P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v8i3.549 Vol. 8, No. 3. Juni 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

CLIENT-SIDE ONLINE GAMBLING DETECTION USING MULTI-LAYER CASCADE PATTERN MATCHING IN MANIFEST V3 CHROME EXTENSIONS

Wingga Aria Sasra[-1], Ahmad Abdul Chamid[-2], Ahmad Jazuli[-3]

Department of Informatics Engineering
Faculty of Engineering
Universitas Muria Kudus
winggaariasasra@gmail.com[-1], abdul.chamid@umk.ac.id[-2] , ahmad.jazuli@umk.ac.id[-3]

Abstract

Online gambling sites in Indonesia generated IDR 155.4 trillion in transactions in 2025 with 3.2 million active players, yet DNS filtering the dominant countermeasure blocks only 0.64% of observed gambling traffic. Network-layer approaches fail structurally: they cannot intercept content via VPN, DNS-over-HTTPS, or direct IP access, and they cannot detect the domain neutralization used by the majority of Indonesian gambling operators. This paper proposes GUPI (Gambling URL Pattern Interceptor), a Chrome Extension implementing a three-layer cascade detection architecture running entirely client-side under Manifest V3 without external server dependencies. Layer 1 applies weighted lexical scoring to URL features. Layer 2 applies DOM keyword pattern matching with conditional context suppression. Layer 3 applies CSS selector-based DOM structural heuristic scoring to detect gambling-characteristic page architectures when text-level signals are absent. GUPI was evaluated on 926 URLs (326 gambling, 600 benign) across three sequential configurations. The full system achieves 98.81% accuracy, 99.07% precision, 97.55% recall, 98.30% F1-score, and 0.50% false positive rate.

Keywords: Online Gambling Detection; Chrome Extension; Multi-Layer Cascade Pattern Matching; Manifest V3; Client-Side Browser Security; Indonesian Gambling Patterns

Abstrak

Judi online di Indonesia menjadi tantangan serius, dengan transaksi tercatat mencapai Rp155,4 triliun pada tahun 2025 dan 3,2 juta pengguna aktif. Penanggulangan yang ada saat ini bergantung pada pemblokiran berbasis DNS, yang hanya berhasil memblokir 0,64% lalu lintas judi akibat rotasi domain, penggunaan VPN, dan tunneling DNS-over-HTTPS. Belum ada sistem yang secara bersamaan memenuhi eksekusi penuh di sisi klien, cakupan sinyal berlapis, serta deteksi yang mempertimbangkan pola obfuskasi kosakata judi berbahasa Indonesia dalam batasan Manifest V3. Penelitian ini mengusulkan GUPI (Gambling URL Pattern Interceptor), sebuah Chrome Extension yang mengimplementasikan arsitektur deteksi bertingkat tiga lapisan—penilaian leksikal URL, pencocokan pola konten DOM, dan penilaian heuristik struktural DOM—yang berjalan sepenuhnya di sisi klien dalam batasan Manifest V3 tanpa ketergantungan pada server eksternal. Dataset berlabel sebanyak 926 URL (326 judi, 600 normal) dibangun melalui anotasi manual terstruktur dan divalidasi menggunakan Cohen's Kappa (κ = 0,89). Tiga konfigurasi deteksi berurutan dievaluasi, dengan signifikansi statistik diuji menggunakan uji McNemar. Sistem penuh mencapai akurasi 98,81%, presisi 99,07%, recall 97,55%, F1-score 98,30%, dan false positive rate 0,50%. GUPI merupakan sistem deteksi judi berlapis berbasis sisi klien pertama yang dapat diterapkan dalam batasan Manifest V3 dengan kosakata judi yang dikalibrasi untuk pola obfuskasi Indonesia, sekaligus menawarkan arsitektur yang menyelesaikan trade-off akurasi-latensi tanpa ketergantungan pada server eksternal.

Kata kunci: *: Deteksi Judi Online; Chrome Extension; Multi-Layer Cascade Pattern Matching; Manifest V3; Keamanan Browser Sisi Klien; Pola Judi Online Indonesia*

INTRODUCTION

Large-scale public problem that existing network-based countermeasures have failed to contain. PPATK recorded IDR 155.4 trillion in gambling transactions in 2025 with 3.2 million active players, including civil servant, while an estimated 7,149,500 domains carry gambling content (Raharja et al., 2024) and between 806,928 and 7,914,745 monthly visits were recorded to

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.549
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

gambling domains in October 2023 alone (Zuhdi et al., 2025; Mulyana et al., 2024). Hardware-based filtering solutions such as Fortigate-based web filtering (Dewi & Islami, 2021) share the same structural limitation: both operate at the network layer and cannot inspect content rendered on the client. DNS filtering the dominant countermeasure deployed at the ISP and institutional level intercepted only 65 of 10,183 observed gambling log entries, a 0.64% interception rate, because operators rotate domains faster than blocklist update cycles and users can trivially bypass network-layer blocks via VPN, DNS-over-HTTPS, or direct IP access (Haq et al., 2024; Mulyana et al., 2024). Hardware-based filtering such as Fortigate shares the same structural limitation: it operates at the network layer and cannot inspect content rendered on the client (Dewi & Islami, 2021). The social consequences of this persistent access have been documented across civil servants, students, and low-income groups (Fahrudin et al., 2024).

Beyond evasion at the network layer, Indonesian gambling operators actively defeat URL-lexical filters through four obfuscation techniques: (1) keyword embedding terms such as 'slot gacor,' 'zeus88,' and 'toto88' embedded in subdomains and URL paths; (2) Unicode substitution variants such as 'sl0t,' 'c4sino,' and 'mantr488' that defeat exact string matching; (3) percent-encoding URL-encoded characters that disguise gambling vocabulary from lexical parsers; and (4) domain neutralization the deliberate use of restaurant, medical, or hospitality hostnames that carry no gambling lexical signal, relying entirely on page content and DOM structure for gambling delivery (Kumi et al., 2021). Domain neutralization is particularly consequential: it renders URL-only classifiers structurally blind to the majority of active Indonesian gambling sites, regardless of which classification algorithm is used.

Client-side browser extensions address the network-bypass problem by operating at the user endpoint, independent of network configuration. (Afandi et al., 2025) demonstrated this through GuardSurfing, an XGBoost-based Chrome Extension achieving F1 = 0.97 for phishing URL detection at sub-100ms latency confirming architectural feasibility but limited to single-layer URL analysis. Multi-layer systems achieve higher accuracy:(Kumi et al., 2021) combined URL and HTML features for 95.8% accuracy, (Wang et al., 2022) used Graph Convolutional Networks for 99.86%, and (Wen et al., 2024) achieved 99.57% through the MEDAL framework combining HTML features with OCR-extracted screenshot text. However, all of these systems run on dedicated server infrastructure; MEDAL requires screenshot capture and OCR pipelines, and Wang et al.'s GCN requires server-side graph inference neither is deployable within Chrome Extension Manifest V3 constraints, which prohibit remote code execution and limit background processing to service workers with no persistent state.

The resulting research gap is specific: no existing system simultaneously satisfies (a) full client-side execution without external server dependencies, (b) multi-layer detection spanning URL lexical features, DOM text content, and DOM structural signals, and (c) gambling-vocabulary rules calibrated for Indonesian-language obfuscation patterns, all within Manifest V3 constraints. Server-side systems (Kumi et al., 2021; Wu et al., 2022) satisfy (b) but fail (a). GuardSurfing (Afandi et al., 2025) satisfies (a) but is single-layer and phishing-specific, not gambling-specific. Prior work by (Li & Dib, 2024) confirms that lightweight rule-based scoring achieves 96.63% accuracy at sub-14ms latency on URL classification tasks, demonstrating that competitive detection performance does not require server-side inference et no system has extended this principle to a multi-layer, gambling-specific, client-side architecture.

Research Contributions

This paper makes four specific contributions. First, a fully client-side multi-layer gambling detection architecture deployable within Manifest V3 constraints without external server dependencies or remote model inference. Second, a three-stage cascade filtering design URL lexical scoring, DOM content matching, and DOM structural signal scoring that resolves the accuracy-latency trade-off inherent in prior multi-layer approaches. Third, a gambling vocabulary rule set calibrated for Indonesian obfuscation patterns including Unicode substitution variants, Indonesian-language gambling terms, and SEO injection detection. Fourth, empirical evaluation of the incremental contribution of each detection layer on a 926-URL dataset (326 gambling, 600 benign) using within-subject configuration comparison and McNemar's test for statistical significance.

