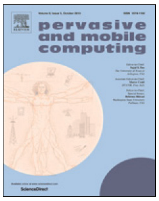
Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc



Utilising the co-occurrence of user interface interactions as a risk indicator for smartphone addiction

Björn Friedrichs [*], Liam D. Turner, Stuart M. Allen

School of Computer Science and Informatics, Cardiff University, Abacws, Senghennydd Road, Cardiff, UK

ARTICLE INFO

Article history:
Received 14 April 2022
Received in revised form 2 August 2022
Accepted 5 August 2022
Available online 13 August 2022

Keywords:
Smartphone addiction
Smartphone usage
User behaviour
Addiction risk

ABSTRACT

The push to a connected world where people carry an always-online device which has been designed to maximise instant gratification and prompts users via notifications has lead to a surge of potentially problematic behaviour as a result. This has lead to a rising interest in addressing and understanding the addictiveness of smartphone usage, as well as for particular applications (apps). However, capturing addiction from usage involves not only assessment of potential addiction risk but also requires understanding of the complex interactions that define user behaviour and how these can be effectively isolated and summarised. In this paper, we examine the correlation of physical user interface (UI) interactions (e.g. taps and scrolls) and smartphone addiction risk using a large dataset of those smartphone events (65,093,343, N=301,024 sessions) collected from 64 users over an 8-week period with an accompanying smartphone addiction survey. Our novel method which reports on the probability of a users addiction risk and in a model case we show how it was be used to identify 57 of 64 users correctly. This supports our observations of UI events during sessions of usage being indicative of addiction risk while improving previous approaches which rely on summative data such as screen on time. Within this we also find that users only exhibit addictive behaviour in a subset of all sessions while using their smartphone.

1. Introduction

Smartphones have found their way into many aspects of modern life, creating unparallelled connectivity between people and information. However, the human psyche can be overwhelmed with the never-ending instant gratification that is produced by those devices [1]. As a result, problematic behaviour such as overuse in the form of Smartphone Addiction (SA) can be induced [2]. In those cases the behavioural issues of SA are also referred to as Problematic Smartphone Use. This has been linked to multiple issues which include time spent on the device [2], certain types of applications [3,4], or amount of interactions taken [5]. However, where previous studies have provided indicators based on those circumstances, they have not considered the potential correlations between SA and behaviour summarised from a direct input source such as user interaction events that occur during usage sessions on devices.

The circumstances that can lead to addictive behaviour with technology is complex [6] and made more so for SA by the large body of research aiming to understand smartphone usage [7,8]. SA has been shown to have a significant effect on productivity and well being, with overuse of the device causing missed work, concentration issues or even physical symptoms [9]. Instead of focusing on these symptoms, in this paper we focus on examining links between interaction

[*] Corresponding author.
E-mail addresses: friedrichsb@cardiff.ac.uk (B. Friedrichs), TurnerL9@cardiff.ac.uk (L.D. Turner), AllenSM@cardiff.ac.uk (S.M. Allen).

https://doi.org/10.1016/j.pmcj.2022.101677

events (such as individual taps, scrolls or keystrokes) and SA. To do this, we utilise the Tymer dataset [5,10] which consists of 60,841,255 app-window, device and user-interaction events collected from 64 users over an 8-week period. SA is quantified through smartphone addiction scale (SAS) scores, which are extracted from the surveys undertaken by the participants.

We develop a methodology that extracts and represents user usage behaviour via UI events within smartphone sessions and use this to correlate with an SA score. From this, we build regression models which indicate the users risk of addiction on a per-session basis. This approach builds on previous event-based studies between smartphone usage and addictive behaviour [5,8] by firstly considering sets of interaction events that occur during a usage session and secondly, considering correlations to SA as a risk indicator. From this, the paper provides the following research contributions:

- We show that UI interactions are a stronger indicator of SA than previous metrics such as application use by time or application switches.
- Furthermore, that co-existing events in a session create a stronger risk evaluation than events considered in isolation.
- The type of application in combination with co-existing events has direct influence on the risk effect.

The end-goal from this study is to supply a framework to passively observe behaviour which correlates with SA that could even be computed on the smartphone itself. We envisage that this could be a stepping stone for tools that can help detect potentially problematic user states in real time through smartphone usage, without the need for exhaustive surveys, life-intrusive equipment and sensors or other problems arising from experience sampling [11].

2. Related work

The study of smartphone use has taken many facets and usually utilises multiple influences such as length of use, signal strength, or battery consumption, among others [12,13]. These high level factors have been captured on devices to build an understanding of how humans use these devices amongst daily life [14]. These findings have been used as the basis to explore potentially related psychological factors and the creation of applications aiming to support usage. This includes using the information to improve smartphone user experience by predicting the next-app the user will open and providing recommendations [15,16] or pre-loading applications in the background [17].

Another body of work examines usage with other psychological and social factors, such as interruptibility detection and prevention mechanisms [18,19], prediction of context such as user location [20]) or session behaviour (e.g. in call, just browsing) [21], or monitoring mood [22,23]. In these works there is a common coupling of real world data (e.g. length of smartphone use) mixed with additional data collection such as individual user surveys that record a users mental states [5]. A common basis within the analyses of these studies is the focus on summary statistics of usage [12], or on types of interaction events in isolation [5]. This creates a notable limitation and opportunity to examine the potential inter-play between different types of interaction behaviour within apps and sessions.

2.1. Smartphone addiction

Generally, addiction is a case of compulsive or obsessive behaviour that continues even when faced with the negative consequences (financially, socially, etc.) of those actions. There are different kinds of addiction and SA is often considered a 'behavioural addiction' [2] where habits get enforced from gratification instead of as a result of e.g. substances. Parallels of this behaviour can be seen in gaming [24] or internet [6]. This kind of addiction has also been linked to some extent with various personality and identity traits [6] such as anti-social behaviour [25]. It has also been suggested that SA's contributing factors are hard to differentiate from those of non-addicted users [26].

Much like other addictions, whether or not, and to which extent someone is affected by SA is not immediately obvious. A common way to measure the severity of SA in the literature is the Smartphone Addiction Scale [9]. The SAS is a self-report survey that was developed to identify the level of addictedness of users. A short version (SAS-SV) [27] was produced which cut the amount of questions from 33 to 1. The participants answered those questions on a six point Likert scale with results ranging from 1 to 6. Part of the SAS-SV is an extension that presents cut-off values for male (31) and female (33) users which when exceeded indicate SA. The SAS has been used to link addiction to multiple facettes of mental health such as social anxiety [28], a need for social acceptance [2] or stress [29].

Additionally, SA is often linked with certain application categories such as social networking and communication [3,30], entertainment [4] or video games [31]. These studies have found that, while there can be many reasons why one is unable to stop using their phone, many times it is down to keeping up their status in their relevant social circles [4] or to relieve anxiety [32]. Additionally, there have been reports of categories of applications or individual applications being identified as a contributor to SA that do not seem immediately obvious (e.g. reading in a bible application [33]).

A main drawback across the previous studies is the absence of a flexible model for embedding behaviour and examining correlations to addiction risk. This motivates the development of a flexible semi-supervised methodology as part of this study.

2.2. Hypotheses

From the previous work surrounding smartphone use and its connections to addictive behaviour we formulate the following hypotheses:

H1 Co-occurring UI events are reflective of addiction and offer a more strongly correlated indicator (in terms of effect size) when assessing SA risk compared to length of use or events in isolation.

H2 The within session usage will not always be uniquely represented by only an addicted or non-addicted group of users. Rather, we expect that sessions by addicted users will sometimes display traits of sessions usually generated by non-addicted users and vice-versa.

H3 Certain application categories show a stronger response than others in context of their interactions, specifically **H3a** social, **H3b** communication, **H3c** entertainment and **H3d** game applications.

3. Methodology

To examine the interaction effects between UI events and SA scores we create multiple models of sessions from the raw data. As done by Friedrichs et al. [34] we create two representations of the user interaction data for each session, one using absolute counts of each event type and another utilising the natural language processing technique, term frequency-inverse document frequency (TF-IDF), to represent the relevance of events in a session, comparable to keywords in a document. This in turn allows us to consider the impact of embedding interaction event information as a set of singular events and as co-existing events in a session. We extend this by introducing a dependent variable, a smartphone addiction level, by means of an accompanying per-user SAS survey. We first examine the correlation of individual UI events with SA before moving on to evaluating their co-existence and the relevance of application types.

3.1. Dataset

The Tymer dataset [5,10], was collected from a similarly named Android application that collected and tracked device interaction and notification data. Participants were asked before and after the usage period to complete a survey which allowed us to place them on the SAS. The average SAS score from each of the 64 users was retrieved. Of those 64 users (30 female, 34 male), 40 million interaction events were captured over the span of 8 weeks (after pruning duplicate events). Users were split using the cutoff points as defined by Kwon et al. [27] and discussed in Section 2.1, which resulted in 13 users being identified as addicted.

