Automated Mobile Sensing Strategies Generation for Human Behaviour Understanding

NAN GAO, Department of Computer Science and Technology, Tsinghua University, China and University of New South Wales (UNSW), Australia
ZHUOLEI YU, Department of Computer Science and Technology, Tsinghua University, China
CHUN YU, Department of Computer Science and Technology, Tsinghua University, China
YUNTAO WANG, Department of Computer Science and Technology, Tsinghua University, China
FLORA D. SALIM, University of New South Wales (UNSW), Australia
YUANCHUN SHI, Department of Computer Science and Technology, Tsinghua University, China

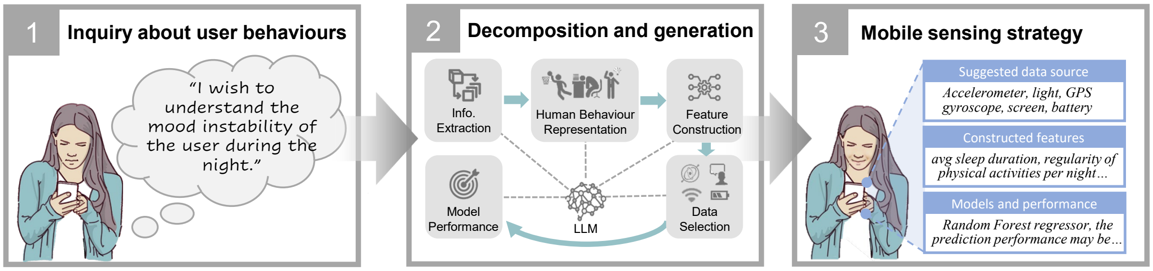

1
Inquiry about user behaviours
"I wish to understand the mood instability of the user during the night."
2
Decomposition and generation
Info. Extraction
Human Behaviour Representation
Feature Construction
Model Performance
LLM
Data Selection
3
Mobile sensing strategy
Suggested data source
Accelerometer, light, GPS gyroscope, screen, battery
Constructed features
avg sleep duration, regularity of physical activities per night…
Models and performance
Random Forest regressor, the prediction performance may be…

Fig. 1. A researcher inquires the system about the mood instability of smartphone users during the night. The system extracts the research objective and decomposes it into multi-granular behaviours. Following the defined design rules, the system utilizes large language models to generate a mobile sensing strategy, including suggested data sources, constructed features, models and estimated performance.

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. To address these challenges, we propose an automated mobile sensing strategy for human behaviour understanding. First, we establish a

Authors' addresses: Nan Gao, Department of Computer Science and Technology, and Tsinghua University, Beijing, China and University of New South Wales (UNSW), Sydney, Australia, 1466, nangao@tsinghua.edu.cn; Zhuolei Yu, Department of Computer Science and Technology, and Tsinghua University, Beijing, China, yuzl21@mails.tsinghua.edu.cn; Chun Yu, Department of Computer Science and Technology, and Tsinghua University, Beijing, China, chunyu@mail.tsinghua.edu.cn; Yuntao Wang, Department of Computer Science and Technology, and Tsinghua University, Beijing, China, yuntaowang@tsinghua.edu.cn; Flora D. Salim, flora.salim@unsw.edu.au, University of New South Wales (UNSW), Sydney, Australia, 1466, flora.salim@unsw.edu.au; Yuanchun Shi, Department of Computer Science and Technology, and Tsinghua University, Beijing, China, shiyc@tsinghua.edu.cn.

2474-9567/2023/8-ART52 $15.00
https://doi.org/10.1145/1122445.1122456

knowledge base and consolidate rules for effective feature construction, data collection, and model selection. 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.

ACM Reference Format:
Nan Gao, Zhuolei Yu, Chun Yu, Yuntao Wang, Flora D. Salim, and Yuanchun Shi. 2023. Automated Mobile Sensing Strategies Generation for Human Behaviour Understanding. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 37, 4, Article 52 (August 2023), 18 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The development of *Internet of Things* (IoT) devices has revolutionised the way we capture and analyse digital traces of an individual's daily life. Mobile sensing, a form of passive sensing, plays a crucial role in this revolution. By leveraging smartphone sensor data, encompassing both software and hardware sensors, mobile sensing enables a comprehensive understanding of human behaviours [11, 40, 44]. Compared with wearable sensing and environmental sensing, mobile sensing has distinct advantages. It allows for unobtrusive data collection in real-world settings over extended periods, providing valuable insights into human behaviours. This passive approach reduces the burden on users and offers convenience without requiring additional devices [25]. Additionally, the utilisation of multiple sensors in mobile devices yields rich and diverse data, facilitating a contextual understanding of the surroundings.

Recently, mobile sensing has become increasingly popular in understanding human behaviours such as affective states [40], academic performance [42], life satisfaction [47], social functioning [44] and even personality [11]. It could be used as an effective *Quantified-Self* [26] tool to improve self-awareness and human well-being, and has the potential to be applied in various fields like health monitoring and personalised services. For instance, Gao et al. [11] predicted Big-5 personality traits of smartphone users using call logs, message logs, and accelerometer data. Wampfler et al. [40] predicted affective states using the smartphone touch data and *Inertial Measurement Unit* (IMU) data. Wang et al. [42] extracted behavioural features such as activity, conversational interaction, mobility, and class attendance to predict the cumulative GPA of college students.

However, conducting mobile sensing research studies to understand human behaviour, particularly complex or high-level behaviours like well-being and traits, poses significant challenges. On one hand, this process necessitates proficiency in multiple domains. Researchers had to have a deep understanding of domain knowledge related to the research objective, such as psychology in [11, 17], medical in [32], or education in [10, 42], to effectively extract relevant features. Additionally, they must be knowledgeable about sensor combinations that optimise battery usage, device settings, and enable the extraction of pertinent features. Furthermore, expertise in data-driven modelling is crucial to construct accurate models. For instance, in a study [44] for modelling the social functioning of individuals with schizophrenia, various mobile sensing data types (e.g., ambient sound/light, GPS, phone locks, and screen time) were utilised, and features related to social functioning (e.g., the number of locations visited, duration of social network app usage, and amplitude of ambient sound) were extracted. Finally, the performance of tree-ensemble models, linear regression, and SVR models were compared. This example highlights the need for expertise in domain knowledge, such as understanding schizophrenia, social functions, and machine learning skills in [44] when utilising mobile sensing to understand human behaviours.

