Are You Killing Time? Predicting Smartphone Users' Time-killing Moments via Fusion of Smartphone Sensor Data and Screenshots

Yu-Chun Chen; Yu-Jen Lee; Kuei-Chun Kao; Jie Tsai; En-Chi Liang; Wei-Chen Chiu; Faye Shih; and Yung-Ju Chang

CHI '23: 2023 CHI Conference on Human Factors in Computing Systems

> **Smartphones and Notifications**
> https://programs.sigchi.org/chi/2023/program/content/96013
> https://www.youtube.com/watch?v=OvrSPuGkhc0

Abstract

Time-killing on smartphones has become a pervasive activity, and could be opportune for delivering content to their users. This research is believed to be the first attempt at time-killing detection, which leverages the fusion of phone-sensor and screenshot data. We collected nearly one million user-annotated screenshots from 36 Android users. Using this dataset, we built a deep-learning fusion model, which achieved a precision of 0.83 and an AUROC of 0.72. We further employed a two-stage clustering approach to separate users into four groups according to the patterns of their phone-usage behaviors, and then built a fusion model for each group. The performance of the four models, though diverse, yielded better average precision of 0.87 and AUROC of 0.76, and was superior to that of the general/unified model shared among all users. We investigated and discussed the features of the four time-killing behavior clusters that explain why the models' performance differ.

Introduction

- Given human beings' limited attentional resources, **a crucial problem for anyone delivering content to phones is how to make it stand out from the feast of other incoming information**.
    - One mainstream approach to achieving this is to predict moments at which users are receptive to such content, e.g., the content related to notifications, questionnaires, and reading material explored in prior studies.
    - Moments of **"attention surplus"** constitute another opportunity for such detection attempts.
        - Pielot et al., for example, **attempted to detect one kind of "attention surplus" state – boredom**. In such situations, some people tend to seek stimulation on their phones to alleviate boredom. Beyond boredom, however, research has shown that mobile phone usage is not always driven by a specific purpose, but is often accompanied by, or is primarily, "time-killing" behavior
- **Time-killing** is not necessarily linked to boredom, Rather, this type of usage can **simply be a result of habit**.
    - However, when the trigger of habitual phone usage is to fill free time, even if it may occasionally include productive tasks such as checking emails and messages, it is typically viewed as aimless, and if it occurs frequently without regulation, as addictive and compulsive.
    - There has been limited effort in detecting the occurrence of such prevalent phone usage intended for time-killing, making the feasibility of detecting this phone usage unclear.

- We **aim to develop a model for effectively detecting phone usage intended for time-killing**.
    - To achieve this aim, we developed an **Android research application** that automatically collected smartphone screenshot and phone-sensor data, and an interface that allowed its **users to efficiently annotate their screenshot data** with time-killing periods and their availability during those periods for viewing notifications.
    - Data collection with 36 participants over 14 days yielded a dataset of 967,466 pairings of annotated phone-sensor data with screenshots, covering 1,343.7 hours of phone usage.

Related Work

- Interruptibility, Breakpoint, and Opportune Moment Prediction
    - Employ **machine-learning techniques** to predict mobile interruptibility or opportune moments for receiving calls and notifications.
    - **Identify opportunities** based on breakpoints
    - Detecting moments when device users want to engage with content
    - **Attention prediction** involves identifying "attention surplus" moments and timing the delivery of specific content and tasks accordingly
- Phone-usage Research
    - Utilize **self-report methods** such as interviews and diaries
        - However, because self-report methods are subject to recall biases, quantitative analysis of phone-usage logs is becoming increasingly popular.
    - Attempts to **construct models of phone usage**
    - Understanding differences in usage features across **distinct user clusters**.
    - Use **screenshots and video recordings**
        - because log data are limited to system events like screen events and app states
    - Use deep-learning models **trained on large amounts of Graphical User Interface (GUI) data** to detect screenshots.

