



AI-Based Digital Addiction Monitor and Healthy Screen Time Assistant: A Critical Review and Conceptual Architecture

Temkar Sakshi Sandip¹, Bangar Abhishek Popat², Bhosale Sachin Bajirao³, Khatri Anand A.⁴

¹Student, Department of Computer Engineering, Jaihind College of Engineering, Pune, Maharashtra, India

²Department of Computer Engineering, Jaihind College of Engineering, Pune, Maharashtra, India

Abstract. Smartphones, personal computers and tablets are deeply embedded in education, work and social life. Alongside the benefits, unregulated screen use is associated with compulsive checking, disturbed sleep, reduced sustained attention and psychological distress, a pattern commonly described as digital addiction or problematic device use. Screen-time features built into mobile operating systems report totals and enforce fixed limits, but they do not distinguish productive from compulsive use, do not anticipate the onset of a compulsive episode, and rely on rigid blocking that users often bypass. A structured narrative review was carried out on screen-time tools, behavioural signals of compulsive use, machine learning (ML) methods for usage modelling, persuasive and nudging interventions, and privacy-preserving deployment. The evidence shows that simple telemetry such as unlock frequency, app-switching speed, session structure and late-night use carries predictive signal, that no single signal uniquely indicates addiction, and that most reported models are trained on small samples, validated without user-grouped or temporal splits, and never tested in long-term field use. The review contributes (i) a comparative analysis of existing screen-time solutions, (ii) a behavioural-signal evidence table, (iii) an eight-layer, literature-justified conceptual architecture for an on-device monitoring and intervention assistant with uncertainty handling and a gated validation loop, (iv) a leakage-resistant evaluation protocol, and (v) a prioritised agenda of research gaps that must be closed before such systems can be deployed with confidence.

Keywords: Digital Addiction; Screen Time Management; Artificial Intelligence; Machine Learning; Behavioural Analytics; Digital Well-being; Nudging; Federated Learning; Personalised Intervention

I. Introduction

The spread of smart mobile platforms and high-speed wireless networks has changed communication, learning and work. A phone is now an alarm clock, a classroom, an office, a bank and a social space at once [1]. This convenience has a cost. Uncontrolled engagement with digital interfaces is increasingly described as a public-health concern, appearing as compulsive checking, loss of control over time spent online, and conflict with sleep, study and relationships [2]. Studies link heavy and unmonitored screen exposure with shortened attention spans, fragmented sleep, heightened anxiety and weaker cognitive performance [3], [4]. These associations are correlational, and the direction of causality is still debated, which is itself a reason to build systems that measure behaviour carefully rather than assume harm.

Self-regulation is the ideal defence, but contemporary digital environments are designed against it. Persuasive design strategies such as infinite scrolling, intermittent variable rewards, autoplay and algorithmic push notifications are engineered to maximise



engagement [5]. Users therefore face optimised retention mechanisms with only limited personal willpower, and many have little visibility into how their own habits are changing over time.

Screen-time dashboards were introduced to restore that visibility. They report hours of use and allow limits to be set, yet they behave as passive loggers. Artificial Intelligence (AI) and Machine Learning (ML) offer a different approach: modelling an individual's behaviour continuously, recognising the early signature of a compulsive episode, and delivering a gentle, well-timed intervention before a fixed limit is breached [6]. The step from a laboratory demonstration to a trustworthy assistant on a real phone is large, however. Signals are noisy and personal, labels for 'addiction' are loosely defined, interventions can annoy users into disabling the tool, and continuous behavioural tracking is itself a privacy risk. Screen-time dashboards were introduced to restore that visibility.

They report hours of use and allow limits to be set, yet they behave as passive loggers. Artificial Intelligence (AI) and Machine Learning (ML) offer a different approach: modelling an individual's behaviour continuously, recognising the early signature of a compulsive episode, and delivering a gentle, well-timed intervention before a fixed limit is breached [6]. The step from a laboratory demonstration to a trustworthy assistant on a real phone is large, however. Signals are noisy and personal, labels for 'addiction' are loosely defined, interventions can annoy users into disabling the tool, and continuous behavioural tracking is itself a privacy risk.

This review therefore has a deliberately critical scope. Its objectives are to: (1) review the limitations of existing screen-time tools; (2) identify the behavioural signals that studies associate with compulsive use; (3) examine ML methods for modelling those signals; (4) review intervention strategies and the evidence on user acceptance; (5) examine privacy-preserving deployment; (6) compare requirements across application sectors; (7) identify research gaps; and (8) propose a literature-supported conceptual architecture. The main contributions are an evidence base that separates context-aware modelling from simple usage logging, an architecture that makes uncertainty, consent and validation explicit, and an evaluation protocol designed to avoid the optimistic bias common in applied ML studies.