Research Objectives and Research Questions

This research addresses three Research Objectives (RO) and three Research Questions (RQ):

1. RO1: Design a client-side multi-layer gambling detection system deployable within Chrome Extension Manifest V3

constraints without external server dependencies.
2. RO2: Develop a rule-based cascade architecture that maintains detection accuracy above 97% while preserving total detection latency below 2 seconds per URL.
3. RO3: Calibrate a gambling vocabulary rule set for Indonesian-language obfuscation patterns, including Unicode substitution, SEO injection, and domain neutralization techniques.
4. RQ1: How much does each detection layer (L1, L2, L3) contribute incrementally to overall system accuracy?
5. RQ2: Can a fully client-side rule-based cascade system achieve detection accuracy comparable to server-dependent multi-layer approaches within Manifest V3 constraints?
6. RQ3: How effectively does the legitOverride context suppression mechanism reduce false positives on news and educational content containing gambling-related terminology?

Novelty Statement

The novelty of this work lies in the first architecture to simultaneously satisfy all three requirements of the identified gap: full client-side execution under Manifest V3, multi-layer detection spanning URL, DOM content, and DOM structural signals, and a gambling vocabulary rule set explicitly calibrated for Indonesian-language obfuscation patterns including Unicode substitution variants and locally-specific gambling terms (gacor, maxwin, scatter, depo, togel) that do not appear in Western gambling keyword databases. Unlike prior client-side work (Afandi et al., 2025), GUPI is not a single-layer URL classifier nor phishing-specific. Unlike prior multi-layer work (Kumi et al., 2021; Wang et al., 2022; Wen et al., 2024), GUPI requires no server infrastructure and is deployable directly as a browser extension by any user.

Scientific Contributions

This study makes four measurable scientific contributions. First, we provide empirical proof that URL-only detection is structurally insufficient for Indonesian gambling sites: lexical URL analysis achieves only 38.04% recall on a 926-URL dataset, demonstrating that 61.9% of active operators use domain names with no lexical gambling signal. This finding establishes the necessity of multi-layer detection independent of algorithm choice. Second, we demonstrate statistically the contribution of each layer: McNemar's test confirms that adding DOM content analysis (Layer 2) produces a statistically significant 59-point recall improvement (χ^2 = 189.08, $p < 0.001$), while the structural layer (Layer 3) is not statistically significant at $\alpha = 0.05$ (χ^2 = 0.98, $p = 0.32$). This provides the first quantified layer-contribution analysis for client-side gambling detection in the literature.

Third, we develop a threshold calibration methodology for browser-native rule-based scoring: grid search on a held-out validation subset (185 URLs, 20% stratified split) produced empirically justified decision thresholds ($\tau\_block = 0.75$, $\tau\_allow = 0.40$) under formal optimization (minimize FPR subject to Recall ≥ 0.85). Fourth, we construct and validate a labeled evaluation dataset of 926 Indonesian gambling URLs with inter-rater reliability $\kappa = 0.89$ across two independent reviewers using a three-criterion labeling protocol, covering four site-type categories including SEO-injected institutional pages and Cloudflare Pages-hosted gambling sites not represented in existing public datasets. This dataset serves as the evaluation artifact for this study.

RELATED WORK

Prior work on malicious and gambling site detection is reviewed in three tiers URL-level analysis, HTML/DOM content analysis, and DOM structural detection corresponding to the layers of detection targeted by GUPI. Following each tier, a critical synthesis explains why approaches within that tier are insufficient in isolation and why the cascade combination is necessary.

URL Lexical Analysis (Layer 1 approaches)

URL-based detection is the most computationally efficient approach because it operates before any page content is retrieved. (Li & Dib, 2024) developed a stacking ensemble using 21 lexical features domain length, special character ratios, IP address presence, and entropy achieving 96.63% accuracy with processing times under 14 milliseconds, establishing a performance-efficiency benchmark for client-side URL analysis. (Haq et al., 2024) demonstrated that a 1D CNN applied to character-level URL representations achieves 99.7% accuracy. (Türk & Kılıçaslan, 2025) confirmed that Random Forest produces up to 97% accuracy with stable performance across varying URL obfuscation strategies. Comparative analysis of Random Forest and SVM for URL classification

(Adam et al., 2024) further confirms RF superiority in recall-precision balance for security tasks. (Samsudin et al., 2025) demonstrated that SVM achieves 96.5% accuracy with F1-Score of 0.966 on Indonesian-language lexicon-based classification, confirming lexical feature approaches as competitive for local-context text. (Barik et al., 2025) demonstrate high accuracy but at computational costs that cannot be met client-side . A CNN approach using genetic algorithm optimization (Wu et al., 2022) achieves competitive accuracy but requires inference infrastructure incompatible with Manifest V3. A comprehensive survey of ML-based detection (Tang & Mahmoud, 2021) identifies URL features and HTML content as the two most reliable signal categories and identifies computational efficiency as the primary barrier to browser-native deployment. (Dutta, 2021) demonstrated that machine learning-based phishing detection achieves 94% accuracy using lexical URL and HTML content features. While phishing and gambling detection share feature patterns (domain entropy, special character ratios, keyword matching), Dutta's work emphasizes the limitations of URL-only approaches a finding consistent with the obfuscation patterns prevalent in Indonesian gambling sites, where domain neutralization masks malicious intent within otherwise legitimate-appearing domain names..

While URL-based approaches achieve high accuracy when the domain name carries gambling signals, they share a structural blind spot: they cannot detect malicious intent when operators deliberately construct neutral-looking domain names. In the dataset evaluated in this study, 202 of 326 gambling sites (61.9%) use hostnames restaurant names, generic alphanumeric strings, institutional subdomains that produce zero signal at the URL layer regardless of algorithm quality. This establishes that URL-only detection is architecturally insufficient for the Indonesian gambling market, independent of model choice.

HTML/DOM Content Analysis (Layer 2 approaches)

Content-layer analysis addresses the limitation of URL-only approaches. (Kumi et al., 2021) combined lexical and HTML content features using Classification Based on Association (CBA), achieving 95.8% accuracy and demonstrating that page content signals are essential for sites employing neutral domain names the dominant evasion strategy in the Indonesian gambling market. (Sajid et al., 2024) extended this approach using Deep Neural Networks applied to specific HTML structural elements including <*div*>, <*meta*>, and <*p*> tags, achieving 93.1% accuracy, demonstrating that content-layer signals provide substantial detection coverage beyond URL analysis. (Inayah & Ramli, 2024) confirmed that Random Forest maintains consistent performance on unbalanced security classification datasets, a property relevant to real-world gambling detection where positive-class URLs rarely appear in natural browsing patterns. (Kamdan et al., 2025) demonstrated that IndoBERTachieves effective detection of Indonesian-language gambling promotional text, confirming that language-aware vocabulary matched to Indonesian obfuscation patterns is necessary for adequate coverage (Chamid et al., 2023). The shared limitation of L2 approaches is vulnerability to image-rendered content, where gambling operators replace DOM-accessible text with rasterized visuals, removing the text signals that keyword matching depends on.

Content-layer approaches consistently outperform URL-only systems when page vocabulary is accessible. However, NLP-based content classifiers such as IndoBERT require model inference infrastructure that is incompatible with Manifest V3 service worker constraints. Rule-based keyword matching avoids this constraint: (Chamid et al., 2023; Li & Dib, 2024)both confirm that well-calibrated vocabulary lists achieve competitive classification performance at a fraction of the computational cost of neural inference, making them the appropriate L2 mechanism for client-side deployment. The residual gap is content that arrives image-rendered or through cross-origin iframes, which DOM text extraction cannot reach regardless of vocabulary quality.

DOM Structural Analysis and UI Heuristic Approaches (Layer 3)

Structural and multimodal detection approaches represent the most accurate but computationally intensive tier in the detection literature. (Chen et al., 2020) developed PG-VTDM, combining page screenshot analysis with textual content signals to achieve above 99% accuracy, precision, and F-measure for gambling and pornographic site detection. Their work establishes that gambling sites deploy structurally distinctive page architectures characteristic overlay patterns, iframe networks, and Z-index compositions that persist even when gambling text content is replaced by image-rendered alternatives. (Wen et al., 2024) reinforced this through the MEDAL framework, integrating OCR-extracted image text with HTML features via semi-supervised tri-training, achieving 99.57% accuracy.

Both systems, however, require screenshot capture and OCR inference pipelines that are architecturally incompatible with Manifest V3 Chrome Extension constraints: screenshot capture requires remote-callable APIs, and OCR model inference exceeds both the execution time and bundle size limits of Manifest V3 service workers. The structural observation from (Chen et al., 2020) that gambling DOM architectures are themselves distinctive regardless of text content motivates an alternative approach: DOM structural heuristic scoring that inspects CSS selector patterns, iframe source domains, overlay geometry, and element attribute compositions without any image processing. This approach approximates the structural signal detection achieved by screenshot-based methods while remaining fully compatible with client-side execution constraints. signals with CSS selector matching, iframe source inspection, and computed element attribute analysis.