Implementation

I. Data Collection

1. Input-data Selection

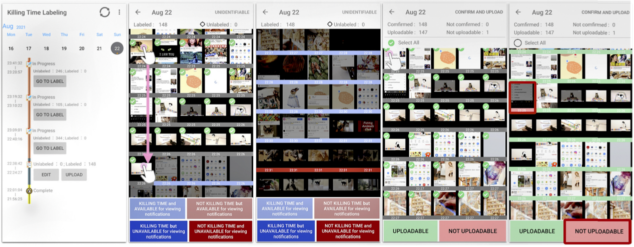
- **Screenshot collection** has become a popular method in HCI research, because it allows researchers to collect **quantitative and qualitative data simultaneously** in high granularity and rich detail.
  ➔ We aimed to leverage **screenshot data, along with phone-sensor interaction information** (including user/phone interaction and phone status)
    - i.e., Android accessibility events, screen status, network connections, phone volume, application usage, and type of transportation

2. Research Instrument

- We developed an Android research application, called **Killing Time Labeling (KTL)**
- During whatever 12+-hour window the user has chosen(default: 10 a.m.~10 p.m.),
    - Phone-sensor data is **collected every five seconds**.
    - Screenshots are also **captured every five seconds**, but only when the phone screen is on.
- KTL also **captures the notifications** its users receive, the times at which they receive them, and how they are dealt with.
- KTL also **sent a reminder at night** and invalidated screenshots that had not received any

annotations after two days, which participants could no longer annotate.

- For each screenshot, participants had five annotation options:
  1. killing time and available for viewing notifications
  2. not killing time but available for viewing notifications
  3. killing time but unavailable for viewing notifications
  4. not killing time and unavailable for viewing notifications
  5. unidentifiable



(a) Main screen (b) Labeling screen (c) Screen after labeling (d) Upload selection (e) Screen after selection

Figure 1: User interfaces for the main functions of the Killing Time Labeling application

3. Study Procedure

- Prior to data collection, due to the COVID-19 pandemic, we allowed our participants to **choose between remotely and physically attending**
  - We told them that we expected them to annotate all screenshots automatically captured by KTL every day, and that 14 days of active participation were needed for their data to be useful to us.
- On their respective final days of participation, to aid future analysis, they completed **four additional questionnaires**
  - measured their boredom proneness, smart-phone addiction, inattention, and perceived acceptability of time-killing detection being deployed on their phone.
- In addition, we invited all participants to **two optional semi-structured interviews**. (after 7 days & after participation completes)
  - We asked them about their labeling processes, time-killing behaviors and preferences, and how they killed time (both typically and during the study).

4. Recruitment and Participation

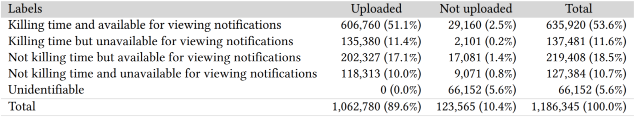
- We **selected participants with various occupations**, expecting they would have different time-killing patterns.
- To ensure that sufficient data were collected, we selected participants who used their mobile phones more than one hour a day, according to their self-reporting in a screening questionnaire.
- As a result, data from 36 people were used for training our time-killing detection model
  - Aged between 20 and 54 (M = 27.4, SD = 6.8),
  - 16 identifying as male and 20 as female.
  - Half were students, and the other half in employment.

5. Data Collection

- In total, 1,186,345 screenshots were annotated (per-participant M = 32,954.0, SD = 15,557.9)
  - represents approximately 1,633.8 hours of phone use.

- 1,062,780 (89.6%) screenshots were uploaded
  - ➔ There were **more time-killing moments** than non-time-killing ones
  - ➔ The participants reported being **available to view notifications more frequently during the time-killing moments (82.2%)** than during non-time-killing ones (63.3%).