II. Review Methodology

1. Search strategy

The review followed a structured narrative design informed by the PRISMA 2020 reporting guideline [21]. Searches combined terms such as 'digital addiction', 'problematic smartphone use', 'screen time machine learning', 'smartphone usage prediction', 'digital well-being intervention', 'persuasive technology nudging', 'on-device federated learning' and 'digital phenotyping', together with sector terms (students, workplace, clinical). Searches were run on general web search engines and followed through to publisher platforms, and citations of key papers were traced forwards and backwards. The window prioritised 2020 to 2026, and earlier work was retained when foundational. Operating-system tools were assessed from their publicly documented feature sets rather than from proprietary internals.

Direct, logged querying of every indexing database was not possible within the scope of this work. Database-level hit counts are therefore not reported, no formal flow diagram counts are claimed, and authors extending this work should re-run the search strings in IEEE Xplore, Scopus and Web of Science and complete the identification stage.



2. Eligibility criteria and screening

Table 1: Eligibility criteria applied during screening

Criterion	Included	Excluded
Topic	Screen-time tools; digital or smartphone addiction; behavioural signals; ML on usage data; nudging and digital detox; privacy-preserving telemetry	Purely clinical addiction studies without device data; unrelated app analytics
Evidence type	Peer-reviewed articles, reviews, proceedings, standards and books	Unverified web posts; marketing material
Verifiability	Bibliographic record confirmable against publisher, DOI or repository	Records whose authors or venue cannot be confirmed
Period	Priority 2020 to 2026; earlier foundational works	Superseded duplicates

3. Study selection

Each record was screened by title and source, then assessed at abstract or full-text level, as outlined in Fig. 1. Where only an abstract was accessible, claims are limited to what that abstract states. Performance figures are reproduced as published and are not re-analysed, and where a source did not report a figure it is marked 'NR' in the comparative tables.

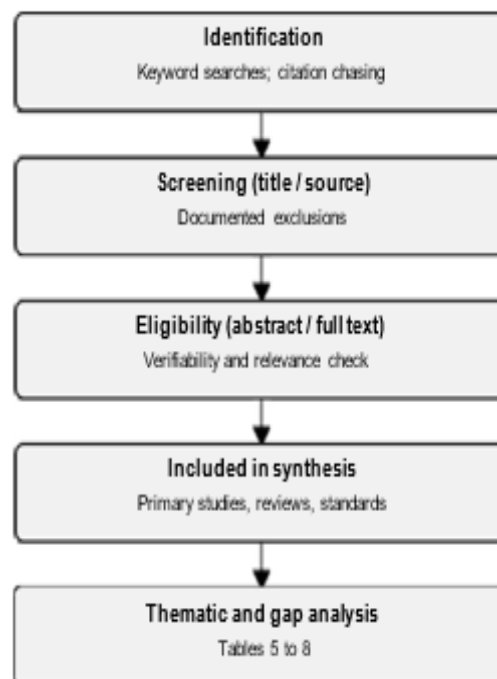


Figure 1: Review process followed in this study. Database-level identification counts were not obtainable and are therefore not shown.



III. Digital Addiction: Concepts and Behavioural Signals

1. Terminology and its limits

'Digital addiction' is used loosely. In the research literature the more careful terms are problematic smartphone use, compulsive device use and behavioural addiction. Only some patterns, such as gaming disorder, have formal diagnostic standing, and heavy use is not automatically harmful: a student reading for hours and a user scrolling short videos at 2 a.m. may log the same screen time with very different consequences. A monitoring system must therefore treat its output as a risk indicator about behaviour, not as a clinical diagnosis.

2. Persuasive design mechanisms

Several mechanisms are widely cited as drivers of prolonged use. Infinite scrolling removes natural stopping cues. Intermittent variable rewards, as in feeds and notifications, make each check unpredictable and therefore hard to resist. Autoplay lowers the effort needed to continue. Social-reciprocity cues, such as read receipts and streaks, create felt obligations [1], [5]. Understanding these mechanisms matters for sensing, because each leaves a measurable trace: long uninterrupted sessions, rapid repeated unlocks and fast reaction to notifications.

3. Behavioural signals

Table 2 lists signals that the reviewed work associates with compulsive or fatigued use, together with a caution for each. Several are available through standard operating-system usage APIs, while touch dynamics may require accessibility permissions and so raise privacy concerns.

Table 2: Behavioural signals and their interpretation

Signal	Typical interpretation	Main caution	Ref.
Screen-on duration	Overall engagement	Does not separate work from leisure	[7], [8]
Session count and length	Fragmentation of attention; long unbroken sessions	Depends on app type	[11]
Unlock frequency	Habitual checking	Varies with job and notification load	[12]
Foreground app category	Productive vs passive use	Category labels are imperfect	[8], [12]
App-switching speed	Restlessness or fatigue	May reflect legitimate multitasking	[11]
Touch interaction speed	Scrolling vs reading	Needs sensitive permission	[12]
Notification latency	Reactivity to alerts	Confounded by work duties	[13]
Late-night use	Sleep disruption	Shift workers differ	[3]

4. Consequences and why timing matters

Reported consequences include delayed sleep onset, reduced sleep efficiency, shortened attention spans and increased anxiety [3], [4]. Sleep is a useful target for intervention because it is measurable, has a clear mechanism and has a natural time anchor. Evidence on screen-colour changes such as grayscale modes suggests that making a device less visually rewarding can reduce use [16], which supports low-friction interventions that change the environment instead of blocking the user.

