

An Explainable Web-Based Model to Identify Dark Pattern on E-Commerce User Interfaces

¹*Tarilatei Mao Claudius, ²N. D. Nwiabu, ³C. G. Igiri

^{1,2,3}Department of Computer Science, Rivers State University, Port-Harcourt, Nigeria

Abstract: Dark patterns are manipulative interface designs that steer users toward unintended purchases, subscriptions, or disclosures, and they are widespread across e-commerce platforms. Existing automated detection approaches are largely single-modality and offer limited transparency into why a pattern was flagged, which restricts their use by regulators, platform developers, and consumers. This study designed, implemented, and evaluated DpNet-v2.4.1-XAI, an explainable multi-modal detection system that fuses visual (CNN), textual (BERT), and DOM-structural signals and pairs them with a dual SHAP and Grad-CAM explainability layer. The model was trained on 5,200 labelled samples across eight dark-pattern categories (Urgency, Scarcity, Misdirection, Social Proof, Forced Action, Obstruction, Sneaking, and Confirmshaming) and delivered as a six-module web application. On a held-out test set of 520 e-commerce interface samples it achieved 92.3% accuracy and an F1-score of 89.9%, outperforming the benchmark system of Vedhapriyavadhana *et al.* (2025) by 3.2 and 2.5 percentage points on accuracy and F1-score respectively, and by 8.0 percentage points on explainability faithfulness (87.0% versus 79.0%). Field testing on 18 live Nigerian and cross-border e-commerce pages detected 98 pattern instances, with confidence scores as high as 93%. The results demonstrate that multi-modal fusion measurably improves both detection performance and explanation faithfulness over a single-modality baseline, offering a deployable, transparent tool for consumers, platform developers, and regulators seeking to identify manipulative design at the point of interaction.

Keywords: Dark Patterns; Explainable AI; E-Commerce; Multi-Modal Deep Learning; SHAP; Grad-CAM; Consumer Protection; Web Application.

I. INTRODUCTION

Dark patterns were formally conceptualised by Brignull as tricks in websites and apps that make users do things they did not mean to [1]. Unlike poor design, which results from ignorance or resource constraints, dark patterns are intentional: they are crafted to manipulate user behaviour for the benefit of the implementing organisation at the expense of the user. Taxonomic studies identify numerous forms, including confirmshaming, roach-motel cancellation barriers, hidden costs, trick questions, disguised advertisements, and misdirection [2, 3]. Their prevalence in e-commerce is well documented. An automated audit of approximately 11,000 shopping websites catalogued 1,818 dark-pattern instances, led by urgency, scarcity, and social-proof cues [3], and an analysis of 240 popular Android applications found dark patterns in over 95% of them [4]. Controlled experiments further show that they change behaviour, with mild dark patterns raising sign-up rates by roughly twenty per cent and aggressive ones by more than forty per cent [5].

Regulators have begun to respond. The EU General Data Protection Regulation requires consent to be freely given,

specific, informed, and unambiguous [6]; the EU Digital Services Act prohibits dark patterns on very large online platforms [7]; and the US Federal Trade Commission has taken action against deceptive subscription practices [8]. Enforcement, however, does not scale, because human auditors cannot keep pace with the dynamic, continuously updated nature of e-commerce interfaces. This creates a strong case for automated, AI-driven detection. Machine-learning approaches have been applied to textual dark patterns in consent banners [9, 10] and to UI screenshots [11]. A persistent limitation is opacity: most deployed models are black boxes that return a label without indicating why a design element was flagged, which hinders adoption in regulatory, legal, and educational settings where findings must be understood, contested, and acted upon. Explainable artificial intelligence (XAI) addresses this through methods such as SHAP, LIME, and Grad-CAM, which attribute a prediction to the features or image regions that drove it [12, 13].

The concern is acute in Nigeria, where platforms such as Jumia and Konga serve a rapidly growing consumer base that often lacks regulatory awareness or recourse. NITDA and the

FCCPC have identified consumer protection in digital commerce as a priority [14], yet, to the best of our knowledge, no dedicated automated tool supports the detection and reporting of dark patterns on Nigerian or wider African e-commerce platforms. In summary, four gaps motivate this work: existing detectors classify without human-interpretable explanations; they cover specific sub-types rather than a multi-category framework; no publicly accessible web tool combines detection with XAI explanations for non-technical stakeholders; and detection tooling and empirical evidence remain scarce for African e-commerce. This study therefore designs, implements, and evaluates an explainable multi-modal web-based system that addresses all four gaps together.

1.1 Aim and Objectives of the Study

The aim of this study is to develop an explainable web-based model to identify dark patterns on e-commerce user interfaces. Specifically, the study set out to: (i) review dark-pattern taxonomies and automated detection approaches to establish a multi-category detection framework; (ii) design and train a deep-learning model that classifies dark patterns from e-commerce pages across multiple categories; (iii) integrate SHAP and Grad-CAM to generate human-interpretable explanations for each detection; (iv) develop a web interface through which users submit e-commerce URLs and receive explained results; and (v) evaluate the model in terms of accuracy, precision, recall, F1-score, and explanation quality. The system targets eight categories on desktop and mobile e-commerce interfaces (product listing, checkout, subscription-management, and consent pages) and is developed as a proof-of-concept web application.

1.2 Organisation of the Paper

The remainder of this paper is organised as follows. Section 2 reviews the literature on dark patterns, explainable AI, and automated detection, and identifies the knowledge gap. Section 3 describes the methodology, comprising the research approach, requirements, system architecture, dataset, model and training procedure, algorithm, implementation, and evaluation procedure. Section 4 presents and discusses the results. Section 5 concludes the paper and outlines future work.

