Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 281-292. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1848

# Optimizing Online Gambling Site Detection via XLM-RoBERTa and ResNet34-Based Early Fusion

Rahmat Nugroho and Denni Kurniawan

Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia

*E-mail: 2311602284@student.budiluhur.ac.id, denni.kurniawan@budiluhur.ac.id*

**Abstract**

The illegal online gambling ecosystem in Indonesia has evolved into a persistent cyber threat, driven by the adaptability of threat actors in manipulating content and network infrastructure. Conventional detection methods relying on domain reputation-based blocking (blacklists) and keyword matching now face systemic failure due to sophisticated evasion techniques such as domain hopping, SEO manipulation, and content cloaking. This study proposes an automated detection framework based on Multimodal Deep Learning that simultaneously integrates semantic, visual, and infrastructure metadata analysis. We employ an Early Fusion strategy by constructing a 1,284-dimensional combined feature vector, consisting of 768 dimensions of text embeddings from the XLM-RoBERTa model, 512 dimensions of global visual features from the ResNet34 architecture, and 4 hybrid technical metadata features. To ensure high-quality ground truth and address previous transparency concerns, this approach is evaluated using a balanced dataset of 3,546 sites constructed via active crawling using the Playwright framework. The data was rigorously verified by a panel of five security experts using a majority voting scheme, achieving a Fleiss' Kappa agreement score of 0.87, which indicates almost perfect consensus. Experimental results demonstrate that the Early Fusion model with Random Forest classification achieved an F1-Score of 95.52%, significantly outperforming the Late Fusion strategy (91.62%) and other unimodal approaches. Furthermore, this study empirically confirms the phenomenon of adversarial adaptation, revealing that traditional metadata features contribute only 0.62% to the classification decision. These findings underscore the urgency of a paradigm shift towards Deep Content Inspection to modernize national cybersecurity filtering infrastructure.

**Keywords:** *Online Gambling Detection, Early Fusion, XLM-RoBERTa, ResNet34, Content Cloaking.*

## 1. Introduction

The exponential growth of online gambling sites in Indonesia has precipitated alarming socio-economic impacts, ranging from psychological addiction to illicit financial flows abroad. A recent report [1] highlights that organized crime in Southeast Asia is increasingly leveraging technological innovations to conduct transnational gambling and money laundering operations. Consequently, while these illegal platforms primarily target the Indonesian demographic, their digital infrastructure, operational syndicates, and content delivery networks are inherently transnational. They are frequently orchestrated from regional Southeast Asian hubs (such as Vietnam, Cambodia, or the Philippines), which inherently injects a complex, multi-lingual dimension into the web content encountered by security crawlers. This threat has been officially recognized as a national emergency. Recently, [2] has explicitly warned that online gambling syndicates have evolved into complex criminal ecosystems, exploiting digital space not only for betting but also for massive data theft and fraud.

At the technical level, [3] found that these syndicates have even infiltrated government domains through hidden script injection techniques, demonstrating the profound vulnerability of digital infrastructure. Regulatory efforts to block millions of domains are often hampered by the characteristics of asymmetric warfare, where site operators possess high flexibility to create thousands of new domains within minutes. This problem is exacerbated by the adoption of increasingly sophisticated evasion techniques.

Existing literature categorizes gambling detection strategies into two primary generations: metadata-based and content-based analysis. Early research predominantly focused on URL lexical features and domain reputation. Studies utilizing features such as URL length, special character frequency, and domain age initially showed

promise in identifying malicious sites [4]. However, these metadata-based methods have proven increasingly ineffective against "domain hopping" tactics, where perpetrators rapidly register thousands of benign-looking domains using legitimate Top-Level Domains (TLDs) to bypass reputation filters [5]. Moreover, the widespread availability of free Automated Certificate Management Environment (ACME) issuers has diminished the value of HTTPS as a trust indicator, allowing gambling operators to easily obtain valid SSL certificates and blend in with legitimate traffic [4].

First-generation detection systems based on URL blacklists and simple keyword searching are now considered obsolete. In a deep exploration of the illegal gambling ecosystem, [6] revealed that actors often use HTML code obfuscation techniques to deceive signature-based detection systems. Furthermore, [6] explained the mechanism of cloaking, a server-side technique that serves benign or blank pages to security crawler bots while real users continue to see gambling content. Additionally, [5] highlighted the phenomenon of infrastructure manipulation, where gambling syndicates purchase aged domains and utilize valid SSL certificates to manipulate network reputation scores, thereby causing traditional metadata features to lose their discriminative power.

To overcome metadata limitations, subsequent research shifted towards content inspection. In the textual domain, Natural Language Processing (NLP) techniques ranging from Keyword Matching to TF-IDF have been widely adopted [3]. While effective for static content, these methods are easily circumvented by obfuscation attacks, where gambling keywords are replaced with homoglyphs or embedded within images [6-7]. Consequently, visual analysis approaches using Convolutional Neural Networks (CNN) emerged to detect gambling-specific visual markers such as slot machine reels, card tables, or betting odds tables. Despite their precision, unimodal visual models often suffer from high computational costs and false positives when analyzing sites with complex, clutter-heavy layouts that resemble legitimate gaming or e-commerce platforms [7]. This limitation underscores the critical need for a fusion strategy that can correlate semantic context with visual cues to distinguish between a benign gaming site and an illegal gambling operation accurately.

Responding to this complexity, the paradigm of malicious content detection is shifting towards a multimodal approach. [8] define multimodal learning as an approach that processes and relates information from various modalities to improve AI model performance. While [4] demonstrated that integrating various data sources improves phishing detection, effective fusion requires rich feature representations from each modality [9]. In the specific context of gambling detection, the most closely related work is [7], which successfully utilized a hybrid multimodal fusion method. However, a critical research gap remains: their framework was limited to integrating only two modalities (textual and visual) and was primarily evaluated on monolingual datasets.