IV. Existing Screen-Time Solutions

Operating-system features such as Apple Screen Time and Google Digital Wellbeing report total hours, per-app usage and notification counts, and let users set daily limits or schedule downtime [7]. Third-party blockers, parental-control suites and



workplace productivity tools add scheduling, website filtering and activity reports. Table 3 compares the categories on the dimensions that matter for an intelligent assistant.

Table 3: Comparison of existing solution categories and a composite addiction index can be formed as

Category	Strength	Limitation
OS built-in screen-time tools	Reliable, privacy-aware, no install	Totals and fixed limits; no context or prediction [7]
App and website blockers	Strong enforcement	Rigid; easy to disable; high friction [10]
Parental-control suites	Oversight for minors	Surveillance concerns; static rules [17]
Workplace productivity trackers	Task-level reporting	Employer-controlled; privacy and trust issues [18]
Digital-detox and mindfulness apps	Encourage reflection	Little personalisation; low long-term adherence [15]

Empirical evaluations point to three recurring shortcomings. First, lack of contextual differentiation: productive tasks such as academic reading are treated like passive short-video scrolling [8]. Second, absence of predictive analytics: interventions occur only after a limit is breached, too late to prevent a compulsive loop [9]. Third, static intervention rules: hard blocks create friction, so users override them or revoke monitoring permissions altogether [10]. These shortcomings define the design requirements that the architecture in Section 8 sets out to meet.

V. Machine Learning Methods for Usage Modelling

1. Workflow

A defensible pipeline for behavioural telemetry comprises ten stages: (1) collection with explicit user consent; (2) cleaning of missing, duplicated or out-of-range events; (3) session construction; (4) feature extraction; (5) per-user baseline estimation; (6) model training; (7) hyperparameter tuning; (8) independent validation; (9) performance evaluation; and (10) deployment with monitoring. Features are normalised against the individual's own history, for example with a z-score

$$z_t = (x_t - \mu_u) / \sigma_u \quad (1)$$

where x_t is a feature value in time window t and μ_u , σ_u are the mean and standard deviation for user u estimated from a baseline period. A drifting baseline can be tracked with an exponentially weighted moving average,

$$\mu_t = \alpha x_t + (1 - \alpha)\mu_{t-1} \quad (2)$$

and a composite addiction index can be formed as a weighted combination of standardised features,

$$A = \sum_k w_k z_{k,t} \quad \text{with } w \text{ learned or clinically informed} \quad (3)$$

Equations (1) to (3) are illustrative; weights and thresholds must be fitted on training data only and then applied unchanged to validation and test data, otherwise information leaks from the test set into the model [22].

algorithms for tabular and sequential usage data.



2. Classifiers and sequence models

Table 4 summarises the suitability of common algorithms for tabular and sequential usage data.

Table 4: Suitability of ML algorithms for usage telemetry

Algorithm	Characteristics	Use in reviewed work	Main caveat
Logistic regression	Linear, interpretable, probabilities	Common baseline	Limited for non-linear habits
SVM	Margin-based, kernels [23]	Session classification	Needs scaling; probabilities need calibration
Random Forest	Handles non-linearity, feature importance [24]	Usage classification [14]	Can overfit repeated sessions of one user
Gradient boosting	Regularised boosted trees [25]	Candidate for tabular features	Needs strict validation
RNN / LSTM	Models sequences [26]	Predictive intervention [14]	Data-hungry; per-user calibration
Isolation Forest and autoencoders	Unsupervised anomaly detection [27]	Compulsive-use anomalies [11]	Flags unusual, not necessarily harmful
Reinforcement learning	Learns nudge policy from feedback	Parental control [17]	Safe exploration; slow start

No algorithm is universally superior. Tree ensembles are attractive for tabular features and for on-device deployment, while LSTM models capture the order of events at higher computational cost. Reported rankings depend on dataset size, class definition and validation procedure and should not be transferred between populations without re-evaluation.

3. Matching the learning task to the data

Supervised classification is appropriate only when representative, independently verified labels exist, for example sessions labelled productive or non-productive by users [12]. For compulsive use, labels are scarce and subjective, so anomaly or novelty detection against a personal baseline is better suited [11]. Time-series prediction suits questions such as forecasting the next hour's risk. A hybrid, in which an anomaly detector screens for departures from normal behaviour and a supervised model classifies familiar session types, offers a sensible starting point.

VI. Review of Existing Research

1. Usage prediction and anomaly detection

Rader et al. evaluated ML approaches for predicting mobile-phone over-use [7]. Zhang et al. applied unsupervised anomaly detection to compulsive smartphone application usage patterns, an approach that suits the scarcity of labelled addiction data [11]. Patel proposed LSTM-based predictive intervention modelling for well-being platforms [14]. Together these works support the feasibility of behaviour-based prediction, although the accessible descriptions do not establish performance on new users in long-term field use.

