently and evaluated based on seven dimensions: *Mental Demand*, *Physical Demand*, *Performance*, *Satisfaction*, *Enhanced Understanding*, *Ease of Use*, and *Willingness to Reuse*. As there are no existing automated mobile sensing strategies for comparison, experts rated the system directly on these dimensions. The usability ratings (Figure 5) showed average values above 3 for all dimensions, indicating good performance. Mental and physical demands were low, and experts expressed a strong willingness to use the system for future research.

Experts may propose varying research objectives. While some behaviours are easier to infer (e.g., smartphone addiction), others are more challenging (e.g., heart attack). Despite this, 5 out of 8 experts found the generated strategies meaningful and expressed a desire to use the system for designing their own experiments. Three experts found the proposed features inspiring and valuable. For instance, Expert 5 remarked, *"I was pleasantly surprised to see that application data was incorporated into the automated strategy. Including participants' usage of grooming software would undoubtedly make the experiment more comprehensive"*.

However, there was one case where an expert felt the system's performance was less satisfactory. This dissatisfaction arose from their research objective, which focused on suggesting changes or improvements in behaviour rather than understanding, modeling, or predicting behaviours. Users should ensure that their research objective aligns with understanding human behaviours through mobile sensing. In another scenario, the strategy suggested the feature *"The number of positive/negative messages sent per day"*, which raised two primary concerns: potential violation of privacy rights and ambiguity in distinguishing positive from negative messages. While some generated sensors/features may be valuable, their real-world applicability could be constrained. Further discussion can be found in Section 7.

7 IMPLICATIONS AND LIMITATIONS

This research proposes an automatic generation system for mobile sensing strategies to understand human behaviour. For researchers, it reduces the burden of designing strategies, offers effective feature suggestions, and aids decision-making based on estimated performance. The system can adapt to different research objectives, providing tailored suggestions and experimental designs. For individuals, it enhances self-awareness by offering an objective method to understand themselves through passive sensing data, potentially improving well-being and quality of life.

However, this study has limitations. Firstly, not all devices have the same sensors, and availability varies. For example, some devices lack barometers or thermometers, and iOS devices generally have more constraints than Android devices. Secondly, the study did not cover parameter tuning and data cleaning, focusing instead on data source selection, feature construction, model building, and performance estimation. Tools like AutoML can manage parameter tuning. Thirdly, the research centers on designing automatic mobile sensing strategies without computing or implementing features. The main contribution is strategy design, saving researchers time and reducing the burden of data selection and feature construction, paving the way for future studies. Lastly, privacy is a concern when collecting data. Although this study does not involve actual data collection, future researchers should implement necessary privacy protection measures. Data processed for automated human behaviour computation would be strictly protected and processed on the user's device, minimizing privacy concerns.

8 CONCLUSION

This paper introduces an automated mobile sensing strategy generation system that allows users to input inquiries related to understanding human behaviours through smartphones. This automation reduces the burden on researchers and provides new insights for mobile sensing strategy design. Future work will explore automatic feature computation to develop intelligent systems that understand human behaviour, ultimately assisting individuals in gaining self-awareness and enhancing well-being.

9 ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (Grant No. 62302252) and the China Postdoctoral Science Foundation (Grant No. 2023M731949).

REFERENCES

[1] Daniel A Adler, Vincent W-S Tseng, Gengmo Qi, Joseph Scarpa, Srijan Sen, and Tanzeem Choudhury. 2021. Identifying mobile sensing indicators of stress-resilience. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–32.

[2] Jakob E Bardram. 2020. The CARP mobile sensing framework–A cross-platform, reactive, programming framework and runtime environment for digital phenotyping. *arXiv preprint arXiv:2006.11904* (2020).

[3] Yu-Chun Chen, Yu-Jen Lee, Kuei-Chun Kao, Jie Tsai, En-Chi Liang, Wei-Chen Chiu, Faye Shih, and Yung-Ju Chang. 2023. Are You Killing Time? Predicting Smartphone Users' Time-killing Moments via Fusion of Smartphone Sensor Data and Screenshots. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[4] Maryam Banitalebi Dehkordi, Abolfazl Zaraki, and Rossitza Setchi. 2020. Feature extraction and feature selection in smartphone-based activity recognition. *Procedia Computer Science* 176 (2020), 2655–2664.

[5] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6.

[6] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-sklearn 2.0: The next generation. *arXiv preprint arXiv:2007.04074* 24 (2020).

[7] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis* 38, 4 (2002), 367–378.

[8] Nan Gao. 2022. *Human behaviour sensing and profiling in the wild.* Ph. D. Dissertation. Ph. D. Dissertation. RMIT University.

[9] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. 2020. n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–26.

[10] Nan Gao, Wei Shao, and Flora D Salim. 2019. Predicting personality traits from physical activity intensity. *Computer* 52, 7 (2019), 47–56.

[11] Kerri A Goodwin and C James Goodwin. 2016. *Research in psychology: Methods and design.* John Wiley & Sons.

[12] Consuelo Granata, Aurelien Ibanez, and Philippe Bidaud. 2015. Human activity-understanding: A multilayer approach combining body movements and contextual descriptors analysis. *International Journal of Advanced Robotic Systems* 12, 7 (2015), 89.

[13] Judith S Heinisch, Nan Gao, Christoph Anderson, Shohreh Deldari, Klaus David, and Flora Salim. 2022. Investigating the Effects of Mood & Usage Behaviour on Notification Response Time. *arXiv preprint arXiv:2207.03405* (2022).

