Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Technical Countermeasures for Deceptive user Interfaces: A Comprehensive Detection and Mitigation Framework

Nial Rojan
Department of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering
(Fr.CRCE) Bandra (W), Mumbai, India

Dr. Smita Sanjay Ambarkar
Associate Professor, Dept. of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering
(Fr.CRCE) Bandra (W), Mumbai, India

Prof. Sangeeta Parshionikar
Department of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering (Fr.CRCE)
Bandra (W), Mumbai, India

***Abstract*—Dark patterns are the deceptive interface designs that manipulate users into harmful actions. Even though regulators now impose very large fines, t e chnical d e fenses still remain limited: recent studies show a 54.5% coverage gap, with only 31 of 68 known dark-pattern types being detected by the tools available. This is due to three main reasons. The datasets being offered in public are small and narrow, typically a few thousand examples covering at most 15 to 20 pattern types. Even though detectors analyze static screenshots or isolated text, they are not acquainted with multi-step flows l i ke R o ach Motel cancellations and hidden subscriptions. Meanwhile, advanced Al offers more personalized deception and weakens defenses through adversarial attacks on NLP and vision models. This work surveys the existing text-based, visual, multimodal, and conversational detection methods, comparing approaches, datasets, metrics, and limitations. Building on this analysis, it proposes a four-engine framework that integrates multiple variable like DOM, visual, linguistic, and behavioral signals, and outlines a 10,000-example, multi-platform dataset aligned with a 245-pattern ontology. Finally, it sketches privacy-preserving and adversarially robust deployment strategies for real-time protection.**

***Index Terms*—dark patterns, deceptive interfaces, machine learning, user journey tracking, adversarial robustness, privacy-preserving countermeasures**

I. Introduction

Dark patterns are interface designs that steer or coerce users into actions they would not take under fully informed, independent choice. Hidden fees, subscription traps, data over-sharing, and loss of trust, and have triggered substantial regulatory fines worldwide are some examples of dark patterns [11], [12]. With digital services becoming more complex and personalized, there is a growing need for technical systems that can detect and mitigate deceptive design on a scale.

Existing research has made significant progress in the same. Extensive searches of shopping sites reveal how widespread dark patterns are [1], while the UX taxonomies and ontologies now describe hundreds of variants of patterns [10]. Machine-learning approaches are detecting dark patterns from UI text [4], screenshots [19], or combined visual-text features [20], and conversational benchmarks probe large language models for manipulative behaviors [22]. Multimodal deception work shows that the integration of behavioral signals can improve detection in controlled settings [9], [34].

However, these efforts still remain fragmented and incomplete. After evaluating against a 68-type taxonomy, current tools detect less than 50 per cent of the patterns that are known, particularly missing many multi-step behaviors like cancellation friction, hidden costs that are revealed late, and subscription escalation [6], [21]. This is because Public datasets are small, domain-specific, and mostly static; very few datasets capture full user journeys or interaction sequences [20], [35]. At the same time, invaders increasingly exploit powerful language-vision models [7], and minute adversarial changes to text or pixels can fool many detectors [31], [32].

This paper addresses these gaps by, first, consolidating the current landscape of dark-pattern detection methods, datasets, and metrics into a structured comparison. Second, it motivates a four-engine architecture that jointly analyzes code structure, visual layout, language, and user behavior over time, designed to capture both static and dynamic patterns across web, mobile, and conversational interfaces. Thirdly, it outlines a large-scale, multi-platform dataset aligned with a rich ontology, and discusses how such a system can be deployed in a privacy-preserving and adversarially robust way.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

II. Literature Survey

The landscape of existing dark-pattern detection methodologies is analyzed across 18 major works. Tables I and II categorize these approaches by their technical engines, metrics, and the inherent limitations that this current research aims to resolve.

Analysis of these works reveals a significant reliance on static data and text-based classification, which very so often fails to capture the dynamic nature of user journeys. While the visual and multimodal approaches like AidUI and ContextDP show promising results, they remain constrained by the single-screen analysis and proprietary datasets. The literature highlights a critical need for frameworks that integrate behavioral signals and maintain robustness against adversarial AI, as current tools leave over 50% of known pattern types undetected.