Our research directly addresses these gaps by introducing key novelties. First, we extend the architecture to encompass three distinct modalities (semantic text, visual features, and infrastructure metadata), providing a more comprehensive defense against adversarial evasion compared to [7]. Second, by employing XLM-RoBERTa [10], our approach is specifically optimized to decode cross-lingual code-mixing and slang obfuscation (e.g., Indonesian, English, and Mandarin) prevalent in localized gambling networks.

Therefore, the primary objective of this study is to propose Early Fusion as the optimal robust solution for online gambling detection, with Late Fusion and unimodal models serving strictly as baseline comparisons. This objective is realized through three key contributions:

1. Developing an automated detection system using an Early Fusion strategy that comprehensively integrates cross-lingual semantics from XLM-RoBERTa, visual features from ResNet34, and technical metadata.
2. Validating the model on a rigorously constructed dataset of 3,546 sites, curated through a Human-in-the-Loop mechanism involving multiple security experts to ensure ground truth quality.
3. Providing empirical evidence of "adversarial adaptation," demonstrating how traditional metadata features (e.g., domain age, SSL) have become heavily limited as standalone indicators in the modern cybercrime ecosystem.

## 2. Related Work

This section reviews the evolution of gambling site detection strategies, transitioning from metadata analysis to content-based inspection. It also introduces the fundamental concepts of multimodal learning that form the basis of our proposed framework.

Furthermore, it highlights the specific methodological gaps in existing approaches, particularly concerning cross-lingual challenges in Southeast Asia and the lack of comparative studies

on fusion architectures, which this study aims to address.

### 2.1 Metadata and Reputation-Based Analysis

Early generation detection systems primarily relied on network properties and domain reputation. Previous studies, such as [5] established foundational work by utilizing features such as domain age, SSL certificate validity, and autonomous system numbers (ASN) to identify malicious hosts. Their approach proved computationally efficient for filtering known threats.

However, the efficacy of these methods has been challenged by the phenomenon of "infrastructure manipulation." As noted in recent observations, threat actors increasingly acquire aged domains and utilize legitimate certificates (e.g., Let's Encrypt) to mimic benign traffic behavior. Consequently, while metadata remains a useful preliminary filter, reliance on these features as primary discriminators has become prone to high false-negative rates in the face of adversarial adaptation.

### 2.2 Unimodal Content Inspection

To address the limitations of metadata, subsequent research shifted focus towards content analysis, typically treating text and visual data as separate modalities. In terms of textual analysis, research by [3] demonstrated the effectiveness of keyword matching and regular expressions in detecting gambling script injections within government domains. While effective for static pages, [6] highlighted that such signature-based methods are increasingly circumvented by HTML obfuscation techniques and the use of homoglyphs. These evasion tactics disrupt standard NLP tokenizers and severely reduce the accuracy of text-only models.

In parallel, computer vision approaches using Convolutional Neural Networks (CNN) have been employed to detect gambling logos or betting tables. However, unimodal visual models often struggle with clutter-heavy interfaces where gambling advertisements are injected into legitimate movie streaming or e-commerce sites. This lack of supporting semantic context often leads to potential misclassification, proving that visual analysis alone is insufficient to handle sophisticated cloaking mechanisms.

### 2.3 Multimodal Deep Learning Approaches

Recognizing the limitations of single-modality systems, the current state-of-the-art has moved towards Multimodal Deep Learning. A recent study [9] emphasized that fusing heterogeneous data sources can capture latent correlations missed by unimodal models.

- In the broader cybersecurity domain, [11] successfully applied a multimodal approach combining HTML DOM graphs and URL features to detect phishing sites, achieving significant accuracy improvements.
- Specifically within the gambling domain, [7] proposed a hybrid framework integrating visual features with OCR-extracted text. Their work represents a significant step forward. However, the study focused primarily on a specific fusion architecture without providing a comprehensive empirical comparison between Early Fusion and Late Fusion strategies. Consequently, the question of which fusion level yields the most robust representation for the specific characteristics of gambling sites, which often employ complex "code-mixing" and visual distractions, remains an open research question.

While previous NLP applications in cybersecurity often rely on monolingual models or standard multilingual BERT (mBERT), these models consistently struggle with extreme 'code-mixing' and out-of-vocabulary (OOV) slang prevalent in Southeast Asian gambling sites. Similarly, earlier visual detection models typically utilize heavier architectures like VGG16 or ResNet50, which, despite high accuracy, pose significant latency bottlenecks for real-time traffic filtering. To overcome these specific limitations observed in prior studies, our research specifically adopts XLM-RoBERTa to decode cross-lingual slang obfuscation and ResNet34 to achieve an optimal empirical balance between visual feature extraction depth and computational efficiency.

### 2.4 Research Gap and Contribution

Despite significant advancements, two critical gaps remain in the existing literature that this study seeks to bridge. First, regarding the comparative analysis of fusion strategies, while multimodal learning is gaining traction, there is a lack of studies that systematically compare the performance of feature-level (Early) versus decision-level (Late) fusion specifically for the gambling domain. This study addresses this gap by implementing and evaluating both strategies to identify the most optimal architecture.

Second, a recurring challenge in prior studies is the reliance on automated crawling with limited ground truth verification, which inadvertently introduces the risk of label noise. To ensure

reproducibility and scientific validity, this study constructs a dataset that undergoes a rigorous Human-in-the-Loop (HITL) validation process. Involving multiple security experts ensures a more reliable benchmark for assessing model performance against sophisticated adversarial evasion techniques.

## 3. Methodology

The research methodology is systematically designed following the Knowledge Discovery in Databases (KDD) framework [12], spanning from data acquisition, pre-processing, and multimodal feature extraction, to data fusion and classification modeling. To ensure reproducibility and validity, the experimental environment was deployed on a high-performance workstation (MacBook Pro M1 Pro, 10-core CPU, 16-core GPU) to handle the computational demands of deep learning processes.