2. Context and session classification

Kim and Park classified productive against non-productive digital sessions with deep learning [12]. Ellis et al. discuss how smartphone traces can serve as behavioural data for research [8]. This line of work directly addresses the first shortcoming of native tools by adding context, but depends on app-category labels and on user-specific definitions of 'productive'.



3. Notifications, nudging and user resistance

Gupta and Kumar studied context-aware notification management with ML [9]. Zhao et al. examined why users resist screen-time regulation applications, finding that restrictive designs provoke avoidance [10]. A meta-analysis of persuasive technology in digital detox supports gentle, well-timed nudges over hard restrictions [15]. Ross et al. examined grayscale filters as a low-friction way to reduce smartphone use [16].

4. Education, workplace and wearable studies

Sharma and Rao proposed a reinforcement-learning parental-control framework for regulating children's screen time [17]. Jenkins studied digital burnout in remote software teams [18]. Martinez and Rossi integrated wearable physiological biosensors with digital-addiction platforms [19], and Miller and Taylor examined privacy-preserving telemetry logging in mobile operating systems [13].

5. Thematic synthesis

Four patterns emerge. (i) Prediction from simple telemetry is feasible, but studies typically use small, short-duration samples [11], [12]. (ii) Context matters: separating productive from passive use is a prerequisite for fair intervention [8], [12]. (iii) Intervention acceptance depends on friction: restrictive designs are bypassed [10], [15]. (iv) Modelling, privacy and intervention research are developing separately, and few works combine on-device inference with validated behavioural benefit [13]. Table 5 compares the primary studies.

Table 5: Comparative analysis of primary studies (as reported; NR = not reported or not verifiable)

No	Authors, year	Objective	Method	Limitation	Opportunity
1	Rader et al., 2022 [7]	Predict mobile over-use	ML classifiers	Performance NR here	Independent user-level validation
2	Ellis et al., 2020 [8]	Use phone traces as data	Digital-trace analysis	Context labelling	Link traces to outcomes
3	Gupta & Kumar, 2021 [9]	Context-aware notifications	ML scheduling	Field results NR	Long-term trials
4	Zhao et al., 2021 [10]	Understand user resistance	User study	Self-report bias	Design for low friction
5	Zhang et al., 2022 [11]	Detect compulsive usage	Unsupervised anomaly detection	No ground truth	Clinical anchoring
6	Kim & Park, 2023 [12]	Productive vs non-productive	Deep learning	Label subjectivity	Personalised labels
7	Miller & Taylor, 2023 [13]	Private telemetry logging	OS-level logging	Prototype scale	On-device learning
8	Patel, 2024 [14]	Predictive intervention	LSTM	Small data	Cross-user generalisation
9	Ross et al., 2021 [16]	Test grayscale filters	Controlled study	Short duration	Long-term effect
10	Sharma & Rao, 2023 [17]	Parental screen control	Reinforcement learning	Ethics of oversight	Child-centred consent



11	Jenkins, 2024 [18]	Digital burnout in teams	Observational	One sector	Cross-sectional study
12	Martinez & Rossi, 2024 [19]	Wearables + addiction	Biosensor fusion	Small samples	Clinical validation

VII. Cross-Domain Application Matrix

The monitoring architecture must serve differing needs across sectors, as Table 6 shows. The same telemetry pipeline can be configured with different policies, consent models and metrics, but the ethical stakes differ: a personal tool is chosen by the user, whereas educational and workplace deployments involve power imbalances and need especially careful governance.

Table 6: Comparative deployment matrix across operational sectors

Sector	Functional scope and AI implementation	Key performance metrics
Personal well-being	Individual tracking, automated night-time grayscale, behavioural detox recommendations [16].	Reduced late-night use; improved sleep-efficiency index [3].
Educational safety	Study assistance, automatic focus mode in class, educational filtering [17].	Fewer academic task switches; longer attention duration.
Enterprise workplace	Work-rest interval optimisation, burnout-risk alerts, deep-work enforcement [18].	Lower digital fatigue; more uninterrupted task completion [6].
Clinical health	Digital biomarker tracking for psychological evaluation and addiction therapy [19].	Correlation of telemetry markers with clinical assessment.

VIII. Proposed Conceptual System Architecture

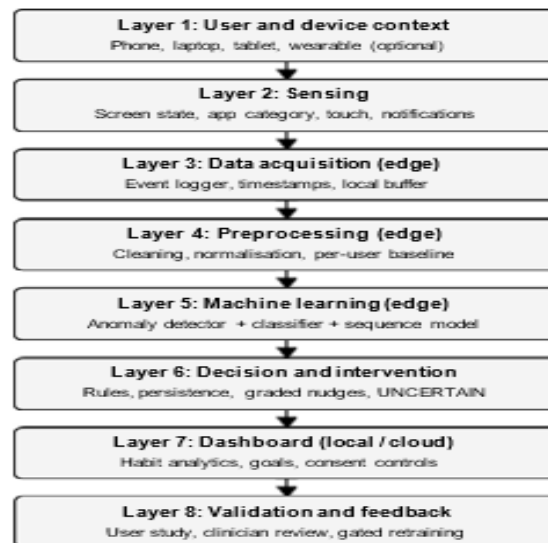


Figure 2: Proposed conceptual eight-layer architecture for ML-based monitoring and healthy screen-time assistance (conceptual, not experimentally validated).