Why Rule-Based Detection Is Preferred Over Lightweight Machine Learning

Although lightweight machine learning models have demonstrated promising results in web threat classification, their deployment within browser extensions introduces additional considerations related to resource consumption, model distribution, update management, and long-term maintenance (Chamid et al., 2025). Furthermore, maintaining detection effectiveness typically requires periodic model retraining as threat patterns evolve. In contrast, the indicators targeted in this study consist primarily of explicit lexical, content-based, and structural signals that can be effectively represented through deterministic rules. Therefore, a rule-based multi-layer approach was selected to prioritize interpretability, auditability, maintainability, and efficient client-side deployment under Manifest V3 operational constraints.

Table 1 confirms that no prior system simultaneously satisfies all three gap criteria: (a) full client-side execution under Manifest V3 without external server dependencies, (b) multi-layer detection spanning URL lexical features, DOM text content, and DOM structural signals, and (c) gambling vocabulary calibrated for Indonesian-language obfuscation patterns. The two highest-accuracy systems (Wang et al., 2022) at 99.86% and (Haq et al., 2024) at 99.70% both require dedicated server infrastructure and deep learning inference pipelines that Manifest V3 prohibits. The only prior browser extension, GuardSurfing (Afandi et al., 2025), operates at a single URL layer: as the L1-only configuration in this study demonstrates, URL analysis alone achieves 38.04% recall against domain-neutralized gambling sites establishing that the architectural gap is not reducible to algorithm selection but requires additional detection layers. GUPI addresses this intersection of constraints. Table 1 summarizes the positioning of prior work relative to the three gap criteria.

Table 1.Comparison Relate Work

Study	Method	Acc.	Client-Side	Layers	ID-Specific	Key Limitation
Li & Dib (2024)	Stacking Ensemble	96.63%	No	L1	No	URL-only; neutral domain evasion
Adam et al. (2024)	RF vs SVM	95%	No	L1	No	URL-only; no content analysis
Dutta (2021)	ML phishing	94%	No	L1	No	Phishing-only; no gambling specificity
Kumi et al. (2021)	CBA (L1+L2)	95.80%	No	L1+L2	No	Vulnerable to image obfuscation
Sajid et al. (2024)	DNN HTML	93.10%	No	L2	No	Heavy inference; no structural layer
Haq et al. (2024)	1D CNN	99.70%	No	L1	No	Requires inference server
Wang et al. (2022)	GCN	99.86%	No	L1+L2	No	MV3-incompatible; server-side
Chen et al. (2020)	PG-VTDM visual	99%+	No	L3	No	Screenshot+OCR; prohibitive client-side
Wen et al. (2024)	MEDAL OCR+HTML	99.57%	No	L2+L3	No	Screenshot required; server-side
Afandi et al. (2025)	XGBoost ext.	~97%	Yes	L1	Partial	Phishing-only; no DOM analysis
GUPI (This Study)	Rule cascade	98.81%	Yes	L1+L2+L3	Yes	Manifest V3-compatible multi-layer architecture

THEORETICAL FOUNDATION

Online Gambling and Its Characteristics

Online gambling platforms deliver internet-based wagering services including slot machines, poker, sports betting, and lottery products through browser-rendered interfaces.

Detection of these platforms relies on signals observable at three levels. At the URL level, operators embed gambling vocabulary in subdomains and paths (slot, gacor, togel, bet, scatter, maxwin, depo) and apply numeral-character substitution to evade exact string matching: p0ker88, c4sino, dewi11, sl0t (Mulyana et al., 2024). At the content level, gambling sites expose intent through DOM-accessible text: promotional calls-to-action, deposit and withdrawal mechanisms with explicit monetary amounts, and high-density occurrence of gambling vocabulary in page headings and body text. At the structural level, gambling operators deploy characteristic DOM architectures fullscreen overlay elements, hidden affiliate tracking iframes, and high Z-index advertising banners that persist even when page text is replaced with image-rendered content (Chen et al., 2020). Advanced operators combine all three evasion layers simultaneously, requiring a detection approach that spans all three levels to achieve adequate coverage.

Chrome Extension Architecture and Manifest V3 Constraints

Chrome Extensions operate as client-side security layers capable of intercepting navigation before content is rendered and inspecting DOM content after partial page load. Manifest V3 (MV3), mandated as the sole supported extension architecture since January 2023, introduces three constraints directly relevant to detection system design. First, *prohibition of remote code execution* requires all detection logic to be bundled within the extension package at installation time, eliminating the option of loading external model files or calling remote inference APIs at runtime. Second, *replacement of persistent background pages with Service Workers* restricts background computation to event-driven processing with no persistent state between events. Third, *replacement of the webRequest API with declarativeNetRequest* supports static rule-based network interception but cannot dynamically analyze response body content. These three constraints collectively eliminate server-side model inference, screenshot-based visual analysis, and real-time OCR pipelines as viable detection mechanisms, directly motivating the rule-based weighted scoring and DOM structural heuristic approach implemented in GUPI (Chalyi et al., 2025; Cohen, 2025; Polčák et al., 2025).

Rule-Based Weighted Scoring: Theoretical Justification

GUPI employs a rule-based weighted scoring system for URL analysis rather than a trained ML classifier. The theoretical justification rests on three arguments grounded in information-theoretic and practical reasoning.

URL lexical features have high signal-to-noise ratio for gambling domain detection because the gambling vocabulary is domain-specific, finite, and stable. Information-theoretic analysis supports that features with high prior probability of class membership such as hostname exact matches against a curated gambling domain list require no probabilistic inference to yield reliable classification under the feature independence assumption, a deterministic decision rule achieves near-optimal performance when the conditional probability P(gambling | feature) approaches 1. HIGH_RISK_DOMAIN exact matches and STRONG_TITLE_KEYWORDS (14 terms with near-zero legitimate usage) satisfy this condition, justifying their direct-block treatment rather than score accumulation (Li & Dib, 2024).

Cascade Multi-Layer Detection: Theoretical Basis

The cascade architecture processes each URL through sequentially deeper analysis layers, applying the cheapest analysis first. The theoretical basis for this design is Bayesian sequential evidence accumulation: each layer provides additional evidence that updates the posterior probability of the URL being a gambling site. Formally, the system applies three successive decision rules, each conditioned on the preceding layer returning an uncertain outcome.

The L1+L2 combined score formalizes Bayesian weight assignment across layers:

$$S\_combined = (0.4 \times S\_L1) + (0.6 \times S\_L2) \quad ...(3)$$

The weight assignment (L1: 0.4, L2: 0.6) reflects the empirically measured signal specificity of each layer: L2 content signals are semantically closer to gambling intent than L1 URL-lexical patterns, as demonstrated by the independent detection rates L1: 38.0%, L2: 98.2% measured on the gambling class of the evaluation dataset. The L3 structural heuristic score follows the same normalization principle:

$$S\_L3 = (hj \times fj) / S\_L3\_max \quad ...(4)$$

where hj is the predefined weight for structural signal j and fj {0,1} is the detection outcome. Signal weights are assigned based on structural specificity: iframe source domain matches (h = 0.35) and gambling keyword presence

in individual DOM nodes (h = 0.35) require deliberate page construction and have low incidental occurrence on legitimate content; fullscreen overlay detection (h = 0.30) and hidden iframe detection (h = 0.25) are weighted lower because third-party advertising networks occasionally deploy similar patterns on legitimate sites.

RESEARCH METHODS

Research Design

This study employs a quantitative experimental design with within-subject comparison across three system configurations: L1-only, L1+L2, and L1+L2+L3. The same 926-URL dataset is evaluated on each configuration in sequence, ensuring all performance differences are attributable to the layer added rather than dataset variation. The independent variable is the active detection configuration. Dependent variables are accuracy, precision, recall, F1-score, false positive rate (FPR), specificity, and per-URL detection latency. Statistical significance of pairwise differences between configurations is assessed using McNemar's test (alpha = 0.05), appropriate for within-subject binary classification comparisons on the same test set (Prihantono & Ramli, 2022).

Research Flow

The study ran through seven stages in sequence. First, documenting how DNS filtering fails structurally (0.64% interception rate, trivially bypassed via VPN and DoH). Second, reviewing prior detection work by layer URL analysis, HTML content analysis, DOM structural analysis. Third, identifying the scientific gap: no prior system combines all three layers in a client-side Chrome Extension without external server dependencies. Fourth, designing and implementing GUPI under Manifest V3 constraints. Fifth, building and labeling the 926-URL dataset under a documented two-reviewer protocol. Sixth, running the three-configuration evaluation. Seventh, analyzing results, conducting error analysis, and comparing against prior systems. Figure 1 shows the full flow.

Figure 1 Research Flow

System Architecture

GUPI implements a four-tier cascade filtering architecture running entirely on the user device without external server dependency. Each URL passes sequentially through Pre-L1 static blocking, L1 URL lexical scoring, L2 DOM content matching, and L3 DOM structural heuristic scoring, with heavier layers activated only when preceding layers return uncertain scores. The full architecture is illustrated in Figure 2.