III. Proposed Solution

Solutions which are currently present suffer from four main limitations: limited coverage of pattern types, small and narrow datasets, predominantly static analysis, and vulnerability to AI-enabled attacks and evasion. To address these issues, this work proposes a unified, multimodal framework for dark-pattern detection built around four complementary engines, supported by a large-scale dataset and privacy-preserving deployment strategy.

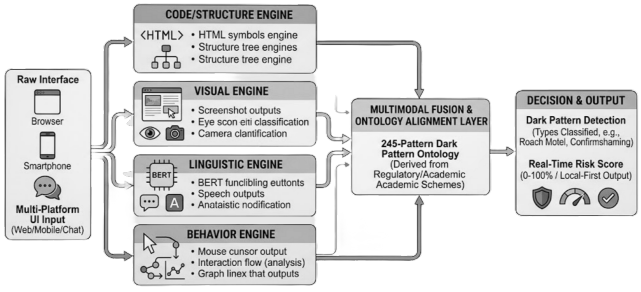

Fig. 1. Multimodal Dark Pattern Detection Architecture

Firstly, the framework starts with a **code/structure engine** that analyzes DOM trees, HTML attributes, JavaScript behaviors, and app navigation graphs. This engine detects the structurally defined patterns like forced registration, prechecked consent, hidden options, and Roach Motel-style asymmetries between entry and exit flows (e.g., 1–2 clicks to subscribe versus many steps to cancel). Second, a **visual engine** operates on the rendered screenshots and GUI element bounding boxes to identify layout-based patterns, including visual asymmetries between "accept" and "reject," disguised ads, countdown timers, and obstruction overlays. Third, a **linguistic engine** based on transformer models processes all the interface text and conversational logs to classify emotional pressure, confirm-shaming, misleading phrasing, scarcity claims, and manipulative chat-bot behaviors. Fourth, a **behavior engine** models full user journeys as sequences or graphs of states and actions (clicks, dwell times, back-tracks), learning the characteristic trajectories for multi-step patterns such as hidden costs, cancellation frictions, and subscription escalations.

The outputs retrieved from these engines are fused in a decision layer that produces pattern scores at an element, screen, and journey level. This late-fusion or multimodal-transformer layer is explicitly designed to align with a rich ontology ($\approx$ 245 patterns) along with a 68-type evaluation taxonomy, with the goal of substantially closing the current coverage gaps. To overcome the data bottleneck, this framework is paired with a multi-platform data-collection pipeline targeting at least 10,000 labeled examples across web, mobile, and conversational interfaces. For each flow of interest (for example: sign-up, checkout, cancellation, consent change), the system records synchronized DOM snapshots, screenshots, and interaction traces, which are then annotated against the unified ontology and relevant regulatory categories.

Finally, the solution incorporates adversarially robust and privacy-preserving deployment. Lightweight, distilled versions of the engines that cann be run on-device (browser, app, or OS) to keep raw UI content and interaction logs local, while federated learning is used to refine models without centralizing user data. Training-time adversarial augmentation (paraphrased text, perturbed layouts) and cross-engine consistency checks increase the robustness against minor modifications intended to evade detection. These detectors can also expose the risk scores to automated web agents, enabling comparatively safer navigation of dark-pattern-rich environments. Together, this proposed framework provides a very concrete, technically grounded path toward comprehensive, dynamic, and privacy-respecting detection of deceptive user interfaces.

IV. Conclusion

Dark-pattern detection methods presently available perform well in narrow settings but do not provide a comprehensive defense. Text-only models achieve comparatively high accuracy on small datasets, and visual or multimodal detectors are able to localize a limited set of patterns on static screens, while conversational benchmarks expose the manipulative behaviors in large language models. However, these approaches share key limitations: restricted taxonomic coverage which is contributed by small and specialized datasets, a focus on static snapshots rather than user journeys, and limited attention to adversarial robustness and privacy.