On the other hand, traditional mobile sensing studies often focus on specific research objectives (e.g., measuring depression and anxiety during the COVID-19 [31], identifying time-killing moments on smartphones [4], predicting weekend nightlife drinking behaviour [29]), which demands a comprehensive study on its own. This results in a mobile sensing framework that lacks generalisability, which makes it challenging to apply the framework to different scenarios and participants, particularly when minor variations in sensor usage occur.

Therefore, we aim to explore the automation of mobile sensing strategies for dynamic and varying research objectives. While there have been successful implementations of automation in traditional modelling tasks, such as *AutoML* [19], *Auto-Sklearn* [7], *Auto-WEKA* [23], *Auto-Pytorch* [50] and closed commercial frameworks, most previous practices have focused on traditional tabular data rather than mobile sensing settings, and none have effectively utilised semantic information in an automated manner. Therefore, our research questions are as follows: *1. What specific types of data should be collected to achieve different research objectives in the context of human behavioural understanding using mobile sensing technologies? 2. How can the collected data be effectively utilised to generate meaningful features that align with the research objective? 3. Which models can be utilised, and what is the estimated performance based on the research objective?*

To address the above questions, we propose an automated mobile sensing strategy generation system (Figure 1). In particular, we conduct a comprehensive review of mobile sensing studies published in top venues within the field of mobile sensing and ubiquitous computing, constructing a knowledge base. From this review, we consolidated rules for effective feature construction, sensor selection, and model suggestions. Additionally, we develop a multi-granular human behaviour decomposition mechanism that allows the complete understanding of behaviours from varying levels. *Large Language Models* (LLMs) were then utilised in assisting the five steps during strategy generation. Finally, the system outputs the automated mobile sensing strategies that dynamically respond to user inquiries. In summary, our contributions are as follows:

- We establish a mobile sensing knowledge base by selecting 55 mobile sensing studies from reputable venues such as CHI and IMWUT. Through this process, we have identified the mobile sensing rules that can greatly enhance the effectiveness of feature construction, sensor selection, and model suggestions. These practices could have great potential to benefit researchers in designing and conducting mobile sensing studies in the future.
- We develop a multi-granular human behaviour representation mechanism for understanding human behaviours in mobile sensing settings. By categorising behaviours at different levels, the mechanism contributes to the effective feature construction for human understanding. This will benefit future researchers in gaining deeper insights and designing more impactful features in mobile contexts.
- We propose an automated mobile sensing strategy that provides suggestions for data selection, feature construction, model building, and performance estimation in response to user inquiries. The effectiveness of the proposed strategy has been validated through blind comparative studies and usability evaluation. This automation significantly reduces the manual configuration burden and enables the mobile sensing strategy to adapt to changing research objectives, ultimately enhancing the overall effectiveness and applicability of mobile sensing research studies.

The remainder of the paper is organised as follows. In Section 2, we present related works on traditional mobile sensing methods for human behavioural understanding and recent progress in automated data modelling tasks. Section 3 discusses the construction of the mobile sensing knowledge base, including data collection, data sources, features, models, and performance summary. Next, in Section 4, we introduce the multi-granular human behavioural decomposition mechanism. Section 5 presents the automated mobile sensing strategies, including feature construction, sensor selection, model building, and performance estimation. In Section 6, we evaluate the proposed strategies through two user studies. Section 7 lists the implications and limitations of our work. Finally, in Section 8, we summarise this research and indicate potential directions for future work.

2 RELATED WORKS

In this section, we begin by reviewing various examples of human behaviours that could be inferred from mobile sensing approaches. Subsequently, we delve into the procedures involved in mobile sensing, encompassing sensor data collection, feature selection, and model building. Lastly, we explore the recent advancements in automated machine learning approaches and large language models.

2.1 Modelling Human Behaviours using Mobile Sensing Technologies

Mobile sensing technologies have revolutionised the field of human behaviour understanding, allowing researchers to gain insights into various aspects of human life. One important application of mobile sensing technologies is the study of human psychological characteristics. Researchers have successfully utilised these technologies to predict personality traits [11], compound emotions [49], depression [45], stress-resilience [1], social anxiety [33], and even schizophrenia [43]. In addition, mobile sensing technologies have also been leveraged to study human daily behaviours. For instance, researchers have explored the relationship between mobile sensing data and alcohol drinking behaviour [29], developed methods to detect physiological and behavioural patterns after job promotion by passive sensing from smartphones [30], and investigated time-killing moments via fusion of smartphone sensor data and screenshots [4]. These studies underscore the diverse range of daily behaviours that can be captured and analysed using mobile sensing technologies. By capturing real-time data in naturalistic settings, mobile sensing technologies offer researchers unprecedented opportunities to understand various aspects of human life.

2.2 General Mobile Sensing Procedures

Though mobile sensing has been utilised to infer various aspects of human behaviours (see Section 2.1), each research objective requires a comprehensive study on its own. Researchers typically invest a significant amount of time in designing mobile sensing strategies, specifically identifying the data to be collected, constructing effective features, and selecting models for comparison.

2.2.1 Smartphone Data Collection. To collect data from smartphones, traditionally methods usually relied on a background app (e.g., SensingKit [22], AWARE [6], AWARE-Light [39], CARP[2]) that continuously runs and captures multiple sources of data. However, researchers often face challenges in determining which data to collect. While some opt for collecting as much data as possible without considering specific research objectives, this approach leads to unused data, drains battery life, and burdens researchers and participants. Moreover, it reduces willingness to participate due to privacy concerns and requires additional post-processing and memory storage [39]. Alternatively, some researchers may choose to collect specific data during the data collection, which may pose challenges in including additional data due to budget and ethical constraints. This limited data collection approach can hinder the comprehensive understanding of human behaviour and restrict the potential insights that could be derived from a broader range of data sources.

To address this issue, it is crucial to determine the specific types of data that should be collected to achieve different research objectives. Therefore, we propose the following research question (RQ1): *What specific types of smartphone data should be collected to achieve different research objectives in the context of human behavioural understanding* Answering RQ1 will provide valuable insights into optimising data collection strategies. By identifying the most relevant data types, researchers can minimise costs, reduce battery consumption, and alleviate the burden on both researchers and participants, while still achieving their research objectives.