To provide a clear overview of the entire research workflow, from initial data collection and pre-processing to the execution of the multimodal Early Fusion strategy, the complete proposed method pipeline is illustrated in Figure 1.

The dataset was constructed using a custom automated crawler utilizing a microservices architecture implemented in Python 3.9. Unlike standard request-based crawlers, we employed the Playwright framework to render dynamic web pages, ensuring the capture of JavaScript-heavy content typical of modern gambling sites. The automated crawler identified new candidate domains by scanning for hyperlinked anchors containing high-frequency gambling-related keywords such as 'slot', 'gacor', 'deposit puls4', 'agen judi', and 'live casino'."

To mitigate geo-blocking and ISP-level filters (e.g., Internet Positif), the crawling traffic was routed through a NordVPN tunnel. Given that the illegal online gambling ecosystem targeting Indonesia operates as a fundamentally cross-border network, the automated crawling and snowballing processes naturally captured authentic, heterogeneous web traffic. This environment includes localized mirrors, multilingual redirect templates, and landing interfaces featuring a mix of regional languages, such as Vietnamese, Mandarin, and English, alongside localized Indonesian betting slang. Retaining this multilingual diversity within the dataset is critical to training a robust classifier capable of operating within real-world, borderless cybercrime networks.

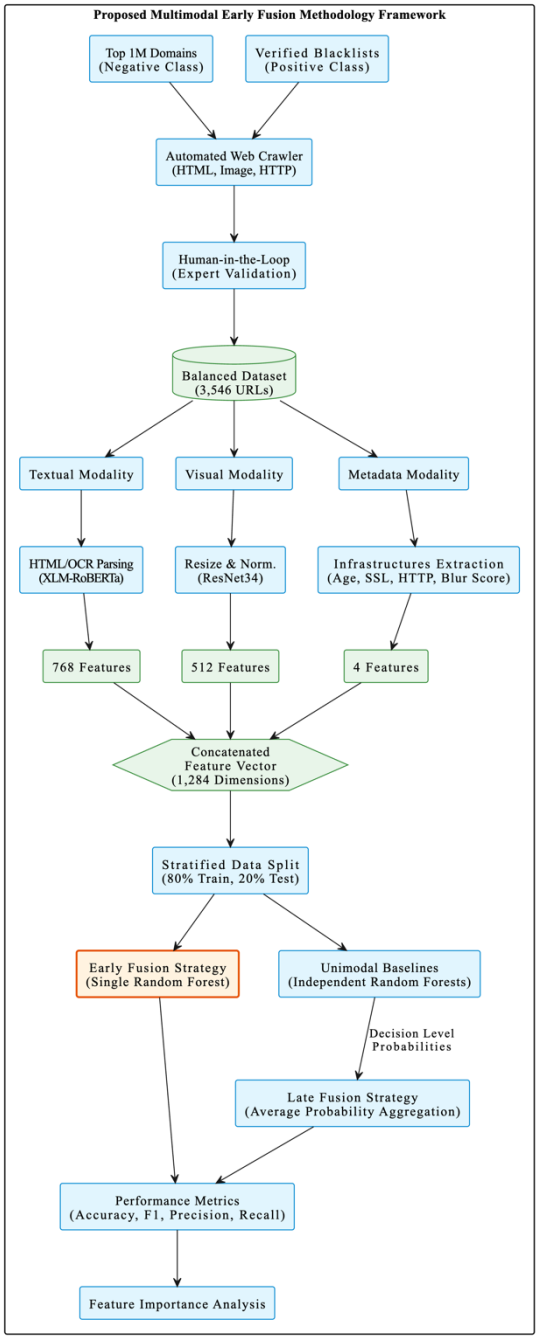


**Figure 1.** Complete pipeline of the proposed multimodal early fusion framework based on the KDD methodology.

We employed a hybrid sampling strategy to collect the initial pool of 12,450 candidate URLs:

1. Purposive Sampling: Focusing on confirmed gambling domains. We extracted exactly 4,000 URLs sourced from verified cybersecurity community blacklists, specifically selecting the gambling extension from the StevenBlack host-file repository [13].
2. Snowball Sampling: Following outbound links from identified gambling sites to discover interconnected domains within the syndicate

network, yielding an additional 3,450 candidate URLs.

3. Benign Sampling: To represent the negative class and ensure diversity, 5,000 negative samples were sourced from the TrustPositif Whitelist and Cloudflare Radar rankings, covering news, education, and e-commerce categories [14], [15].

### 3.1 Expert-Validated Ground Truth (Human-in-the-Loop)

To ensure input quality and address the limitations of automated labeling, this study established a Human-in-the-Loop (HITL) validation mechanism. The ground truth was not determined by a single annotator. Instead, a panel of five (5) independent information security practitioners manually inspected the samples.

It is important to note that the panel of experts did not manually review all 12,450 initial URLs. To manage the annotation workload and eliminate dead links, an automated pre-filtering stage was applied first. By deploying a Python-based asynchronous script, we filtered out inactive domains, parked pages, and server errors (e.g., HTTP 404, 500). This automated stage successfully reduced the pool to approximately 4,200 active and reachable URLs. Consequently, only these ~4,200 active URLs were forwarded to the five independent practitioners for manual Human-in-the-Loop (HITL) validation.

The annotation protocol followed a strict Majority Voting scheme:

- **Independent Inspection**: Each URL was inspected independently by all validators to prevent confirmation bias.
- **Consensus Requirement**: A class label (y=1 for Gambling, y=0 for Benign) was considered valid only if it achieved a clear consensus of at least 3 out of 5 votes. To account for complex evasions, annotators were also permitted to flag a site as "Ambiguous" if they could not confidently determine its nature.
- **Removal of Ambiguous Samples**: Samples failing to reach a valid class consensus due to high ambiguity or contentious content (e.g., highly obfuscated pages causing expert disagreement or multiple "Ambiguous" votes) were permanently excluded to minimize noise in the ground truth.