The data used in this paper includes all detected screen on, screen off, and boot events, as well as the following events: *Tap*, *Long Tap*, *Text input*, *App switch* (e.g. opening an application), and *Scrolling*, along with pseudo-events to capture the periods between interaction events in a session — *Short idle* (1 s) and *Long idle* (45 s, based on [35]). One limitation of the data collection was the capture of applications which were built using pure 'canvas' screens (gaming applications frequently make use of these). Only inputs of the Android interface were captured, this means that interactions with the canvases are not considered. Multiple approaches of creating sessions have been proposed across the literature, including screen events [36], application boundaries [37] or non-interaction timeouts [35]. As we are looking to detect problematic behaviour every time someone picks up their phone we decided to proceed with the screen event based approach, similarly used by Oulasvirta et al. [38] and Hintze et al. [39].

To create the sessions, all events are assigned to their respective user and then sorted based on their timestamp. All events between screen on and screen off form a session. Some screen on events were not followed by screen off events, which could be due to a number of reasons including the battery depleting or issues in the stability of the data collection mechanism from the original study. To account for this, we inferred the endpoints for these sessions as the timestamp of the last interaction event before a number session began. The total number of sessions in the dataset was 316,072 (per user Std. = 3427.48). The inherent distribution of session usage is strongly skewed towards very short bursts of interaction, which resulted in n = 15,048 sessions having no UI interactions at all. We argue that these sessions will not be suitable for analysis in this study as the events will be used as features for comparison with SAS, so were removed.

Each session was then labelled with the SAS label of the user using the remaining N = 301,024 sessions, resulting in n = 79,354 addicted (13 users) and n = 221,670 non-addicted (51 users) sessions. The aggregated set of sessions is used as the basis for analysis, rather than compressed traits for individual users, as the objective of the study is to isolate interaction behaviour in sessions that could be indicative of an addicted user. We hypothesise (H2) that a labelled addicted user will not exhibit correlating behaviour in every usage session.

3.2. Representing usage behaviour and addiction

To represent behaviour, previous work [34] has shown that embedding interaction events to represent behaviour reveals common 'types' of smartphone usage sessions. These events also reveal types of usage sessions that are not apparent using summative features such as the length of a usage session. As a parallel, in this paper we examine correlations between SA and both, UI event feature embeddings and summative features (such as session length) to observe the additional utility that interaction events can bring. We analyse the individual interaction of features with

SA by probing their separability between users. We then move on to logistic regressions that combine all of the above features to build models that can predict a users addiction risk based on their usage session behaviour. We report multiple metrics, as the standard results for Mann–Whitney U (U and *p*-value) and logistic regression (error rates and confidence intervals) are not directly comparable. For comparisons we will rely on effect sizes since p-values by themselves might indicate significance but fail to capture the strength of the prognostic capabilities on the outcome variable [40].

3.3. Effect size

The Mann–Whitney U (MWU) test is a common nonparametric test to check for significant differences in size between two samples by comparing their medians. Our session data can be separated by the respective addiction label supplied by the user. These tests will provide a baseline of general separability between the addicted and non-addicted samples. Due to the nature of statistical tests, even small deviations will be reported as significant for large sample sizes, so we also report the effect sizes as an area under curve (AUC) in addition to test statistics and p-values. AUC is a standard way to describe effect sizes [41] and will also provide a basis for comparison in the logistic-regression task.

The AUC score ranges from 0 to 1, where given two sets of data it describes the predictive capabilities of a chosen variable or model. The bounding values 0 and 1 correspond to a strong (negative or positive) diagnostic ability and .5 to no diagnostic ability. While AUC values have no strict boundaries they can be categorised by rule of thumb, we follow [42] label them as poor for $.5 \leq AUC < .7$, acceptable for $.7 \leq AUC < .8$, excellent for $.8 \leq AUC < .9$ and outstanding for $AUC \geq .9$. Formally the AUC can be derived from the MWU statistic [43,44] by letting U be the test statistic result, n_x be the size of a sample and ϕ the normal cumulative distribution function, so that the score can be calculated as:

$$z = \frac{U - \frac{n_1 \times n_2}{2} - .5}{\sqrt{\frac{n_1 \times n_2 \times (N+1)}{12}}} \tag{1}$$

$$r = \frac{|z|}{\sqrt{N}} \tag{2}$$

$$d = \frac{2 \times r}{\sqrt{1 - r^2}} \tag{3}$$

$$AUC = \phi \frac{d}{\sqrt{2}} \tag{4}$$

4. Assessing risk indicators

In this section we will explore the utility of using smartphone interaction behaviour as the basis to predict probabilities relating to SA. Due to the sensitive nature of the topic there are implications of hard misclassifications if represented as strict binary classification labels (addicted when not addicted or vice versa). As a consequence we will report our findings in terms of potentially correlating factors and probabilistic risk, rather than treating it as a strict classification problem. Then we add only suggestions as to how they could be used in practice. We first investigate the effectiveness of high-level features separating the samples of addicted and non-addicted users. We then show how those results are improved by utilising UI and introducing application categories.

4.1. Separability by summative, high-level features

SA has been connected to summative, high-level features such as time spent in applications [33] or application changes in fragmented use [4]. Therefore, in addition to creating session embeddings from UI events (e.g., taps and scrolls) we also calculate the session length, number of switches between applications, and the total number of events to provide a comparable baseline. Each session is labelled with according to the SAS of the user that generated them (discussed further in Section 3.1).

Table 1 shows that the median session length (in seconds) for sessions that were labelled addicted (Mdn = 2.38 s, M = 101.59 s, SD = 276.71 s) is slightly longer than for those non-addicted (Mdn = 19.98 s, M = 12.58 s, SD = 324.05 s), U = 1700735811, p ≤ .001, AUC = .526. We extend this by examining the time spent in different categories of applications per session. We extracted the Google Play Store category for the app the event occurred in, with all apps not on the Google Play Store placed in an *Other* category, resulting in a total of 45 categories. When considering time spent in specific categories, simulation games showed the strongest differentiation and none of the categories in our hypotheses H3a-c show a significant difference (other categories were omitted). Statistically these are significant based on the *p*-value threshold, but the effect size shows that there is limited potency. Additionally, we can observe how the categories with the strongest effect sizes have low cross-user representation (e.g. game simulation apps are represented by only 4 of 64 users). This limits confidence in the use of session length for specific types of applications across a population.

Sessions by addicted users typically included more application switches (Mdn = 4, M = 5.9, SD = 14.61) than non-addicted users (Mdn = 2, M = 7.75, SD = 34), with a slightly stronger response than overall session length, U = 1666167040, p ≤ .001, AUC = .533. However the effect sizes are also small. Lastly, the total number of interactions

Table 1
MWU statistic of length of use in seconds for overall sessions and the top 5 scores of usage time within specific application categories. $p \leq .001$ applies for all results.

Category	AUC	U	Addicted			Non addicted		
			Users	Mdn	SD	Users	Mdn	SD
All	.527	1 659 087 399	13	81.5	395.2	51	70.3	429.6
Game simulation	.686	11 973	1	119.6	113.5	3	315.9	474.4
Education	.682	9 758	4	89.1	141.4	16	16.1	154.0
Sports	.681	28 752	2	4.0	106.7	7	24.2	91.1
Tools	.675	53 666 516	13	11.7	207.4	51	2.0	131.3
Finance	.672	179 604	6	11.5	60.1	24	30.9	92.2

Table 2
MWU tests using the count and TF-IDF scores of user interaction events in sessions. $p \leq .001$ for each feature. The results are sorted by effect size.

Type	Feature	AUC	U	Addicted			Non addicted		
				Users	Mdn	SD	Users	Mdn	SD
Count	Text input	.568	481 378 196	13	22	172.2	51	11	146.5
	Long idle	.542	131 478 681	13	2	16.9	51	3	18.7
	Long tap	.535	969 989	13	1	1.2	51	1	2.3
	Scrolling	.534	1 022 691 388	13	7	68.5	51	5	448.4
	Short idle	.524	1 664 563 262	13	66	919.6	51	66	732.4
	Tap	.523	1 196 571 408	13	3	29.8	51	4	29.5
	App switch	.510	1 609 207 477	13	5	11	51	5	1.9
TF-IDF	Tap	.556	1 322 548 432	13	.309	.137	51	.342	.140
	Scrolling	.554	1 000 752 988	13	.465	.185	51	.424	.177
	Long idle	.542	141 631 150	13	.366	.192	51	.409	.193
	Long tap	.534	993 693	13	.365	.148	51	.339	.139
	Text input	.527	557 535 949	13	.641	.186	51	.608	.184
	App switch	.509	1 716 098 754	13	.332	.135	51	.339	.138
	Short idle	.508	1 768 056 702	13	.464	.147	51	.466	.156

has previously been used as a basis for correlating with SAS scores over time [5]. We find that while there is a significant link between the total number of interactions produced by addicted users (Mdn = 26, M = 226.23, SD = 844.87) and non-addicted (Mdn = 28, M = 225.38, SD = 725.54) within sessions, the effect size is poor (U = 1666065715, $p \leq .001$, AUC = .533).