II. LITERATURE REVIEW

2.1 Dark Patterns: Taxonomy, Prevalence, and Impact

Gray et al. derived a five-category taxonomy (nagging, obstruction, sneaking, interface interference, and forced action)

from practitioner-authored repositories [2]. Mathur *et al.* extended the field computationally, identifying 15 dark-pattern types across 11,000 shopping websites and finding a positive correlation between site popularity and prevalence, which suggests that commercial pressure drives adoption [3]. Luguri and Strahilevitz supplied causal evidence of behavioural impact through randomised online experiments [5]. Bongard-Blanchy et al. showed that users who are aware of dark patterns often remain susceptible to them, a phenomenon they term informed helplessness, so awareness alone is insufficient protection and system-level, automated interventions are needed [15]. From a normative perspective, Ahuja and Kumar found that most frameworks treat autonomy as an individual cognitive capacity and underemphasise its structural dimensions, leading to interventions that favour disclosure over design prohibition [16].

2.2 Cognitive, Regulatory, and Ethical Context

Dark patterns exploit documented cognitive biases: the scarcity heuristic [17], loss aversion [18], and the default effect [19]. Bounded rationality [20] explains why users under time pressure or cognitive load are especially vulnerable, which implies that XAI outputs must themselves be cognitively accessible. The regulatory context has evolved quickly. The Irish Data Protection Commission fined Meta EUR 390 million in 2023 partly on dark-pattern grounds [21]; the Digital Services Act (Article 25) is the first direct legislative prohibition of dark patterns in EU law [7]; the UK Competition and Markets Authority has issued guidance on subscription traps [22]; and the US "Click to Cancel" rule targets roach-motel cancellation designs [23]. These developments underline the urgency of scalable, explainable detection tooling.

2.3 Dark-Pattern Categories Targeted

This study adopts a consolidated typology of eight categories prevalent in e-commerce, drawn principally from Mathur *et al.* and Gray *et al.* [3, 2]. *Urgency* manufactures time pressure, for example through countdown timers that reset on reload. *Scarcity* signals limited availability, genuine or fabricated, such as "Only 2 left!". *Misdirection* draws attention toward or away from information by visual, linguistic, or interactive means, for example a prominent "Accept All" button beside a low-contrast reject option. *Social Proof* displays real or fabricated indicators of other users' activity or approval. *Forced Action* requires an unrelated task, such as account creation, before the user can achieve their original goal. *Obstruction* makes a disfavoured action, most often cancellation, unreasonably difficult. *Sneaking* conceals or delays disclosure of

decision-relevant information, such as fees revealed only at final checkout. *Confirmshaming* frames the decline option in emotionally manipulative language.

2.4 Machine Learning and Explainable AI for UI Analysis

Convolutional neural networks achieve strong results on UI analysis tasks such as element detection and layout classification [24], and the visual modality matters for dark patterns because many are implemented through colour contrast, button size, and visual salience. Because annotated dark-pattern datasets are small, transfer learning from pre-trained models such as ResNet [25], EfficientNet [26], and Vision Transformers [27] is standard practice, and attention mechanisms [28] offer a built-in interpretability signal. XAI methods can be classified as local or global, ante-hoc or post-hoc, and model-agnostic or model-specific [29]. Grad-CAM localises discriminative image regions from the gradients of the class score at the final convolutional layer [30], with Grad-CAM++ [31] extending it to multi-instance images; these model-specific methods are computationally efficient and suit real-time web deployment. SHAP applies Shapley values from cooperative game theory to attribute a prediction to input features, is model-agnostic, and has strong axiomatic guarantees [32]. LIME fits a local interpretable model [33] but has been criticised for instability [34], which makes SHAP preferable for high-stakes use.

2.5 Related Work on Automated Detection

Mathur *et al.* combined crawling, natural-language processing, and visual heuristics to identify dark patterns at scale, but the approach was limited to textual patterns and offered no explanation of individual detections [3]. Raju *et al.* built a user-facing detector based on natural-language processing and sentiment analysis for confirmshaming, trick questions, and disguised advertisements, which cannot capture visually implemented patterns [35]. Nazarov and Baimukhambetov used unsupervised clustering of interface elements and DOM structures to discover pattern categories without labels, but the outputs required expert interpretation [36]. Yada *et al.* released a benchmark dataset and identified data scarcity as a principal bottleneck [37]. Mansur *et al.* proposed AidUI, which recognises dark patterns from UI element detection and screenshot analysis for developer testing workflows, without explainability [38]. Srinivas Babu *et al.* developed SURFER, a multi-stage pipeline of crawling, DOM analysis, and visual feature extraction, whose outputs were limited to category labels [39]. Król compared browser-based tools with LLM-based approaches on university

websites and found that LLM-based detection achieved higher recall at significantly higher computational cost [40].

The closest prior work to the present study is that of Vedhapriyavadhana *et al.*, who fine-tuned BERT to classify the text of retail pages as dark pattern or not and reported high precision and recall, but the approach is text-only and cannot detect patterns implemented purely through visual means such as colour contrast, button sizing, or layout [41]. Related work widens the picture further: DECEPTICON shows that dark patterns also mislead AI web agents [42]; temporal analysis of Amazon's cancellation flow introduces dark patterns that unfold across sessions [43]; users with ADHD are significantly more susceptible to several categories [44]; and organisational enablers of dark-pattern adoption include competitive pressure, weak enforcement, and short-term revenue incentives [45]. Ethical analyses stress that detection tools should be transparent and independently verifiable [46].