Table 1: Summary of data collection

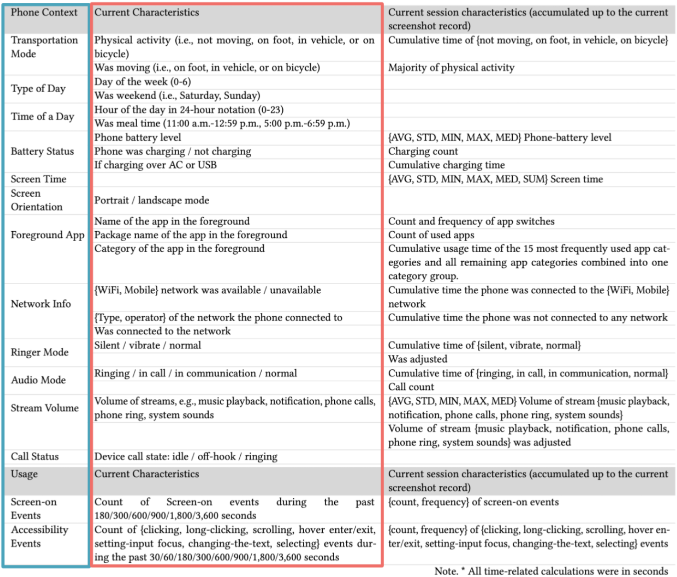


Labels	Uploaded	Not uploaded	Total
Killing time and available for viewing notifications	606,760 (51.1%)	29,160 (2.5%)	635,920 (53.6%)
Killing time but unavailable for viewing notifications	135,380 (11.4%)	2,101 (0.2%)	137,481 (11.6%)
Not killing time but available for viewing notifications	202,327 (17.1%)	17,081 (1.4%)	219,408 (18.5%)
Not killing time and unavailable for viewing notifications	118,313 (10.0%)	9,071 (0.8%)	127,384 (10.7%)
Unidentifiable	0 (0.0%)	66,152 (5.6%)	66,152 (5.6%)
Total	1,062,780 (89.6%)	123,565 (10.4%)	1,186,345 (100.0%)

6. Feature Selection and Extraction

- To predict time-killing moments, we extracted **two kinds of feature sets** from the phone-sensor data: **phone context and user interactions**.
- For each of these feature sets, we created **two temporal ranges**,
  1. **Describing the phone at the moment** when a screenshot was taken
  2. **Describing the characteristics of the phone-use session**[Note] during which it was taken.
- As a result, the final dataset for developing the model consisted of **967,466 annotated screenshots**, from which 183 features were derived.

Table 2: The sensor features used in the study



Phone Context	Current Characteristics	Current session characteristics (accumulated up to the current screenshot record)
Transportation Mode	Physical activity (i.e., not moving, on foot, in vehicle, or on bicycle)	Cumulative time of {not moving, on foot, in vehicle, on bicycle}
	Was moving (i.e., on foot, in vehicle, or on bicycle)	Majority of physical activity
Type of Day	Day of the week (0-6)	
	Was weekend (i.e., Saturday, Sunday)	
Time of a Day	Hour of the day in 24-hour notation (0-23)	
	Was meal time (11:00 a.m.-12:59 p.m., 5:00 p.m.-6:59 p.m.)	
Battery Status	Phone battery level	{AVG, STD, MIN, MAX, MED} Phone-battery level
	Phone was charging / not charging	Charging count
	If charging over AC or USB	Cumulative charging time
Screen Time		{AVG, STD, MIN, MAX, MED, SUM} Screen time
Screen Orientation	Portrait / landscape mode	
Foreground App	Name of the app in the foreground	Count and frequency of app switches
	Package name of the app in the foreground	Count of used apps
	Category of the app in the foreground	Cumulative usage time of the 15 most frequently used app categories and all remaining app categories combined into one category group.
Network Info	{WiFi, Mobile} network was available / unavailable	Cumulative time the phone was connected to the {WiFi, Mobile} network
	{Type, operator} of the network the phone connected to	Cumulative time the phone was not connected to any network
	Was connected to the network	
Ringer Mode	Silent / vibrate / normal	Cumulative time of {silent, vibrate, normal}
		Was adjusted
Audio Mode	Ringing / in call / in communication / normal	Cumulative time of {ringing, in call, in communication, normal}
		Call count
Stream Volume	Volume of streams, e.g., music playback, notification, phone calls, phone ring, system sounds	{AVG, STD, MIN, MAX, MED} Volume of stream {music playback, notification, phone calls, phone ring, system sounds}
		Volume of stream {music playback, notification, phone calls, phone ring, system sounds} was adjusted
Call Status	Device call state: idle / off-hook / ringing	
Usage	Current Characteristics	Current session characteristics (accumulated up to the current screenshot record)
Screen-on Events	Count of Screen-on events during the past 180/300/600/900/1,800/3,600 seconds	{count, frequency} of screen-on events
Accessibility Events	Count of {clicking, long-clicking, scrolling, hover enter/exit, setting-input focus, changing-the-text, selecting} events during the past 30/60/180/300/600/900/1,800/3,600 seconds	{count, frequency} of {clicking, long-clicking, scrolling, hover enter/exit, setting-input focus, changing-the-text, selecting} events

Note. * All time-related calculations were in seconds

> ⓘ **Note**
>
> **Phone-use session**: continuous use of the phone during which any brief screen-of interval was not longer than 45 seconds.