Following this strict protocol, the experts confirmed exactly 1,773 active illegal gambling sites (Positive Class). The remaining 2,427 active URLs in the pre-filtered pool consisted of verified benign sites and a minority of excluded ambiguous samples.

To quantify the reliability of this manual annotation process, we calculated the Fleiss' Kappa coefficient [16]. The resulting score was 0.87, indicating an "Almost Perfect" level of agreement among the five experts. This high level of consensus ensures that the final labels are robust and minimize subjective bias during model training.

The class distribution before balancing was highly skewed, with 1,773 verified gambling sites against the larger pool of benign active URLs. To prevent bias toward the majority class during model training, we applied a Stratified Random Undersampling technique. From the pool of verified normal sites, we randomly selected exactly 1,773 URLs to perfectly match the number of gambling sites. This process resulted in a perfectly balanced final dataset of 3,546 verified websites (1,773 gambling and 1,773 benign), ensuring a robust foundation for evaluating the multimodal fusion models.

Post-balancing, the dataset was split into training (80%) and testing (20%) sets using Stratified Random Sampling. This technique ensures that class proportions in the training set $S_{train}$ and testing set $S_{test}$ remain consistent with the original population, thereby minimizing variance bias during model evaluation and guaranteeing class distribution consistency according to classification evaluation standards [17].

### 3.2 Pre-processing and Feature Extraction

Before entering the feature extraction pipelines, the raw data underwent a rigorous pre-processing phase to ensure input quality and consistency. For the textual modality, raw HTML content was first parsed using the Beautiful Soup library to strip tags and isolate human-readable text. Additionally, text embedded in image banners was extracted using Tesseract OCR. We applied a filtering mechanism to remove non-ASCII characters and excessive whitespace, while deliberately retaining specific punctuation marks (such as currency symbols like "Rp" or "$") which often serve as strong indicators in gambling contexts. The text was then tokenized using the SentencePiece tokenizer pre-trained on the XLM-RoBERTa corpus, which effectively handles out-of-vocabulary words by breaking them down into sub-word units. This is particularly crucial for Indonesian gambling slang which frequently modifies standard words to evade keyword filters (e.g., writing "g4cor" instead of "gacor").

For the visual modality, all screenshots were resized to a standard resolution of 224x224 pixels to match the input specifications of the ResNet34 architecture. We applied standard normalization

using the mean and standard deviation values of the ImageNet dataset to facilitate faster convergence during the feature extraction process. Furthermore, to handle the variance in aspect ratios of web pages, a center-crop strategy was employed, focusing on the middle region of the page where prominent content and betting tables are typically displayed.

The system implements three parallel feature extraction pipelines to capture semantic, visual, and metadata representations.

### 3.2.1 Textual Modality (XLM-RoBERTa)

Data derived from the Title, Meta Description, HTML Body, and OCR results were processed using the XLM-RoBERTa Base model. The specific selection of XLM-RoBERTa over standard multilingual BERT (mBERT) or monolingual IndoBERT is strongly driven by its superior cross-lingual representation capabilities [10].

Indonesian gambling syndicates heavily employ linguistic code-mixing by blending standard Indonesian, regional slang, English technical terms, and Mandarin artifacts, alongside deliberate out-of-vocabulary modifications (e.g., character substitution tactics like 'g4cor' or 'puls4') to bypass signature filters. XLM-RoBERTa's extensive pre-training on 2.5TB of filtered CommonCrawl data across 100 languages [10] allows it to retain high semantic density and capture nuanced contextual dependencies in mixed-language obfuscations where other models typically suffer from sub-word fragmentation. This Transformer-based architecture employs a Self-Attention mechanism to capture contextual dependencies, mathematically defined for query $Q$, key $K$, and value $V$ as equation (1).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Where $d_k$ is the dimension of the key vector. The last hidden states were averaged to generate a textual feature vector $V_{text}$ with 768 dimensions.

### 3.2.2 Visual Modality (ResNet34)

Simultaneously, visual features were extracted from site screenshots using the ResNet34 architecture [18], selected for its ability to mitigate the vanishing gradient problem through Residual Blocks. Furthermore, the selection of the ResNet34 architecture provides an optimal empirical balance between feature extraction depth and computational throughput [18].

While deeper residual networks (such as ResNet50 or ResNet152) offer marginally higher visual abstraction, they introduce massive parameter inflation and excessive computational overhead. In a real-world deployment scenario, where automated national firewalls or ISP-level web crawlers must inspect thousands of active domains concurrently, computational efficiency is paramount. ResNet34 possesses sufficient depth to accurately capture structural, localized visual markers (e.g., slot machine reels, betting odds tables, and promotional banners) while maintaining a lightweight footprint to prevent latency bottlenecks. Each residual block processes input x through a transformation function $F(x)$ and adds it to the identity input itself (skip connection), formulated as equation (2).

$$y = F(x, \{W\_i\}) + x \quad (2)$$

Where $F(x, \{W_i\})$ represents the learned residual mapping. Features were extracted from the final Global Average Pooling (GAP) layer, yielding a visual feature vector $V_{vis}$with 512 dimensions, effectively reducing spatial dimensions while maintaining invariance to visual object location.

### 3.2.3 Metadata Modality

Additionally, four metadata features were extracted to detect infrastructure anomalies: Domain Age, SSL Validity, HTTP Status, and Image Blur Score. Numeric features such as Domain Age were normalized using Min-Max Scaling to the range [0, 1] to accelerate model convergence, utilizing the formula in equation (3).

$$x_{norm} = \frac{x - x_{min}}{x - x_{max}} \quad (3)$$

To counter cloaking techniques involving blank pages, an Image Blur Score was calculated using the variance of the Laplacian operator $\nabla^2$ convolved with the image $I(x, y)$. A low variance value indicates a scarcity of edges, implying the image is blurry or blank as shown in equation (4).

$$BlurScore = \sum x \sum y \, (\nabla^2 I(x,y) - \mu)^2 \quad (4)$$

Where μ is the mean intensity of the Laplacian convolution result. This Laplacian Variance method was adopted from [19] due to its computational efficiency.