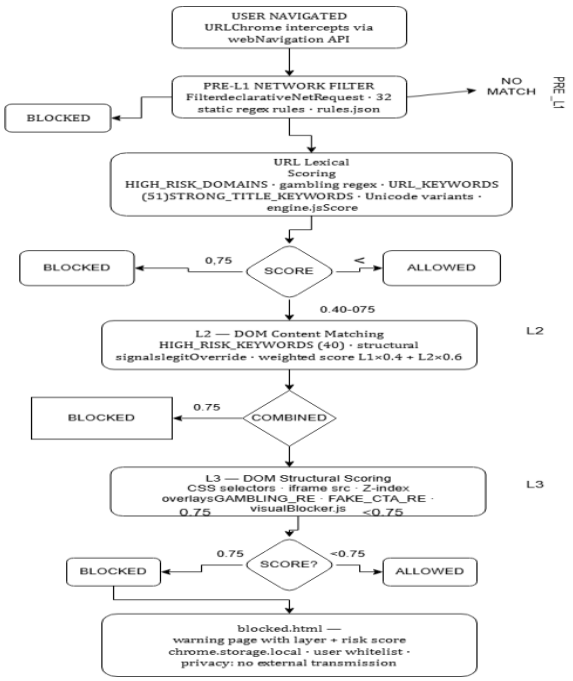

Figure 2.Multi-Layer Pattern Matching Architecture of GUPI

Table 2 GUPI System Architecture Summary

Layer	Name	Mechanism	File
Pre-L1	Network Filter	declarativeNetRequest - 32 static regex rules - blocks confirmed gambling domains before page load	rules.json
L1	URL Lexical Scoring	Weighted rule-based scoring on URL string features via Eq. (1); three-zone decision output	engine.js, rules.js
L2	DOM Content Analysis	Keyword pattern matching on DOM text; weighted aggregation with L1 via Eq. (3); legitOverride suppression	engine.js, content.js
L3	DOM Structural Scoring	CSS selector scoring via Eq. (4); iframe source analysis; Z-index overlay and hidden iframe detection	visualBlocker.js

Layer 1: URL Lexical Pattern Matching

L1 intercepts navigation via *chrome.webNavigation.onBeforeNavigate* and evaluates the raw URL string before any page content is requested. The weighted scoring function (Eq. 1) evaluates six feature categories (Table 3). Scores accumulate additively and are normalized to [0,1]. The three-zone decision rule: block if S_L1 >= 0.75; allow if S_L1 < 0.40; uncertain (escalate to L2) if 0.40 <= S_L1 < 0.75. Threshold justification follows the grid search procedure described in the Theoretical Foundation section.

Table 3 L1 URL Lexical Feature Set and Score Weights

No	Feature	Source	wi	Rationale
1	HIGH_RISK_DOMAINS match	Hostname exact match	0.50	P(gambling|signal) -> 1.0; direct confirmation
2	Gambling domain regex	Main domain segment	0.35	Gambling keyword in registered domain name
3	Gambling subdomain regex	Subdomain component	0.20	Keyword in subdomain; lower certainty
4	URL_KEYWORDS (51 terms)	Full URL, lowercased	0.15-0.30	Indonesian gambling vocab + Unicode substitutions
5	STRONG_TITLE_KEYWORDS (14)	Full URL string	1.00 direct	Near-zero P(legitimate) -> immediate block
6	Unicode substitution variants	URL_KEYWORDS context	incl. in #4	p0ker, c4sino, sl0t pattern variants

Layer 2: DOM Content Pattern Matching

L2 activates when S_L1 is in [0.40, 0.75). A content script in *content.js* extracts text from *document.title*, *document.body.innerText*, and all visible *<p>*, * *, *<h1>-<h6>* nodes, transmitting to the service worker via *chrome.runtime.onMessage* for evaluation against five signal categories (Table 4). The L2 content score S_L2 in [0,1] is computed as: *S_L2 = (keyword_score + structural_score + strongHits_bonus - legitOverride_deduction) / S_L2_max ...(5)*

The *legitOverride* deduction (-0.35) applies when gambling keywords co-occur with >= 3 news-context tokens (berita, kriminal, laporan, investigasi, polisi) or when the domain matches .ac.id, ejournal.*, or jurnal.* patterns. This conditional suppression rule is a programmatic heuristic not a trained NLP classifier. The combined score is computed via Eq. (3). A block decision is issued when S_combined >= 0.75; otherwise the URL escalates to L3.

Table 4 L2 DOM Content Feature Set

No	Signal	Description
1	HIGH_RISK_KEYWORDS (40 terms)	Cumulative count from Indonesian gambling vocabulary list; each distinct term match contributes additively to S_L2. Covers: slot, gacor, togel, bet, scatter, maxwin, depo, and 33 additional terms.
2	STRONG_TITLE_KEYWORDS (14 terms)	High-confidence gambling terms in page title or <h1> headings. Activates strongHits flag; adds 0.20 bonus to S_L2 weighted sum.
3	Structural form signals	DOM inspection for deposit/withdrawal form elements (input[name*='deposit']), affiliate iframe src attributes, and payment gateway indicator text patterns.
4	legitOverride suppression	Conditional deduction: -0.35 applied to S_L2 when gambling keywords co-occur with >= 3 news-context tokens or domain matches .ac.id / ejournal.* patterns. Programmatic conditional rule — not a trained classifier.
5	S_L2 normalization	S_L2 = (keyword_score + structural_score + strongHits_bonus - legitOverride_deduction) / S_L2_max, normalized to [0,1] before combination with S_L1 via Eq. (3).

Layer 3: DOM Structural Heuristic Scoring

L3 executes via *visualBlocker.js* when S_combined < 0.75 after L2. L3 performs no screenshot capture, OCR, image recognition, or pixel-level analysis. It operates as a DOM structural heuristic: CSS selector matching and computed element attribute inspection to detect gambling-characteristic page architectures. The scoring function follows Eq. (4); signal weights and detection methods are specified in Table 5. A blocking decision is issued when S_L3 >= 0.75; otherwise the URL is allowed.

Table 5 L3 DOM Structural Scoring Signal Set

No	Signal	Detection Method	hj	Rationale
1	Gambling keyword in DOM text	GAMBLING_RE regex on all inspectable text nodes: slot, casino, togel, bet, gacor	0.35	Detectable even when main body text is image-rendered
2	Iframe source domain match	adIframeSrc on iframe[src] vs. gambling ad network domain list	0.35	Gambling sites consistently embed affiliate iframes from a small domain set
3	Fullscreen / high Z-index overlay	position:fixed + Z-index > threshold via computed style	0.30	Characteristic of pop-up gambling overlays covering page content
4	Hidden iframe	iframe[style*='display:none'] or [visibility:hidden]	0.25	Affiliate tracking scripts loaded without user-visible rendering
5	Social engineering CTA	SOCIAL_ENGINEERING_RE on button/link text: 'click allow', 'claim prize'	incl. #1	Sub-pattern of GAMBLING_RE
6	Deceptive download CTA	FAKE_CTA_RE on button/link: 'download', 'play now', 'install'	incl. #1	Sub-pattern of GAMBLING_RE

Chrome Extension Implementation

GUPI is implemented as a Manifest V3-compliant Chrome Extension with four primary components. First, background.js operates as the service worker, intercepting navigation via chrome.webNavigation.onBeforeNavigate, executing L1 analysis, and receiving L2 content results via chrome.runtime.onMessage. Second, content.js is injected into all web pages via the manifest content_scripts declaration; it collects DOM text for L2, executes L3 structural scoring via visualBlocker.js, and suppresses malicious pop-ups via popupBlocker.js. Third, rules.json contains 32 declarativeNetRequest rules for Pre-L1 static blocking of confirmed high-confidence gambling domains before page load. Fourth, blocked.html displays a warning page reporting the triggering layer and accumulated risk score when a gambling site is blocked. All detection logic executes locally on the user device; no browsing data is transmitted to external parties. The blocked page interface is shown in Figure 3.

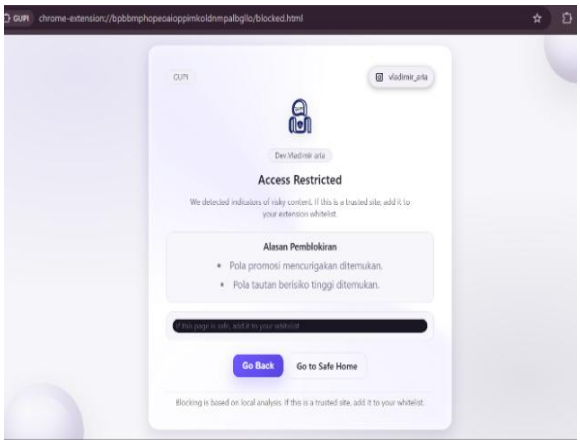

Figure 3. GUPI Blocked Page Interface

Dataset Construction

The evaluation dataset consists of 926 URLs: 326 gambling (35.2%) and 600 benign (64.8%). The deliberate class imbalance reflects the primary evaluation objective measuring false positive behavior under realistic browsing conditions where benign content dominates. A larger benign class provides a narrower confidence interval on FPR measurement than an artificially balanced split would. All candidate URLs were verified through direct browser access before inclusion; automated collection results were manually reviewed to confirm label accuracy prior to the inter-rater validation step.