2.2.2 Feature Construction Methods. Feature engineering is the process of creating new features from raw input data [24]. Among the steps involved in mobile sensing modelling, feature construction is one of the most time-consuming tasks [15]. Traditionally, effective feature construction from smartphones heavily relies on human practitioners with expertise in multiple domains, which greatly impacts model performance. Poorly constructed features could yield poor model results, while well-designed features enhance even the simplest models. Although some conventional features such as statistical features (standard deviation, mean, maximum) are commonly used [5], their effectiveness is limited due to the complex correlations and variability in human behaviour. Constructing effective features requires significant time, effort, and domain knowledge expertise, hindering scalability and generalisability for those unfamiliar with multiple domains. To tackle this issue, it is necessary to construct effective features from smartphone data for human behavioural understanding. Therefore, we propose the following research question (RQ2): *How can the collected data be effectively utilised to generate meaningful features that align with the research objective?* Addressing RQ2 will enable researchers to develop automated feature construction methods, reducing reliance on human expertise, and streamlining data collection by providing rationales for sensor data selection in response to RQ1.

2.2.3 Prediction Models and its Performance. To understand human behaviour, it is essential to develop prediction models. The modelling is generally considered a regression [10, 11, 44] or classification task [4, 28]. Traditional machine learning models, such as *Random Forest* (RF) [36], *Gradient Boosting* (GB) [8], *K-Nearest Neighbor* (KNN) [37], *Naive Bayes* (NB) [34], and *Support Vector Machine* (SVM) [3], are commonly employed in this context. However, the application of neural networks and deep learning techniques in human-centred studies is often limited due to the small sample of participants. Before conducting a mobile sensing study, researchers need to have a basic understanding of the estimated performance of the chosen model. For instance, in emotion understanding, the performance may be not highly accurate due to the subjective and complex nature of emotions [27] and researchers should have realistic expectations to guide their research objectives. Thus, it is crucial to select appropriate models and comprehend their performance characteristics. In light of this, we propose the following research question (RQ3): *Which models can be effectively utilised, and what is their estimated performance based on the research objective?* By addressing RQ3, researchers can identify suitable models for their specific research objectives and gain insights into the expected performance. This knowledge will assist in making informed decisions to select which behaviours to explore and understand the limitations of the prediction.

2.3 AutoML and Large Language Models

Auto Machine Learning (AutoML) [19] has emerged as a powerful tool in the data science landscape, offering automated solutions for identifying efficient machine learning pipelines. Notable successes in this field include AutoSklearn [7], Auto-WEKA [23], and Auto-Pytorch [20]. However, these existing AutoML approaches primarily focus on traditional statistical and mathematical features, leaving the potential of semantic information largely untapped. Recently, LLMs have experienced a series of breakthroughs. These models, pretrained on expansive textual data, have demonstrated exceptional performance in a variety of natural language processing tasks [38]. Importantly, LLMs encapsulate a wealth of domain knowledge, offering a promising avenue for automating data science tasks that require a deep understanding of context.

Hollmann [16] made a noteworthy contribution in this space by proposing CAAFE, an automated feature engineering method. CAAFE leverages LLMs to iteratively suggest semantically meaningful features for datasets based on their descriptions. However, Hollmann's work predominantly focuses on simple tabular datasets, leaving a gap in the application of such methods to more complex data types, such as sensing data, and to address more complex research objectives, such as understanding human behaviour. The prevalence of LLMs has opened up new possibilities for understanding human needs, particularly through exploring correlations between variables related to human behaviours. The vast domain knowledge embedded in LLMs can be harnessed to automate a wide range of data science tasks, especially those involving intricate contextual information. This intersection of AutoML and LLMs presents a promising direction for future research.

3 CONSTRUCTION OF KNOWLEDGE BASE

3.1 Data Collection

To construct our knowledge base, we focused on papers that solely utilised mobile sensing without the use of other sources such as wearables or earables. To ensure the reliability and relevance of the selected papers, we specifically searched for articles published in two top venues in the field of mobile sensing and ubiquitous computing: CHI (Conference on Human Factors in Computing Systems) and IMWUT (Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies). Specifically, we utilised the following method:

1. Access 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 have collected a total of 121 papers, comprising 42 papers from IMWUT, 22 papers from CHI, and 57 papers from Ubicomp. Subsequently, we meticulously reviewed each full paper, ensuring that only those exclusively utilising mobile sensing were retained. Additionally, we eliminated any papers that focused on topics unrelated to human behavioural understanding, such as speech enhancement [48] and sensor calibration [13]. As a result of this rigorous selection process, we were left with a final set of 55 papers.

3.2 Overview of Data Sources

After evaluating the set of papers mentioned above, we have identified the commonly used sensors for mobile sensing studies. We excluded sensing data collected in less than or equal to 2 papers, as this may be due to difficulties in collecting that type of sensing data, such as privacy concerns. Additionally, the set of sensors may vary for specific devices, as discussed in Section 7. Therefore, we focused on the most commonly used sensors in previous mobile sensing studies.

Furthermore, we observed that the names of data sources are inconsistent across papers. For the same type of sensing data, different papers may use different names. Additionally, there are similar but slightly different names for certain sensing data. For example, the original names for 'Screen' could be 'screen', 'screen status', or 'screen interaction data'. Similarly, for 'Application', the original names could be 'application', 'app', or 'APP'. Moreover, some papers used 'GPS' for location, while others did not specify the measurement method. To address these inconsistencies and to prioritise the frequently used data, we have integrated the names of data sources and identified the following commonly utilised data sources in smartphones:

- **Hardware Sensors**. Hardware sensors are integral components in mobile devices that monitor physical activities. These devices measure the tangible repercussions of processes that enable the transmission and processing of information in cyberspace. Commonly used hardware sensors are: *[Accelerometer, Gyroscope, Light, Magnetometer, Gravity, Temperature, Humidity, Orientation, Barometer, Proximity, Microphone, Bluetooth, WiFi]*.
- **Software Senors**. Software sensors emerge when hardware sensors are amalgamated with specific software models. This integration deduces new variables from measures, which might otherwise be tough or unfeasible to directly capture. Commonly used software sensors are: *[Application, Calls, Message, GPS/Location, Notification, Keyboard]*.
- **Contextual Information**. Contextual information refers to data that provides insight into the surrounding circumstances or environment in which a device operates or a user interacts. This information is paramount

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

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
	Relationship	Correlation Ranking	Pearson, Spearman, Kendall tau correlation The most frequent place visited
	Diversity	Diversity of the 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)

for understanding user behaviour, preferences, and the state or conditions of device usage. It often serves as the backdrop against which other data, like sensor readings, can be interpreted more accurately and meaningfully. Commonly used contextual information is as below: *[Screen, Time, Date, Battery]*.