## 3.3 Early Fusion and Classification Strategy

Following feature extraction, an Early Fusion strategy was employed, as feature-level fusion has been proven effective for heterogeneous data [20].

This strategy involves concatenating feature vectors from all three modalities into a single combined vector representation $V_{total}$ prior to the learning process. Defining the concatenation operation as $\oplus$, the final input vector $X$ for the $i$-th sample is equation (5).

$$X^{(i)} = V_{text}^{i} \oplus V_{vis}^{i} \oplus V_{meta}^{i} \quad (5)$$

To ensure numerical stability and prevent modality dominance during the learning process, a feature scaling step is applied to the resulting 1,284-dimensional vector. Since the 768-dimensional XLM-RoBERTa embeddings , 512-dimensional ResNet34 features , and 4 metadata features are derived from different architectures and scales, we applied Layer Normalization. This procedure ensures that the high-dimensional textual embeddings do not overshadow the visual or technical metadata due to magnitude differences, allowing the classifier to treat each feature space's contribution based on its discriminative power rather than its raw scale. This high-dimensional combined vector was classified using the Random Forest algorithm, an ensemble method constructing $N$ decision trees with $N_{estimators} = 100$ . Random Forest performs implicit feature selection at each node splitting by minimizing Gini Impurity. The final prediction is determined via a majority voting mechanism from all decision trees in the forest. Random Forest was chosen due to its robustness against overfitting on high-dimensional data and its capability to provide feature importance scores necessary for model interpretability analysis.

The selection of Random Forest as the classifier over other popular algorithms like Support Vector Machines (SVM) or Gradient Boosting was driven by two theoretical considerations relevant to this high-dimensional dataset. First, the 'bagging' (bootstrap aggregating) mechanism inherent in Random Forest significantly reduces the variance of the model without increasing the bias [21], which is essential when dealing with a 1,284-dimensional feature vector where the risk of overfitting is prominent. Second, Random Forest demonstrates superior resilience to the "curse of dimensionality" compared to distance-based classifiers like k-Nearest Neighbors (KNN), as it operates based on hierarchical rule splits rather than Euclidean distances [22]. This capability allows the model to effectively isolate the small subset of informative features (the specific BERT embedding dimensions) from the noise, providing a robust decision boundary even when a large portion of the feature space specifically the infrastructure metadata contains non-discriminative information due to adversarial adaptation by the attackers.

## 4. Results and Analysis

This section presents a comprehensive evaluation of the proposed detection system's performance, covering quantitative metric analysis, a comparative study of fusion strategies, interpretation of feature importance, and prediction error investigation. The primary evaluation was conducted on a test dataset consisting of 710 samples, with a balanced proportion between positive (gambling sites) and negative classes. Table 1 summarizes the model performance based on Accuracy, F1-Score, Precision, and Recall metrics.

**Table 1.** Performance comparison of multimodal vs. unimodal models.

| Model / Scenario | Accuracy | F1-Score | Precision | Recall |
|---|---|---|---|---|
| ***Early Fusion*** | ***95.49%*** | ***0.9552*** | ***0.9499*** | ***0.9606*** |
| Late Fusion | 91.55% | 0.9162 | 0.9086 | 0.9239 |
| Text (XLM-R) | 95.21% | 0.9518 | 0.9573 | 0.9465 |
| Unimodal Visual | 83.52% | 0.8475 | 0.7888 | 0.9155 |
| Unimodal Meta | 74.79% | 0.7604 | 0.7245 | 0.8000 |

Experimental results confirm the superiority of the Early Fusion approach with an F1-Score of 95.52%. This strategy consistently outperforms Late Fusion by a margin of +3.9%. While previous studies, such as [7], have reported exceptionally high accuracies (approximately 99%) using hybrid fusion on monolingual datasets, a rigorous structural comparison between Early and Late Fusion remains critically necessary to advance the state of the art in highly evasive environments. This advantage is strictly attributed to the Early Fusion model's ability to learn non-linear correlations between features in the combined vector space (1,284 dimensions) before classification decisions are made. Conversely, Late Fusion merely averages final probabilities, failing to compensate for weakened modalities. Under aggressive cloaking, a near-zero text probability can drastically suppress Late Fusion's final score. Early Fusion structurally prevents this vulnerability through holistic evaluation by learning complex non-linear correlations across all inputs. Meanwhile, XLM-RoBERTa dominated the unimodal models (F1 95.18%), proving that semantic traces like promotional keywords remain the primary discriminator, while non-textual features robustly complement it against evasion tactics.

**Table 2.** Sample texts with highest activation on BERT_85 feature (Vietnamese context).

| Rank | Activation Score | Text Snippet | Semantic Pattern Analysis |
|---|---|---|---|
| 1 | 0.1485 | *"Số Đỏ Casino NHẬN KHUYẾN MÃI CẦN CHÚ Ý... đăng ký nhận ưu đãi trong mục Nạp Tiền..."* | Detected entity "Casino" combined with financial terminology "Nạp Tiền" (Deposit) and "KHUYẾN MÃI" (Promotion). |
| 2 | 0.1409 | *"QH88... Cách nạp tiền tại QH88 Phương thức n..."* | Detected explicit instructions regarding Payment Methods and deposit procedures. |
| 3 | 0.1392 | *"TK88... Khám Phá Tk88 : Đế Chế Cá Cược..."* | Detected specific gambling keywords "Cá Cược" (Betting). |