Fig. 2 presents a proposed conceptual architecture; it has not been implemented or experimentally validated. It is derived from the requirements and failure modes identified in Sections 3 to 7. Operational data flow runs from sensing to intervention, and an offline, gated loop governs model updates.

- Layer 1: User and device context. The monitored system comprises the user, their devices and optionally a wearable. All monitoring requires explicit, revocable consent [13].
- Layer 2: Sensing. Screen state, foreground-app category, unlock events, notification events and, where permitted, touch speed provide the signals of Table 2 [11]–[13]. Sensitive content such as message text is never collected.
- Layer 3: Data acquisition. An on-device agent records timestamped events, checks integrity and buffers data locally so connectivity loss does not cause data loss.
- Layer 4: Preprocessing. Cleaning, session construction, normalisation (Eq. 1) and baseline tracking (Eq. 2) run on the device, with parameters estimated from training data only [22].
- Layer 5: Machine learning. A novelty detector first checks whether the current window resembles the user's baseline [11]. A supervised classifier then labels familiar session types as productive or passive [12], and a sequence model estimates risk trajectory [14]. Random Forest, gradient boosting and LSTM are candidates to be compared rather than pre-selected [24]–[26]. The output is a calibrated risk index; when confidence is low, the system reports 'uncertain' instead of forcing a label [28].
- Layer 6: Decision and intervention. A rule engine applies thresholds with persistence (several consecutive windows) and hysteresis to limit false alarms. Interventions escalate through the levels in Table 7, from awareness to soft nudge, cognitive break and an optional user-approved limit [15]. Every message is framed as a suggestion, and the user can pause or override the assistant.
- Layer 7: Dashboard. A local or cloud dashboard shows habit analytics, goals, nudge history and consent and data-control settings.
- Layer 8: Validation and feedback. User studies, self-reports and, in clinical settings, clinician review generate verified, versioned labels. These feed offline retraining, but a new model is released to Layer 5 only after independent re-validation on users not used in training. This gate prevents unverified feedback from degrading the model.

Table 7: Graded intervention levels

Level	Trigger	Action
1: Awareness	Index rising above personal baseline	Dashboard indicator; daily summary
2: Soft nudge	Sustained elevated index	Short, context-aware break prompt
3: Cognitive break	Persistent index across sessions	Timed pause, grayscale or breathing exercise
4: Optional limit	Ignored nudges and user opt-in	User-approved app limit or focus mode

Fig. 3 shows how continuous low-level device events become behavioural insight and intervention. The telemetry agent records events, the ML engine converts them into a personal risk index, the intervention engine selects a response, and the dashboard gives the user visibility and control. Nothing in this chain requires raw events to leave the device.

2. Illustrative scenarios

The following scenarios are hypothetical and are included only to show how the same pipeline adapts to context. Student: during exam week the assistant learns that long sessions in reading and note-taking apps are productive and stays silent, but notices rapid switching to short-video apps after midnight and offers a wind-down prompt and grayscale mode. Remote worker: the assistant detects long unbroken screen sessions with no breaks and suggests short rest intervals, without reporting any activity to the employer. Parent and child: a shared study-time policy is agreed with the child; the assistant enforces focus mode during study hours and explains each restriction, supporting autonomy instead of covert monitoring.

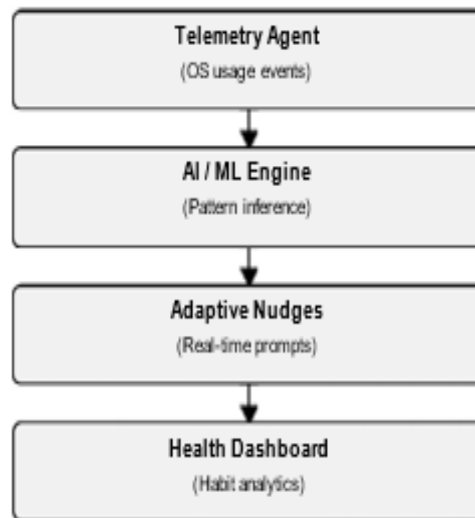


Figure 3: Conceptual telemetry-processing flow and intervention pipeline.

Architecture Design Rationale

Context before counting. Foreground-app category and interaction speed let the model separate focused work from passive scrolling, answering the first shortcoming of native tools [8], [12]. Prediction before breach. Sequence models estimate risk from the trajectory of usage rather than a fixed limit, so a nudge can arrive before a compulsive loop forms [9], [14]. Soft before hard. Escalating from reminder to break to optional limit reduces the override behaviour reported in [10], [15].