3.3 Overview of Features

After a thorough review of the 55 mobile sensing studies, we have identified several key findings related to feature construction. Firstly, it is evident that the features used in various mobile sensing studies lack consistency and coherence. Some papers solely rely on statistical features without considering the underlying meaning of these features. On the other hand, certain papers incorporate meaningful features but fail to establish a clear and organized granularity of human behaviour. Often, researchers tend to select behaviours they believe to be relevant without considering the complete spectrum of human behaviour, resulting in incomplete feature extraction. Furthermore, while some papers mention the time span of features, they often apply the same time span to all features, rendering it meaningless. Additionally, there are instances where the time span of a feature is not mentioned at all. In summary, our analysis reveals that this area lacks a clear guiding principle for designing effective features.

After analyzing current mobile sensing studies, we have found that effective features for human-centred mobile sensing typically consist of three components: the time span of the sensing data, the metrics used for

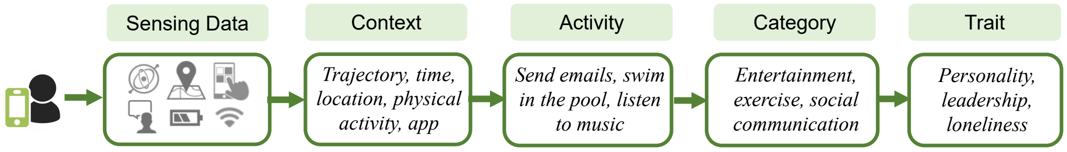

Sensing Data
Context
Trajectory, time, location, physical activity, app
Activity
Send emails, swim in the pool, listen to music
Category
Entertainment, exercise, social communication
Trait
Personality, leadership, loneliness

Fig. 2. Multi-granular human behaviour representation

measurements, and the specific human behaviours being studied. To ensure effectiveness and relevance, a well-designed feature should incorporate all three components. For instance, consider the feature *"Duration of screen time per weeknight"*. In this case, the metric is the *"Duration of time"*, the atomic behaviour is the *"screen"* and the time span is the *"weeknight"*. Table 1 provides a summary of the commonly used time spans and metrics.

3.4 Model and Performance

Based on the analysis results from 55 mobile sensing studies, we have found that the majority of research utilizes similar machine learning models, coming from the list: *[Random Forest, Gradient Boosting Machine, Linear Regression, Gaussian Mixture Model, Support Vector Machine, Naive Bayes, K-nearest Neighbour, and Logistic Regression]*. It is worth noting that some models can function as both regressors and classifiers, such as *Random Forest Regressor* and *Random Forest Classifier*. The primary purpose of using these machine learning models is to evaluate the effectiveness of features in predicting the research objective, rather than determining which model performs the best. However, it is still useful to provide users with some suggestions. For example, *Random Forest* and *Gradient Boosting* methods may outperform the others in certain cases due to their ability to handle complex relationships and capture non-linear patterns. They are robust to outliers and effective for high-dimensional data, which are often recommended to be used in scenarios involving complex and non-linear data. Additionally, knowing the approximate level of performance for the research objective can also be helpful. We have discussed the importance of estimating performance in Section 2.2.3.

4 MULTI-GRANULAR HUMAN BEHAVIOUR REPRESENTATION

A deep understanding of human behaviours is key to effective feature construction and successful mobile sensing studies. The translation of sensing signals, smartphone usage and context descriptors into a specific human behaviour (e.g., emotion, alcohol drinking) remains a challenge due to the inherently complex nature of human behaviours [9, 14]. Unfortunately, many mobile sensing studies overlook this crucial aspect. Instead, they merely extract seemingly relevant data, such as statistical features or trajectory data, without considering the broader context. This approach not only wastes time in extracting these features but also lacks a comprehensive explanation of features. Consequently, it would not cover all facets of human behaviour.

Human behaviour refers to the potential and expressed capacity for physical, mental, and social activity, responding to internal and external stimuli throughout human life [21]. It has been explored by various fields, such as psychology, sociology, ethology, and human-centred design. While there are many different facets of human behaviour, no single definition or field of study can fully encapsulate its entirety. For example, human behaviour could be decomposed from various perspectives, such as the temporal phase of development (prenatal life, infancy, childhood, adolescence, adulthood, and old age) [21], reactive mode (reactive and deliberative behaviours) [35], dimensions (actions, cognition and emotion), etc.

To understand how human behaviours could be captured through smartphones, we propose a multi-granular human behaviour representation mechanism. This mechanism serves as a foundation for constructing meaningful

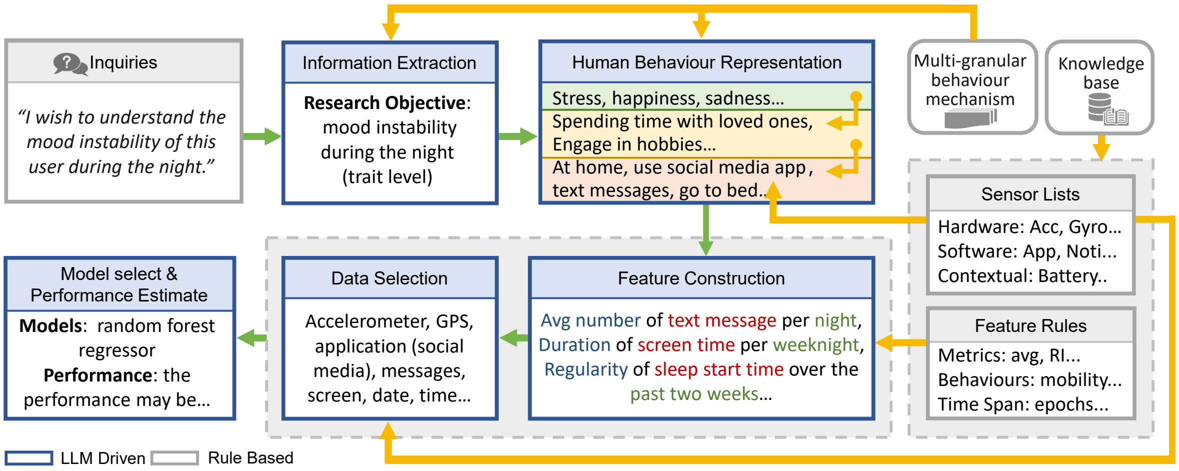

Inquiries
"I wish to understand the mood instability of this user during the night."
Information Extraction
Research Objective: mood instability during the night (trait level)
Human Behaviour Representation
Stress, happiness, sadness...
Spending time with loved ones, Engage in hobbies...
At home, use social media app, text messages, go to bed...
Multi-granular behaviour mechanism
Knowledge base
Sensor Lists
Hardware: Acc, Gyro...
Software: App, Noti...
Contextual: Battery..
Feature Rules
Metrics: avg, RI...
Behaviours: mobility...
Time Span: epochs...
Model select & Performance Estimate
Models: random forest regressor
Performance: the performance may be...
Data Selection
Accelerometer, GPS, application (social media), messages, screen, date, time...
Feature Construction
Avg number of text message per night, Duration of screen time per weeknight, Regularity of sleep start time over the past two weeks...
LLM Driven
Rule Based