This analysis reveals that while sessions do have statistically significant differences between their medians of session length, application switches, and number of interaction events, the effect sizes are small enough to make their predictive capabilities unsuitable. Overall, we can conclude that the results of summative, high level features, are unlikely to be suitable for the goal of isolating SA. We explore user interactions based on specific events next.

4.2. Separability by user interactions

As a comparison to the summative, high-level features (e.g., session length) we examine different representations of user interaction activity within sessions and observe correlations with SA.

For the first representation, we count the occurrences of each event type per session. Table 2 shows that while significant, very weak effects can be observed for individual event types. For the second representation, we consider sessions from all users as a 'corpus', and use this to calculate the *relevance* of each event using TF-IDF. We define the relevance as the TF-IDF score for each event type e in session d as:

$$\begin{aligned} \text{TF-IDF}_{ed} &= \text{tf}_{ed} \times \text{idf}_e \\ \text{tf}_{ed} &= 1 + log\,(freq(e,d)) \\ \text{idf}_e &= log\left(\frac{1+n}{1+\text{df}_e}\right) \end{aligned}$$

where $freq(e,d)$ is the number of events of type e that occurred in session d, n is the total number of sessions and df_e is the number of sessions that contain an event of type e. Each session d is then represented by a feature vector $f_d = (\text{TF-IDF}_{e_1 d}, \ldots, \text{TF-IDF}_{e_7 d})$. We then test the difference of the medians between the session samples for every event type using MWU.

Table 2 also shows the results when considering TF-IDF scores rather than the number of times events occur. TF-IDF shows equally low effect sizes to counts. However, these findings indicate that considering some interaction events may provide slightly stronger predictive power in comparison to overall session length, number of application switches, and number interaction events. From here we examine interaction events within specific categories of apps using counts and

Table 3
Results of a MWU test for the top ten features (count of each event type in an app category) with the highest effect sizes. N refers to the number of users in each group.

Category	Feature	AUC	p	U	Addicted			Non addicted		
					N	Mdn	SD	N	Mdn	SD
Game trivia	Scrolling	.815	.009	1	3	19	20.6	2	17.5	21.2
Game trivia	Tap	.806	.004	10	4	41	187.7	3	30	174.6
Game simulation	Long idle	.794	<.001	10 238	1	3	4	3	10.5	26.8
Music and audio	Long tap	.788	.032	62	3	1	.37	2	2	2.6
Sports	App switch	.776	<.001	5 883	2	1	2	7	1	1.5
Game simulation	App switch	.767	<.001	16 069	1	1	0.4	3	2	2.9
Education	Text input	.760	<.001	2 240	2	16	12.8	9	3	9.4
Education	Long idle	.747	<.001	1 724	4	1	2.1	9	3	10.2
Launcher	Long idle	.732	<.001	2 156 723	12	1	4.3	45	2	9.9
Sports	Scrolling	.701	<.001	3 818	2	7	21.9	6	3	129.3

Table 4
Results of a MWU test for the top ten features (TF-IDF score of each event type in an app category) with the highest effect sizes. N refers to the number of users in each group.

Category	Feature	AUC	p	U	Addicted			Non addicted		
					N	Mdn	SD	N	Mdn	SD
Education	Text input	0.934	<.001	433	2	16	12.8	9	3	9.4
Game trivia	Scrolling	0.837	.007	0	3	19	20.6	2	2	0.5
Game trivia	Tap	0.832	.002	7	4	41	187.7	3	3	4.5
Music & audio	Long tap	0.814	0.02	66	3	1	0.4	2	2	2.6
Sports	Scrolling	0.780	<.001	2 699	2	7	21.9	6	3	129.3
Sports	Short idle	0.747	<.001	12 310	2	12	327.5	7	38	228.1
Tools	Short idle	0.725	<.001	24 947 084	13	14	1371.9	51	4	192.1
Game word	Short idle	0.715	<.001	1 459	3	14	63.8	8	41	563.5
Sports	App switch	0.709	<.001	6 481	2	2	2.1	7	1	0.7
Finance	Text input	0.707	<.001	6 604	5	6	30	15	3	19.7

TF-IDF scores. This will provide a comparable basis to the time spent in categories of apps which produced stronger effect sizes in comparison to time spent across all apps, however with a limited amount of users in the dataset.

4.3. Influence of application categories

We extend the analysis to consider the potential efficacy of UI events as features by examining the events for specific categories of applications, rather than generally. From this, given the set of all event types $E = \{e_1, \dots, e_7\}$ and the set of all categories $C = \{c_1, \dots, c_{45}\}$, we construct a feature combination vector $f_{c_{315}d}$ based on $E \times C$, containing information on each event type for each category. After constructing the final vector it was reduced from 315 to 278 features by removing 37 combinations of events and app categories which did not occur in the dataset. We formalise this as an expansion of the original feature vector f_d as $f_{c_{278}d} = (\text{TF-IDF}_{e_1 d}, \dots, \text{TF-IDF}_{e_{278} d})$.

We first consider the features based on the number of times they occurred in a session. Of the 278 features, 108 are statistically significant and those with the largest effect sizes can be seen in Table 3. The effect sizes, with the largest being $AUC = .815$ (Scrolling in trivia games), resemble a reasonable indicator for the utility of considering interaction events relative to the type of app they occurred in, in comparison to generally and the summative, high-level features in Section 4.1.

We repeat the analysis using vectors comprised of TF-IDF scores to determine whether a metric that considers the co-occurrence of events has stronger effects than event counting. 109 features were statistically significant and Table 4 shows the features with the largest effect sizes, with the largest being .924 (*Text input* in *education*).

We attribute this jump of performance to two factors. Firstly, the test statistic (U values) are very low for some of the results. These low values mean there were fewer sessions to evaluate from. This is also reflected by the samples under-representing total users where for the most part less than a dozen of users contributing. Secondly, the differing results between the top features of count and TF-IDF can be accounted to the scaling that takes place during the TF-IDF vectorisation. TF-IDF's ability to distinguish nuances in usage might be compromised by the very low vocabulary of only 7 features (i.e. event types). While TF-IDF may show a slightly higher effect sizes (AUC values) for some features, overall they perform similarly. The cross product of features and categories created a better model for TF-IDF itself.

Some of the MWU scores for isolated user interaction events showed an improvement over session length (e.g. *Text input* in education or *scrolling* in trivia games). This suggests that when all data points apart from a single category-feature combination were dismissed, it showed better results across the board for those sessions that remained. As we can only use sessions that those combinations occurred in, we lose many sessions for the evaluation. To be able to use all sessions we want to consider all features together, this can be achieved via a logistic regression and forms the next focus of our

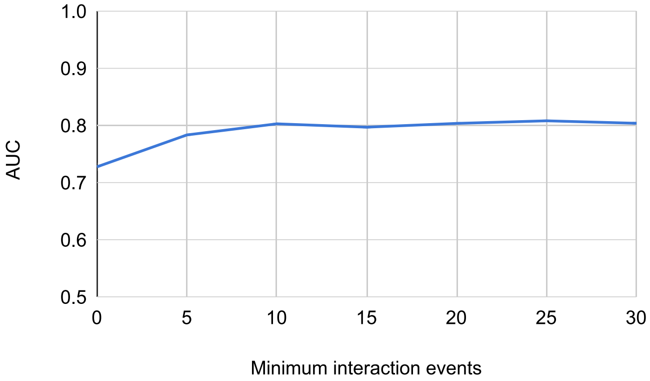


Fig. 1. Increase in effect size (AUC) for testing the model as the minimum of required events increases.

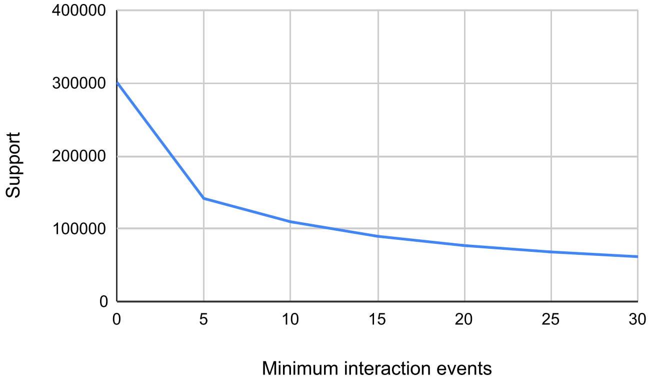


Fig. 2. Reduction of available sessions user for testing the model as the minimum of required events increases.

analysis. The higher performance of the TF-IDF results that already consider the co-existence of events among others, and this motivates that the interplay of multiple features may produce strong results when applied to a trained model.