**Table 3.** Sample texts with highest activation on BERT_85 feature (English and Indonesian context).

| Rank | Activation Score | Text Snippet | Semantic Pattern Analysis |
|---|---|---|---|
| 1 | 0.1472 | "Restricted Page Dafabet Sportsbook is an online sports betting site in Asia offering live betting on a variety of sports events. Find football betting, NBA..." | Detected landing templates of localized sportsbook mirror sites, specifically identifying geoblocking messages ("Restricted Page") and betting event listings. |
| 2 | 0.0911 | "m8winvnd – Ilmu Pengetahuan Industri Judol Permainan casino online memberikan cara taruhan yang menarik dan mudah untuk dinikmati melalui smartphone Android iOS..." | Captures localized Indonesian gambling slang ("Judol" / Judi Online) and transactional verbs like "taruhan" combined with mobile-specific vectors. |
| 3 | 0.0861 | "Perusahaan Penyedia Asuransi di Indonesia \| Prudential Indonesia Perlindungan asuransi jiwa & kesehatan dari Prudential Indonesia. Solusi finansial terpercaya..." | Demonstrates structural template overlaps, where the formal corporate terms ("Penyedia", "Perlindungan") activate the feature as a baseline non-judi sample. |

However, using text alone is vulnerable to cloaking. Integration with visual and metadata features in Early Fusion provides additional robustness, especially for detecting sites with ambiguous text but striking visual patterns. These results align with [11], demonstrating that multimodal integration substantially improves detection accuracy in adversarial web environments.

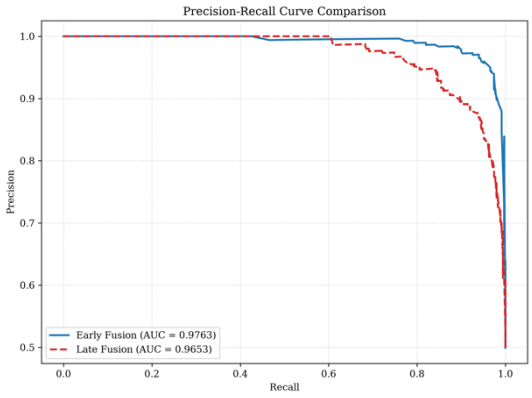


**Figure 2.** Precision-recall curve comparison: Early Fusion (solid line) shows a larger area under curve (AUC) compared to Late Fusion (dashed line), indicating better model stability across various sensitivity levels.

Given the high security risk cost of False Negatives and user inconvenience from False Positives, Precision-Recall (PR) curve analysis is crucial. Figure 2 visualizes the comparison between Early Fusion and Late Fusion strategies.

The Early Fusion model maintains high precision even when forced to achieve recall approaching 1.0. Based on curve analysis, the optimal decision threshold was set at 0.45. This value, slightly lower than the standard 0.5, aims to prioritize Recall (detection power) without significantly sacrificing Precision. At this operating point, the system successfully captures 96% of all gambling sites with a False Alarm rate of less than 5%.

One of the most critical findings in this study is the feature importance distribution generated by the Random Forest classifier. The algorithm calculates the intrinsic importance of each of the 1,284 features using the Mean Decrease in Impurity (Gini Importance) metric across all decision trees. To determine the holistic contribution of each modality before analyzing individual features, we aggregated these values. Specifically, we summed the raw importance scores of the 768 textual dimensions (XLM-RoBERTa), the 512 visual dimensions (ResNet34), and the 4 metadata features. These three aggregate sums were subsequently normalized to 100%.

This quantitative aggregation reveals a stark hierarchy: Textual features overwhelmingly

contribute 86.79% to the decision boundary, followed by Visual features at 12.59%, while Metadata contributes a mere 0.62%. Following this aggregate modality analysis, Figure 3 visualizes the Top-15 most discriminative individual features within the combined vector space.

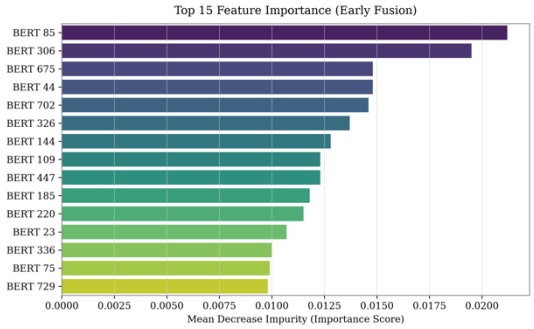


**Figure 3.** Top-15 feature importance: entirely dominated by text embedding vector dimensions (Text_Emb) and visual features, without infrastructure metadata features present in the top ranks.

As seen in Figure 3, the highest-ranking individual features are entirely dominated by XLM-RoBERTa embedding dimensions. This empirical finding, combined with the mere 0.62% overall metadata contribution, directly confirms the phenomenon of Adversarial Adaptation. Cyber threat actors now heavily invest in aged domains and valid SSL certificates to intentionally manipulate network reputation scores, rendering these parameters heavily limited as primary detection features. This provides strong evidence that modern detection paradigms must shift from metadata analysis toward Deep Content Inspection.

To understand the specific semantic patterns captured by the model, we conducted a granular inspection of the top-weighted feature: BERT Dimension 85. Analysis reveals that this feature functions as a latent "Transaction Intent Detector," correlating high activation with transactional imperatives rather than just keywords.

While Table 2 illustrates the model's capability in detecting promotional structures within the Vietnamese context, our dataset encompasses a broader Southeast Asian syndicate network that heavily targets Indonesia. To specifically address the localized threat landscape and demonstrate XLM-RoBERTa's cross-lingual robustness, Table 3 presents representative examples of English and Indonesian code-mixed texts, alongside a benign baseline.

To understand the practical implications of these metrics and specifically address the efficacy of multimodal fusion in handling evasion tactics, a qualitative analysis was conducted based on the Confusion Matrix shown in Figure 4.