Fig. 3. The generation process of mobile sensing strategies. Green arrows indicate the data flow of the user's inquiry in natural language. Yellow arrows are the data flow of designed rules. They eventually merge into the generated mobile sensing strategies.

features in mobile sensing research. Specifically, it comprises four dimensions that encompass human behaviours at varying levels of granularity: contexts, activities, categories and traits (see Figure 2).

- *Context:* At the context level, we consider the information that could be directly inferred or easily calculated from smartphone sensors. These atomic activities include location/trajectory (GPS), physical activity (Android Activity Recognition API [2]), time, screen usage, and more.
- *Activity:* At the activity level, we zoom in to identify specific activities or behaviours exhibited by individuals. These activities may include sending emails, swimming in the pool, listening to music, and so on.
- *Category:* Moving to the category level, we group similar behaviours together based on shared characteristics or attributes of activities. This higher-level abstraction allows us to identify commonalities and patterns across activities, enabling a more holistic understanding of human behaviours. Categories may include entertainment, exercise, communication, social, etc.
- *Trait:* At the trait level, we consider enduring characteristics or traits that are intrinsic to individuals and reflect their behaviour patterns. These traits may include personality traits, social abilities, leadership, loneliness, and more.

For example, if we wish to investigate someone's mood instability, which exists at the trait level, we can identify relevant categories associated with mood, such as stress, happiness, and sadness. Within these categories, various activities may contribute to mood fluctuations, such as work-related tasks, spending time with loved ones, or engaging in hobbies that impact emotional well-being. Moving to the context level, we can examine specific atomic activities like using the smartphone, opening social media apps, texting messages, 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 and enable a deeper understanding of complex phenomena like mood instability.

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

5 AUTOMATED MOBILE SENSING FRAMEWORK

In this section, we propose a feature construction strategy and discuss the identification of the best sensor sources based on the selected features. We also suggest an automated ML model and provide an estimation of its performance. To achieve this, we introduce several prompting techniques to adapt LLMs for generating automated mobile sensing strategies. LLMs support in-context few-shot learning through *prompting*, eliminating the need for fine-tuning or re-training the model for each new task. By prompting an LLM with a few input and output data exemplars from the target tasks, we could leverage its capabilities effectively. Moreover, we explore *Chain-of-Thought* [46] techniques, which is a widely used method for prompting LLMs to elicit logical reasoning. This approach has shown effectiveness in generating intermediate results before producing the final output.

5.1 Design Rules for Mobile Sensing Strategies

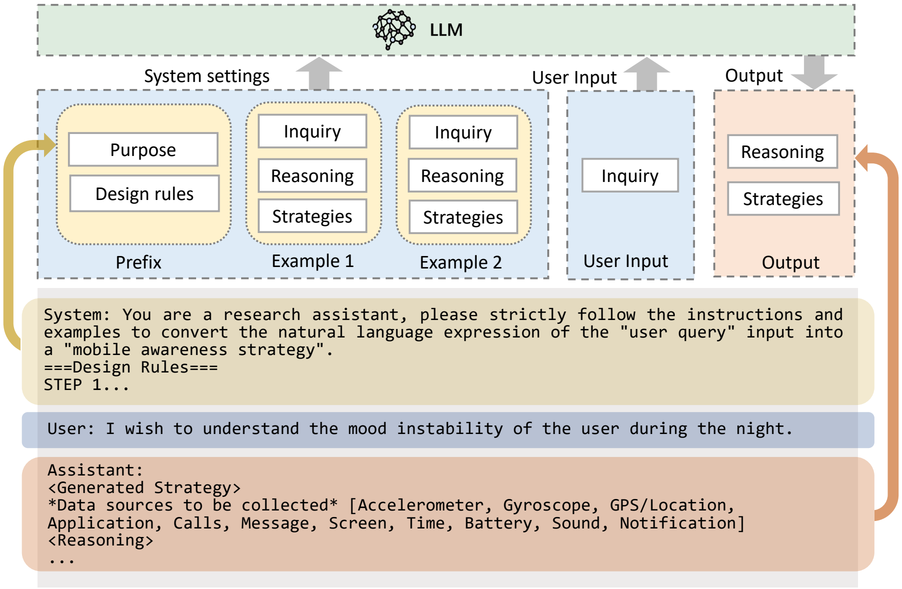
To achieve effective mobile sensing strategies, there are five main steps (see Figure 5): *Information Extraction*, *Human Behaviour Representation*, *Feature Construction*, *Data Selection* and *Model Selection & Estimated Performance*. After completing these steps, the system outputs the mobile sensing strategy based on the user's input inquiry.

5.1.1 Information Extraction (Step 1). The user initiates an inquiry to the system, such as *"I wish to understand the mood instability of this user during the night."* The system then extracts the research objective to be modelled in the mobile sensing research. In this case, the research objective is *"mood instability during the night"*. Next, the system needs to define the level of human behaviour from the options provided (i.e., *trait*, *category*, *activity*, *context levels*), based on the multi-granular human behaviour mechanism described in Section 4. Since *"mood instability during the night"* is a characteristic or trait that is intrinsic to individuals and affects their behaviour patterns, it is considered at the *trait level.*

5.1.2 Human Behavior Representation (Step 2). In this step, the system extracts the multi-granular human behaviours based on the research objective and the defined level from Step 1 (*trait level* in our case). The behaviours are extracted in a hierarchical manner, starting from the highest level (*category level*), then moving to the *activity level*, and finally to the *context level.* The system continues this extraction process until it reaches the lowest level. The 'context' level behaviours can be easily computed or directly inferred from smartphone sensing data. It is important to note that the system considers the sensor lists extracted from the knowledge base (see Section 3.2) to generate meaningful 'context' behaviours. Therefore, these behaviours should be both related to the research objective and able to be inferred from smartphone data.