5. Combination of features

We extend our statistical analysis of individual event types to examine H1 further by building Logistic Regression models that embed multiple event types. We trained the models with 90% (n = 270,922) of the available sessions and then tested with the remaining 10% (n = 30,102) and performed 10-fold cross-validation to limit any selection bias and report the mean values. Before building the models we need to address high collinearity issues by pruning the vectors defined in Section 4.3 ($f_{c_{278}d}$) which removes the features that show a correlation higher than .9, creating a final vector of 256 event type pairs ($f_{c_{256}d}$). To evaluate these models we reuse the AUC of the receiver-operator-characteristic (ROC) as a measure of performance of the true positive rate against false positive rate.

The performance of the models built from the counts of event types per app category is similar to the single event effect sizes in Table 3 (AUC = .68, p<.001, SD = 0.002). However, for models built from the TF-IDF value equivalent vectors, the model performance improves by almost 5%, AUC = .73, p<.001, SD = 0.002. This achieves a reasonable result in terms of effect size as defined in 3.3 and shows that considering multiple types of UI events (and their weighting relative to each other) provides the improved separability over singular events — supporting H1. This method also allows us to captures all 64 users, similar to the tests without category restrictions while improving on their efficiency by more than 20%.

As discussed in Section 3.1, short sessions can be an issue when trying to predict the label as these typically contain few events which consequently could inhibit the ability to distinguish between an addicted user and not. To investigate the effects of this, we removed sessions that had less than 5 user interactions of any type. While this reduces the amount of sessions to 45% (302,734 to 141,588 sessions) it retains sessions for each user (M = 2212.31, MD = 2126, Std = 1365.3). Doing so improves the performance by 5% to AUC = .773. We also examined alternative event count thresholds and the effects of this on model performance (Fig. 1) and the number of sessions this removes (Fig. 2), which shows support for considering a minimum, however the benefits reduce the larger this is.

Table 5
FDR corrected significant categories and feature combinations for classifying addiction. Negative coefficients are a stronger indicator for non addiction. Switch = App switch.

Category	Feature	β	AUC	p
Communication	Scrolling	1.38	.704	<.001
	Text input	.13	.52	.012
	Short idle	−.55	.415	<.001
	Tap	−1.01	.35	<.001
	Long idle	−1.19	.322	<.001
Education	Text input	2.15	.799	.041
Entertainment	Short idle	−1.36	.298	.003
Finance	Switch	2.01	.784	.013
	Text input	1.62	.736	.013
	Short idle	−1.23	.315	.001
Adventure[a]	Switch	3.27	.899	<.001
	Short idle	.58	.59	<.001
	Long idle	−.75	.385	.001
Casual[a]	Switch	3.64	.922	<.001
	Long idle	2.28	.813	<.001
Puzzle[a]	Long idle	1.96	.777	.016
Sports[a]	Long idle	3.49	.913	.004
Strategy[a]	Short idle	−3.67	.076	.003
Health and Fitness	Switch	3.63	.921	<.001
	Short idle	−3.02	.12	<.001
Launcher	Scrolling	2.68	.852	<.001
Lifestyle	Switch	4.00	.94	<.001
	Text input	2.59	.844	<.001
	Scrolling	1.28	.691	<.001
	Long idle	−1.05	.341	.001
	Short idle	−1.19	.321	<.001
Music and audio	Text input	1.24	.685	.005
	Tap	.52	.58	.013
	Scrolling	.457	.571	.048
	Long idle	−2.19	.197	<.001
News and magazines	Scrolling	2.14	.798	<.001
	Switch	1.72	.749	<.001
	Short idle	−.52	.42	<.001
	Tap	−.803	.38	<.001
	Long idle	−2.75	.142	<.001
Web browser	Tap	3.56	.917	<.001
	Text input	2.47	.833	<.001
	Switch	1.70	.746	.002
	Long idle	−1.92	.227	.049
	Short idle	−2.75	.142	<.001
Navigation	Short idle	−1.28	.309	.006
Personalisation	Switch	−4.20	.051	.041
	Short idle	−4.95	.027	.031
Photography	Scrolling	1.45	.714	<.001
	Switch	1.14	.672	.002
	Short idle	−.77	.381	<.001
Productivity	Switch	5.93	.99	<.001
	Scrolling	.85	.63	<.001
	Text input	−.48	.426	.037
	Tap	−1.51	.28	<.001
	Short idle	−2.49	.166	<.001
	Long idle	−2.91	.128	<.001
Social	Switch	1.44	.712	<.001
	Text input	.79	.621	<.001
	Short idle	.37	.558	<.001
	Scrolling	−.31	.452	<.001
	Tap	−.429	.43	.001
	Long idle	−1.10	.334	<.001
Sports	Short idle	−2.83	.135	<.001
Tools	Short idle	4.82	.97	<.001
	Long tap	1.92	.773	.049
	Scrolling	.74	.613	<.001
	Text input	−1.25	.313	<.001
	Tap	−1.87	.23	<.001
	Long idle	−3.07	.116	<.001
	Switch	−9.23	0	<.001
Travel and local	Scrolling	1.498	.72	<.001
	Switch	1.03	.656	.006
	Text input	1.00	.652	.011
	Short idle	−1.64	.261	<.001
Video players	Switch	−1.63	.263	.009
	Scrolling	−2.41	.174	<.001
Weather	Short idle	−5.77	.012	<.001
Other	Short idle	1.44	.712	<.001
	Long tap	1.34	.699	.001
	Scrolling	.89	.626	<.001
	Long idle	−.99	.35	<.001
	Switch	−1.33	.302	<.001
	Text input	−1.36	.298	<.001
	Tap	−2.48	.17	<.001

[a] Game category.

5.1. Categories as a predictor

From the Logistic Regression models we can extract feature importance by utilising the models' coefficients. The AUC in this case is related to the coefficient and is added as a way to show the effect size. Where e is the Eulers number, β the regression coefficient and OR is the odds ratio, then formally [45]:

$$OR = e^{\beta} \tag{5}$$

$$d = \ln OR \times \frac{\sqrt{3}}{\pi} \tag{6}$$

$$AUC = \phi \frac{d}{\sqrt{2}} \tag{7}$$

Table 5 shows the event category, coefficient β, effect size (AUC) for features where p<.05. The features are sorted by the strength of their significant coefficients β in descending order. In this section we will focus on only reporting effect sizes that show moderate to high correlations. The first observation is that significant features for addiction are spread across almost all categories and their contributing magnitude or even polarity is not always matching. We note that since we did not select features and there are many of them, there is the chance that some of the features (even with low p-values after false discovery rate correction) are false positives or (not-featured) negatives. While we will discuss the observations, the main focus of this section is the magnitude and differences of effect size for features across categories and how they differ from each other.

Another aspect to consider is, given the transformation by TF-IDF, these coefficients do not necessarily hint towards the frequency of the events occurring in a session, but rather towards the relevance of the event. For example, when *app switches* are the most relevant in lifestyle applications, they are positively correlated towards addiction, whereas *long idles* are negatively correlated. Sessions can have fewer *app switches* than *long idles* but if they are more important in context of the session the evaluation will swing towards addiction risk.

App switches (AUC = .712, p<.001) in social apps are moderately contributing towards risk of addiction. This observation could be part of the fragmented use of frequent switches, especially surrounding social media apps, as referenced previously [4]. This would be further supported by *long idles* being a negatively correlating factor (AUC = .334, p<.001), which could hint towards more considered use in social apps.

Scrolling (AUC = .704, p<.001) in communication apps is moderately positively correlated with addiction which would reflect the findings in [5]. *Long idle* (AUC = .322, p<.001) and *tap* (AUC = .347, p<.001) events are on the edge of moderate negative correlation. This could be connected to factors such as scrolling being a more engaged action than simply waiting for responses or tapping various interface elements to reply.

The presence of 6 and 5 events respectively for social and communication apps being significant with effect sizes ranging from moderately negative to moderately positive is indicative that we can observe separability based on social and communication apps and therefore shows support for H3a and H3b. Overall, social and communication apps have lower effect sizes than gaming or some of the other categories (all AUCs<.8). This may be reflective of them being common activities across all smartphone users [46].

Entertainment apps only show *short idles* (AUC = .298, p = .003) as the only significant feature, being negatively correlated. With only a single variable for this category it is hard to argue for stronger separability using entertainment apps compared to others. Our data shows that other categories (e.g. games or communication) are more widely used, the additional data could lead to better separability and therefore stronger correlation coefficients. From this we conclude that events in the entertainment category are not a strong indicator for distinguishing the addiction risk in users and we reject H3c.

Multiple event types in game apps also show correlations to SA. *App switches* are strongly positively correlated in casual and adventure games (AUC≈.9, p<.001). In casual, sports and puzzle games *long idles* have a moderate to strong effect towards SA (.75<AUC < .95, p<.05). *Long idles* in adventure games (AUC = .385, p = .001) show a slight negative correlation with addiction risk and were not significant for all other game categories. *Short idles* are strongly negatively correlated for strategy games (AUC = .076, p = .003) and vaguely so in adventure games (AUC = .385, p = .001). This shows support for H3d as playing games for a variety of sub-categories (i.e., adventure, casual, puzzle, sports and strategy). Compared to other categories, coefficients are high, specifically switching to and idling (potentially because of capture limitations discussed in Section 3.1) with the games open show strong correlations.

Significant coefficients of the same events between categories also reveal notable observations. *App switches* display moderate to very high AUCs for almost all categories, outliers are made up of personalisation, tools, video players, and the 'other' category. This signifies that when switches are the most relevant interactions, a session is more likely to be connected to SA. At least the 'other' category could be explained by the fact that it will include a wide range of applications and switching to any of those would not always be considered neutral. Conversely, *short and long idles* are almost exclusively a negative predictor for SA apart from some of the game categories (specifically casual, sports and puzzle games). This pattern exists for all events, a predominant polarity of coefficients with few outliers, including a gradient of effect sizes from poor to excellent. This supports prior findings where singular events are generally an influence [5] on SA but simultaneously also support our assumptions that no single event can evaluate SA in isolation across all categories.

We were able to observe that events within these categories (e.g. game, social and communication applications) are able to separate addicted from not addicted behaviour. Moreover, there are tangible differences in predictive capabilities and magnitude for the same event types across categories. Some categories (e.g. dating or shopping) did not display any significant features at all. This supports our hypothesis H3 in that problematic behaviour is not uniform for all application categories.

6. Applying risk

Labelling someone as addicted due to a usage session(s) should be handled with care due to the risk of false positives or negatives being potentially harmful and any model could be argued as only to be used as decision support tool with human oversight. In the literature, regression models are often evaluated based on some score or classification rule. This includes defining a threshold in the probability range 0 to 1 and then classifying data and points higher or lower than this threshold accordingly [47]. In a balanced dataset this threshold corresponds to .5 but to account for the imbalance

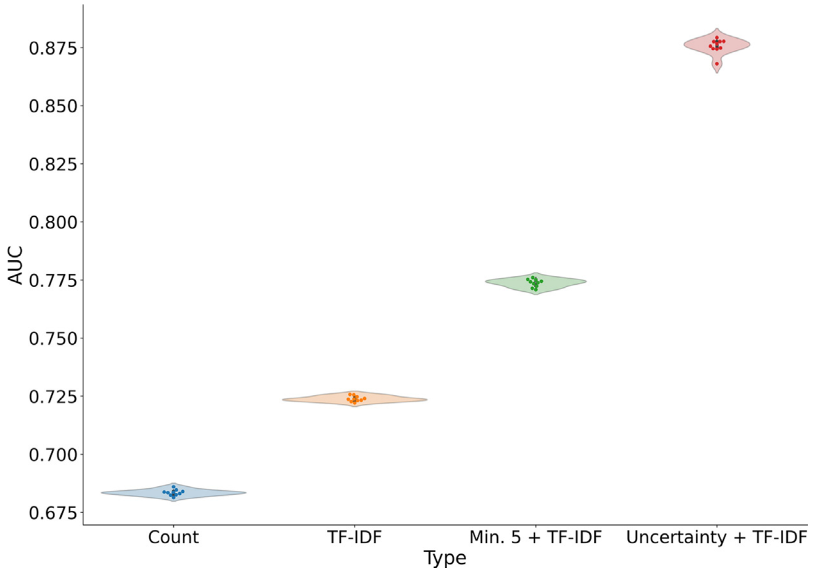


Fig. 3. Improvements of AUC when detecting SA using 10-fold cross-validation with (left to right) count, TF-IDF, TF-IDF with 5 minimum events, TF-IDF when applying an uncertainty area of .15 at the threshold of .23.

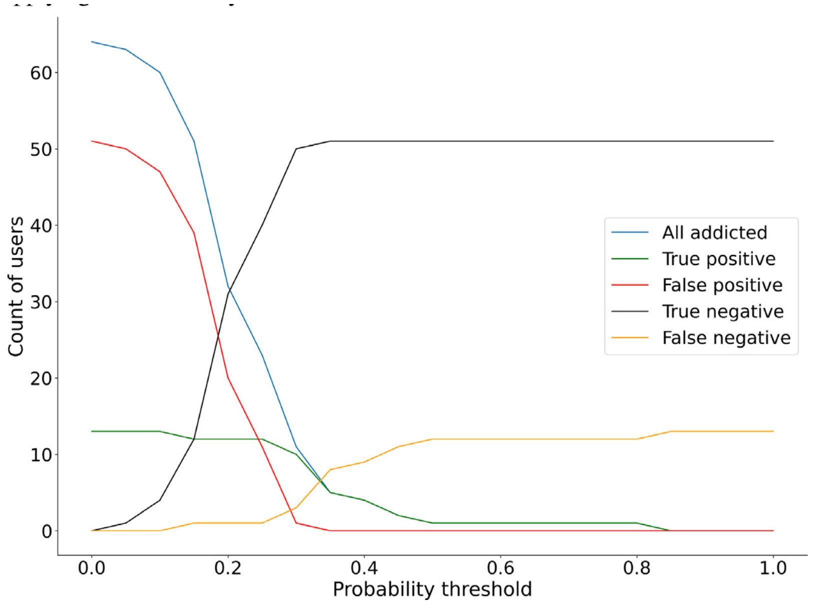


Fig. 4. User count of true positives, false positives, true negatives and false negative based on mean probabilities.

of addicted compared to non-addicted users and potentially different priorities in minimising false positives and false negatives, we explore changes to regression performance for different thresholds.

We can uncover this threshold by calculating the maximum Youden index [48] to optimise the break-even point between the false positive and true positive rate.

$$sensitivity = \frac{\text{true positives}}{\text{true positive} + \text{false negatives}} \tag{8}$$

$$specificity = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \tag{9}$$

$$J = sensitivity + specificity - 1 \tag{10}$$

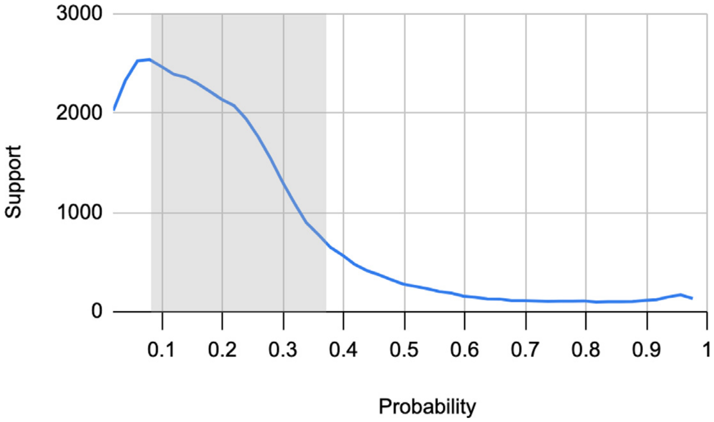


Fig. 5. A .15 uncertainty range around a .23 threshold based on their prediction probability. These sessions would be considered *uncertain* and be excluded.

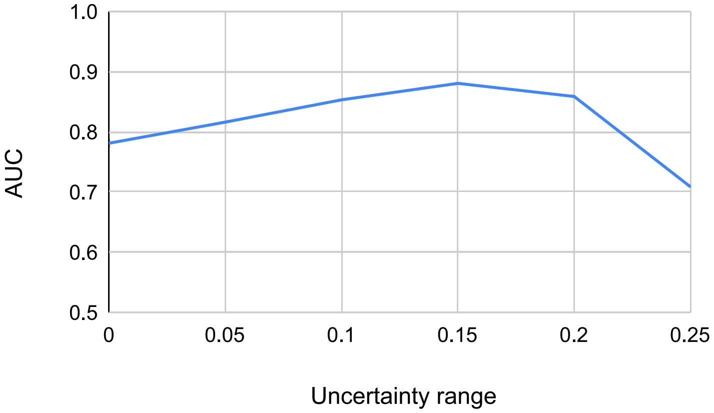


Fig. 6. AUC of a logistic regression model where *uncertain* sessions inside of the uncertainty range are not considered as part of the scoring.