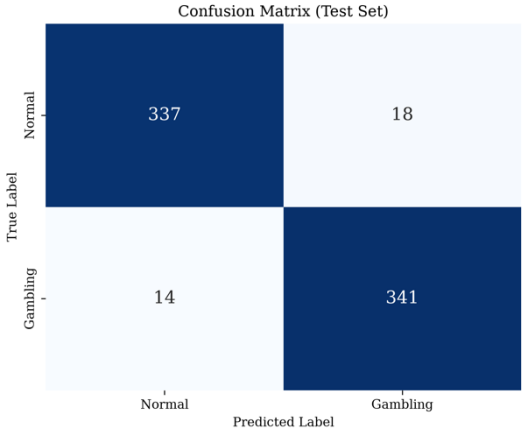


**Figure 4.** Confusion matrix on test data (total 710 samples).

Out of 710 test samples, the model erred on only 32 samples (4.5% Error Rate). Deep analysis revealed two primary patterns: False Positives (18 cases) and False Negatives (14 cases).

The False Positives primarily occurred when benign sites were classified as gambling. Manual investigation showed most of these were illegal movie streaming sites (e.g., LK21, IndoXXI) and pirated software download forums. These sites are typically saturated with aggressive gambling banner ads (pop-ups). The ResNet34 visual model accurately captured "casino chip" or "slot" patterns in these embedded ads, causing the visual modality to dominate the classification decision and override the benign primary text content (movies/software).

To demonstrate whether multimodal fusion actually improves cloaking detection compared to unimodal models, we must investigate the False Negatives. In these 14 cases, gambling sites successfully evaded detection. However, it is crucial to distinguish between basic "Login Wall Cloaking" and extreme "Advanced Server-Side Cloaking." As illustrated in Figure 5, the Early Fusion mechanism proved its necessity in overcoming basic visual and structural cloaking that completely deceived the Unimodal Text baseline.

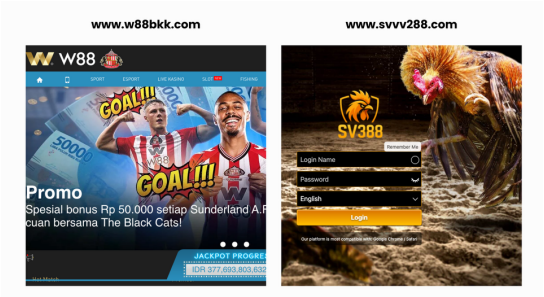


**Figure 5.** Empirical evidence of Early Fusion successfully detecting cloaked websites (www.svvv288.com and www.w88bkk.com) that were missed by the unimodal text model.

In the cases shown in Figure 5, the syndicates

employed a "Login Wall" tactic. The HTML source code is aggressively stripped of any explicit gambling keywords. For instance, the actual text successfully extracted by the scraper for www.*svvv288.com* was merely: "*SV388 Remember Me English 中文 Tiếng Việt Indonesia Login Install Web APP*". Similarly, *www.w88bkk.com* only revealed generic registration instructions: "*Belum mempunyai account? Klik di sini untuk daftar sekarang! Jika Anda mengalami masalah saat login, silakan hubungi Customer Service kami untuk bantuan lebih lanjut*". Because these text snippets are entirely devoid of betting jargon, the NLP-only model extracted near-zero gambling probabilities, leading to a False Negative. However, the Early Fusion model successfully classified these sites as positive. By concurrently analyzing the visual modality, the ResNet34 pipeline captured subtle structural patterns, such as the background layout of fighting arenas (SV388) or promotional sports banners (W88), acting as a critical "safety net" that compensated for the neutralized text modality.

Despite this substantial improvement over unimodal baselines, it is important to objectively acknowledge that extreme "Advanced Cloaking" cases remain unresolved within the 14 False Negative samples. In scenarios where the server completely blocks the crawler's IP, employs geo-fencing, or serves a completely blank/benign page based on user-agent detection, neither text nor visual modalities can extract meaningful features. Addressing these advanced server-side cloaking tactics will require future integration of dynamic traffic analysis and behavioral mimicking during the crawling phase.

Beyond the quantitative metrics, the practical implications of this study are significant for national cybersecurity strategies. The proposed Early Fusion framework offers a robust solution for regulatory bodies to modernize their filtering infrastructure. Currently, many national firewalls rely heavily on blacklist databases which are reactive. By deploying this automated crawler-classifier system, regulators can scan newly registered domains in real-time and identify gambling content with 95.52% reliability before it becomes widely accessible to the public. The system's ability to ignore irrelevant metadata (like domain age) and focus on deep content inspection ensures that new "disposable domains" created by syndicates are detected just as effectively as established ones.

A critical observation must be addressed regarding the performance gap between the Multimodal Early Fusion model (Accuracy 95.49%) and the Unimodal Text baseline (Accuracy 95.21%). From a purely statistical perspective, this margin may appear relatively small. However, in an adversarial cybersecurity context, relying solely on a unimodal text system creates a highly vulnerable "single point of failure." If regulatory firewalls utilize text-only detection, gambling syndicates will inevitably adapt by shifting their operations entirely to text-in-image obfuscation or the aforementioned "Login Wall" tactics, rendering the 95.21% text accuracy entirely obsolete. Therefore, the integration of the visual modality via Early Fusion is not merely about achieving marginal metric gains. Rather, it is about establishing a crucial "safety net." This structural robustness ensures the system does not completely fail when attackers successfully neutralize the semantic layer.

To systematically justify the substantial computational cost of running both XLM-RoBERTa and ResNet34 concurrently for this added security layer, a tiered deployment strategy is highly recommended for real-world ISP environments. Lightweight metadata scanners and basic NLP filters can act as a preliminary sieve (Tier 1) to quickly process massive volumes of internet traffic, immediately discarding obviously benign sites or blocking explicitly vulgar gambling domains. The computationally expensive Multimodal Early Fusion architecture (Tier 2) is then selectively deployed strictly as a specialized inspector for suspicious or ambiguous domains that trigger uncertainty in Tier 1. This architectural trade-off ensures national-level scalability without compromising the rigorous multimodal precision required to defeat advanced evasions and minimize the over-blocking of legitimate websites.

## 5. Conclusion

This study successfully validates the effectiveness of an online gambling site detection framework based on Multimodal Deep Learning utilizing an Early Fusion strategy. Based on a comprehensive series of experiments, it is concluded that the integration of semantic vector representations from XLM-RoBERTa and visual features from ResNet34 into a single high-dimensional feature space yields superior detection performance, achieving an F1-Score of 95.52%. This result consistently outperforms the Late Fusion approach (91.62%) and unimodal baselines. The analysis demonstrates that while textual features are dominant, the visual modality (contributing 12.59% to the decision importance) acts as a critical "safety net," capturing implicit patterns in ambiguous sites that rely heavily on graphical banners, which are often overlooked by conventional detection methods.

A significant theoretical contribution of this

research is the empirical confirmation of the Adversarial Adaptation phenomenon within the cybercrime ecosystem. The finding that infrastructure metadata features (e.g., domain age, SSL validity) contributed only 0.62% to the classification decision provides compelling evidence that while metadata is not entirely obsolete, its standalone effectiveness is heavily limited within our specific dataset. Cyber threat actors have successfully manipulated these trust indicators, necessitating a shift in cybersecurity defense lines towards Deep Content Inspection. This shift focuses on the site's intent (transactional semantics) rather than its technical identity, utilizing metadata merely as a supplementary validation signal rather than a primary identifier.

Despite the system demonstrating high accuracy, practical challenges remain in handling adversarial content injected into third-party sites. The occurrence of False Positive errors on illegal movie streaming sites indicates that the current visual model struggles to differentiate between main page content and aggressive gambling advertisements. Consequently, future research directions are strongly recommended to integrate Object Detection mechanisms, such as YOLO (You Only Look Once), or semantic segmentation techniques. This approach is anticipated to isolate advertisement elements from the main page structure prior to classification, thereby enhancing system precision without compromising its detection capability against pure gambling sites employing high-level cloaking techniques.

## References


[1] United Nations Office on Drugs and Crime, "Transnational Organized Crime and the Convergence of Cyber-Enabled Fraud, Underground Banking and Technological Innovation in Southeast Asia: A Shifting Threat Landscape," 2024. Accessed: May 01, 2025. [Online]. Available: https://www.unodc.org/roseap/uploads/documents/Publications/2024/TOC_Convergence_Report_2024.pdf

[2] Komdigi, "Kementerian Komunikasi dan Digital." Accessed: Jan. 22, 2026. [Online]. Available: https://www.komdigi.go.id/berita/siaran-pers/detail/dirjen-wasdig-minta-masyarakat-waspada-penipuan-permintaan-data-pribadi-terkait-judi-online

[3] M. Nurseno, U. Aditiawarman, H. Al Qodri Maarif, and T. Mantoro, "Detecting Hidden Illegal Online Gambling on .go.id Domains Using Web Scraping Algorithms," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 365–378, 2024, doi: 10.30812/matrik.v23i2.3824.

[4] M. S. Alzboon, M. Subhi Al-Batah, M. Alqaraleh, F. Alzboon, and L. Alzboon, "Phishing Website Detection Using Machine Learning," *Gamification and Augmented Reality*, vol. 3, p. 81, 2025, doi: 10.56294/gr202581.

[5] M. Min and D. A. Lee, "Illegal Online Gambling Site Detection using Multiple Resource-Oriented Machine Learning," *J. Gambl. Stud.*, Dec. 2024, doi: 10.1007/s10899-024-10337-z.

[6] H. Yang *et al.*, "Casino Royale: A deep exploration of illegal online gambling," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2019, pp. 500–513. doi: 10.1145/3359789.3359817.

[7] C. Wang, M. Zhang, F. Shi, P. Xue, and Y. Li, "A Hybrid Multimodal Data Fusion-Based Method for Identifying Gambling Websites," *Electronics (Switzerland)*, vol. 11, no. 16, Aug. 2022, doi: 10.3390/electronics11162489.

[8] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," Feb. 01, 2019, *IEEE Computer Society*. doi: 10.1109/TPAMI.2018.2798607.

[9] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications," *Computers, Materials and Continua*, vol. 80, no. 1, pp. 1–35, 2024, doi: 10.32604/cmc.2024.053204.

[10] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 8440–8451, Jul. 2020. doi: 10.18653/v1/2020.acl-main.747.

[11] J. H. Yoon, S. J. Buu, and H. J. Kim, "Phishing Webpage Detection via Multi-Modal Integration of HTML DOM Graphs and URL Features Based on Graph Convolutional and Transformer Networks," *Electronics (Switzerland)*, vol. 13, no. 16, 2024, doi: 10.3390/electronics13163344.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," vol. 17, 1996. Accessed: Jun. 24, 2025. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230/1131

[13] S. Black, "Consolidating and extending hosts files from several well-curated sources. Optionally pick extensions for porn, social media, and other categories." Accessed: May 21, 2025. [Online]. Available: https://github.com/StevenBlack

[14] Alsyundawy, "Kompilasi Raw TrustPositif dan beberapa blacklist buatan alsyundawy." Accessed: May 21, 2025. [Online]. Available: https://github.com/alsyundawy

[15] Cloudflare, "Domain Rankings Worldwide - Cloudflare Radar," 2024. Accessed: May 25, 2025. [Online]. Available: https://radar.cloudflare.com/domains

[16] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, pp. 378–382, Nov. 1971, doi: 10.1037/h0031619.

[17] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[19] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martínez, and J. Fernández-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proceedings of the 15th International Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 314–317, 2000. doi: 10.1109/ICPR.2000.903548.

[20] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective Techniques for Multimodal

Data Fusion: A Comparative Analysis," *Sensors*, vol. 23, no. 5, Mar. 2023, doi: 10.3390/s23052381.

[21] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: with Applications in Python*. Cham: Springer International Publishing, 2023.

[22] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, 4th ed. Cambridge: Morgan Kaufmann, 2023.