II. Model Design

1. Training details

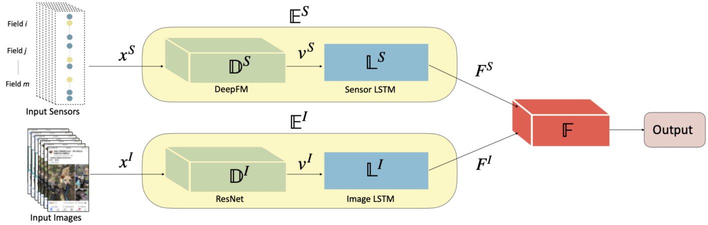


Figure 2: Illustration for the architecture of our proposed model, which takes the input composed of the phone-sensor data and the screenshots (collected within a certain time window, e.g., 30 seconds) and predicts the users' time-killing moments.

- We adopted a stage-wise training procedure, in which **we first trained the encoders,** $\mathbb{E}^S$ **and** $\mathbb{E}^I$**, independently**.
    - After pre-training both encoders till they converged, we removed the attached fully connected layers and fixed the weights of encoders. Then we trained the fusion subnetwork $\mathbb{F}$
- Instead of concatenating two features and utilizing a simple classifer to perform a multimodal fusion, we introduced an additional multi-fusion layer that takes both features as inputs to predict the reweighting coefficients α^S and α^I
    - The reweighted features, denoted as $\tilde{\mathbb{F}}^S = \alpha^S \otimes \mathbb{F}^S$ and then concatenate with the original $\mathbb{F}^S$

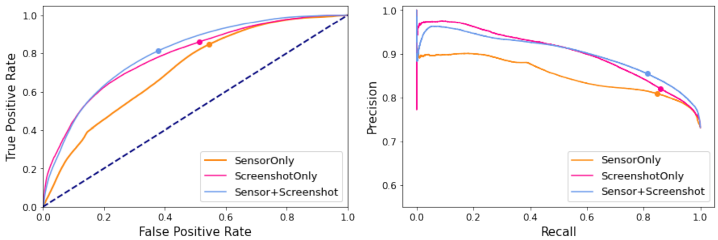
2. Experiment

- Dataset
    1. To predict whether a screenshot was labeled as time-killing or non-time-killing, we used **features derived from the screenshots and their paired sensor data 30 seconds (i.e., seven screenshots) prior to** the predicted one.
    2. To prevent our model from being overly biased towards particular participants, **sampled 20,000 screenshots from each participant**.
    3. ❶ data collected on both weekends and weekdays, and ❷ exactly equal numbers of time-killing and non-time-killing instances.
- Evaluation Metrics
    1. **Testing dataset** had more time-killing instances than non-time-killing ones, **in the ratio 7:3**.
    2. ROC(Receiver Operating Characteristics)-curve & PR(Precision Recall)-curve
- Model Evaluation
    1. **3-fold cross-validation** on the dataset.
    2. Fusion model vs. SensorOnly model vs. ScreenshotOnly model

3. Result

Table 3: The three models' time-killing prediction task performance

Model	Accuracy	Precision	Recall	AUROC	Specificity
Fusion (Sensor+Screenshot)	0.76	0.83	0.81	0.72	0.62
SensorOnly	0.74	0.8	0.85	0.65	0.45
ScreenshotOnly	0.76	0.81	0.86	0.67	0.49



(a) ROC Curves

(b) PR Curves

Figure 3: Two performance measurements of our proposed fusion model (i.e., *Sensor+Screenshot*), its variants (i.e., *SensorOnly* and *ScreenshotOnly*). Note. Point on the curves represents a classification threshold equal to 0.5.