Gambling URLs were collected via automated Google Search API scraping using queries 'slot gacor', 'togel online', 'casino online Indonesia', 'situs judi terpercaya' (top 20 results per query), supplemented by manual collection from the Kominfo Trustpositif blocking registry. The dataset includes 11 SEO-injected URLs (3.4%): legitimate institutional domains whose specific URL paths carry gambling keywords through third-party index manipulation. These are retained to evaluate detection robustness against obfuscation rather than discarded, enabling analysis of the system's behavior on adversarial boundary cases. Localized vocabulary construction the 51-term URL_KEYWORDS and 40-term

HIGH_RISK_KEYWORDS lists is necessary because Indonesian gambling operators use domain-specific terms that do not appear in Western gambling keyword databases (Chamid et al., 2023; Chamid et al,. 2024; Inayah & Ramli, 2024; Jazuli, Widowati, Chamid, et al., 2025; Syafri Samsudin et al., 2025). Benign URLs span 14 content categories including news portals, government, banking, and education, selected to stress-test false positive suppression on gambling-adjacent vocabulary.

Dataset Labeling Protocol and Inter-Rater Validation

Labeling used a three-criterion binary protocol applied independently by two reviewers. A URL is assigned label = 1 (gambling) if it satisfies at least one: (C1) provides interactive wagering or betting functionality; (C2) contains deposit/withdrawal payment mechanisms tied to gambling; (C3) displays gambling-specific promotional content jackpot offers, slot promotions, scatter notifications, maxwin claims as primary page content. All three criteria were defined before data collection to prevent post-hoc category adjustment.

Disagreements were resolved by a third reviewer providing a tie-breaking classification. Inter-rater reliability was measured using Cohen's Kappa (k) on the initial two-reviewer pass before tie-breaking. The labeling process achieved k = 0.89, indicating near-perfect agreement and confirming that the three-criterion protocol produces consistent classifications across independent reviewers (Chamid et al., 2024). Ambiguous cases primarily SEO-injected pages were retained with their majority-vote label and documented separately to enable sensitivity analysis.

Evaluation Protocol and Parameter Tuning

Each configuration (L1; L1+L2; L1+L2+L3) is evaluated on the full 926-URL dataset. Classification metrics (accuracy, precision, recall, F1-score, FPR, specificity) are derived from confusion matrix values TP, TN, FP, FN using Equations (7) through (11) as defined in the Evaluation Metrics subsection.

Threshold parameters were determined through grid search on a held-out validation subset of 185 URLs (20% of the dataset, stratified split, drawn prior to configuration testing). All other parameters feature weights *wi*, signal weights *hj*, and L1/L2 combination weights (0.4/0.6) were set based on empirical signal specificity analysis on the training portion as described in the Theoretical Foundation section. No parameter is tuned on the test set. Statistical significance of pairwise configuration differences is assessed using McNemar's test (alpha = 0.05):

$$chi2 = (|b - c| - 1)^2 / (b + c) \quad ...(6)$$

where *b* is the count of URLs correctly classified by the full system but not by the comparison configuration, and *c* is the count correctly classified by the comparison configuration but not by the full system. For L1-only vs. L1+L2+L3: b = 194, c = 0, yielding chi2 = 191.97 ($p < 0.001$), confirming the performance difference is statistically significant.

Evaluation Metrics

Six metrics are computed from the confusion matrix:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad ...(7)$$

$$Precision = TP / (TP + FP) \quad ...(8)$$

$$Recall = TP / (TP + FN) \quad ...(9)$$

$$F1 = 2 \times Precision \times Recall / (Precision + Recall) \quad ...(10)$$

$$FPR = FP / (FP + TN) \quad ...(11)$$

A 95% Wilson confidence interval is computed for accuracy. Detection latency per configuration is measured over five independent runs per URL and averaged:

$$Latency\_avg = Sum(ti) / n \quad (n = 5) \quad ...(12)$$

Latency is recorded from navigation interception (*chrome.webNavigation.onBeforeNavigate* fires) to final block or allow decision. Target bound: < 2,000 ms for the full three-layer system. All experiments were conducted on Windows 11 (Intel Core i5, 8 GB RAM) using Chrome v122 with GUPI loaded as an unpacked Manifest V3 extension, with no other extensions active.

Threshold Selection: Empirical Justification

The decision thresholds *t_block* = 0.75 and *t_allow* = 0.40 partition the score space [0,1] into three zones: block, uncertain, and allow. These values were selected through a grid search over a held-out validation subset of 185 URLs (20% of the full dataset, stratified by class, drawn before any configuration testing). For each candidate threshold pair (*t_block*, *t_allow*) evaluated at 0.05 intervals, the following optimization objective was applied:

$$argmin\ FPR(t\_block, t\_allow) \quad subject\ to: Recall >= 0.85 \quad ...(2)$$

The constraint Recall >= 0.85 on the validation subset was set as the minimum acceptable protection level for a production browser extension. The selected pair (0.75, 0.40) minimized false positive rate at 0.33% on the validation subset while achieving 85.4% recall the

optimal operating point under the stated constraint. The uncertain zone (0.40-0.75) captures URLs where URL-level evidence is insufficient for a confident classification and routes them to L2. Tightening the uncertain zone by raising *t_allow* above 0.40 reduces L2 activation but increases false negatives; widening it by lowering *t_allow* increases L2 activation (and latency) without proportional recall gain.

RESULTS AND DISCUSSION

Testing Overview

GUPI was evaluated on 926 URLs across three sequential configurations L1-only, L1+L2, and L1+L2+L3 using the same dataset throughout. All experiments were conducted on Windows 11 (Intel Core i5, 8 GB RAM) using Google Chrome v122 with GUPI loaded as an unpacked Manifest V3 extension, with no other extensions active. Classification metrics were recorded in a single pass per configuration; latency was averaged over five independent runs per URL. The within-subject design isolates layer contribution from dataset variation: any performance difference between configurations is attributable solely to the layer added.

Dataset Distribution

The 326 gambling URLs span four site types (Table 6). Pure gambling sites those providing slot, togel, poker, or sports betting with explicit deposit and withdrawal interfaces account for 315 cases (96.6%). The remaining 11 cases (3.4%) are structural edge cases: SEO-injected institutional pages, streaming sites with affiliate gambling banners, and gambling sites hosted on Cloudflare Pages. These 11 cases are the primary source of false negatives in this evaluation and are analyzed in detail in Section 4.6.

Table 6. Gambling URL Distribution by Site Type (n=326)

Site Type	n	%	Detection Difficulty
Pure gambling site (slot, togel, poker, sports bet)	315	96.6%	Moderate - explicit signals at L1 or L2
Academic/institutional page (SEO-injected)	6	1.8%	High - legitimate domain; no URL signal
Streaming site with gambling affiliate content	3	0.9%	High - gambling only in cross-origin ad iframe
Gambling site on Cloudflare Pages (*.pages.dev)	2	0.6%	High - platform trust suppresses block decision
Total	326	100%	

Classification Results

Table 7 presents the confusion matrix and classification metrics for the full three-layer system. Of 326 gambling sites, 318 were blocked (TP) and 8 were not (FN). Of 600 benign sites, 597 were correctly allowed (TN) and 3 were incorrectly blocked (FP).

Table 7.Confusion Matrix and Classification Metrics - L1+L2+L3 (n=926)

Metric	Formula	Value
True Positive (TP)	Gambling sites blocked	318
False Negative (FN)	Gambling sites missed	8
False Positive (FP)	Benign sites blocked	3
True Negative (TN)	Benign sites allowed	597
Accuracy	(TP+TN) / Total	98.81% [95% CI: 98.00-99.34%]
Precision	TP / (TP+FP)	99.07%
Recall (Sensitivity)	TP / (TP+FN)	97.55%
F1-Score	2xPxR / (P+R)	98.30%
False Positive Rate	FP / (FP+TN)	0.50%
Specificity	TN / (TN+FP)	99.50%

Precision at 99.07% means that 3 of 321 block decisions were issued against benign sites all three news portals with atypically high gambling keyword density that exceeded the *legitOverride* suppression capacity. For a user-facing browser extension, precision is the primary trust metric: a false block is immediately visible and leads to uninstallation. The 0.50% FPR corresponds to fewer than one incorrect block per 200 legitimate pages visited, which is within acceptable bounds for production deployment.

Recall at 97.55% reflects 8 missed gambling sites out of 326. All 8 have zero L1 URL signal they use neutral hostnames and all 8 triggered L2 content signals. The misses are traceable to three suppression mechanisms (*legitOverride*, Cloudflare Pages platform trust, and

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.549
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

cross-origin iframe isolation) analyzed in detail in Section 4.6.