[14] Noah Hollmann, Samuel Müller, and Frank Hutter. 2023. GPT for Semi-Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering. *arXiv preprint arXiv:2305.03403* (2023).

[15] Juyoung Hong, Jiwon Kim, Sunmi Kim, Jaewon Oh, Deokjong Lee, San Lee, Jinsun Uh, Juhong Yoon, and Yukyung Choi. 2022. Depressive symptoms feature-based machine learning approach to predicting depression using smartphone. In *Healthcare*, Vol. 10. MDPI, 1189.

[16] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M Walker. 2011. Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX). *IIE transactions on healthcare systems engineering* 1, 2 (2011), 131–143.

[17] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated machine learning: methods, systems, challenges.* Springer Nature.

[18] Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. 2021. PyTorch. *Programming with TensorFlow: Solution for Edge Computing Applications* (2021), 87–104.

[19] Jerome Kagan, Marc H Bornstein, and Richard M Lerner. 2020. human behaviour. Encyclopedia Britannica.

[20] Kleomenis Katevas, Hamed Haddadi, and Laurissa Tokarchuk. 2014. Poster: Sensingkit: A multi-platform mobile sensing framework for large-scale experiments. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 375–378.

[21] Lars Kotthoff, Chris Thornton, Holger H Hoos, Frank Hutter, and Kevin Leyton-Brown. 2019. Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA. *Automated machine learning: methods, systems, challenges* (2019), 81–95.

[22] Max Kuhn and Kjell Johnson. 2019. *Feature engineering and selection: A practical approach for predictive models.* Chapman and Hall/CRC.

[23] Francisco Laport-López, Emilio Serrano, Javier Bajo, and Andrew T Campbell. 2020. A review of mobile sensing systems, applications, and opportunities. *Knowledge and Information Systems* 62, 1 (2020), 145–174.

[24] Victor R Lee. 2014. What's happening in the" Quantified Self" movement? *ICLS 2014 proceedings* (2014), 1032.

[25] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–32.

[26] Lakmal Meegahapola, Florian Labhart, Thanh-Trung Phan, and Daniel Gatica-Perez. 2021. Examining the social context of alcohol drinking in young adults with smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–26.

[27] Subigya Nepal, Shayan Mirjafari, Gonzalo J Martinez, Pino Audia, Aaron Striegel, and Andrew T Campbell. 2020. Detecting job promotion in information workers using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–28.

[28] Subigya Nepal, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Alex Dasilva, Meghan Meyer, and Andrew Campbell. 2022. COVID student study: A year in the life of college students during the COVID-19 pandemic through the lens of mobile phone sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[29] Mikio Obuchi, Jeremy F Huckins, Weichen Wang, Alex Dasilva, Courtney Rogers, Eilis Murphy, Elin Hedlund, Paul Holtzheimer, Shayan Mirjafari, and Andrew Campbell. 2020. Predicting brain functional connectivity using mobile sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 1 (2020), 1–22.

[30] Haroon Rashid, Sanjana Mendu, Katharine E Daniel, Miranda L Beltzer, Bethany A Teachman, Mehdi Boukhechba, and Laura E Barnes. 2020. Predicting subjective measures of social anxiety from sparsely collected mobile sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–24.

[31] Irina Rish et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. 41–46.

[32] Bernard Schmidt. 2000. *The modelling of human behaviour.* Vol. 132. Society for Computer Simulation International.

[33] Mark R Segal. 2004. Machine learning benchmarks and random forest regression. (2004).

[34] Alexander Tornede, Difan Deng, Theresa Eimer, Joseph Giovanelli, Aditya Mohan, Tim Ruhkopf, Sarah Segel, Daphne Theodorakopoulos, Tanja Tornede, Henning Wachsmuth, et al. 2023. AutoML in the Age of Large Language Models: Current Challenges, Future Opportunities and Risks. *arXiv preprint arXiv:2306.08107* (2023).

[35] Niels van Berkel, Simon D'Alfonso, Rio Kurnia Susanto, Denzil Ferreira, and Vassilis Kostakos. 2023. AWARE-Light: a smartphone tool for experience sampling and digital phenotyping. *Personal and Ubiquitous Computing* 27, 2 (2023), 435–445.

[36] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R Schinazi, Markus Gross, and Christian Holz. 2022. Affective state prediction from smartphone touch and sensor data in the wild. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[37] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[38] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 295–306.

[39] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, et al. 2017. Predicting symptom trajectories of schizophrenia using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.

[40] Weichen Wang, Shayan Mirjafari, Gabriella Harari, Dror Ben-Zeev, Rachel Brian, Tanzeem Choudhury, Marta Hauser, John Kane, Kizito Masaba, Subigya Nepal, et al. 2020. Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[41] Weichen Wang, Subigya Nepal, Jeremy F Huckins, Lessley Hernandez, Vlado Vojdanovski, Dante Mack, Jane Plomp, Arvind Pillai, Mikio Obuchi, Alex Dasilva, et al. 2022. First-gen lens: Assessing mental health of first-generation students across their first year at college using mobile sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 2 (2022), 1–32.

[42] Onur Yürüten, Jiyong Zhang, and Pearl HZ Pu. 2014. Predictors of life satisfaction based on daily activities from mobile sensor data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 497–500.