Examples

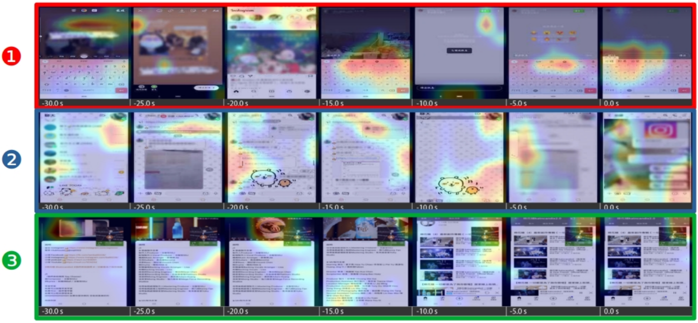


Figure 4: Example attention maps, produced by Grad-CAM [79] and the *ScreenshotOnly* model, comprising a sequence of time-killing screenshots in the top row, and two sequence of non-time-killing ones in the middle and bottom row. The sensitive content on the images have been blurred for privacy reasons.

1. ✓ ScreenshotOnly ✗ SensorOnly ✓ Fusion (time-killing case)
    - SensorOnly: because **a series of text-changed events were detected**, despite in an Instagram application, which was more likely to occur when not killing time.
    - ScreenshotOnly: it recognized the layout of the user interface of Instagram's Story feature.
2. ✗ ScreenshotOnly ✓ SensorOnly ✓ Fusion (time-killing case)
    - ScreenshotOnly: recognized **the layout of LINE**.
    - SensorOnly: by the call status and the change of the call volume.
3. ✗ ScreenshotOnly ✗ SensorOnly ✓ Fusion

    (In this non-time-killing case, the participant was sitting in a moving vehicle and was using a navigation app while simultaneously watching a video on Youtube.)

  - SensorOnly: mistook this situation as time-killing **based on the application and physical-activity information**.
  - ScreenshotOnly: **Youtube's layout** but ignored the navigation map, possibly due to the similar color of the video screen to that of the map.
  - Fusion: correctly identified this moment, possibly because it considered **both the physical-activity information and the use of the navigation functionality**.

Evaluation

---

1. Clustering Participants Based on their Phone-usage Behavior

- Clustering to identify **distinct phone-usage behavioral patterns**.
  - **9 features (inspired by interview)**: 1) session duration, 2) screen-switching frequency, 3) application-switching frequency, 4) scroll-event frequency, 5) text-change event frequency, 6) maximum and 7) minimum gap durations for scroll events, and 8) maximum and 9) minimum gap durations for text-change events.
  - Applied k-means, and used the Elbow method to determine the number of clusters. **(K=5)**
- **Clustered participants** according to how often their phone use belonged to each of the identified phone-usage patterns
  - We calculated the proportions of each user's five outcome behaviors, and used those proportions to cluster users.
  - Elbow methods: the resulting **k-value was 4**
    - the number of participants being 11, 11, 9, and 5 respectively.
    - time-killing:non-time-killing=13:6, 3:1, 81:19, and 3:2

2. Overall Performance of the Cluster-based Models

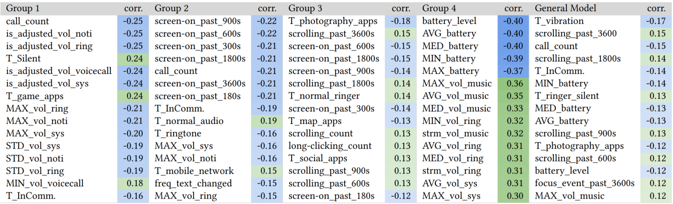
Table 4: Experimental Results: Clustering Participants by Behavioral and Temporal Characteristics

Group	Accuracy			Precision			Recall			AUCROC			Specificity		
Group 1	0.70	0.73	0.73	0.81	0.81	0.88	0.85	0.81	0.80	0.68	0.76	0.77	0.38	0.54	0.59
Group 2	0.75	0.77	0.77	0.85	0.89	0.91	0.84	0.87	0.84	0.70	0.75	0.78	0.46	0.44	0.55
Group 3	0.80	0.77	0.78	0.91	0.95	0.93	0.87	0.82	0.82	0.68	0.75	0.72	0.39	0.48	0.50
Group 4	0.72	0.74	0.77	0.71	0.74	0.77	0.78	0.78	0.79	0.70	0.73	0.77	0.63	0.69	0.74
Average	0.74	0.75	0.76	0.82	0.84	0.87	0.83	0.82	0.81	0.69	0.75	0.76	0.47	0.54	0.60
General model	0.74	0.76	0.76	0.80	0.81	0.83	0.85	0.86	0.81	0.65	0.67	0.72	0.45	0.49	0.62

Note. The white, light gray, and dark gray backgrounds indicate the results for *SensorOnly*, *ScreenshotOnly*, and Fusion (*SensorOnly+ScreenshotOnly*) models, respectively.