5.1.3 Feature Construction (Step 3). This step focuses on constructing comprehensive features for modelling the research objective. As discussed in Section 3, effective features for human-centred mobile sensing typically consist of three components: the time span of the sensing data, the metrics of the measurements, and the specific human behaviours. The choice of time span and metrics depends on the specific scenarios and human behaviours involved. To achieve effective feature construction, the system utilises the *context level* behaviours extracted in Step 2 and selects appropriate metrics and time spans based on domain knowledge from LLMs. For example, in the case of modelling the research objective *"mood instability during the night"*, one of the constructed features may be the *"regularity of sleep start time over the past two weeks"*. By analysing the sleep start time data over a two-week period, the system could identify patterns or irregularities in the user's sleep schedule, which may indicate mood instability during the night.

5.1.4 Data Selection (Step 4). In this step, the system determines the data that needs to be collected to compute the features extracted in Step 3. The sensing data source is selected from the identified sensing data sources in Section 3.2, which includes the hardware sensor data, software sensor data, and contextual information. For example, in the case of the feature *"regularity of sleep start time over the past two weeks"*, both time data and



```
System: You are a research assistant, please strictly follow the instructions and
examples to convert the natural language expression of the "user query" input into
a "mobile awareness strategy".
===Design Rules===
STEP 1...
```

```
User: I wish to understand the mood instability of the user during the night.
```

```
Assistant:
<Generated Strategy>
*Data sources to be collected* [Accelerometer, Gyroscope, GPS/Location,
Application, Calls, Message, Screen, Time, Battery, Sound, Notification]
<Reasoning>
...
```

Fig. 4. An example illustrating the proposed prompt structure

sleep data need to be computed. The sleep activities can be identified using sensors such as the accelerometer and gyroscope. This step ensures that the necessary data sources are included to calculate the desired features accurately.

5.1.5 Model and its Estimated Performance (Step 5). In this step, the system suggests a machine learning model that can achieve the best prediction performance based on three types of information: (1) the research objective extracted in Step 1, (2) the selected sensing data source extracted in Step 4, and (3) the constructed features from Step 3. The choice of the model depends on these factors because the research objective, features, and data sources all impact prediction performance. Especially, good features play a crucial role in achieving superior performance, while limited data sources may hinder the smartphone's ability to accurately understand human behaviour. Once the suggested ML models are determined, the system estimates the prediction performance using natural language and provides some reasons for the estimation. However, it is important to note that the estimated performance serves as a basic understanding. For instance, modelling psychological traits often results in lower prediction performance compared to physical activities. This discrepancy is due to the inherent complexity and variability associated with psychological traits.

5.2 Prompt Structure

Our prompt structure is inspired by [41], and consists of three components as below: (1) **Prefix**. The prompt begins with a clear and concise introduction that outlines the purpose of the prompt and the design rules to be followed. These design rules are discussed in detail in Section 5.1. The prefix provides a high-level overview of the prompt and sets the context for the examples that follow. (2) **Examples**. The prompt includes a series of N (where $N = 2$) examples. Each example is designed to address a specific inquiry related to the topic. The examples are divided into three parts: a) *Inquiry*. Each example starts with an inquiry that is expressed in natural language to enhance understanding. For example, an inquiry could be *"I wish to understand the mood instability of this user during the night"*. The inquiries should be clear and specific to facilitate effective reasoning and mobile sensing strategy formulation. b) *Reasoning*. To elicit logical reasoning from LLMs, we provide a clear and coherent chain of thought for each example. This chain of thought follows the design rules mentioned in the prefix and explains the reasoning behind each design decision. The reasoning process is well-structured, providing step-by-step explanations and justifications. c) *Mobile Sensing Strategy*: For each example, we outline the chosen mobile sensing strategies. This includes specifying the data to be collected, the features to be constructed, the models to be built and an estimation of the performance. (3) **User Input**. The user is expected to provide their own inquiry related to their objective. The user inquiry is expressed in natural language.

Overall, our prompt consists of a prefix, three examples, and the user inquiry. An example of our prompt is illustrated in Figure 4. By structuring the prompt in this way, we aim to provide a comprehensive framework that guides the LLMs in reasoning and formulating mobile sensing strategies based on the given inquiries. It is worth mentioning that, the number of examples can be adjusted according to the users' requirements.

6 EVALUATION

In the experiment, we chose gpt-3.5-turbo as our primary LLM for its robust computational capabilities and its proven efficacy in understanding complex prompts to generate coherent, context-rich responses. This fulfils our experimental requirements. Moreover, to optimise the interactions, we crafted specific prompts for contextual rule creation. By equipping the model with detailed prompts, users can guide it towards generating the intended response. To initiate a conversation with the model, a prefixed indicator "INPUT" can be used. For example, if a researcher wish to model the mood instability of the smartphone users, they could directly type: *INPUT: I wish to understand the mood instability of the user during the night.*

6.1 Expert Evaluators

Distinguishing from other user studies that typically rely on easily recruited ordinary participants, our research is specifically designed to offer valuable insights into human behaviours for researchers in the field of mobile sensing. To achieve this, we have employed a combination of word-of-mouth recommendations and email invitations to enlist the participation of experts. Specifically, we have invited 8 experts who possess significant experience in modelling human behaviours using mobile sensing technologies. On average, these experts boast 4.25 years of research experience in the field. While we acknowledge the limitation in terms of the number of expert evaluators, it is crucial to emphasize that these experts have a profound understanding of mobile sensing techniques and can provide substantial insights into the system.

6.2 Procedure

We conducted two evaluation studies: a comparative study and a usability study. The purpose of the comparative study was to evaluate the effectiveness of the automated mobile sensing strategy in comparison to the existing strategy. To ensure a fair comparison, we adopted the *Blind Comparison* method [12]. For the usability study, we asked experts to type any inquiry they wanted and then complete a survey and participate in an interview. The procedure for our studies will be described below.