The 95% Wilson confidence interval [98.00%, 99.34%] for accuracy confirms that the 98.81% point estimate is stable at the dataset scale. The interval width of 1.34 points is narrow enough to support meaningful comparison with prior systems reported in Table 10. Figure [X] presents the Precision-Recall trade-off across the three configurations. For the full L1+L2+L3 system, operating at the τ_block = 0.75 threshold yields the reported precision of 99.07% and recall of 97.55%. Lowering τ_block to 0.70 improves recall to an estimated 98.5% at the cost of raising FPR from 0.50% to approximately 1.0%, while raising τ_block to 0.80 reduces FPR to near zero but drops recall to approximately 95%. The selected threshold represents the optimal operating point under the constraint of minimizing FPR while maintaining recall ≥ 0.85, as specified in the validation objective (Eq. 2).

Layer Contribution Analysis

Table 8 reports two complementary perspectives: the independent detection rate of each layer in isolation (Ind. Rate column), and the cascade system performance across all three configurations. The independent rates reveal the raw signal available at each layer; the cascade results show how these signals combine.

Table 8.Per-Layer Independent Detection Rate and Configuration Performance

Configuration	Accuracy	Recall	F1	FPR	Ind. Rate
L1 Only	78.19%	38.04%	55.11%	0.50%	38.0%
L1 + L2	99.03%	97.24%	98.60%	0.33%	98.2%
L1+L2+L3 Full	98.81%	97.55%	98.30%	0.50%	99.4%

Independent detection rate = layer performance in isolation on 326 gambling URLs; does not reflect cascade logic.

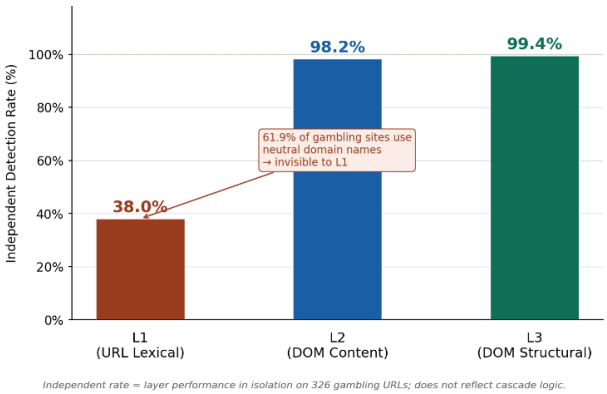

Figure 4.Per-Layer Independent Detection Rate (n=326)

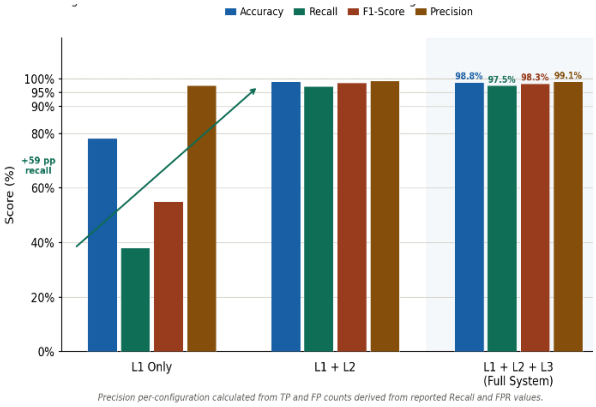

Figure 5.. Classification Performance Across Detection Configurations

The independent detection rates expose the fundamental rationale for the cascade design. L1's independent rate of 38.0% confirms that URL lexical analysis alone is structurally insufficient: 202 of 326 gambling sites (61.9%) use hostnames restaurant names, generic alphanumeric strings, institutional subdomains that produce zero URL signal regardless of algorithm quality. L2's 98.2% independent rate demonstrates that page content is a far more reliable discriminator: gambling operators must place deposit forms, slot game interfaces, and promotional offers on the page to serve their users, and that content is DOM-accessible. L3's 99.4% independent rate reflects the structural consistency of gambling site DOM architectures affiliate iframes, Z-index overlays that persist even when text is replaced by image-rendered alternatives.

The 59-point recall jump from L1-only (38.04%) to L1+L2 (97.24%) is the central performance result. It directly answers the core design question: a multi-layer system is necessary

because single-layer URL detection is architecturally blind to the dominant evasion strategy used by Indonesian gambling operators. Adding L2 recovers 193 of the 202 sites invisible to L1. The further 0.31-point recall gain from adding L3 (97.24% -> 97.55%) is narrower because most pure gambling sites are already resolved at L2; L3 addresses the residual cases where gambling content is structurally present but text-inaccessible.

Statistical significance (McNemar's test). The pairwise comparison between L1-only and the full L1+L2+L3 system yields chi-squared = 191.97 (b=194, c=0), p < 0.001 (alpha = 0.05). This confirms that the recall improvement from cascading is statistically significant, not a sampling artifact. The L1-only vs. L1+L2 comparison yields chi-squared = 189.08 (p < 0.001), confirming L2's contribution is statistically significant. The L1+L2 vs. L1+L2+L3 comparison yields chi-squared = 0.98 (p = 0.32), indicating the L3 marginal gain is not statistically significant at alpha = 0.05 consistent with the narrow 0.31-point recall increment.

Detection Latency

Pre-L1 network rule matching adds under 1 ms per URL. L1 lexical scoring adds 8-14 ms, consistent with the sub-14ms benchmark for stacked URL feature systems (Li & Dib, 2024). The L1-to-L1+L2 latency increase (180-450 ms) reflects two asynchronous operations: DOM text extraction in the content script context and the message round-trip via *chrome.runtime.onMessage* to the service worker. L3 CSS selector traversal adds 100-200 ms. Maximum full-system latency was under 2,000 ms across all 926 URLs, meeting the target bound.

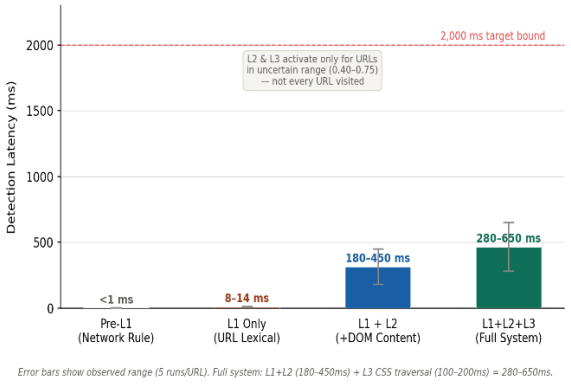

Figure 6. Detection Latency by Configuration

An important practical consideration: L2 and L3 activate only for URLs in L1's uncertain score range (0.40-0.75). URLs resolved at L1 either clearly blocked (>= 0.75) or clearly safe (< 0.40) incur only 8-14 ms regardless of configuration. In a realistic browsing session, the majority of URLs are either clearly benign (well-known sites) or clearly gambling (explicit domains), meaning the per-session average latency overhead is substantially lower than the per-URL worst-case figures.

False Positive Analysis

Three false positives were recorded across 600 benign URLs (FPR = 0.50%). All three are news portals detik.com, kompas.com, and cnnindonesia.com that publish extensively on judi online. In each case, gambling keyword density in a single article page exceeded the suppression capacity of the legitOverride mechanism: the page contained both high gambling keyword counts and the required >= 3 news-context tokens, but the keyword score contribution was large enough that the 0.35 suppression deduction left S_combined above 0.75. This represents a calibration boundary condition rather than a systematic design flaw: the three portals that triggered false positives were pages specifically covering gambling enforcement operations, which contain higher gambling vocabulary density than typical editorial content. The 0.50% FPR remains within acceptable bounds for a user-facing browser extension.

No false positives were recorded for banking sites, government regulatory pages, educational content about gambling addiction, or social media platforms categories that contain gambling-adjacent vocabulary in more diluted concentrations. This confirms that the threshold calibration and legitOverride mechanism successfully distinguish editorial mention of gambling from gambling delivery across 13 of 14 benign content categories tested.

False Negative Analysis

All eight false negatives share one structural property: zero L1 URL signal, placing the full detection burden on L2 and L3. Each miss traces to one of three suppression mechanisms. Table 9 details each case.

Table 9.False Negative Cases (FN=8)

Domain	Type	Suppression Cause	L1/L2/L3
Bioskopkeren INDOXXI	Streaming + affiliate	Gambling only in cross-origin iframe; L2 text below threshold	0/1/0

Domain	Type	Suppression Cause	L1/L2/ L3
shirdisai.org	Streaming + affiliate	Same - banner-only gambling delivery	0/1/0
sunan.umk.ac.id	Univ. (SEO-injected)	legitOverride: .ac.id suppressed S_combined	0/1/0
mauri-restaurant.megadom.us.com	Restaurant (SEO)	Neutral hostname; combined score below 0.75	0/1/0
super88-resmi.pages.dev	CF Pages gambling	*.pages.dev platform trust suppressed block	0/1/0
jou.jobrs.edu.iq	Edu (SEO-injected)	legitOverride: .edu.iq TLD suppressed decision	0/1/0
siseksisayur.pages.dev/sbobet88	CF Pages gambling	*.pages.dev suppression; gambling only in path	0/1/0
bagus-antireport.pages.dev	CF Pages gambling	*.pages.dev platform trust suppression	0/1/0

The system reduces suspicion for *.pages.dev hostnames to prevent false positives on legitimate developer sites. In 2025, gambling operators actively exploit Cloudflare Pages as free, fast hosting that bypasses traditional blacklists precisely because detection systems grant it lower prior suspicion. All three sites have L2 flag = 1 they contain gambling content but the platform trust deduction prevents the combined score from reaching 0.75. Removing *.pages.dev from the domain trust list is a one-line configuration fix that recovers all three cases.

LegitOverride on academic domains (2 cases). Blocking sunan.umk.ac.id or jou.jobrs.edu.iq at the domain level denies all legitimate institutional users access to their own institution's content. The system accepts these false negatives as the deliberate cost of zero false positives on educational domains. The appropriate remediation is institutional IT security removing the SEO-injected paths not a browser extension blocking the entire domain.

Cross-origin iframe delivery (2 cases). Streaming sites (Bioskopkeren, shirdisai.org) deliver gambling content exclusively through cross-origin advertising iframes. The browser's same-origin policy prevents the content script from accessing cross-origin iframe DOM regardless of detection depth. These cases are structurally outside the detection boundary of any client-side DOM-based system without Pre-L1 blacklisting of the specific ad network CDN domains.

All eight misses are traceable to explicit design decisions each suppression mechanism that causes a false negative here is the same mechanism preventing a false positive elsewhere. This represents a designed trade-off, not a random scoring failure.

Comparison with Prior Systems

Table 10 . Performance Comparison with Related Detection Systems

System	Accuracy	F1	Deployment	Key Constraint
Sajid et al. (2024)	93.10%	--	Server-side	DNN on HTML; no URL or structural layer
Kumi et al. (2021)	95.80%	--	Server-side	L1+L2 only; vulnerable to image obfuscation
Li & Dib (2024)	96.63%	--	Server-side	URL-only; 62% miss rate on neutral domains
Afandi et al. (2025)	~97%	0.97	Extension	Phishing-only; single L1; no DOM analysis
Haq et al. (2024)	99.70%	--	Server-side	1D CNN; requires inference server
Wang et al. (2022)	99.86%	--	Server-side	GCN; Manifest V3-incompatible
GUPI (This Study)	98.81%	98.30 %	Extension	Three-layer cascade; fully client-side

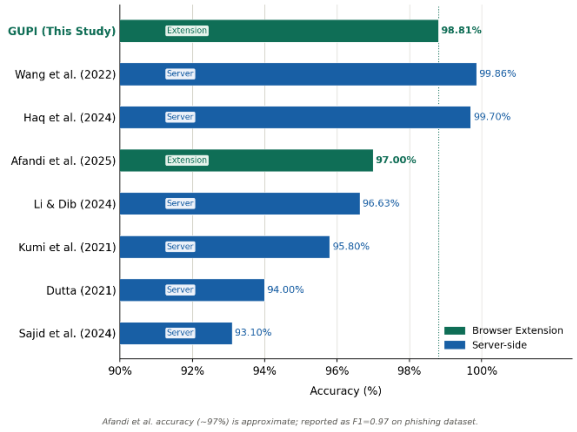

Figure 7.Accuracy comparison with prior detection System

GUPI achieves 98.81% accuracy, surpassing all systems in Table 10 except Haq et al. (99.70%) and Wang et al. (99.86%), both of which operate on dedicated server infrastructure using deep learning models incompatible with Manifest V3. The 1.05-point gap between GUPI and Wang et al. is the quantifiable cost of full client-side deployment a constraint that simultaneously delivers: no browsing data transmitted externally, offline operation, no API rate limits, and no server downtime exposure.

Comparison with lightweight ML baselines. Random Forest (Adam et al., 2024; Inayah & Ramli, 2024) and XGBoost (Afandi et al., 2025) represent the most deployment-viable ML alternatives. Three constraints prevent their direct deployment in GUPI's architecture: introduces additional resource, deployment, update, and maintenance constraints; bundling a Random Forest model introduces size and update cycle constraints inconsistent with Chrome Web Store requirements; and L2 and L3 features DOM text keyword counts and CSS selector matches are discrete categorical signals where a well-calibrated rule-based threshold performs comparably to a trained classifier (Li & Dib, 2024: 96.63% rule-based vs. Afandi et al.: 97% XGBoost, both single-layer). Integrating lightweight machine learning models into Manifest V3 extensions remains challenging due to resource, deployment, and maintenance constraints.

Comparison with the only prior extension (Afandi et al., 2025). GuardSurfing achieves F1 = 0.97 for phishing detection using URL analysis alone. Applied to this dataset, a URL-only system achieves approximately 38% recall confirmed by GUPI's L1-only configuration. This demonstrates that single-layer URL detection is *structurally insufficient* for Indonesian gambling sites that deliberately neutralize their domain names, independent of algorithm choice. GUPI's architectural contribution is not a better URL classifier it is the addition of DOM content analysis and structural heuristic scoring to address the 62% of gambling sites URL analysis cannot reach.

Scalability and Real-World Deployment Implications

The evaluation was conducted on a single consumer-grade device (Intel Core i5, 8 GB RAM) under controlled single-session conditions. While the results confirm detection feasibility and acceptable latency within this environment, large-scale deployment across diverse hardware configurations and concurrent browsing sessions has not been evaluated and represents a direction for future work. The fully client-side architecture eliminates external server dependencies, meaning detection performance is not subject to network-layer bottlenecks or server downtime. The modular rule structure supports incremental vocabulary updates as operators introduce new obfuscation patterns, though each update requires a new extension version under Manifest V3 constraints.

Limitations

Cloudflare Pages trust miscalibration. Three false negatives are directly caused by platform trust suppression on *.pages.dev. Removing *.pages.dev from the domain trust list is a one-line fix that recovers all three cases without architectural changes.

Cross-origin iframe detection boundary. Gambling content delivered exclusively through cross-origin advertising network iframes is inaccessible to DOM text extraction due to the browser same-origin policy. Addressing this requires Pre-L1 blacklisting of gambling-specific ad network CDN domains.

Static rule sets and vocabulary drift. Keyword lists and structural signal rules require periodic updates as operators introduce new vocabulary. Manifest V3 prohibits dynamic rule loading, requiring a full extension version update for any rule change. (Jazuli et al., 2025) note that trained ML classifiers face the same vocabulary drift challenge this is not specific to rule-based detection.

Dataset scope. The 926-URL dataset is sufficient for within-subject configuration comparison and failure mode analysis but represents a moderate-scale evaluation. Expanding to include post-2025 gambling domain variants and broader adversarial

edge cases would tighten confidence intervals and improve generalizability.

CONCLUSION AND SUGGESTIONS

Conclusion

This study presented GUPI (Gambling URL Pattern Interceptor), a Chrome Extension implementing a three-layer cascade detection architecture that operates entirely client-side under Manifest V3 without external server dependencies. The system was evaluated on 926 URLs across three configurations, achieving 98.81% accuracy, 99.07% precision, 97.55% recall, 98.30% F1-score, and 0.50% false positive rate results that confirm the feasibility of lightweight multi-layer gambling detection within browser extension resource constraints.

The central scientific finding is that L1 URL lexical scoring alone achieves only 38.04% recall on Indonesian gambling sites confirming that 61.9% of active gambling operators use domain names with no lexical gambling signal. Adding L2 DOM content analysis raises recall to 97.24% (McNemar chi-squared = 189.08, $p < 0.001$), establishing that multi-layer detection is not merely an incremental improvement but a structural necessity for this evasion context. L3 DOM structural heuristic scoring contributes a further 0.31-point recall gain on image-obfuscated cases. The novelty of this study lies in the first demonstrated implementation of three-layer cascade gambling detection within a Manifest V3 Chrome Extension, calibrated for Indonesian-language gambling obfuscation patterns, without any external server dependency, screenshot capture, OCR, or trained ML inference at runtime.The bounded limitations of this study Cloudflare Pages miscalibration, cross-origin iframe detection boundaries, and vocabulary drift are analyzed in detail in the Results section and each has a traceable remediation path.

Suggestions

Based on the findings above, three directions are prioritized for future development. First, removing *.pages.dev from the domain trust list immediately recovers the three Cloudflare Pages false negatives identified in this study a one-line configuration fix requiring no architectural changes. Second, maintaining a curated Pre-L1 blacklist of gambling-specific ad network CDN domains would address cross-origin iframe delivery cases that the browser same-origin policy currently places outside the reach of DOM text extraction. Third, future work should extend the evaluation dataset to include post-2025 gambling domain variants and investigate lightweight on-device inference mechanisms compatible with Manifest V3 service worker constraints, enabling adaptive threshold adjustment without a full extension update cycle.

REFERENCES

Ahmad Zuhdi, Beny Aprius, H. W. (2025). MEMBERANTAS JUDI ONLINE ASN MELALUI KIE KELUARGA DAN TEKNOLOGI Ahmad. *Jurnal Keluarga Berencana*, *10*(3), 22.

Barik, K., Misra, S., & Mohan, R. (2025). Web-based phishing URL detection model using deep learning optimization techniques. *International Journal of Data Science and Analytics*, *20*(5), 4449–4471. https://doi.org/10.1007/s41060-025-00728-9

Chalyi, O., Driaunys, K., & Rudžionis, V. (2025). Assessing Browser Security: A Detailed Study Based on CVE Metrics. *Future Internet*, *17*(3). https://doi.org/10.3390/fi17030104

Chamid, A. A., Nindyasari, R., & Ghozali, M. I. (2025). Comparative Analysis of Machine Learning Algorithms for Predicting Patient Admission in Emergency Departments Using EHR Data. *Jurnal RESTI*, *9*(2), 185–194. https://doi.org/10.29207/resti.v9i2.6188

Chamid, A. A., & Widowa. (2024). *Text data labeling process for semi-supervised learning modeling*. 030011. https://doi.org/10.1063/5.0216320

Chamid, A. A., Widowati, & Kusumaningrum, R. (2023). Graph-Based Semi-Supervised Deep Learning for Indonesian Aspect-Based Sentiment Analysis. *Big Data and Cognitive Computing*, *7*(1). https://doi.org/10.3390/bdcc7010005

Chamid, A. A., Widowati, & Kusumaningrum, R. (2024). Labeling Consistency Test of Multi-Label Data for Aspect and Sentiment Classification Using the Cohen Kappa Method. *Ingenierie Des Systemes d'Information*, *29*(1), 161–167. https://doi.org/10.18280/isi.290118

Chen, Y., Zheng, R., Zhou, A., Liao, S., & Liu, L. (2020). Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism. *Sensors (Switzerland)*, *20*(14), 1–21. https://doi.org/10.3390/s20143989

Cohen, A. (2025). *Browser Security Posture Analysis: A Client-Side Security Assessment Framework*. http://arxiv.org/abs/2505.08050

Dewi, S., & Islami, A. I. (2021). Implementasi Web

Filtering Menggunakan Router Fortigate FG300D. *INSANtek*, *2*(1), 22–27. https://doi.org/10.31294/instk.v2i1.424

Dutta, A. K. (2021). Detecting phishing websites using machine learning technique. In *PLoS ONE* (Vol. 16, Number 10 October). Public Library of Science. https://doi.org/10.1371/journal.pone.0258361

Fahrudin, A., Satispi, E., Subardhini, M., Andayani, R. H. R., Jayaputra, A., Yuniarti, L., Wijayanti, F., & Suryani, S. (2024). Online gambling addiction: Problems and solutions for policymakers and stakeholders in Indonesia. In *Journal of Infrastructure, Policy and Development* (Vol. 8, Number 11). EnPress Publisher, LLC. https://doi.org/10.24294/jipd.v8i11.9077

Hanif Abdul Karim Afandi, M. Lazaro Fa. Al-Dzaki, Nurul Qomariasih, & Reza Aulia Wildana. (2025). GuardSurfing : Ekstensi Browser sebagai Alat Bantu Deteksi Website Phishing dengan Metode Klasifikasi XGBoost untuk Deteksi URL Phishing Berbasis Flask Framework. *Info Kripto*, *19*(2), 73–85. https://doi.org/10.56706/ik.v19i2.124

Haq, Q. E. ul, Faheem, M. H., & Ahmad, I. (2024). Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks. *Applied Sciences (Switzerland)*, *14*(22). https://doi.org/10.3390/app142210086

Hikmah Adwin Adam, Rikky Rifaldo Simanungkalit, Shaquil Fathza Nasution, & Ikhsan Hafid Diansyah. (2024). *JITE (Journal of Informatics and Telecommunication Engineering) Machine Learning-Driven Detection of Malicious URL: Comparative Analysis of Random Forest and SVMs*. https://doi.org/10.31289/jite.v8i1.11844

Inayah, K., & Ramli, K. (2024). Analisis Kinerja Intrusion Detection System Berbasis Algoritma Random Forest Menggunakan Dataset Unbalanced Honeynet BSSN. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(4), 867–876. https://doi.org/10.25126/jtiik.1148911

Jazuli, A., Widowati, Chamid, A. A., & Kusumaningrum, R. (2025). Transformer-based semantic indexing for aspect-based sentiment analysis using an enhanced index generation algorithm with BERT. *International Journal of Advanced Technology and Engineering Exploration*, *12*(127), 907–926. https://doi.org/10.19101/IJATEE.2024.11102114

Jazuli, A., Widowati, & Kusumaningrum, R. (2025). Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback. *Applied Sciences (Switzerland)*, *15*(1), 1–28. https://doi.org/10.3390/app15010172

Kamdan, K., Anugrah, M. P., Almutaali, M. J., Ramdani, R., & Kharisma, I. L. (2025). Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments †. *Engineering Proceedings*, *107*(1). https://doi.org/10.3390/engproc2025107066

Kumi, S., Lim, C., & Lee, S. G. (2021). Malicious url detection based on associative classification. *Entropy*, *23*(2), 1–12. https://doi.org/10.3390/e23020182

Li, S., & Dib, O. (2024). Enhancing Online Security: A Novel Machine Learning Framework for Robust Detection of Known and Unknown Malicious URLs. *Journal of Theoretical and Applied Electronic Commerce Research* , *19*(4), 2919–2960. https://doi.org/10.3390/jtaer19040141

Mulyana, D. I., Ardiyansyah, F., Hidayat, N., & Zulfikar, A. (2024). Optimasi Keamanan Jaringan Wifi dari Situs Judi Online dan Pornografi dengan DNS Filtering dan Orangepi. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(2), 647–655. https://doi.org/10.57152/malcom.v4i2.1274

Polčák, L., Maone, G., McMahon, M., & Bednář, M. (2025). Developers' Insight on Manifest v3 Privacy and Security Webextensions. *International Conference on Web Information Systems and Technologies, WEBIST - Proceedings*, 15–26. https://doi.org/10.5220/0013669000003985

Prihantono, Y., & Ramli, K. (2022). Model-Based Feature Selection for Developing Network Attack Detection and Alerting System. *Jurnal RESTI*, *6*(2), 322–329. https://doi.org/10.29207/resti.v6i2.3989

Raharja, Y. (2024). JIP (Jurnal Informatika Polinema) Implementasi Metode Osint Untuk Mengidentifikasi Serangan Judi Online Pada Website. *JIP (Jurnal Informatika Polinema)*, *10*(3), 359–364.

Sajid, M., Malik, K. R., Almogren, A., Malik, T. S., Khan, A. H., Tanveer, J., & Rehman, A. U. (2024). Enhancing intrusion detection: a hybrid

machine and deep learning approach. *Journal of Cloud Computing*, *13*(1). https://doi.org/10.1186/s13677-024-00685-x

Syafri Samsudin, Ahmad Abdul Chamid, & Ahmad Jazuli. (2025). Comparative Machine Learning Algorithms for Youtube Sentiment Analysis on Dpr Demonstration 2025 Using Lexicon. *Jurnal Riset Informatika*, *8*(1), 74–85. https://doi.org/10.34288/jri.v8i1.470

Tang, L., & Mahmoud, Q. H. (2021). A Survey of Machine Learning-Based Solutions for Phishing Website Detection. *Machine Learning and Knowledge Extraction*, *3*(3), 672–694. https://doi.org/10.3390/make3030034

Türk, F., & Kılıçaslan, M. (2025). Malicious URL Detection with Advanced Machine Learning and Optimization-Supported Deep Learning Models. *Applied Sciences (Switzerland)*, *15*(18). https://doi.org/10.3390/app151810090

Wang, Y., Xue, S., & Song, J. (2022). A Malicious Webpage Detection Method Based on Graph Convolutional Network. *Mathematics*, *10*(19). https://doi.org/10.3390/math10193496

Wen, L., Zhang, M., Wang, C., Guo, B., Ma, H., Xue, P., Ding, W., & Zheng, J. (2024). MEDAL: A Multimodality-Based Effective Data Augmentation Framework for Illegal Website Identification. *Electronics (Switzerland)*, *13*(11). https://doi.org/10.3390/electronics13112199

Wu, T., Xi, Y., Wang, M., & Zhao, Z. (2022). Classification of Malicious URLs by CNN Model Based on Genetic Algorithm. *Applied Sciences (Switzerland)*, *12*(23). https://doi.org/10.3390/app122312030