- Build the same fusion model for each of the four user groups.
- **It is beneficial to divide users into groups** according to their phone-usage behavior and build a time-killing prediction model for each such user group.

Table 5: The 15 non-category features most highly correlated (either positively or negatively) with time-killing moments, by user group



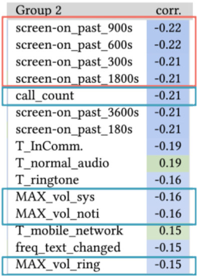
Group 1	corr.	Group 2	corr.	Group 3	corr.	Group 4	corr.	General Model	corr.
call_count	-0.25	screen-on_past_900s	-0.22	T_photography_apps	-0.18	battery_level	-0.40	T_vibration	-0.17
is_adjusted_vol_noti	-0.25	screen-on_past_600s	-0.22	scrolling_past_3600s	0.15	AVG_battery	-0.40	scrolling_past_3600	0.15
is_adjusted_vol_ring	-0.25	screen-on_past_300s	-0.21	screen-on_past_600s	-0.15	MED_battery	-0.40	call_count	-0.15
T_Silent	0.24	screen-on_past_1800s	-0.21	screen-on_past_1800s	-0.15	MIN_battery	-0.39	scrolling_past_1800s	0.14
is_adjusted_vol_voicecall	-0.24	call_count	-0.21	screen-on_past_900s	-0.14	MAX_battery	-0.37	T_InComm.	-0.14
is_adjusted_vol_sys	-0.24	screen-on_past_3600s	-0.21	scrolling_past_1800s	0.14	MAX_vol_music	0.36	MIN_battery	-0.14
T_game_apps	0.24	screen-on_past_180s	-0.21	T_normal_ringer	0.14	AVG_vol_music	0.35	T_ringer_silent	0.13
MAX_vol_ring	-0.21	T_InComm.	-0.19	screen-on_past_300s	-0.14	MED_vol_music	0.33	MED_battery	-0.13
MAX_vol_noti	-0.21	T_normal_audio	0.19	T_map_apps	-0.13	MIN_vol_ring	0.32	AVG_battery	-0.13
MAX_vol_sys	-0.20	T_ringtone	-0.16	scrolling_count	0.13	strm_vol_music	0.32	scrolling_past_900s	0.13
STD_vol_sys	-0.19	MAX_vol_sys	-0.16	long-clicking_count	0.13	AVG_vol_ring	0.31	T_photography_apps	-0.12
STD_vol_noti	-0.19	MAX_vol_noti	-0.16	T_social_apps	0.13	MED_vol_ring	0.31	scrolling_past_600s	0.12
STD_vol_ring	-0.19	T_mobile_network	0.15	scrolling_past_900s	0.13	strm_vol_ring	0.31	battery_level	-0.12
MIN_vol_voicecall	0.18	freq_text_changed	-0.15	scrolling_past_600s	0.13	AVG_vol_sys	0.31	focus_event_past_3600s	0.12
T_InComm.	-0.16	MAX_vol_ring	-0.15	screen-on_past_180s	-0.12	MAX_vol_sys	0.30	MAX_vol_music	0.12

Note. The T prefix indicates the cumulative time; the green and blue backgrounds indicate positive and negative correlations, respectively, with darker colors indicating higher correlations.

- The correlations between time-killing moments and phone-sensor features.
- Doing so revealed **features correlated with time-killing moments specifically for certain participants**, which would not have been revealed had they not been divided into groups.

3. Model Performance and Behavior by User Group

- Group2: **ScreenshotOnly model achieved better** performance than its SensorOnly model for all metrics **except specificity**
  - ScreenshotOnly was accurate in predicting time-killing moments but less so in predicting non-time-killing moments.