6.2.1 Comparative Study. In the comparative study, we selected two mobile sensing tasks that have been published in top venues in the field and have more than 100 citations in total. From these studies, we extracted the research objectives, selected data sources, and constructed features from the existing description. For each of the selected studies, we applied the sensing strategy generation system and entered the research objective as the user input. However, we did not extract the selected model and performance. This was done to prevent any potential bias that could arise from revealing whether the strategy was from the existing strategy or auto-generated, ensuring a fair comparison. Furthermore, we maintained a consistent format and style when describing the data and features. We also excluded other details about the features, such as the sensors used and the category they belong to. Overall, the only difference between the existing strategy published and the automatically generated strategies lies in the type of sensing data used and the features themselves.

Each expert participating in the study was presented with the existing strategy and the auto-generated strategy for both studies. The order in which the two studies were presented was randomised to account for any potential *Order Effects* [12]. During the comparison process, the experts were instructed to imagine an assistant generating two strategies based on the given research objective. They were then asked to compare the two strategies and determine which one they would prefer to employ to achieve their research objective. Specifically, they were asked to assess the effectiveness, interpretability, relevance of data/features to the research objective, and completeness. Each scored with a 5 Likert scale where 1 indicates very negative and 5 indicates positive. This assessment was conducted through semi-structured interviews. These procedures were repeated for each expert until they had completed both tasks.

6.2.2 Usability Study. In this usability study, experts were instructed to use the proposed automatic sensing strategies system on their own. They were encouraged to type user inquiries into the system, focusing on any human behaviour they wanted to understand through smartphones. Upon entering the user inquiry, the system would generate strategies and provide a reasoning process. The generated strategies included the data sources to be collected, features to be constructed, and models to be built, along with estimated performance. The system also displayed the reasoning process step by step.

After using the system, we employed an adapted version of the NASA-TLX [18] evaluation method to assess the generated strategies for each expert. We excluded questions such as temporal demand or effort in our evaluation because they were not relevant to our system. Our system only incurred waiting time for system response, rather than requiring any significant temporal demand or effort from the participants. Participants rated the following aspects on a 5-point Likert scale, with 1 being the most negative experience and 5 the most positive:

- Mental demand: *How mentally demanding was the task?*
- Physical demand: *How physically demanding was the task?*
- Performance: *How successful were you in accomplishing what you plan to do?*

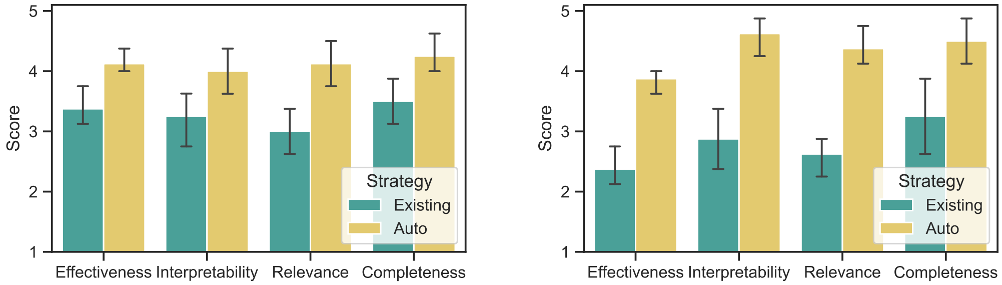
We also ask participants to evaluate their overall experience and answered several additional questions on a 5-point Likert scale, where 1 indicates 'not at all' and 5 indicates 'very much':

- Satisfaction: *How satisfied are you with the automated generated strategy?*
- Enhanced Understanding: *Does the automated strategy enhance your understanding of the research objective?*
- Ease of use: *How easy was the system to use?*
- Willingness to reuse: *How likely are you to use this assistant again in the future?*

To gain a deeper understanding of the experts' thoughts about the automated mobile sensing strategies, a concluding interview was conducted. This interview aimed to gather insights into their perception of the system's effectiveness, its impact on their research process, and any suggestions or improvements they may have.

6.3 Result and Discussion

6.3.1 Comparative Performance Analysis. To compare the automated and existing strategies in two selected mobile sensing studies, Figure 5 presents the evaluation results for two selected mobile sensing studies, referred to as Study A and Study B, where A aims to predict *Brain Functional Connectivity* and study B aims to understand *Compound Emotion.* We calculated the average scores for *Effectiveness*, *Interpretability*, *Relevance*, and *Completeness* based on expert opinions. The scores for the automated strategy ('Auto') proposed in the research were compared to the scores for the existing strategy ('Existing') in the original study. The results for *Effectiveness* were as follows:



(a) Task: Study A

(b) Task: Study B

Fig. 5. The evaluation results for both studies from experts

'Auto' had an average score of 4.0 with a standard deviation (STD) of 0.37, while 'Existing' had an average score of 2.88 with an STD of 0.72. In terms of *Interpretability*, 'Auto' scored 4.31 (STD = 0.60) compared to 'Existing' with a score of 3.06 (STD = 0.77). For *Relevance*, 'Auto' scored 4.25 (STD = 0.58) while 'Existing' scored 2.81 (STD = 0.54). Lastly, for *Completeness*, 'Auto' scored 4.375 (STD = 0.5) and 'Existing' scored 3.375 (STD = 0.81). Overall, our proposed automated strategies outperformed the existing strategies in all dimensions.

Although our methods performed better than the baseline existing studies in all dimensions, the performance varied between the two studies. This discrepancy can be attributed to the different research objectives of each study. For instance, Study B extracted many basic statistical features, e.g., *"Longitude, Altitude, Latitude of GPS"*, which are low-level features and do not consider many semantic relations to the research objective. As a result, experts found these features to be less relevant to the research objective, while our proposed features, such as *"Distance travelled per day/weeknight"*, were deemed more meaningful. Of the eight experts consulted, two raised concerns about the feasibility of computing the features within our proposed strategy. The existing strategy's low-level statistical features are simpler to compute than our more intricate ones. Nonetheless, the features in the automated strategy (e.g., locations, and physical activities.) can still be easily computed through many mature algorithms.

Additionally, three experts mentioned that they got inspiration from our proposed features, as they covered aspects that they had not previously considered but believed to be meaningful. For example, Expert 2 stated, *"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. To assess the usability of the proposed automated system, experts were asked to test the system freely and provide their evaluation 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 were directly asked to rate the system on these dimensions. The usability ratings provided by the experts are depicted in Figure 6. Overall, the average values for all dimensions were above 3, indicating that the proposed strategy performs well. Notably, the mental demand and physical demands were found to be very low, and experts expressed a strong willingness to utilise the system for their mobile sensing research in the future.

Especially, it is worth noting that experts may propose varying research objectives. While our strategy aims to enable understanding of any human behaviour using smartphones, certain behaviours may be easier to infer (e.g., smartphone addiction behaviour) compared to others that are more challenging (e.g., heart attack). Despite this, 5 of 8 experts expressed that the generated strategies are meaningful, and even if they may not directly adopt

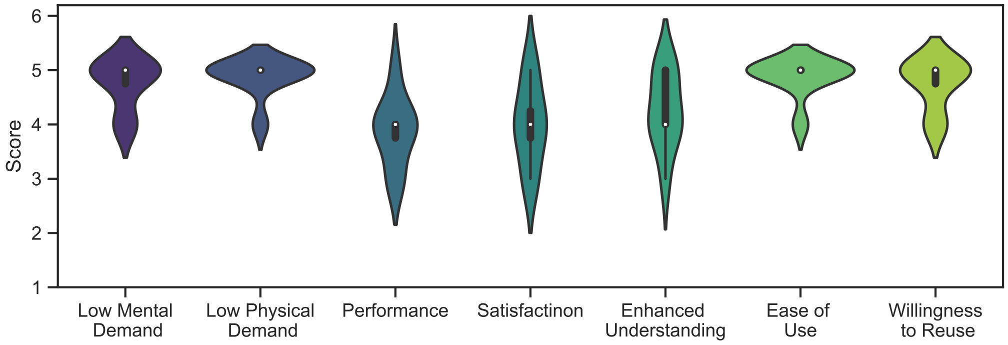


Fig. 6. Ratings for the automated mobile sensing strategy from experts

the strategy, they have a strong desire to have the system assist them in designing their own systems. Three experts indicated that our proposed features inspired them, covering elements they hadn't previously taken into account but found 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 than satisfactory. Upon investigation, we discovered that this dissatisfaction arose from their research objectives focused on asking for suggestions to change or improve human behaviour rather than on understanding, modelling or predicting human behaviours. Overall, users of the system should make sure that the research objective is to use mobile sensing to understand human behaviours. In a different scenario, the automated generated strategy suggests the feature *"The number of positive/negative messages sent per day"*, which raises two primary concerns: first, the potential violation of users' privacy rights due to such data collection; and second, the ambiguity in distinguishing what constitutes a positive message from a negative one. While some generated sensors/features may indeed offer value, their real-world applicability could be constrained. Further discussion could be found in Section 7.

7 IMPLICATIONS AND LIMITATIONS

The research presented proposes an automatic generation of mobile sensing strategies for understanding human behaviour, with potential implications for the development of a human behaviour computing system that can be tailored to any research objective. This system offers benefits to both researchers and individuals. For researchers, it not only reduces the burden of designing mobile sensing strategies but also provides effective feature suggestions and aids in decision-making based on estimated performance. Furthermore, the system can quickly adapt to different research objectives, offering suggestions and experimental designs. For individuals, the system may enhance self-awareness by providing a more objective method of understanding themselves through passive sensing data. Thus, in turn, may improve their well-being and overall quality of life.

However, the study has a few limitations. Firstly, not all devices are equipped with every type of sensor. While the most commonly collected data from smartphones have been reviewed in Section 3.2, the availability of sensors varies across devices. For example, some devices may lack barometers or thermometers, and the IOS system typically has more constraints compared to Android devices, with data more readily accessible from the latter.

Secondly, certain procedures in mobile sensing, such as parameter tuning and data cleaning, were not considered in this study. The focus was primarily on critical procedures for generating automatic mobile sensing strategies, such as data source selection, feature construction, model building, and performance estimation. Other tasks, like parameter tuning, can be facilitated using existing tools such as AutoML.

Thirdly, the research primarily focuses on generating automatic mobile sensing strategies without computing or implementing the features. The main contribution lies in the design of the mobile sensing strategy, which holds potential for further research in understanding human behaviour. By saving researchers time and reducing the burden of designing data selection strategies and constructing effective features, our research can pave the way for future studies.

Lastly, privacy is a notable concern when collecting data from individuals. However, our research primarily focuses on generating mobile sensing strategies and does not involve data collection. Future researchers can adapt our approach while implementing necessary privacy protection measures. It is important to note that all data processed for automated human behaviour computation would be strictly protected and processed only on the user's own device, minimising privacy concerns.

8 CONCLUSION

This paper introduces an automated mobile sensing strategy generation system that allows users to input inquiries related to understanding any human behaviours through smartphones. To achieve this, we have designed a multi-granular human behaviour mechanism that decomposes research objectives into different levels of human behaviours. This enables the extraction of atomised behaviours that can be easily inferred or extracted from smartphone sensing data. By leveraging a knowledge base and utilising the large language models, our system could automatically generate the sensing strategy, including the sensing data to be collected, features to be constructed, and suggested models along with their estimated performance. This automation significantly reduces the burden on researchers who would otherwise have to manually craft features, while also providing new insights for effective mobile sensing strategy design. In future work, we plan to explore the automatic computation of features from smartphones to support the development of intelligent systems that can understand human behaviour. Ultimately, this research has the potential to assist individuals in gaining objective insights into themselves, thereby improving self-awareness and enhancing overall well-being.

9 ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their valuable service to the CHI community.

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and colour spaces.

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] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* 408 (2020), 189–215.

[4] 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.

[5] 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.

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

[7] 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).

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

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

[10] 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.

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

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

[13] Andreas Grammenos, Cecilia Mascolo, and Jon Crowcroft. 2018. You are sensing, but are you biased? a user unaided sensor calibration approach for mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[14] 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.

[15] Jeff Heaton. 2016. An empirical analysis of feature engineering for predictive modeling. In *SoutheastCon 2016*. IEEE, 1–6.

[16] 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).

[17] 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.

[18] 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.

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

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

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

[22] 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.

[23] 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.

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

[25] 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.

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

[27] Iris B Mauss and Michael D Robinson. 2009. Measures of emotion: A review. *Cognition and emotion* 23, 2 (2009), 209–237.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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

[37] Yunsheng Song, Jiye Liang, Jing Lu, and Xingwang Zhao. 2017. An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing* 251 (2017), 26–34.

[38] 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).

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[47] 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.

[48] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–30.

[49] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. 2018. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–30.

[50] Lucas Zimmer, Marius Lindauer, and Frank Hutter. 2021. Auto-pytorch: Multi-fidelity metalearning for efficient and robust autodl. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 9 (2021), 3079–3090.

Received 10 September 2023; revised 12 November 2023; accepted 15 January 2024