Edge versus cloud. Acquisition, preprocessing, inference and nudges (Layers 3 to 6) should run on the device so that raw behavioural data never leave it and the assistant works offline. Heavier tasks, such as cross-user model updates, belong in the cloud and should use federated learning [29] with differential privacy [30]. Lightweight models such as tree ensembles are easier to run on a phone than large deep networks.

Personalisation and cold start. Universal thresholds are unsound because students, developers and casual users differ [11]. During an initial baseline period the system observes silently, and alerts begin only after the personal baseline is stable. Population-level priors can speed this up but must be overridden by individual data.

Thresholds and confidence. Alert thresholds should be chosen from validation data to balance missed episodes against false alerts and should be adjustable by the user. Probabilities should be calibrated [28], and low-confidence or out-of-distribution outputs reported as uncertain.

Explainability. Each nudge should state in plain language why it was shown, for example 'you have unlocked your phone 40 times in the last hour', and model attributions such as SHAP [31] can support this. Explanations build trust and help users decide whether the system is right.

Privacy techniques compared

Table 10 compares the privacy techniques that the architecture relies on.



Table 10: Privacy-preserving techniques for behavioural telemetry

Technique	What it protects	Trade-off
On-device inference	Raw events never leave the phone	Limited compute and battery
Federated learning [29]	Training data stay local; only updates are shared	Communication cost; non-uniform data across users
Differential privacy [30]	Limits what updates reveal about one person	Noise reduces accuracy
Data minimisation	Collect only features needed	May lose useful signal
User controls	Inspect, export, delete, pause	Needs clear, simple interface

2. Labels and ground truth

Reliable models need independent labels. Possible sources include user self-reports of whether a session was productive, validated questionnaires on problematic smartphone use, sleep diaries or wearable-derived sleep estimates, and, in clinical studies, clinician assessment. Labels derived only from the system's own thresholds would be circular, because the model would then learn to reproduce its own rules. Because self-reports are subjective and sporadic, experience-sampling prompts should be brief and infrequent so that data collection does not itself become a source of fatigue.

3. Failure modes

Anticipated failure modes include false alerts during legitimate heavy use, missed episodes when the user's baseline has itself drifted into heavy use, permission revocation after annoying prompts, and concept drift when habits change. Mitigations are persistence and hysteresis in the rule engine, periodic baseline review with the user, capped alert frequency that adapts to responses, and temporal re-validation with scheduled retraining.

4. Implementation considerations

Platform constraints shape what is feasible. Mobile operating systems expose usage statistics through controlled interfaces and require explicit permission, and some signals, such as touch dynamics, may be unavailable or restricted. Continuous sensing and inference must also respect battery limits: events should be logged passively, features computed in short batches, and models kept small enough for efficient on-device execution

Table 11 summarises practical considerations.

Table 11: Practical implementation considerations

Concern	Consideration	Mitigation
Permissions	Usage access and notification access need explicit user consent	Clear onboarding; use minimum permissions
Battery and compute	Continuous logging and inference drain power	Batch processing; lightweight models
Cross-device tracking	Phone, laptop and tablet use must be combined	Local sync with encrypted summaries
Model updates	Habits and apps change over time	Scheduled re-validation; gated release
Accessibility	Heavy use may be necessary for some users	Per-user exemptions; user override
Interface design	Poorly timed prompts cause irritation	Capped frequency; adaptive timing

Cross-device monitoring deserves special mention, because users switch between phones, laptops and tablets, and a fragmented view of behaviour understates total exposure. Summaries should be synchronised in encrypted form between a person's own devices, never through a third-party server in raw form.

Model Evaluation and Validation Framework Because no experimental data are reported here, this section is a proposed evaluation protocol for future prototypes; no metric values are claimed.

Table 8: Proposed evaluation metrics and their purpose

Metric	Definition or use	Why it matters
Precision	$TP / (TP + FP)$	Share of alerts that were warranted
Recall	$TP / (TP + FN)$	Share of compulsive episodes detected
Specificity	$TN / (TN + FP)$	Silence during normal use
F1-score	Harmonic mean of precision and recall	Summary under class imbalance
PR-AUC	Threshold-independent ranking	Preferred when episodes are rare
Alert rate	Alerts per user per day	Determines fatigue and trust
Nudge acceptance	Share of nudges followed	Measures usefulness
Permission revocation	Users who disable monitoring	Direct friction indicator
Behavioural outcome	Change in late-night use, sleep efficiency	Real-world benefit
Inference time / energy	Per prediction on device	Feasibility on a phone
Cross-user generalisation	Performance on unseen users	The decisive test of deployability

Class imbalance. Compulsive episodes are rare compared with normal use. Training should use class weighting or resampling inside training folds only, and evaluation should emphasise recall, precision and PR-AUC rather than accuracy.