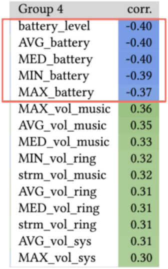


Group 2	corr.
screen-on_past_900s	-0.22
screen-on_past_600s	-0.22
screen-on_past_300s	-0.21
screen-on_past_1800s	-0.21
call_count	-0.21
screen-on_past_3600s	-0.21
screen-on_past_180s	-0.21
T_InComm.	-0.19
T_normal_audio	0.19
T_ringtone	-0.16
MAX_vol_sys	-0.16
MAX_vol_noti	-0.16
T_mobile_network	0.15
freq_text_changed	-0.15
MAX_vol_ring	-0.15

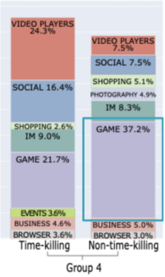
- Implies that these participants' non-time-killing moments were **more often associated with making calls**.
- The prior research has reported a **high association between quiet ringer modes and proactive phone-checking behaviors**

➔ the Group 2 behaviors we observed could have indicated participants checking their phones frequently to avoid missing calls and/or notifications.

- Group4: **Fusion model** performed particularly **well in specificity**
  - Fusion model was better at predicting its members' non-time-killing moments.



Group 4	corr.
battery_level	-0.40
AVG_battery	-0.40
MED_battery	-0.40
MIN_battery	-0.39
MAX_battery	-0.37
MAX_vol_music	0.36
AVG_vol_music	0.35
MED_vol_music	0.33
MIN_vol_ring	0.32
strm_vol_music	0.32
AVG_vol_ring	0.31
MED_vol_ring	0.31
strm_vol_ring	0.31
AVG_vol_sys	0.31
MAX_vol_sys	0.30

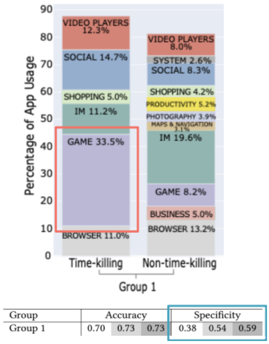


- predominantly **battery related** ones, which were **negatively correlated with time-killing moments**.
- **played games much more often during non-time-killing** moments than during time-killing ones (37.2% vs. 21.7%)

➔ Possibly because of the large quantity of this distinctive behavior during non-time-killing moments, the Group 4 fusion model's true negative rate was particularly high.

- Group1: Fusion model was better at predicting its members' time-killing moments

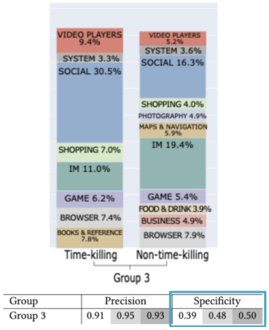
- Fusion model achieved significantly higher precision.



Group	Accuracy			Specificity		
Group 1	0.70	0.73	0.73	0.38	0.54	0.59

- Spent considerable time **playing games**, but in contrast to the Group 4 members, they were much more likely to do so **during time-killing moments**.
- SensorOnly model achieved much poorer specificity than its ScreenshotOnly model, suggesting that the fusion model **relied heavily on screenshot data to recognize non-time-killing moments**.

- Group3
  - Notably, correlations between features and time-killing moments were also lowest for Group 3, suggesting that its members' time-killing behaviors tended to be diverse.



Group	Precision			Specificity		
Group 3	0.91	0.95	0.93	0.39	0.48	0.50

- Despite having the highest precision among the four groups, it had a particularly **low true-negative rate**.

➔ might be attributed to it **having the most unbalanced dataset**: 80% of the instances were time-killing moments.

Limitations

- There may be an overlap between moments when people are killing time and moments when they feel bored, but the study didn't use a method to collect data on participants' boredom, making it unable to quantify the differences between these moments.
- Because of the sensitive data collected, participants might have altered their phone usage behavior during the study, impacting the results. Additionally, participants had limited control over sensor data collection, which could also affect their behavior.
- The data collected relied on participants' annotations, which may not always be reliable due to recall errors and mislabeling.
- The study did not consider additional features that could help infer users' activity on the phone screen.
- The dataset was small and limited to a specific demographic, raising questions about generalizability.
- Privacy concerns were not addressed adequately, and future research should consider encryption and other privacy protection techniques.
- The study did not analyze demographic characteristics, occupations, or notification-attendance behavior, which should be considered in future studies.