In our case this is .23 (visualised in Fig. 4), where the false positives decrease drastically but the true positives remain. In terms of risk to be addicted, the probability evaluation of sessions could be accompanied by a scale that places users into a more understandable risk system. A basic example would be defining t such that when p is the addiction probability, low ($p \leq .23 - t$), medium ($.23 - t > p > .23 + t$) and high ($p > .23 + t$) can categories can be understood. The understanding of a scale that has one or more cases which define a medium (or uncertain) case is explored in more detail in the following section.

6.1. Uncertainty range

Sessions are labelled based on the user's addiction label as a binary value. This assumes that every session of an addicted user will be distinct from that of non-addicted users. We propose through H2 that this is not the case and that a subset of sessions will have distinct characteristics and a subset will be similar. For example, the dataset contains sessions that are short and have limited interaction events where correlating characteristics with addiction may be prevalent. To examine this and the impact on the modelling, we utilise the probability range from the previous section to evaluate using three instead of two prediction classes. The intention of this is to isolate types of usage sessions that are common to both addicted and non-addicted users.

To account for uncertain cases of interaction, we continue with training a binary classifier, but then evaluate and adjust the classification of the training data to account for uncertainty in the model. Interpreting ranges in logistic probabilities in as uncertain cases has been discussed in the literature before [49]. In these cases, a low confidence area around the classification threshold is created which besides the labels (in our case, *addicted* and *not addicted*) creates a third label, *uncertain*.

As can be seen in Fig. 5, a uncertainty range is extended to both sides of the threshold yielding the excluded sessions. When applied it will remove sessions from the support but greatly benefit the classification power. Similar to limiting sessions to a minimum event count, Fig. 6 shows that as we increase the threshold and remove uncertain cases from

Table 6
Confusion matrix for addiction classification per user based on a Youden's index of .23 and by utilising the each users mean session probability. Based on 141,588 sessions and probabilities without an uncertainty range.

	Negatives	Positives	
True	36	12	48
False	1	15	16
	37	27	

Table 7
Confusion matrix for addiction classification per user based on a Youden's index of .405 and by utilising the each users mean session probability. Based on 141,588 sessions and 48,832 probabilities with an uncertainty range of .15.

	Negatives	Positives	
True	46	11	57
False	2	5	7
	48	16	

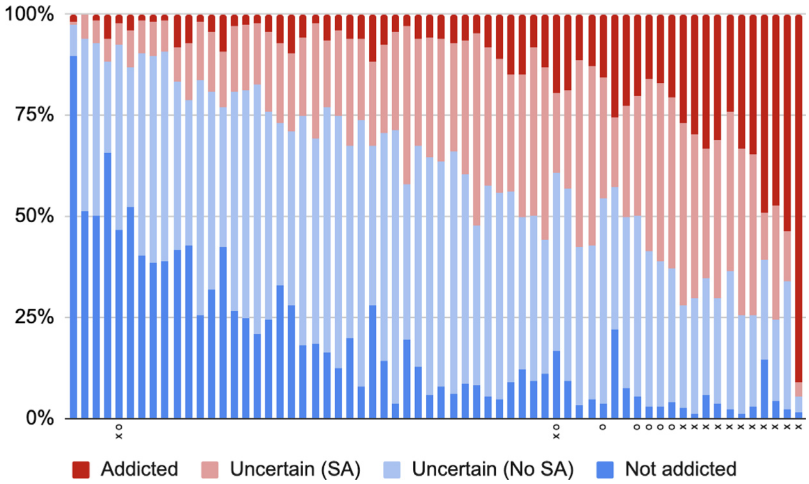


Fig. 7. The distribution of session that are considered addicted, non addicted or uncertain for every user. Uncertain is split into session that would have been classified addicted or not addicted at the .23 threshold. The users are sorted by their mean addiction probability of all their sessions. Symbols along the x-axis: 'x' is a user that is addicted according to the SAS. 'o' shows that if a user is either a false positive or negative (depending on 'x') when considering a Youden index of .405.

the binary classification, the model starts to become more accurate for the remaining cases. At a threshold of .15, this increases the AUC up to .875 and 92,756 sessions are removed from the evaluation. Sessions for all users are retained (M = 763, MD = 621.5, Std = 684.81). This means that we can observe the improved results for the remaining 48,832 sessions. Youden's index for these sessions is .405.

While this approach loses information in terms of sessions, all 64 users still produced sessions which were represented by the *uncertain* label. Also this approach greatly boosts the accuracy of the remaining sessions once the attached users are considered. Table 6 shows that without an uncertainty range, only one false negative is perceived but an issue can be perceived when 15 false positives come into play. In contrast Table 7 shows 2 false negatives but reduces the false positives (15 to 5) drastically.

Fig. 7 visualises how the sessions removed by the uncertainty range affect considered sessions. Every user produced sessions of each category but either sides of the scale show that users predominantly create sessions with their respective label. Additionally, it shows how sessions that were falsely classified mostly created session of the opposite label.

This demonstrates how excluding *uncertain* sessions can transform assessment of addiction risk. Once users have generated enough data points, this strategy identifies sessions which are too generic to judge a users addiction risk and by removing the range of data including these sessions improves results when trying to evaluate users themselves instead of individual sessions. Additionally, as touched upon in the introduction, this paper is developing a methodology to detect SA through continuous use of a smartphone. When viewing it in context of an actual devices, continued use would enable a steady stream of created sessions which eventually would lead to a risk assessment.

7. Discussion

Our analysis shows that user interaction behaviour (e.g., through taps, scrolls, etc.) as part of usage sessions have predictive power for determining whether smartphone usage was undertaken by an individual scoring high on the SAS or not (Section 4). Crucially, the results show that while high-level, summative features such as session length are also somewhat linked (Section 4.1), the effect sizes are substantially smaller. This adds to the growing body of evidence in the literature [50,51] that user interaction behaviour provides important granularity for applications examining and characterising smartphone usage.

Observations made by Nöe et al. [5] which utilised the same dataset focused on the link between the summed number of events and SA. This was an important step to establish the existence of a relationship with user generated events. To extend on it, we considered how the net of complex behaviour could be limited by generalising every event to carry the same weight. We show that the co-occurrence of these different UI events results in a more representative form of influential features and appears to model the usage behaviour more closely. Additionally, we took the consideration of user interaction behaviour a step further than previous studies by comparing the effects of interaction events in isolation to one another, as well as together, through modelling interaction behaviour using TF-IDF and through regression models. The results highlight stronger effects and model performance where interaction events are considered together (Section 5), which strongly supports our hypothesis, H1.

Furthermore, instead of considering events over the course of a day (or any other arbitrary time window) we find that grouping events into sessions that are bounded by interactions with the device allows to capture more nuances of usage on an individual level. This means that the increase in available information can be used to build a predictive model which takes dozens to hundreds of data points per day into account, rather than just one.

The logistic regression models, particularly those that break down event types into the categories of apps they occurred in, perform reasonably well. However, as the goal is prediction of potential addiction, the priorities of the model may involve minimising false positives or false negatives, rather than both equally. This is further supported by the presence of common types of usage sessions irrespective of addiction scores, particularly where there are few user interaction events in a session (Section 6.1). Fig. 3 shows the improvement we were able to achieve using differing strategies such as pruning the training set and accounting for confidence in classification labels. The models utilising TF-IDF scores beat the performance of models that utilise event counting.

We discovered that an uncertainty range between addicted and non-addicted behaviour caused by the inherent diversity in smartphone behaviour exists. Sessions in this range are difficult to evaluate and including them leads to lowered separability and accuracy. We can evidence partial support for H2 under the condition that discarding uncertain cases is viable (enough data points generated) and the per-session risk is not as important as the user's overall risk.

Lastly, categories did show distinguishable features between themselves. Gaming, social and communication use has been presented in the literature before [3,4,24,30,31] which influenced our expectations set in H3a, H3b and H3c. From our analysis, we reject H3c, entertainment applications being a strong indicator, because of only a single significant feature. Also, other categories (e.g. health and fitness) had surprisingly high coefficients hinting towards discerning features of smartphone addiction and motivates further study. This also reflects prior research that found surprising correlations between some applications categories and addiction (e.g. reading the bible [33]). We argue that H3 as a whole could be supported on the grounds that some but not all sub-hypotheses were supported and the overall observation that different features were found to be important across different app categories. This differed to an extent to the findings by Nöe et al. [5] who found that the relationship of addictive behaviour and UI events can vary for specific applications but was not present for any of these categories in general. We argue that the distinction of category specific event types will have revealed associations that would not have been present by simply summing all event counts (e.g. *scrolling* and *long idle* events in communication being opposing influential factors as seen in Table 5).

8. Conclusions and future work

In this paper, we use the co-occurrence of UI events to highlight their predictive power of addiction in users, which were classified as addicted using the SAS. We are able to use our models to show how users can display a positive, neutral, or negative connection to addiction session-by-session, but when combining all session behaviour the users addiction level matches with our risk prediction. We also show how the same interactions in behaviour can cause different (or even inverse) magnitudes of correlation with addiction across application categories.

In the process we discovered that high level features such as session length are not sufficient to accurately test for addiction in smartphone usage. Results from a previously discovered approach to extract term weightings from user behaviour offered improved risk detection. While single features only showed poor risk prediction capabilities for SA, similarly to high level features, we were able to create models that can assess addiction risk from co-occurrence UI events with more accuracy. The results are improved if the model is trained with sessions with sufficient length (10 or more interactions by a user) or when considering uncertainty in the classification which filters out uncertain sessions.

We were also able to show differences in behaviour between categories that are the highest contributors to detected addicted use. In some categories there are distinctions between behaviour that actively contributed to detecting SA and interactions that were less likely to be taken by addicted users. This supports our assumption that not all use in every app is to be treated equal for detecting addiction and rather that it is required to pay closer attention what a user is actually doing. In respect to previous literature, there were some matching occurrences of use surrounding entertainment and social apps, but the communication category of which app are frequently referenced as an outlet for addicted behaviour there was no notable difference to be observed (there was no statistically significant positive or negative correlation for any feature).

The work presented has shown a considerable relationship between UI events and the ability to predict addicted use. From this point onward there are a few open questions that we would like to see addressed. One key assumption in this work is the identification of in-session behaviour and our models show good predictability based around features gathered from interactions in those. However, it is possible that implementing additional parameters such as session interplay, regularity, or task detection could provide additional useful features to further build on the approach of this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Henry H. Wilmer, Jason M. Chein, Mobile technology habits: patterns of association among device usage, intertemporal preference, impulse control, and reward sensitivity, Psychon. Bull. Rev. (ISSN: 1531-5320) 23 (5) (2016) 1607–1614, http://dx.doi.org/10.3758/s13423-016-1011-z.

[2] Anne-Linda Camerini, Tiziano Gerosa, Laura Marciano, Predicting problematic smartphone use over time in adolescence: A latent class regression analysis of online and offline activities, New Media Soc. (ISSN: 1461-4448) (2020) 1461444820948809, http://dx.doi.org/10.1177/1461444820948809.

[3] Sei Yon Sohn, Philippa Rees, Bethany Wildridge, Nicola J. Kalk, Ben Carter, Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and Young people: A systematic review, meta-analysis and GRADE of the evidence, BMC Psychiatry (ISSN: 1471-244X) 19 (1) (2019) 356, http://dx.doi.org/10.1186/s12888-019-2350-x.

[4] Tao Deng, Shaheen Kanthawala, Jingbo Meng, Wei Peng, Anastasia Kononova, Qi Hao, Qinhao Zhang, Prabu David, Measuring smartphone usage and task switching with log tracking and self-reports, Mob. Media Commun. (ISSN: 2050-1579) 7 (1) (2019) 3–23, http://dx.doi.org/10.1177/2050157918761491.

[5] Beryl Noë, Liam D. Turner, David E.J. Linden, Stuart M. Allen, Bjorn Winkens, Roger M. Whitaker, Identifying indicators of smartphone addiction through user-app interaction, Comput. Hum. Behav. (ISSN: 0747-5632) 99 (2019) 56–65, http://dx.doi.org/10.1016/j.chb.2019.04.023.

[6] Aviv Weinstein, Laura Curtiss Feder, Kenneth Paul Rosenberg, Pinhas Dannon, Chapter 5 — Internet addiction disorder: Overview and controversies, in: Kenneth Paul Rosenberg, Laura Curtiss Feder (Eds.), Behavioral Addictions, Academic Press, San Diego, ISBN: 978-0-12-407724-9, 2014, pp. 99–117, http://dx.doi.org/10.1016/B978-0-12-407724-9.00005-7.

[7] Chakajkla Jesdabodi, Walid Maalej, Understanding usage states on mobile devices, in: Proc. UbiComp '15, ACM, ISBN: 978-1-4503-3574-4, 2015, pp. 1221–1225, http://dx.doi.org/10.1145/2750858.2805837.

[8] Simon L. Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves, Vassilis Kostakos, Revisitation analysis of smartphone app use, in: Proc. UbiComp '15, ACM, ISBN: 978-1-4503-3574-4, 2015, pp. 1197–1208, http://dx.doi.org/10.1145/2750858.2807542.

[9] Min Kwon, Joon-Yeop Lee, Wang-Youn Won, Jae-Woo Park, Jung-Ah Min, Changtae Hahn, Xinyu Gu, Ji-Hye Choi, Dai-Jin Kim, Development and validation of a smartphone addiction scale (SAS), PLOS ONE (ISSN: 1932-6203) 8 (2) (2013) e56936, http://dx.doi.org/10.1371/journal.pone.0056936.

[10] Beryl Noë, Liam D. Turner, David E.J. Linden, Stuart M. Allen, Gregory R. Maio, Roger M. Whitaker, Timing rather than user traits mediates mood sampling on smartphones, BMC Res. Notes (ISSN: 1756-0500) 10 (1) (2017) 481, http://dx.doi.org/10.1186/s13104-017-2808-1.

[11] Niels van Berkel, Denzil Ferreira, Vassilis Kostakos, The experience sampling method on mobile devices, ACM Comput. Surv. (ISSN: 0360-0300) 50 (6) (2017) 93:1–93:40, http://dx.doi.org/10.1145/3123988.

[12] Joon-Myung Kang, Sin-seok Seo, James Won-Ki Hong, Usage pattern analysis of smartphones, in: Proc. APNOMS'11, 2011, pp. 1–8, http://dx.doi.org/10.1109/APNOMS.2011.6077030.

[13] Valentino Servizi, Francisco C. Pereira, Marie K. Anderson, Otto A. Nielsen, Mining user behaviour from smartphone data: A literature review, 2020, p. 1, arXiv:1912.11259 [cs, stat].

[14] Tapio Soikkeli, Juuso Karikoski, Heikki Hammainen, Diversity and end user context in smartphone usage sessions, in: Proc. NGMAST 2011, 2011, pp. 7–12, http://dx.doi.org/10.1109/NGMAST.2011.12.

[15] Ke Huang, Chunhui Zhang, Xiaoxiao Ma, Guanling Chen, Predicting mobile application usage using contextual information, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, in: UbiComp '12, Association for Computing Machinery, New York, NY, USA, ISBN: 978-1-4503-1224-0, 2012, pp. 1059–1065, http://dx.doi.org/10.1145/2370216.2370442.

[16] Hong Cao, Miao Lin, Mining smartphone data for app usage prediction and recommendations: A survey, Pervasive Mob. Comput. (2017) 22.

[17] Abhinav Parate, Matthias Böhmer, David Chu, Deepak Ganesan, Benjamin M. Marlin, Practical prediction and prefetch for faster access to applications on mobile phones, in: Proc. UbiComp '13, ACM, ISBN: 978-1-4503-1770-2, 2013, pp. 275–284, http://dx.doi.org/10.1145/2493432.2493490.

[18] Liam D. Turner, Stuart M. Allen, Roger M. Whitaker, Interruptibility prediction for ubiquitous systems: Conventions and new directions from a growing field, in: Proc. UbiComp '15, ACM, ISBN: 978-1-4503-3574-4, 2015, pp. 801–812, http://dx.doi.org/10.1145/2750858.2807514.

[19] Liam D. Turner, Stuart M. Allen, Roger M. Whitaker, Reachable but not receptive: Enhancing smartphone interruptibility prediction by modelling the extent of user engagement with notifications, Pervasive Mob. Comput. (ISSN: 1574-1192) 40 (2017) 480–494, http://dx.doi.org/10.1016/j.pmcj.2017.01.011.

[20] Juuso Karikoski, Tapio Soikkeli, Contextual usage patterns in smartphone communication services, Pers. Ubiquit. Comput. (ISSN: 1617-4909) 17 (3) (2013) 491–502, http://dx.doi.org/10.1007/s00779-011-0503-0, 1617-4917.

[21] A. Rahmati, C. Shepard, C. Tossell, L. Zhong, P. Kortum, Practical context awareness: measuring and utilizing the context dependency of mobile usage, IEEE Trans. Mob. Comput. (ISSN: 1558-0660) 14 (09) (2012) 1932–1946, http://dx.doi.org/10.1109/TMC.2014.2365199.