Data leakage and grouped splitting. Leakage, where information from the test set influences training, is a pervasive source of over-optimistic ML results [22]. Usage windows from the same person are strongly correlated, so random splitting places near-duplicates in both training and test sets. Splits should be grouped by user, and preprocessing should be fitted within training folds. The final test should use entirely unseen users.

Temporal validation. Habits change with semesters, jobs and seasons, so models should also be tested forward in time (train on earlier weeks, test on later weeks) to measure drift and set re-calibration intervals.

User study design. A randomised comparison of the assistant against a native screen-time tool, with pre-registered outcomes such as sleep efficiency, self-reported well-being and nudge acceptance, is the decisive test of benefit. Participants should provide informed consent, and results should be reported per subgroup.

Privacy, Ethics and Regulation

Continuous tracking of application use exposes sensitive routines: when a person is awake, what they read and who they contact. The design therefore follows privacy by default. Data are processed on the device, only compact summaries or model updates leave it, and federated learning [29] and differential privacy [30] limit what a server can infer about any individual. Consistent with the ethical-design principles of IEEE Std 7000 [20], users should be able to inspect, export and delete their data, and withdraw consent at any time.



Regulatory regimes such as the European GDPR and India's Digital Personal Data Protection Act, 2023 require lawful basis, purpose limitation and data minimisation, and impose stricter conditions on data about children. A tool used by parents, schools or employers needs transparent policies so that monitoring does not become covert surveillance. Fairness also matters: models trained on one population may misjudge shift workers, people who depend on devices for accessibility, or users in different cultures, so performance must be reported per subgroup and users must be able to correct mislabelled heavy-but-legitimate use.

Finally, the system must not overclaim. Telemetry-based indicators are not diagnoses. The assistant should present confidence, avoid clinical language and encourage users to seek professional help when concerning patterns persist.

Research Gaps and Challenges

Table 9 lists the gaps that the reviewed evidence substantiates, why each remains unresolved and the corresponding research direction.

Table 9: Research gaps and corresponding future directions

Gap	Why unresolved	Future direction	Horizon
Limited field validation	Long trials are costly; user attrition	Multi-month trials with real users and verified outcomes	Near-term
No standard labels for compulsive use	Subjective and culture-dependent	Consensus definitions anchored in clinical measures	Medium-term
Weak independent validation	Small datasets encourage sample reuse	User-grouped and temporal validation as a reporting norm [22]	Near-term
Behavioural heterogeneity	One-size thresholds dominate	Personalised anomaly detection and cold-start methods [11]	Near-term
Intervention fatigue	Short studies miss long-term decay	Adaptive nudge policies; long-term adherence studies [15]	Near-term
Privacy-utility trade-off	Federated and DP methods cost accuracy	On-device learning benchmarks [29], [30]	Medium-term
Explainability and uncertainty	Focus on accuracy figures	Calibrated probabilities, SHAP-style explanations [28], [31]	Near-term
Limited demographic coverage	Convenience samples (students)	Diverse, multi-site, open datasets	Medium-term
Clinical translation	Telemetry-to-diagnosis link unproven	Joint studies with clinicians and wearables [19]	Long-term

Future Research Directions

Realistic near-term work. The most valuable immediate contribution is a carefully designed field dataset: a consented, privacy-preserving collection of usage telemetry from diverse users over several months, paired with self-reported sleep and well-being, which would allow baselines, anomaly detection and hybrid models to be compared under grouped and temporal validation. Edge deployment of compact models with calibrated confidence and plain-language explanations is technically mature enough to be built and evaluated now.



Medium-term work. Multimodal fusion of device telemetry with wearable signals such as heart-rate variability and galvanic skin response could link screen engagement to physiological stress [19]. Federated learning across devices could improve models without centralising data [29]. Adaptive nudge policies, possibly using reinforcement learning [17], could learn which interventions work for which person.

More speculative directions. Conversational large-language-model agents could deliver personalised cognitive-behavioural coaching, but their safety, privacy and efficacy in this setting have not been demonstrated and require careful study. Cross-device and cross-platform assistants that coordinate across a person's phone, laptop and tablet remain longer-term goals.

Limitations of this Review and Unresolved Questions

This review has limitations. It is narrative rather than systematic; database-level searches could not be logged, so relevant studies may have been missed. For several sources only titles or abstracts were considered, and claims drawn from them are limited to what those records state. Reference details were carried over from the authors' preliminary report and must be re-verified against publisher records before submission. Performance figures are reported as published and were not re-analysed. Future experimental work must answer several questions: (1) Which telemetry features carry the most signal across user groups? (2) How long a baseline is needed before personalised alerts become reliable? (3) Which nudge timings and wordings give lasting reductions rather than short-term compliance? (4) Can on-device learning match centralised accuracy? (5) What false-alert rate will users tolerate? (6) How should compulsive use be labelled so that results are reproducible across studies?