This work has synthesized these strengths into a unified view of approaches along with datasets and metrics, and used that analysis to encourage a four-engine framework that integrates DOM, visual, linguistic, and behavioral signals. It has also pushed for requirements of a larger, more diverse dataset aligned with a 245-pattern ontology, and highlighted implementation strategies that put user privacy as a priority while hardening models against attacks. Together, these contributions provide a technical insight for moving from isolated detectors towards integrated, end-to-end dark-pattern defense.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE I
LITERATURE SURVEY PART I: NLP, CLUSTERING, AND MULTIMODAL SNAPSHOT ANALYSIS

#	Method / Paper	Approach	Metrics (Reported)	Key Limitations
1	Umar et al. [4]	Log-regression on BoW text features.	Accuracy 92%, F1 93%.	Targets Text-only; while missing layout and multi-step flows.
2	Mathur et al. [1]	Text clustering + manual labeling.	Reports 11.1% prevalence.	Real-time detection not possible; only static text segments.
3	AidUI (Mansur [19])	CV + NLP for element localization.	F1 0.65; IoU $\approx$ 0.84.	Only single-screen analysis possible; no journey modeling.
4	DarkDialogs [23]	Rule-based detection on banners.	Mixed results; precision vs recall bias.	Very Domain-specific; rigid to layout updates.
5	Khanna et al. [25]	General ML + NLP framework.	High internal experimental metrics.	Proprietary: no adversarial robustness analysis.
6	Naive Bayes [26]	Naive Bayes on dark UI text segments.	Baseline classification accuracy.	Conceptually limited; no visual content.
7	Ramteke et al. [24]	BERT-based text classification.	Concept/Prototype stage.	Primarily textual dataset; no multi-step modeling.
8	DarkBench [22]	Prompt benchmark for LLM manipulation.	Success rate metrics.	Limited to conversational bots.
9	DPDGPT [33]	Multimodal LLM (Vision + Text).	Precision $\approx$ 0.86, F1 $\approx$ 0.88.	Relies heavily on proprietary LLMs: static focus.

TABLE II
LITERATURE SURVEY PART II: VISUAL IDENTIFICATION AND DYNAMIC JOURNEY MODELING

#	Method / Paper	Approach	Metrics (Reported)	Key Limitations
10	YOLOV5-MGC [5]	Object detector for GUI element labels.	$\approx$ 89.8% identification accuracy.	Element recognition only; lacks DP logic.
11	UIGuard (Chen [20])	CV + NLP extraction of UI properties.	90%+ Acc/F1 on mobile.	Mobile-only; misses exploration of app flows.
12	AppRay (Chen [21])	LLM exploration dynamic detector.	Strong performance on 18 DP types.	Android-restricted; lacks rich behavioral logs.
13	Roach Motel Ext. [29]	Rule-based subscription tracing.	Qualitative identification success.	Single-pattern focus; brittle to UI changes.
14	CogniModal-D [9]	7 modalities (EEG, Gaze, GSR, Video).	Accuracy $\approx$ 79% (Multimodal).	Hardware-heavy; not scalable for web crawls.
15	ContextDP [19]	Visual-text screenshot analysis.	F1 0.65, IoU $\approx$ 0.84.	Manual annotation bottleneck; static focus.
16	Ebusiness -DL [36]	DL model for e-commerce interfaces.	Improved Acc vs baseline ML.	Benchmark not public; architecture nondynamic-.
17	DECEPTICON [32]	Tasks for autonomous web agents.	Manipulation rate $>$ 70%.	Not an end-user detector; focused on AI.
18	IJISAE Cluster [25]	CNN/RNN + rule-based screen features.	80–90% Acc on narrow subsets.	Minuscule datasets; lacks journey sequence modeling.

V. FUTURE SCOPE

The future work can advance this agenda along several directions:

- **Dataset expansion:** The dataset can grow much beyond 10,000 examples to cover more platforms (desktop apps, games, smart devices), more languages, and richer labels (journey graphs, timing, user-impact annotations).
- **Model innovation:** Multiple sequence and graph models can be developed that explicitly represent user intent and multi-step flows, and multimodal transformers that jointly reason over DOM, pixels, and text.
- **Robustness and evaluation:** Systematically multiple stress-test detectors with adversarial text, layout, and interaction perturbations, and extend benchmarks like DarkBench and DECEPTICON to more pattern types and longer interactions can be implemented.
- **Deployment and tooling:** Build more efficient client-side components,browser/OS integrations, and APIs for warning users and supporting regulators.
- **Interdisciplinary collaboration:** Align technical work with HCI, law and policy to define harm more precisely and accurately, connect patterns to regulations, and support the proactive guidelines that reduce incentives to deploy dark patterns.

REFERENCES

[1] A. Mathur et al., "Dark patterns at scale," Proc. ACM HCI, vol. 3, 2019.
[2] C. M. Gray et al., "The dark side of UX design," Proc. CHI '18, 2018.
[3] J. Luguri and L. I. Strahilevitz, "Shining a light," J. Legal Anal., 2021.
[4] A. Umar et al., "Detecting dark patterns," arXiv:2412.14187, 2024.
[5] L. Cheng et al., "YOLOv5-MGC," IEEE Trans. Softw. Eng., 2022.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[6] S. Nie et al., “A comprehensive study,” arXiv: 2412.09147, 2024.
[7] J. S. Park et al., “AI deception,” Artif. Intell. Rev., 2024.
[8] H. Meadi et al., “Ethical challenges,” JMIR Mental Health, 2025.
[9] G. Joshi et al., “Multimodal ML,” Sci. Rep., 2025.
[10] C. M. Gray et al., “Ontology of dark patterns,” Proc. CHI ’24, 2024.
[11] FTC, Bringing Dark Patterns to Light, Staff Report, 2022.
[12] EDPB, Guidelines 3/2022, 2022.
[13] M. T. Khan et al., “Detecting mental manipulation,” in Proc. ACL, 2021.
[14] T. B. Brown et al., “Few-shot learners,” NeurIPS, 2020.
[15] J. Devlin et paper., “BERT,” Proc. NAACL, 2019.
[16] J. Redmon and A. Farhadi, “YOLOv3,” arXiv: 1804.02767, 2018.
[17] J. Pearl, The Book of Why. Basic Books, 2018.
[18] B. Goodman, “EU regulations,” AI Mag., 2017.
[19] S. M. H. Mansur et al., “AidUI,” Proc. ICSE ’23, 2023.
[20] J. Chen et al., “Unveiling the tricks,” arXiv:2308.05898, 2023.
[21] J. Chen et al., “Exploration to revelation,” arXiv: 2411.18084, 2024.
[22] Apart Research, “DarkBench,” Proc. ICLR, 2025.
[23] D. W. Woods et al., “DarkDialogs,” EuroS&P, 2023.
[24] A. R. Ramteke et al., “Deceptive patterns,” arXiv:2406.01608, 2024.
[25] S. Shahriar et al., “AI-driven approach,” IJISAE, 2024.
[26] S. A. Muthukumarasamy, “Automated detection,” NCI, 2025.
[27] N. Gunawan et al., “Comparative study,” Privacy Con, 2022.
[28] ODGDT, Addressing Dark Patterns, Australia, 2024.
[29] L. Di Geronimo et al., “Where to find them,” Proc. CHI ’20, 2020.
[30] DAPDE Project, “Obstacles,” 2022.
[31] EJ-AI Authors, “Adversarial attacks,” Eur. J. Artif. Intell., 2025.
[32] Z. Chen et al., “DECEPTICON,” arXiv:2512.22894, 2025.
[33] Lin et al., “DPDGPT,” Sci. Direct, 2025.
[34] H. Meadi et al., “Datasets review,” Sci. Rep., 2024.
[35] Y. Yada et al., “Big Data dataset,” Proc. IEEE, 2022.
[36] A. R. Gunasekera, “DL evaluation,” Proc. ACM, 2025.