[22] Aku Visuri, Zhanna Sarsenbayeva, Jorge Goncalves, Evangelos Karapanos, Simon Jones, Impact of mood changes on application selection, in: Proc. UbiComp '16, ACM, ISBN: 978-1-4503-4462-3, 2016, pp. 535–540, http://dx.doi.org/10.1145/2968219.2968317.

[23] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, Mirco Musolesi, Mytraces: investigating correlation and causation between users&#x2019; emotional states and mobile phone interaction, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3) (2017) 83:1–83:21, http://dx.doi.org/10.1145/3130948.

[24] Sung-Man Bae, The relationship between the type of smartphone use and smartphone dependence of Korean adolescents_National survey study, Child. Youth Serv. Rev. (2017) 5.

[25] Erika Pivetta, Lydia Harkin, Joel Billieux, Eiman Kanjo, Daria J. Kuss, Problematic smartphone use: An empirically validated model, Comput. Hum. Behav. (ISSN: 0747-5632) 100 (2019) 105–117, http://dx.doi.org/10.1016/j.chb.2019.06.013.

[26] You Jin Jeong, Bongwon Suh, Gahgene Gweon, Is smartphone addiction different from Internet addiction? comparison of addiction-risk factors among adolescents, Behav. Inf. Technol. (ISSN: 0144-929X) 39 (5) (2020) 578–593, http://dx.doi.org/10.1080/0144929X.2019.1604805.

[27] Min Kwon, Dai-Jin Kim, Hyun Cho, Soo Yang, The smartphone addiction scale: development and validation of a short version for adolescents, PLOS ONE (ISSN: 1932-6203) 8 (12) (2013) e83558, http://dx.doi.org/10.1371/journal.pone.0083558.

[28] Jon D. Elhai, Jason C. Levine, Robert D. Dvorak, Brian J. Hall, Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use, Comput. Hum. Behav. (ISSN: 0747-5632) 63 (2016) 509–516, http://dx.doi.org/10.1016/j.chb.2016.05.079.

[29] Maya Samaha, Nazir S. Hawi, Relationships among smartphone addiction, stress, academic performance, and satisfaction with life, Comput. Hum. Behav. (ISSN: 0747-5632) 57 (2016) 321–325, http://dx.doi.org/10.1016/j.chb.2015.12.045.

[30] Xiang Ding, Jing Xu, Guanling Chen, Chenren Xu, Beyond smartphone overuse: Identifying addictive mobile apps, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, in: CHI EA '16, Association for Computing Machinery, New York, NY, USA, ISBN: 978-1-4503-4082-3, 2016, pp. 2821–2828, http://dx.doi.org/10.1145/2851581.2892415.

[31] Chun-Hao Liu, Sheng-Hsuan Lin, Yuan-Chien Pan, Yu-Hsuan Lin, Smartphone gaming and frequent use pattern associated with smartphone addiction, Medicine (Baltimore) (ISSN: 0025-7974) 95 (28) (2016) e4068, http://dx.doi.org/10.1097/MD.0000000000004068.

[32] Jon D. Elhai, Jason C. Levine, Robert D. Dvorak, Brian J. Hall, Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use, Comput. Hum. Behav. (ISSN: 0747-5632) 69 (2017) 75–82, http://dx.doi.org/10.1016/j.chb.2016.12.023.

[33] J.A. Roberts, L.H.P. Yaya, C. Manolis, The invisible addiction: Cell-phone activities and addiction among male and female college students, J. Behav. Addict. 3 (4) (2014) 254–265, http://dx.doi.org/10.1556/JBA.3.2014.015.

[34] Björn Friedrichs, Liam D. Turner, Stuart M. Allen, Discovering types of smartphone usage sessions from user-app interactions, in: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops), 2021, pp. 459–464, http://dx.doi.org/10.1109/PerComWorkshops51409.2021.9431034.

[35] Niels van Berkel, Chu Luo, Theodoros Anagnostopoulos, Denzil Ferreira, Jorge Goncalves, Simo Hosio, Vassilis Kostakos, A systematic assessment of smartphone usage gaps, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, San Jose California USA, ISBN: 978-1-4503-3362-7, 2016, pp. 4711–4721, http://dx.doi.org/10.1145/2858036.2858348.

[36] Nikola Banovic, Christina Brant, Jennifer Mankoff, Anind Dey, ProactiveTasks: The short of mobile device use sessions, in: Proc. MobileHCI '14, ACM, ISBN: 978-1-4503-3004-6, 2014, pp. 243–252, http://dx.doi.org/10.1145/2628363.2628380.

[37] Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus, Anind K. Dey, Contextual experience sampling of mobile application micro-usage, in: Proc. MobileHCI '14, ACM, ISBN: 978-1-4503-3004-6, 2014, pp. 91–100, http://dx.doi.org/10.1145/2628363.2628367.

[38] Antti Oulasvirta, Tye Rattenbury, Lingyi Ma, Eeva Raita, Habits make smartphone use more pervasive, Pers. Ubiquitous Comput. (ISSN: 1617-4909) 16 (1) (2012) 105–114, http://dx.doi.org/10.1007/s00779-011-0412-2.

[39] Daniel Hintze, Rainhard D. Findling, Sebastian Scholz, René Mayrhofer, Mobile device usage characteristics: the effect of context and form factor on locked and unlocked usage, in: Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia, ACM, Kaohsiung Taiwan, ISBN: 978-1-4503-3008-4, 2014, pp. 105–114, http://dx.doi.org/10.1145/2684103.2684156.

[40] Gail M. Sullivan, Richard Feinn, Using effect size—or why the P value is not enough, J. Grad. Med. Educ. (ISSN: 1949-8349) 4 (3) (2012) 279–282, http://dx.doi.org/10.4300/JGME-D-12-00156.1.

[41] Francisco Melo, Area under the ROC curve, in: Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, Hiroki Yokota (Eds.), Encyclopedia of Systems Biology, Springer, New York, NY, ISBN: 978-1-4419-9863-7, 2013, pp. 38–39, http://dx.doi.org/10.1007/978-1-4419-9863-7_209.

[42] Assessing the fit of the model, in: Applied Logistic Regression, John Wiley & Sons, Ltd, ISBN: 978-1-118-54838-7, 2013, pp. 153–225, http://dx.doi.org/10.1002/9781118548387.ch5, chapter 5.

[43] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology (ISSN: 0033-8419) 143 (1) (1982) 29–36, http://dx.doi.org/10.1148/radiology.143.1.7063747.

[44] Michael Borenstein, Larry V. Hedges, Julian P.T. Higgins, Hannah R. Rothstein, Converting among effect sizes, in: Introduction to Meta-Analysis, Wiley, Chichester, ISBN: 978-0-470-05724-7, 2009, pp. 45–49.

[45] Julio Sánchez-Meca, Fulgencio Marín-Martínez, Salvador Chacón-Moscoso, Effect-size indices for dichotomized outcomes in meta-analysis, Psychol. Methods (ISSN: 1939-1463) 8 (4) (2003) 448–467, http://dx.doi.org/10.1037/1082-989X.8.4.448, 1082-989X.

[46] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, Deborah Estrin, Diversity in smartphone usage, in: Proc. MobiSys '10, ACM, ISBN: 978-1-60558-985-5, 2010, pp. 179–194, http://dx.doi.org/10.1145/1814433.1814453.

[47] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, Ying Ju, Finding the best classification threshold in imbalanced classification, Big Data Res. (ISSN: 2214-5796) 5 (2016) 2–8, http://dx.doi.org/10.1016/j.bdr.2015.12.001.

[48] W.J. Youden, Index for rating diagnostic tests, Cancer (ISSN: 1097-0142) 3 (1) (1950) 32–35, http://dx.doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[49] Damjan Krstajic, Ljubomir Buturovic, Simon Thomas, David E. Leahy, Binary classification models with "Uncertain" predictions, 2017, p. 1, arXiv:1711.09677 [stat].

[50] Aleksandar Matic, Martin Pielot, Nuria Oliver, Boredom-computer interaction: Boredom proneness and the use of smartphone, in: Proc. UbiComp '15, in: UbiComp '15, ACM, ISBN: 978-1-4503-3574-4, 2015, pp. 837–841, http://dx.doi.org/10.1145/2750858.2807530.

[51] Beryl Noë, Liam D. Turner, Roger M. Whitaker, Smartphone interaction and survey data as predictors of snapchat usage, in: Proc. UbiComp/ISWC '19 Adjunct, ACM, ISBN: 978-1-4503-6869-8, 2019, pp. 438–445, http://dx.doi.org/10.1145/3341162.3349298.