2.6 Knowledge Gap

The reviewed work confirms that dark patterns are prevalent, harmful, and increasingly regulated, and that automated detection is feasible. However, no current system integrates (i) comprehensive multi-category dark-pattern detection, (ii) XAI explanations at both the region level (Grad-CAM) and the feature-importance level (SHAP), and (iii) a web-deployable interface accessible to non-technical stakeholders within a single framework. Nor has such an integrated approach been applied to e-commerce platforms in the Nigerian or broader African market. This study addresses that gap.

III. METHODOLOGY

3.1 Research Approach

The study adopts constructive research within an applied computer-science paradigm [47], integrating the iterative design, implementation, and evaluation cycles of Design Science Research. The constructive process had four phases: (1) problem conceptualisation and requirements elicitation from the literature and a stakeholder analysis; (2) system design, covering architecture, dataset curation, and model design; (3) implementation and training; and (4) evaluation using quantitative performance metrics and XAI quality assessment. System design followed Object-Oriented Analysis and Design (OOAD) [48, 49], chosen because the system decomposes naturally into image preprocessing, a detection model, an XAI engine, a web interface, and a report generator, each mapping onto classes, encapsulation, and inter-object messaging. The

design produced use-case, sequence, and class diagrams; the class model comprises seven classes (URLs, DetectionModel, ModelWeights, FastAPIBackend, GradCAMEngine, SHAPExplainer, and PostgreSQLDatabase).

3.2 Existing Systems and Requirements

Manual audits and heuristic evaluations are accurate but unscalable. Automated systems fall into NLP text-analysis approaches [9, 10] and computer-vision classifiers [11], while browser extensions rely on rule-based heuristics with limited accuracy and no explanations. Analysis of these systems identified five limitations: absence of explainability; narrow detection coverage; limited geographic representativeness, since corpora are drawn overwhelmingly from US and European platforms; accessibility barriers, since most high-performing systems are research prototypes needing programmatic access; and brittle static-rule dependencies.

Requirements were elicited from three stakeholder groups: consumer advocates and regulators, platform developers and UX auditors, and academic researchers. Users must be able to submit a page URL without capturing a screenshot, receive labelled detections with confidence scores within five seconds, see a highlighted visual explanation and a plain-language description, download a PDF report, and use the interface on desktop and mobile without technical expertise. Functionally, the system crawls and renders the page with a headless browser; preprocesses it for each model branch; classifies it against the eight categories as a multi-label task; generates Grad-CAM maps

and SHAP attributions (100 background samples) for flagged categories; and provides a JSON report, a REST endpoint for batch submission, and PDF export. Quality targets included end-to-end inference and explanation within five seconds on 2 vCPU and 4 GB RAM, a minimum F1-score of 0.80, and explanation faithfulness of at least 0.85 on the Infidelity and Sensitivity metrics [50]; further targets covered scalability, usability, reliability, and data retention.

3.3 System Architecture

The system is a three-tier web application: a React 18 and Tailwind CSS presentation tier, a Python/FastAPI application tier, and a model and data tier holding the trained PyTorch model, the SHAP and Grad-CAM modules, and a PostgreSQL store. It is realised as five microservices deployed with Docker and Kubernetes (Fig. 1). The API Gateway (Nginx reverse proxy with JWT-based sessions) authenticates, rate-limits, and routes requests. The Preprocessing Service resizes and normalises the screenshot with ImageNet statistics, tokenises extracted UI text to a maximum of 512 tokens, and encodes the DOM tree into structural feature vectors. The *Detection Model Service* hosts DPNet-v2.4.1-XAI, returns a probability for each of the eight categories through a sigmoid output layer, and exposes the feature maps and attention weights needed for explanation. The *XAI Service* computes Grad-CAM maps and SHAP attributions for both the visual and textual branches. The *Report Generation Service* compiles the screenshot, overlays, token highlights, category labels, confidence scores, and a plain-language explanation into a web view and a downloadable PDF.

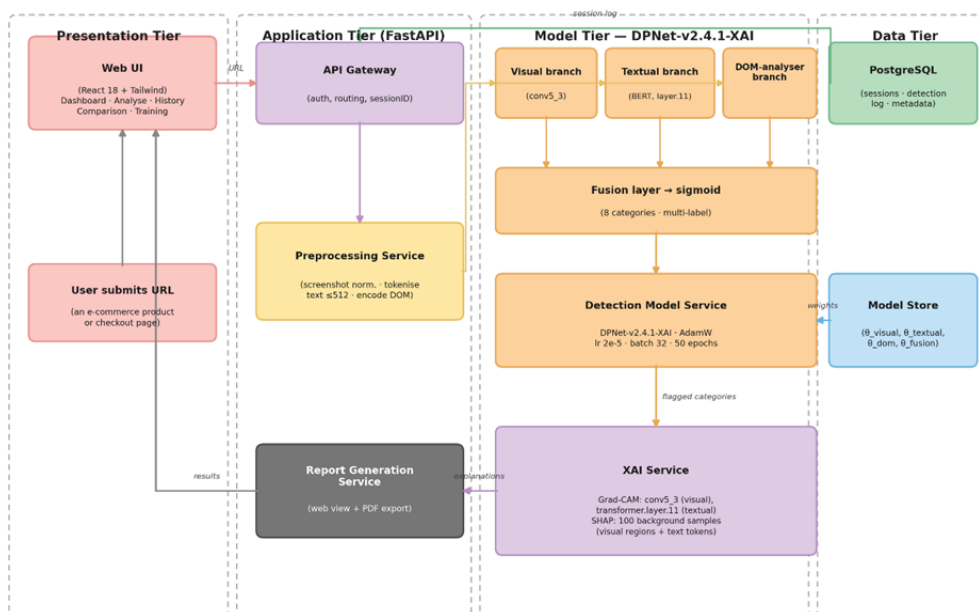


Figure 1: Architecture of the proposed explainable dark-pattern detection system (DPNet-v2.4.1-XAI)

3.4 Dataset

The model was trained on a composite dataset of 5,200 labelled samples assembled from four sources: e-commerce screenshots from Nigerian and pan-African platforms (Jumia Nigeria, Konga, Paystack Checkout, Flutterwave Payment Pages, Slot Nigeria, Supermart, PriceCheck, Jiji Nigeria, Temu Nigeria, and Shein Nigeria); DOM trees extracted and annotated from the same pages; synthetic samples produced by programmatically injecting known manipulative elements into clean page templates, used to augment under-represented categories; and manually labelled pages from a dual-annotator coding process using a codebook derived from the Gray *et al.* taxonomy [2]. Inter-annotator agreement was assessed with Cohen's kappa, and a threshold of $\kappa \geq 0.75$ was required for label inclusion. Because a page may exhibit several categories at once, the data are annotated for multi-label classification, and the absence of all eight labels is the implicit clean case. Random crop and colour jitter (visual branch), text perturbation (textual branch), and DOM structure variation (DOM branch) were applied during training. The data were split 80/10/10 with all categories proportionally represented (Table 1), and no source page appears in both the training and test sets, mitigating data leakage.

Table 1: Dataset split

Subset	Samples	Proportion
Training	4,160	80%
Validation	520	10%
Test	520	10%
Total	5,200	100%

3.5 Model Architecture and Training

DPNet-v2.4.1-XAI fuses three parallel branches. The *textual branch* is a pre-trained BERT encoder fine-tuned on UI text from labels, buttons, banners, and countdown timers. The *DOM-analyser branch* processes structural features (element hierarchy, tag composition, and positional relationships) that are not always visible in a static screenshot. The *visual branch* is a CNN backbone, benchmarked against EfficientNet-B4 [26], ResNet-50, and ViT-Base on a 20% validation subset and retained for its validation F1-score, inference latency below the five-second target, and compatibility with Grad-CAM at its final convolutional block (conv5_3). Branch embeddings are concatenated at a fusion layer that feeds a fully connected head with eight sigmoid outputs, reflecting the multi-label task. The visual backbone was initialised from ImageNet-21k weights [51] and the textual branch from a pre-trained BERT checkpoint,

while the DOM branch was trained from scratch. All branches were trained jointly for 50 epochs with AdamW [52] using the settings in Table 2, on four NVIDIA A100 (80 GB) GPUs in 47 hours 22 minutes. The loss is the mean binary cross-entropy over the eight outputs:

$$L = -(1/C) \sum_{j=1}^C [y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)] \quad (1)$$

where $C = 8$ is the number of categories, $y_j \in \{0,1\}$ is the ground-truth label for category j , and $\hat{y}_j$ is the predicted probability, after sigmoid activation, that category j is present.

Table 2: Training hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Optimiser	AdamW	Learning rate	2×10^{-5}
Batch size	32	Epochs	50
Weight decay	0.01	Warm-up steps	500
Dropout	0.3	Max. sequence length	512 tokens
Gradient clipping	1.0	Label smoothing	0.1

The XAI pipeline runs after inference. Grad-CAM [30] is computed from the gradients of the predicted class score with respect to the visual branch's final convolutional block (conv5_3), with an analogous gradient-based attention map at layer 11 of the BERT encoder. SHAP values [32] are computed for both branches using 100 background samples, attributing importance to image regions and text tokens. Explanation faithfulness is assessed with the Infidelity and Sensitivity metrics [50] against a quality threshold of 0.85.

3.6 Algorithm

Algorithm 1 summarises the training and per-URL inference procedure.

Algorithm 1: DPNet-v2.4.1-XAI - Training and Inference Input: D_{train} (4,160), D_{val} (520), 8 category labels {Urgency, Scarcity, Misdirection, Social Proof, Forced Action, Obstruction, Sneaking, Confirmshaming} Output: weights $\theta = \{\theta_{visual}, \theta_{textual}, \theta_{dom}, \theta_{fusion}\}$ Training 1: Initialise θ_{visual} from ImageNet-pretrained CNN 2: Initialise $\theta_{textual}$ from pretrained BERT 3: Initialise θ_{dom} randomly 4: for epoch = 1 to 50 do 5: for each mini-batch B (size 32) in D_{train} do 6: $x_{vis} \leftarrow \text{screenshot}(B)$ 7: $x_{txt} \leftarrow \text{tokenise}(\text{text}(B), \text{max_len} = 512)$ 8: $x_{dom} \leftarrow \text{encode_DOM}(B)$ 9: $h_{vis} \leftarrow \text{CNN}(x_{vis}; \theta_{visual})$ // conv5_3 10: $h_{txt} \leftarrow \text{BERT}(x_{txt}; \theta_{textual})$ // layer 11 for XAI 11: $h_{dom} \leftarrow \text{DOMEncoder}(x_{dom}; \theta_{dom})$ 12: $h_{fused} \leftarrow \text{Fuse}(h_{vis}, h_{txt}, h_{dom})$ 13: $\hat{y} \leftarrow \text{sigmoid}(\text{FC}(h_{fused}; \theta_{fusion}))$ // 8-dim
--

```

14: L ← BinaryCrossEntropy( $\hat{y}$ , y) // Eq. (1)
15:  $\theta \leftarrow \text{AdamW\_update}(\theta, \nabla L; \text{lr} = 2 \times 10^{-5}, \text{wd} = 0.01,$ 
    warm-up = 500, clip = 1.0)
16: end for
17: Evaluate F1-score on D_val; log loss/accuracy
18: end for
19: return  $\theta$ 
Inference (per submitted URL)
20: page ← CrawlAndRender(url)
21:  $x_{\text{vis}}, x_{\text{txt}}, x_{\text{dom}} \leftarrow \text{Extract}(\text{page})$ 
22:  $\hat{y} \leftarrow \text{Forward}(x_{\text{vis}}, x_{\text{txt}}, x_{\text{dom}}; \theta)$ 
23: flagged ← { j :  $\hat{y}_j > \text{confidenceThreshold}$  }
24: for each j in flagged do
    25: heatmap_j ← GradCAM( $\theta_{\text{visual}}, \text{conv5\_3}, j$ )
    26: attn_j ← GradCAM( $\theta_{\text{textual}}, \text{transformer.layer.11}, j$ )
    27: shapVis_j, shapTxt_j ← SHAP( $\theta, x_{\text{vis}}, x_{\text{txt}}, \text{bg} = 100, j$ )
28: end for
29: return flagged, heatmaps, attn, shapVis, shapTxt, confidence scores

```

3.7 Implementation and Web Application

The system was implemented in Python 3.11 with PyTorch 2.1 and Hugging Face Transformers 4.36, SHAP, and the pytorch-grad-cam library [53], a FastAPI backend, a React 18 and Tailwind CSS frontend, PostgreSQL, and Docker/Kubernetes. The web application has six modules. *Analyse* accepts a URL and returns the pattern count, overall risk level, inference time, elements analysed, and a per-detection breakdown (flagged text, category, severity, and confidence). *Dashboard* shows total scans, total instances, high-risk pages, and a rolling 30-day trend by category. *History* logs each analysed URL with its pattern count and risk classification. *Comparison* benchmarks the system against Vedhapriyavadhana et al. [41]. *Training* exposes the architecture, dataset configuration, training curves, hyperparameters, and XAI settings, and *Architecture* renders the component diagram in the interface.

3.8 Evaluation Procedure

The 520 test samples were held out until training was complete and used to compute accuracy, precision, recall, and F1-score across the eight categories. Explanation quality was assessed with Infidelity and Sensitivity [50]. The system was benchmarked against Vedhapriyavadhana et al. [41] on accuracy, F1-score, and XAI faithfulness. Finally, black-box functional testing was carried out on live, publicly accessible e-commerce pages submitted through the Analyse module exactly as an end user would, and the History and Dashboard modules recorded the outcomes.

IV. RESULTS AND DISCUSSION

4.1 Detection Performance

On the held-out test set of 520 samples across eight categories, DPNet-v2.4.1-XAI achieved an accuracy of 92.3% and an F1-score of 89.9%, above the 0.80 F1 target set in Section 3.2. Table 3 gives per-category results derived from the confusion matrix in Fig. 2 (support is the row total, precision is the diagonal divided by the column total, and recall is the diagonal divided by the row total). Performance is strongest for Urgency (F1 92.0%), Confirmshaming (91.5%), and Forced Action (90.6%) and weakest for Misdirection (83.8%) and Obstruction (85.9%), which are therefore priorities for further data collection.

Table 3: Per-category precision, recall, and F1-score

Category	Support	Precision	Recall	F1-score
Urgency	342	92.9%	91.2%	92.0%
Scarcity	287	87.8%	88.2%	88.0%
Misdirection	198	83.8%	83.8%	83.8%
Social Proof	256	86.3%	86.3%	86.3%
Forced Action	312	91.2%	90.1%	90.6%
Obstruction	183	85.0%	86.9%	85.9%
Sneaking	239	88.5%	90.0%	89.2%
Confirmshaming	206	91.3%	91.7%	91.5%

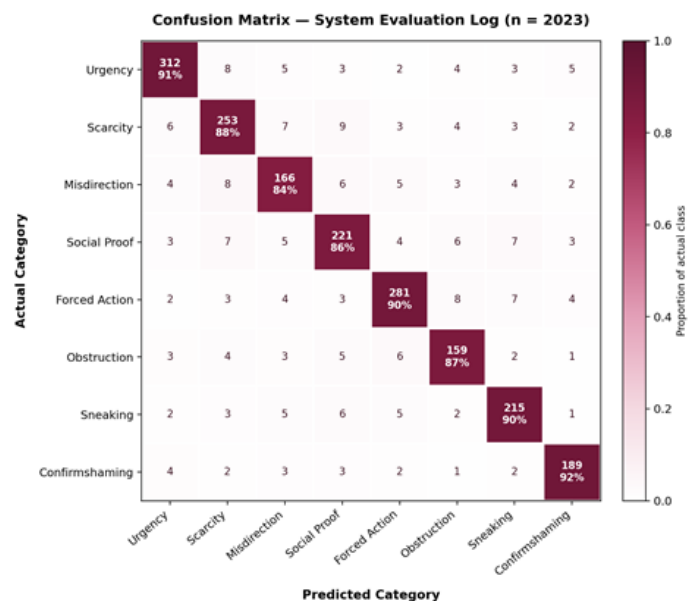


Figure 2: Confusion matrix, actual category (rows) versus predicted category (columns)

4.2 Training Behaviour

Training loss decreased steadily across the 50 epochs, while validation loss plateaued and drifted slightly upward from around epoch 25, so the gap between the two widened. This pattern is typical of a model with enough capacity to fit its training distribution closely. An early-stopping checkpoint around epochs 25–30 may therefore generalise at least as well as the final epoch while reducing training cost.

4.3 Explainability Evaluation

The system achieved an XAI faithfulness score of 87.0%, evaluated with the Infidelity and Sensitivity metrics against the explanation quality threshold of 0.85. This indicates that the SHAP and Grad-CAM explanations attribute importance, on the whole, to the image regions and text tokens that genuinely drive the model's prediction.

4.4 Comparison with Existing Work

Table 4 compares the system with the text-only BERT classifier of Vedhapriyavadhana et al. [41]. The proposed system is ahead on all three axes, with the largest margin on XAI faithfulness (+8.0 percentage points). This reflects the advantage of grounding an explanation in visual, textual, and structural evidence simultaneously rather than in text alone. Table 5 summarises the qualitative differences.

Table 4: Comparative performance

Metric	Proposed	Vedhapriyavadhana et al.	Difference
Accuracy	92.3%	89.1%	+3.2 pp
F1-score	89.9%	87.4%	+2.5 pp
XAI faithfulness	87.0%	79.0%	+8.0 pp

Table 5: Qualitative feature comparison

Aspect	Proposed system	Vedhapriyavadhana et al.
Detection approach	Multi-modal deep learning (CNN + BERT + DOM)	Text-only BERT classifier
XAI techniques	SHAP + Grad-CAM (visual and textual)	Not reported
Input method	Direct URL submission	Pre-scraped text input
Categories	8 (comprehensive taxonomy)	Per source taxonomy

Visual explanations	Grad-CAM heatmaps on page regions	Not applicable (text-only)
Real-time processing	Yes (instant URL analysis)	Not specified
Accessibility	Web-based, no setup required	Not specified
Model transparency	Dual-branch XAI with human-readable explanations	Limited to classification output

4.5 Field Results

Beyond the test set, the Dashboard and History modules provide a field-level view of the system on live pages (Table 6). Across 18 scans the system detected 98 dark-pattern instances, an average of 5.4 per page, and classified all 18 pages as high or critical risk. Among the sites scanned, Temu.com returned 7 patterns (Critical), Jiji.ng returned 5 (High on the first scan and Critical on a re-scan), and Shein.com returned 4 (High).

Table 6: Field-testing summary

Metric	Value
Total scans	18
Total patterns detected	98
Pages flagged high or critical risk	18 (100%)
Average patterns per page	5.4

Table 7 reproduces the full breakdown returned for one Jiji.ng product page: five flagged instances, one from each of five categories, with the responsible text, a severity rating, and a confidence score. Confidences range from 76% to 93%, all above the system's threshold. The two highest (Sneaking, 93%; Scarcity, 91%) concern High-severity detections, while the two Medium-severity detections (Social Proof, Misdirection) score lowest, consistent with more overt manipulative language being easier to attribute confidently than subtler wording.

Table 7: Detection breakdown for a single test page (Jiji.ng)

Category (type)	Flagged text	Severity	Confidence
Social Proof (Fake Activity)	"12,847 verified reviews"	Medium	76%
Scarcity (Low Stock Warning)	"only 3 left in stock – order soon!"	High	91%
Misdirection (False Emphasis)	"FREE shipping (minimum order \$150)"	Medium	83%

Urgency (Limited Time Offer)	"Flash Deal – only available today!"	High	90%
Sneaking (Sneak into Basket)	"Protection plan added to your order"	High	93%

The system's pattern co-occurrence analysis shows the strongest pairing between Urgency and Scarcity (13), followed by Urgency with Sneaking (12) and Scarcity with Sneaking (11). Together with the Jiji.ng breakdown, this suggests that these categories often appear as a bundled manipulative strategy rather than in isolation.

4.6 Discussion

The results address each objective. A review of taxonomies and detectors (Section 2) established the eight-category framework; a multi-modal model was designed, trained, and evaluated (92.3% accuracy, 89.9% F1-score); SHAP and Grad-CAM were integrated with a faithfulness score of 87.0%, above the 0.85 threshold; a six-module web interface delivers explained results from a submitted URL; and the benchmark comparison shows gains over a single-modality baseline on every axis reported.

Several limitations qualify these results. First, the field test contained no known-clean pages: all 18 scanned pages were flagged high or critical, so the false-positive rate has not been established. Second, Jiji.ng received different risk classifications (High, then Critical) on two scans despite an identical pattern count, so scan-to-scan consistency should be investigated before reliability is reported as a formal metric. Third, the DOM-structural branch is the only one of the three inputs without its own attribution technique. Fourth, the comparison is against a single text-only baseline. Fifth, the evaluation reported here covers detection performance, explanation faithfulness, and field behaviour; the usability, concurrency, and uptime targets set in Section 3.2 are not reported. Finally, the system analyses single page loads and the languages represented in its training data, and so cannot yet capture temporal patterns that unfold across sessions [43].

V. CONCLUSIONS

This study presented the design, implementation, and evaluation of DPNet-v2.4.1-XAI, an explainable multi-modal dark-pattern detection system for e-commerce interfaces. By fusing visual (CNN), textual (BERT), and DOM-structural signals into one classifier and explaining it through dual SHAP

and Grad-CAM, the system addresses the narrow single-modality coverage and limited transparency of existing detectors. It achieved 92.3% accuracy and an F1-score of 89.9% on a held-out test set of 520 samples across eight categories, with an XAI faithfulness score of 87.0%, outperforming a text-only baseline by 3.2 points in accuracy, 2.5 points in F1-score, and 8.0 points in faithfulness. Field testing on 18 live Nigerian and cross-border pages detected 98 pattern instances with confidence scores as high as 93%, demonstrating multi-label detection under real operating conditions.

The study contributes a multi-modal detection architecture, a dual explainability layer applied across the visual and textual branches, and field evidence from a regional e-commerce context that the reviewed literature identified as under-represented. By restoring the transparency that informed consumption depends on, the system also supports Sustainable Development Goal 12 (Target 12.8). We recommend integrating it into pre-launch design review, consumer-facing checking tools, and regulatory evidence-gathering, for example by bodies such as Nigeria's FCCPC, and validating it further on balanced field panels that include known-clean pages and fixed re-scan intervals.

Future work will extend coverage to temporal and multi-session dark patterns and to Nigerian Pidgin English and other locally prevalent languages; collect targeted data for Misdirection and Obstruction, the two weakest categories; develop a dedicated explanation method for the DOM-structural branch; and deploy the system as a real-time browser extension with an automated remediation layer suggesting compliant alternatives. A stakeholder study with regulators, developers, and consumers, together with an investigation of the legal admissibility of the system's output, would clarify how much of its recommended real-world impact is realised in practice.

REFERENCES

- [1] Brignull, H. (2010). Dark patterns: Deception vs. honesty in UI design. *A List Apart*. <https://alistapart.com/article/dark-patterns-deception-vs-honesty-in-ui-design/>
- [2] Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. <https://doi.org/10.1145/3173574.3174108>
- [3] Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping

- websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 81:1–81:32. <https://doi.org/10.1145/3359183>
- [4] Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., & Bacchelli, A. (2020). UI dark patterns and where to find them: A study on mobile applications. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM.
- [5] Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1), 43–109. <https://doi.org/10.1093/jla/laaa006>
- [6] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). (2016). *Official Journal of the European Union*, L 119, 1–88.
- [7] European Parliament. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services (Digital Services Act). *Official Journal of the European Union*, L 277, 1–102.
- [8] Federal Trade Commission. (2022). FTC takes action against companies that use "dark patterns" to trap people into subscriptions. <https://www.ftc.gov/news-events/news/press-releases/2022/11/ftc-takes-action-against-companies-use-dark-patterns-trap-people-subscriptions>
- [9] Hidaka, S., Kobayashi, T., & Hamamura, T. (2023). Automatic detection of dark patterns in cookie consent banners using BERT. In *Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM.
- [10] Soe, T. H., Nordberg, O. E., Guribye, F., & Slavkovik, M. (2023). Circumvention by design: Dark patterns in cookie consent interfaces. In *Proceedings of the 18th European Conference on Computer-Supported Cooperative Work* (pp. 1–24). *European Society for Socially Embedded Technologies*.
- [11] Nguyen, H., Do, T., & Nguyen, T. T. (2022). Detecting dark patterns in e-commerce using deep learning. In *Proceedings of the 2022 IEEE International Conference on Big Data* (pp. 3214–3223). IEEE.
- [12] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- [13] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>
- [14] National Information Technology Development Agency. (2021). *Nigeria digital economy policy and strategy*. NITDA. <https://nitda.gov.ng/wp-content/uploads/2021/01/NigeriaDigitalEconomyPolicyandStrategy.pdf>
- [15] Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., & Lenzini, G. (2021). I am definitely manipulated, even when I am aware of it. It's ridiculous! Dark patterns from the end-user perspective. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (pp. 763–776). ACM. <https://doi.org/10.1145/3461778.3462086>
- [16] Ahuja, S., & Kumar, J. (2022). Conceptualizations of user autonomy within the normative evaluation of dark patterns. *Ethics and Information Technology*, 24(4), Article 52. <https://doi.org/10.1007/s10676-022-09672-9>
- [17] Cialdini, R. B. (2001). *Influence: Science and practice (4th ed.)*. Allyn & Bacon.
- [18] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>
- [19] Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, 302(5649), 1338–1339. <https://doi.org/10.1126/science.1091721>
- [20] Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1), 99–118. <https://doi.org/10.2307/1884852>
- [21] Irish Data Protection Commission. (2023). Decision on the Meta Ireland complaint. <https://www.dataprotection.ie/en/news-media/press-releases/dpc-announces-decision-in-meta-ireland-inquiry>
- [22] Competition and Markets Authority. (2022). Online choice architecture: How digital design can harm competition and consumers. CMA. <https://www.gov.uk/government/publications/online-choice-architecture-how-digital-design-can-harm-competition-and-consumers>
- [23] Federal Trade Commission. (2023). FTC finalizes "click to cancel" rule making it easier for consumers to end subscriptions. <https://www.ftc.gov/news-events/news/press-releases/2023/10/ftc-finalizes-click-to-cancel-rule-making-it-easier-consumers-end-subscriptions>
- [24] Swearngin, A., Myers, B. A., & Landay, J. A. (2019). Modeling mobile interface tappability using

- crowdsourcing and deep learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. <https://doi.org/10.1145/3290605.3300305>
- [25] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>
- [26] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR.
- [27] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- [28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates.
- [29] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [30] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). <https://doi.org/10.1109/ICCV.2017.74>
- [31] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision* (pp. 839–847). IEEE. <https://doi.org/10.1109/WACV.2018.00097>
- [32] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates.
- [33] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
- [34] Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv*. <https://doi.org/10.48550/arXiv.1806.08049>
- [35] Raju, S. H., Waris, S. F., Adinarayana, S., Jadala, V. C., & Subba Rao, G. (2021). Smart dark pattern detection: Making aware of misleading patterns through the intended app. In S. Shakya, V. E. Balas, S. Kamolphiwong, & K.-L. Du (Eds.), *Sentimental analysis and deep learning: Proceedings of ICSADL 2021* (pp. 933–948). Springer.
- [36] Nazarov, D., & Baimukhambetov, Y. (2022). Clustering of dark patterns in the user interfaces of websites and online trading portals (E-commerce). *Mathematics*, 10(18), Article 3219. <https://doi.org/10.3390/math10183219>
- [37] Yada, Y., Feng, J., Matsumoto, T., Fukushima, N., Kido, F., & Yamana, H. (2022). Dark patterns in e-commerce: A dataset and its baseline evaluations. In *Proceedings of the 2022 IEEE International Conference on Big Data* (pp. 3015–3022). IEEE. <https://doi.org/10.1109/BigData55660.2022.10020800>
- [38] Mansur, S. M. H., Salma, S., Awofisayo, D., & Moran, K. (2023). AidUI: Toward automated recognition of dark patterns in user interfaces. In *2023 IEEE/ACM 45th International Conference on Software Engineering* (pp. 1958–1970). IEEE. <https://doi.org/10.1109/ICSE48619.2023.00166>
- [39] Srinivas Babu, K., Nithin, B. S., Raghuram, E., & Bharath, A. (2025). Dark Surfer – A dark pattern analyzer & classifier. *ITM Web of Conferences*, 74, Article 01018. <https://doi.org/10.1051/itmconf/20257401018>
- [40] Król, K. (2026). Automated detection of deceptive design patterns on university websites: A comparative analysis of browser-based tools and LLM-based approaches. *Applied Sciences*, 16(9), Article 4543. <https://doi.org/10.3390/app16094543>
- [41] Vedhapriyavadhana, R., Bharti, P., & Chidambaramathan, S. (2025). Detecting dark patterns in shopping websites: A multi-faceted approach using Bidirectional Encoder Representations from Transformers (BERT). *Enterprise Information Systems*, 19(5–6), Article 2457961. <https://doi.org/10.1080/17517575.2025.2457961>

- [42] Cuvin, P., Zhu, H., & Yang, D. (2025). DECEPTICON: How dark patterns manipulate web agents. *arXiv*. <https://doi.org/10.48550/arXiv.2512.22894>
- [43] Gray, C. M., Mildner, T., & Gairola, R. (2025). Getting trapped in Amazon's "Iliad Flow": A foundation for the temporal analysis of dark patterns. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3706598.3713828>
- [44] Mildner, T., Fidel, D., Stefanidi, E., Woźniak, P. W., Malaka, R., & Niess, J. (2025). A comparative study of how people with and without ADHD recognise and avoid dark patterns on social media. In *CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM. <https://doi.org/10.1145/3706598.3713776>
- [45] Singh, V., Vishvakarma, N. K., & Kumar, V. (2024). Tracing the origins of manipulation: Modeling the enablers behind dark patterns usage in e-commerce through TISM and MICMAC analysis. Global Knowledge, Memory and Communication. *Advance online publication*. <https://doi.org/10.1108/GKMC-10-2023-0386>
- [46] Ruohonen, J., Koskinen, J., Klausen, S. H., & Gerdes, A. (2025). A scenario analysis of ethical issues in dark patterns and their research. *arXiv*. <https://doi.org/10.48550/arXiv.2503.01828>
- [47] Järvinen, P. (2007). Action research is similar to design science. *Quality & Quantity*, 41(1), 37–54. <https://doi.org/10.1007/s11135-005-5427-1>
- [48] Booch, G., Maksimchuk, R. A., Engle, M. W., Young, B. J., Conallen, J., & Houston, K. A. (2007). Object-oriented analysis and design with applications (3rd ed.). Addison-Wesley Professional.
- [49] Rumbaugh, J., Jacobson, I., & Booch, G. (2004). The unified modeling language reference manual (2nd ed.). Addison-Wesley.
- [50] Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., & Ravikumar, P. (2019). On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates.
- [51] Ridnik, T., Ben-Baruch, E., Noy, A., & Zelnik-Manor, L. (2021). ImageNet-21K pretraining for the masses. In *Advances in Neural Information Processing Systems* (Vol. 35). Curran Associates.
- [52] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.
- [53] Gildenblat, J., & contributors. (2021). PyTorch library for CAM methods [Computer software]. GitHub. <https://github.com/jacobgil/pytorch-grad-cam>

Citation of this Article:

Tarilatei Mao Claudius, N. D. Nwiabu, & C. G. Igiri. (2026). An Explainable Web-Based Model to Identify Dark Pattern on E-Commerce User Interfaces. *Journal of Artificial Intelligence and Emerging Technologies (JAJET)*. 3(10), 14-24. Article DOI: <https://doi.org/10.47001/JAIET/2026.310003>

*** End of the Article ***