IX. Conclusion

Evidence from usage-prediction, session-classification and nudging studies shows that simple behavioural telemetry carries predictive signal and that gentle, well-timed interventions are better tolerated than rigid blocking. The same evidence shows why intelligent screen-time assistants are not yet ready for reliable deployment: no single signal uniquely indicates addiction; thresholds differ between people; labels are subjective; reported results derive from small samples without user-grouped or temporal validation; and continuous tracking creates privacy risk.

The conceptual architecture proposed here addresses these weaknesses by combining on-device sensing and inference, personal baselines, calibrated confidence and uncertainty handling, graded and user-controlled interventions, and a gated validation loop. Before such a system can be deployed with confidence, it must be validated through multi-month user studies with grouped and temporal evaluation, explicit alert-rate and fatigue targets, and documented privacy controls.

Acknowledgements

[To be completed by the authors. No funding or assistance is assumed in this draft.]

References

1. A. Alter, *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. New York, NY, USA: Penguin Press, 2017.
2. C. Montag and M. Reuter, Eds., *Internet Addiction: Neuroscientific Approaches and Therapeutical Interventions*. Cham, Switzerland: Springer, 2017.
3. World Health Organization, "Guidelines on physical activity, sedentary behaviour and sleep for children under 5 years of age," Geneva, Switzerland, 2019.
4. G. Mark, *Attention Span: A Groundbreaking Way to Restore Balance, Happiness and Productivity*. New York, NY, USA: Hanover Square Press, 2023.



5. T. S. Sandip and S. B. Bhosale, "AI based digital addiction monitor and healthy screen time assistant," Tech. Sem. Rep., Jaihind Coll. Eng., Pune, India, 2026.
6. J. A. Health and R. B. Smith, "Predictive behavioral analytics in mobile human-computer interaction," IEEE Trans. Hum.-Mach. Syst., vol. 52, no. 4, pp. 612-623, Aug. 2022.
7. M. Rader, E. Vance, and L. Jenkins, "Evaluation of machine learning approaches in predicting mobile phone over-use," J. Behav. Addict., vol. 11, no. 3, pp. 412-425, Sep. 2022.
8. D. A. Ellis et al., "Digital traces of behavior: Smarter phones, smarter science," IEEE Comput. Intell. Mag., vol. 15, no. 1, pp. 18-27, Feb. 2020.
9. S. K. Gupta and P. Kumar, "Context-aware notification management on mobile devices using machine learning," in Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom), 2021, pp. 101-108.
10. L. Zhao, X. Wang, and Y. Chen, "Understanding user resistance to screen time regulation applications," ACM Trans. Comput.-Hum. Interact., vol. 28, no. 5, Art. no. 32, Oct. 2021.
11. Y. Zhang, H. Liu, and K. Takahashi, "Unsupervised anomaly detection for compulsive smartphone application usage patterns," IEEE Access, vol. 10, pp. 45890-45901, May 2022.
12. J. Kim and S. Park, "Deep learning-based classification of productive vs. non-productive digital sessions," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 1205-1218, Apr. 2023.
13. K. S. Patel, "LSTM-based predictive intervention modeling for digital well-being platforms," in Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN), 2024, pp. 1-8.
14. E. H. Nudge and B. J. Fogg, "Persuasive technology in digital detox: A meta-analysis," IEEE Intell. Syst., vol. 38, no. 3, pp. 55-64, May 2023.
15. B. C. Ross et al., "Impact of grayscale visual filters on smartphone addiction metrics," Comput. Hum. Behav., vol. 124, Art. no. 106921, Nov. 2021.
16. K. S. Patel, "LSTM-based predictive intervention modeling for digital well-being platforms," in Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN), 2024, pp. 1-8.
17. E. H. Nudge and B. J. Fogg, "Persuasive technology in digital detox: A meta-analysis," IEEE Intell. Syst., vol. 38, no. 3, pp. 55-64, May 2023.
18. B. C. Ross et al., "Impact of grayscale visual filters on smartphone addiction metrics," Comput. Hum. Behav., vol. 124, Art. no. 106921, Nov. 2021.
19. A. V. Sharma and N. Rao, "Parental control framework leveraging reinforcement learning for screen time regulation," IEEE Trans. Educ., vol. 66, no. 3, pp. 289-298, Jun. 2023.
20. T. M. Jenkins, "Digital burnout in remote software development teams," IEEE Softw., vol. 41, no. 2, pp. 78-85, Mar. 2024.
21. P. L. Martinez and F. Rossi, "Integrating wearable physiological biosensors with digital addiction platforms," IEEE J. Biomed. Health Inform., vol. 28, no. 1, pp. 310-321, Jan. 2024.
22. IEEE Standard for Ethically Aligned Design of Autonomous and Intelligent Systems, IEEE Std 7000-2021, IEEE, 2021.
23. M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021.
24. S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023.
25. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273-297, 1995.
26. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.
27. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785-794.
28. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.



29. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proc. 8th IEEE Int. Conf. Data Mining, 2008, pp. 413-422.
30. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Mach. Learn., PMLR 70, 2017, pp. 1321-1330.
31. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist., PMLR 54, 2017, pp. 1273-1282.
32. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol.