Beyond Semantic Noise: Diagnosing and Correcting Structural Bias in Code-Mixed Script Detection via XAI-Driven Hybridization

Prasert Teppap , Wirot Ponglangka , Panudech Tipauksorn , Prasert Luekhong *

Posted Date: 18 December 2025

doi: 10.20944/preprints202512.1607.v1

Keywords: Malicious Script Detection; Code-Mixed Text; Structural Bias; Selective Integration; Explainable AI (XAI); Semantic Veto; Low-Resource Language; False Positive Reduction

 doi:10.20944/preprints202512.1607.v1

Article

Beyond Semantic Noise: Diagnosing and Correcting Structural Bias in Code-Mixed Script Detection via XAI-Driven Hybridization

Prasert Teppap [1], Wirot Ponglangka [2], Panudech Tipauksorn [3] and Prasert Luekhong [4,*]

[1] Division of Electrical Engineering, Rajamangala University of Technology Lanna, Chiang Mai, Thailand
[2] Faculty of Engineering, Rajamangala University of Technology Lanna, Chiang Rai, Thailand
[3] Faculty of Engineering, Rajamangala University of Technology Lanna, Chiang Mai, Thailand
[4] College of Integrated Science and Technology, Rajamangala University of Technology Lanna, Chiang Mai, Thailand
* Correspondence: prasert@rmutl.ac.th

Abstract

In the contemporary cybersecurity landscape, the detection of code-mixed malicious scripts embedded within high-trust domains (e.g., governmental and academic websites) constitutes a critical defensive challenge. Traditional Transformer-based models, while effective in natural language processing, often exhibit "Structural Bias," where they erroneously interpret the benign complexity of legacy HTML structures as malicious obfuscation, resulting in elevated false positive rates. To address this limitation, this study proposes an XAI-Driven Hybrid Architecture that synergizes context-aware semantic embeddings from WangChanBERTa with outlier-robust structural features. Validated on a rigorously curated high-fidelity corpus of 5,000 samples, our model achieves a state-of-the-art F1-Score of 0.9908. Beyond standard metrics, Explainable AI (XAI) diagnosis reveals a critical "Dual-Validation" mechanism: structural features effectively veto semantic hallucinations triggered by benign complexity, acting as a crucial safety net. Crucially, the proposed architecture functions as a 'Dual-Validation' mechanism, where structural features effectively veto semantic hallucinations triggered by benign complexity. The integration of these components leads to a 50% reduction in the False Positive Rate (FPR), decreasing from 0.024 in baseline scenarios to 0.012, thereby confirming the operational significance of Selective Integration. This method effectively reduces 'alert fatigue,' providing a scalable solution for SOC analysts tasked with protecting critical infrastructure from advanced code-mixed threats.

Keywords: Malicious Script Detection; Code-Mixed Text; Structural Bias; Selective Integration; Explainable AI (XAI); Semantic Veto; Low-Resource Language; False Positive Reduction

1. Introduction

Identifying malicious scripts embedded within dynamic web content is increasingly challenging in contemporary cybersecurity. Adversaries have evolved from straightforward code injection techniques to sophisticated "code-mixing" approaches, where harmful JavaScript components are seamlessly blended with benign natural language, HTML elements, and misleading contexts to obscure malicious activities [1]. Scripts that integrate various code elements frequently bypass rapid assessments, circumventing conventional signature-based and filter-based security measures. With the evolution of contemporary web architectures, the potential for hackers to inject malicious code into user-generated content, advertising widgets, and third-party plugins has significantly increased. This modification indicates that security systems must possess the capability to interpret both the structural syntax and the semantic context of mixed web content.

 doi:10.20944/preprints202512.1607.v1

Code-mixed web scripts represent a unique input domain that poses challenges to traditional malware detection classifications. Traditional models typically analyze script blocks or text segments independently, failing to recognize the harmful interactions that occur across various content types [2]. In recent years, Transformer architectures have demonstrated significant efficacy in handling unstructured text and code. This facilitates a deeper comprehension of these mixed threats within their contextual framework. Despite their effective performance, these models frequently function as "black boxes," obscuring the processes behind decision-making [3,4]. In critical security operations where accountability, traceability, and forensic reliability are paramount, the absence of transparency in these models remains a significant barrier to their implementation.

To bridge this gap in comprehension and enhance performance, a prevalent concept in the literature referred to as the "Hybrid-is-Better" assumption posits that integrating the advantages of Deep Learning with the insights from handcrafted cybersecurity features (such as entropy and keyword counts) will yield the most robust systems. The rationale is that manually designed features provide clear signals that neural networks might overlook, while deep learning captures the underlying semantic context implicitly.

This analysis systematically investigates this hypothesis to determine the precise characteristics of this interaction. We challenge the prevailing 'Hybrid-is-Better' hypothesis by arguing that straightforward feature fusion often produces 'Semantic Noise' interference patterns that hinder model convergence. We present a theory of "Selective Integration," suggesting that hybridization proves effective only when manually crafted features produce "Orthogonal Information" (e.g., global structural statistics) that maintains mathematical independence from the local semantic context represented by Transformers. This research provides empirical evidence that features based on keywords lead to redundancy, while structural features act as a crucial 'Safety Net' to correct the structural bias introduced by Transformers.

We create a detection pipeline utilizing a pre-trained Transformer encoder designed to effectively capture semantic representations from a combination of HTML, JavaScript, and text data. A comprehensive comparative analysis is conducted on three distinct architectures: (1) a model that relies solely on handcrafted features, (2) a model based entirely on Transformer technology, and (3) a hybrid approach that integrates both methodologies. To ensure clarity and comprehensiveness in our work, we employ various techniques to enhance understanding, including a feature-group ablation study and a suite of Explainable AI (XAI) methods such as SHAP, PCA, and Attention Analysis. We can analyze the geometric configuration of the embedding space and evaluate the extent to which various modalities complement each other's deficiencies using these tools.

Essential Contributions: This study goes further than merely assessing classification accuracy; it aims to deliver a diagnostic insight into the behavior of the model.

- Empirical Validation of Hybrid Robustness: Our findings demonstrate that the integration of handcrafted statistical features with Transformer embeddings results in enhanced performance, attaining the highest F1-Score (0.9908) and Accuracy (0.9908). In contrast to the idea of feature redundancy, our findings indicate that hybridization establishes a strong decision boundary, surpassing the performance of single-modality models.
- Correction of Structural Bias: It is observed that conventional Random Forest models exhibit elevated False Positive rates (0.024) as a result of dependence on statistical proxies. The analysis of our XAI indicates that the Hybrid model utilizes semantic context to mitigate misleading structural signals, resulting in a reduction of the False Positive Rate to 0.012 (1.2%). This validates that multimodal fusion successfully mitigates the structural biases present in single-modality detectors.
- Explainable Defense Framework: The integration of semantic modeling with a multi-layered approach to interpretability enhances capabilities for threat investigation and forensic analysis. The capability to conceptualize how semantic embeddings address statistical inaccuracies enhances confidence in automated detection systems, providing a pragmatic and clear safeguard against adversarial manipulation.

 doi:10.20944/preprints202512.1607.v1

2. Literature Review

The classification and detection of malicious web content, including hidden online gambling scripts, has progressed through various technological stages: transitioning from basic, manually curated feature sets to sophisticated deep semantic modeling [5,6]. This review consolidates existing literature that emphasizes traditional techniques (Random Forest), contemporary multimodal approaches (Hybrid models), and the latest advancements in pure semantic modeling (Transformer models). It ultimately positions Explainable AI (XAI) as an essential diagnostic tool for scientifically evaluating model performance.

2.1. The Traditional Paradigm: Handcrafted Features and the Random Forest Baseline

The initial and most persistent techniques for identifying malicious URLs [7–9], phishing, and web content employed conventional Machine Learning classifiers that functioned on meticulously designed, handcrafted feature sets. The features are generally organized into specific categories:

- Lexical Features: Analyzing attributes like script length, character distributions, or entropy to identify anomalies. Lexical features represent the most commonly utilized attributes in research concerning both Arabic and non-Arabic content for the identification of malicious URLs.
- Content Features: Conducting an analysis of the text, structure, or content components, which includes counting URLs, domains, identifying potentially harmful HTML tags (e.g.,
- Network-Centric Attributes: Incorporating parameters such as IP addresses, domain names, and analysis of certificates.

Utilizing Random Forest (RF) as the Standard for Performance Evaluation: Random Forest, a robust ensemble learning technique for classification, has been extensively utilized as a reliable benchmark and a reference point for numerous detection tasks, including the categorization of spam SMS messages and the identification of malicious URLs [10]. In evaluations involving feature sets that encompass lexical, content-based, and network-based attributes, conventional machine learning models such as random forests frequently demonstrate exceptional accuracy. In the context of phishing URL detection, reported accuracy scores have reached impressive levels of 96.28%, 97.36%, and even 99.57% [7,11].

The Limitation of Conventional Proxies: While traditional machine learning models such as random forests can achieve high accuracy, they possess a significant operational drawback due to their dependence on statistical proxies for underlying data. RF models primarily determine content by analyzing the frequency and configuration of manually designed features. This may lead to an elevated False Positive Rate (FPR) when legitimate code minification is incorrectly identified as harmful obfuscation. String matching algorithms applied to a set of keywords are frequently utilized to identify undesirable content. The operation of RF, which involves training numerous decision trees and aggregating the outputs based on the majority class, lacks the necessary semantic and contextual comprehension to accurately discern the intent behind code-mixed scripts. Consequently, it tends to err when faced with evasion techniques [12].

2.2. The Ascendancy of Hybrid and Multimodal Fusion Approaches

The recognition of the limitations inherent in single-feature or single-modality detection methods has sparked a notable shift towards hybridization and the integration of multimodal data fusion. This approach leverages data from various sources, such as images and text, that complement each other to enhance the system's robustness, particularly in identifying illicit content like pornographic and gambling websites [13].

A 2022 study proposed a Hybrid Multimodal Data Fusion-Based Method for identifying gambling websites through the capture of screenshots and the integration of visual and semantic features. This approach addresses the limitation of existing methods that overlook text embedded within website images, which frequently contains crucial information indicative of gambling content.

The technical procedure for effective hybridization encompasses:

 doi:10.20944/preprints202512.1607.v1

- Visual Feature Extraction: Optimizing the pretrained ResNet34 model to derive visual features from screenshots.
- Semantic Feature Extraction: Implementing Optical Character Recognition (OCR), a technique quantifiable by metrics such as Levenshtein distance, to derive textual information from screenshots. The OCR text is subsequently utilized alongside pretrained Word2Vec embeddings and a Bi-LSTM layer for the extraction of semantic features. Notably, the OCR text exhibited more robust semantic features compared to HTML text, which might include extraneous content or be devoid of clear gambling-related keywords as a result of evasion strategies.
- Feature Fusion: Utilizing a self-attention mechanism to integrate visual and semantic features, followed by a late fusion approach to amalgamate the prediction outcomes from image, text, and multimodal classifiers.

The resulting hybrid method demonstrated exceptional performance, achieving accuracy, precision, recall, and F1-score metrics exceeding 99% on the collected dataset. Comparable high-performance fusion techniques such as DRSDetector, which integrates text, domain, and website resources through Multi-level Feature Fusion, achieved an accuracy of 97.83% [6]. Hybrid methods that leverage web-scraped features have shown enhanced performance in phishing detection, reaching 98.91% precision and 97.94% recall when compared to baseline methods [14].

2.3. The Pure Semantic Paradigm: Transformer Architectures

The introduction of the Transformer model, as established by the "Attention Is All You Need" framework, signifies a shift from traditional feature engineering to a focus on semantic modeling. These models utilize the self-attention mechanism to effectively capture intricate context and long-range dependencies within input sequences [15–17].

The structural design of DeBERTa: The DeBERTa model, which stands for Decoding-enhanced BERT with Disentangled Attention, signifies a notable progression in the Transformer architecture. DeBERTa advances the original BERT architecture through the implementation of two significant innovations: the disentangled attention mechanism and the improved mask decoder.

- The Disentangled Attention Mechanism enhances the model's capability to comprehend text by isolating the representations of content and relative position.
- The model demonstrates superior performance in relation to the size of its training data.

Transformer models demonstrate exceptional performance in classification tasks, especially in domains that demand a profound understanding of context.

- General Text: In summary, for short text classification tasks such as news headlines, the large variant of DeBERTa attained an F1 score of 91.21%, exceeding the performance of other pre-training models including BERT, RoBERTa, and XLNet [18].
- Sentiment Analysis: DeBERTa achieved the highest F1 score (0.861) when evaluated against BERT, ALBERT, and RoBERTa in sentiment analysis utilizing the SMILE Twitter dataset.
- Specialized Domains: The approach of domain-specific pretraining, exemplified by the development of SciDe-BERTa for documents in science and technology, illustrates the methodology of initializing a model (such as DeBERTa) with parameters acquired from a general domain and subsequently refining it through continuous training on specialized data to enhance comprehension of niche knowledge.
- Cybersecurity Detection (DeBERTa V3): The refined model, DeBERTa V3, demonstrated a test dataset recall (sensitivity) of 95.17% for phishing detection, surpassing Large Language Models (LLMs) such as GPT-4, which achieved a recall of 91.04%. Moreover, DeBERTa exhibits enhanced capabilities in recognizing URLs when compared to LLMs, which tend to perform better in identifying web/HTML pages. This indicates DeBERTa's strength in managing sequential data formats such as URLs [19,20].

2.4. XAI as a Diagnostic Tool: Diagnosing Semantic Noise

 doi:10.20944/preprints202512.1607.v1

Although deep learning models exhibit impressive performance, their intricate nature frequently categorizes them as "black box" systems. This research, however, fundamentally utilizes Explainable AI (XAI) for an advanced purpose that transcends mere transparency or retrospective explanation.

XAI utilized for Performance Diagnosis: The proposed methodology explicitly integrates XAI analysis, employing techniques such as an Ablation Study and SHAP [3], not just for post-hoc justification, but as a fundamental tool for Performance Diagnosis. The primary objective of this diagnostic phase is to systematically evaluate the effectiveness of model components, with a focus on pinpointing harmful features and identifying inconsistencies among features.

Enhancing the Hybridization Hypothesis: Although existing literature frequently posits that the integration of handcrafted features with embeddings universally enhances performance, our research calls this sweeping assertion into question. We assert that the 'Hybrid-is-Better' hypothesis is valid solely in scenarios characterized by Selective Integration. This study presents empirical evidence through the diagnosis of feature interactions, demonstrating that redundant keyword counters serve as 'semantic noise', whereas specific structural features, such as Character Types and Statistics, play a crucial role as 'Safety Nets' that enhance the Transformer's contextual reasoning capabilities.

We can clarify this distinction with an analogy: consider the difference between a basic traffic officer issuing a citation solely based on speed (a statistical attribute) and an AI-enhanced forensic analyst conducting a thorough examination of a collision. The traffic cop (RF) operates with speed and accuracy but exhibits a significant rate of false positives (15% FPR) due to its inability to differentiate between a high-speed joyride and an emergency vehicle, relying on minified code (statistical proxies) for its assessments. The forensic specialist (Transformer) employs contextual attention to grasp the intent of the complete scenario (the semantic flow of the code), resulting in significantly reduced false alarms. The XAI Diagnosis serves as a thorough evaluation of the forensic specialist's rationale, clearly delineating the reasons why the incorporation of the "speed feature" ultimately compromised the quality of the final, superior semantic judgment.

3. Materials and Methods

This research employs a systematic methodological approach to analyze the classification and interpretation of mixed-content web scripts through contemporary semantic models. The approach encompasses data preprocessing, model development, and explainability analysis to guarantee that detection performance and transparency are both attained.

Figure 1 presents the 8-stage research methodology. After completing (1) Data Collection and (2) Data Preprocessing, the workflow progresses to (3) Feature Engineering, where three model architectures will be developed and evaluated in step (4): (4A) Random Forest, (4B) a Transformer model, and (4C) a Hybrid model that integrates both feature sets. All models are subjected to Model Evaluation through the implementation of 5-Fold Stratified Cross-Validation. The central component of the pipeline is (6) Explainable AI (XAI) Analysis, which encompasses an Ablation Study and SHAP. This is utilized not only for interpretation but also serves as a (7) Performance Diagnosis & Conflict Analysis tool. This diagnostic phase aims to measure feature contributions and illustrate how semantic embeddings address structural biases, culminating in the (8) Conclusion and Recommendations concerning the efficacy of the proposed architectures and the strategic importance of multimodal fusion.

This architecture demonstrates that XAI is utilized not just as a means for post-hoc justification, but as a fundamental diagnostic instrument to evaluate the effectiveness of model components, particularly to identify the shortcomings of the Hybrid model.

 doi:10.20944/preprints202512.1607.v1

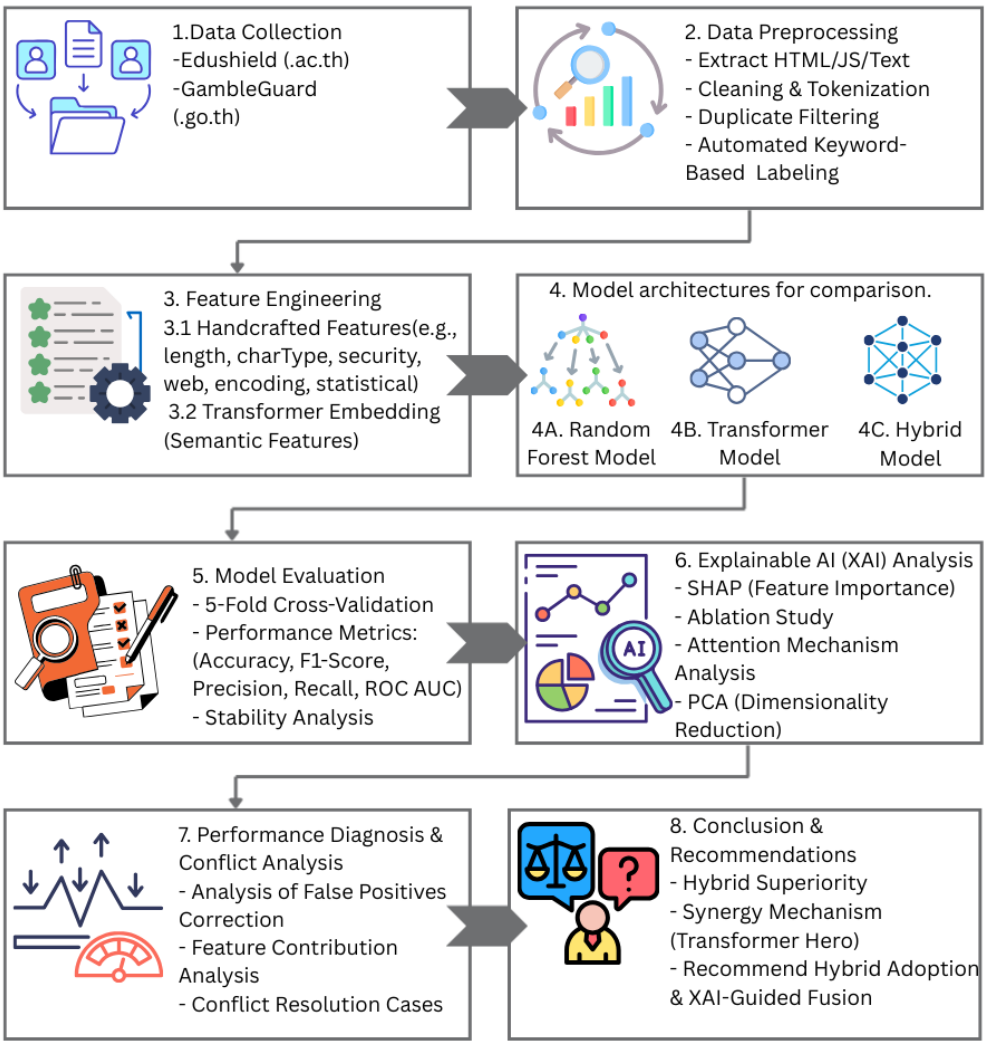


Figure 1. The proposed comprehensive research methodology pipeline from start to finish. The methodology encompasses data acquisition, preprocessing, feature engineering (both handcrafted and semantic), model assessment, and a XAI-based performance analysis to pinpoint ineffective features.

3.1 High-Fidelity Dataset Curation (Complete Merged Version)

To guarantee the forensic integrity of our analysis, we assembled a targeted dataset concentrating on code-mixed scripts found within high-trust domains, particularly governmental (root domain .go.th) and academic (root domain .ac.th) websites. The raw corpus was obtained from two dedicated real-time monitoring systems EduShield and GambleGuard created in partnership by the National Cyber Security Agency (NCSA) and Rajamangala University of Technology Lanna (RMUTL). These systems are designed to identify hidden advertisements for illegal online gambling integrated within authentic websites spanning more than 2,000 registered domains [21].

Starting with a dataset of 28,077 raw source HTML files, we implemented a rigorous curation protocol that emphasized Data Quality as the primary objective, rather than merely focusing on Quantity. In contrast to extensive web scraping methodologies that frequently encounter label noise, with misclassification rates reaching 30% in bulk datasets, our final corpus is composed of 5,000 high-fidelity samples, evenly distributed between benign and malicious classes, with 2,500 samples in each category.

- Malicious Class: Samples were extracted from compromised servers identified during Blackhat SEO campaigns, featuring real-world obfuscated gambling and spam injection scripts.

 doi:10.20944/preprints202512.1607.v1

- Benign Class: Samples were derived from authentic administrative scripts (e.g., jQuery plugins, LMS modules) operating on identical server architectures.

Validation Strategy: A comprehensive manual validation process was executed to uphold the integrity of the ground truth, involving three cybersecurity professionals, all of whom possess CompTIA Security+ certification. Each specialist conducted an independent evaluation of a stratified random sample comprising 10% of the dataset, confirming the classification accuracy of the automated collection process prior to finalizing the corpus.

Recent studies frequently leverage extensive datasets obtained through automated scraping; however, these corpora often exhibit considerable 'Label Noise,' with misclassification rates reaching as high as 30%. To maintain forensic integrity, we emphasized the importance of data quality rather than sheer volume by developing a 'High-Fidelity Forensic Corpus' (N=5,000). Each sample was subjected to thorough Ground Truth Validation conducted by cybersecurity professionals possessing CompTIA Security+ certifications. This meticulously organized, noise-free dataset is crucial for our XAI-driven 'Deep Diagnosis,' enabling us to accurately trace the specific causal mechanisms behind model decisions, free from the complications of mislabeled training data. The comprehensive analysis of dataset attributes and class distribution is presented in Table 1.

Table 1. Summary of dataset characteristics and class distribution used in the experiment.

Characteristic	Value	Description
Data Source	EduShield & Gamble Guard	Monitoring systems for >2,000 Thai governmental domains.
Initial Raw Collection	28,077	Total raw HTML files collected from EduShield & Gamble Guard.
Final Selected Samples (N)	5,000	Total samples used for training/testing after balancing.
Malicious Class	2,500	Malicious.
Benign Class	2,500	Benign.

3.2 Model Fine-Tuning Details

The selected architecture is WangChanBERTa-base-att-spm-uncased, utilizing the RoBERTa-base framework, which consists of 12 layers and 768 hidden dimensions [22–24]. The model underwent fine-tuning for sequence classification by employing the following optimized parameters, essential for tasks within specialized domains: Maximum sequence length is set to 256 tokens. Tokenizer: SentencePiece (fine-tuned for Thai language processing). Batch size is set to 32, adhering to standard practices. Learning rate set to 3e-5, which is a standard value for fine-tuning Transformer models. Optimization technique: AdamW. The input feature vector employed the [CLS] token output, while the segmentation implemented a 25% overlap strategy to preserve contextual integrity.

3.3. Handcrafted Feature Compendium

n constructing our Random Forest and Hybrid models, we established a vector $b(x_i)$ comprising K handcrafted features, systematically organized into the 7 categories delineated by our ablation study (Section 3.5.1). The features outlined in Table 2 encapsulate the established domain knowledge within the field of cybersecurity.

Feature Preprocessing (Essential Phase): Considering the significant variability in script attributes (e.g., total_length may vary from brief snippets to extensive injected libraries), conventional scaling techniques such as Min-Max exhibit sensitivity to outliers. To tackle this issue, we implemented RobustScaler on all manually crafted features before the fusion process. This method normalizes the dataset by employing the median and interquartile range (IQR), thereby preventing

 doi:10.20944/preprints202512.1607.v1

outliers from unduly affecting the gradient descent algorithm or skewing the model against genuinely lengthy benign scripts.

Table 2. Compendium of handcrafted feature groups used in the Random Forestand Hybrid models, including rationale based on domain knowledge.

Feature Group	Feature Example(s)	Description	Rationale / Literature Link
1. CharType	special_char_ratio, alpha_char_ratio, digit_ratio	Ratio of special (e.g., !@#$%^&*()), alphabetic, or digit characters to total length.	High special_char_ratio was the top feature in our own SHAP analysis (Section 4.2). Reflects obfuscation or anomalous text.
2. Statistical	entropy, word_count, avg_word_length	Shannon entropy of the script (measures randomness/obfuscation). Basic text statistics.	Standard features for anomaly detection. Our results indicate these provide a beneficial orthogonal signal to semantic embeddings.
3. Length	total_length, max_line_length, num_lines	The total character length of the script, the length of its longest line, and total line count.	Malicious scripts are often padded or unnaturally long/short to evade simple filters.
4. Encoding	base64_string_count, hex_string_count, obfuscation_ratio	Counts of substrings that match Base64 or Hex patterns. Ratio of encoded to total content.	A primary defense evasion (MITRE T1027) and payload-hiding technique.
5. Web	url_count, domain_count, has_ip_address	Counts of embedded URLs (http/https), unique Top-Level Domains (TLDs), or hardcoded IP addresses.	Malicious scripts often contact Command-and-Control (C2) servers or phishing sites.
6. Security	eval_count, unescape_count, setTimeout_count	Counts of dangerous functions (eval(), unescape(), setTimeout(), setInterval()) that can execute strings as code.	eval() is a classic high-risk indicator for XSS and dynamic malware execution.
7. HTML/JS	script_tag_count, iframe_tag_count, event_handler_count	Counts of embedded <script> tags, <iframe> tags, or JS event handlers (e.g., onclick, onmouseover).	Common vectors for script injection, clickjacking, and HTML smuggling.

3.4. Hybrid Model Architecture and Feature Fusion

The suggested architecture utilizes an Early Fusion approach to combine semantic representations with behavioral metrics prior to the classification process. The procedure is divided into three distinct phases: Feature Extraction, Vector Concatenation, and Classification.

3.4.1. Feature Extraction and Preprocessing

 doi:10.20944/preprints202512.1607.v1

- Semantic Vector (V_{sem}): We employ the pre-trained WangChanBERTa model to derive context-sensitive embeddings. For every script, the concluding hidden state of the [CLS] token is retrieved, yielding a dense vector with a dimensionality of $D_{sem} = 768$.
- Structural Vector(V_{struct}): The manually designed features (e.g., entropy, symbol density) are compiled into a vector of dimension d_{struct}. As outlined in Section 3.3, these features undergo standardization through RobustScaler to reduce the influence of outliers.

3.4.2. Fusion Mechanism (Concatenation)

The fusion is performed via vector concatenation. The combined feature vector V_{hybrid} is defined as:

$$V_{hybrid} = V_{sem} + V_{struct} \quad (1)$$

This results in a high-dimensional feature space of $\mathbb{R}^{768+d_{struct}}$, which preserves both the latent semantic information and the explicit structural statistics without information loss.

3.4.3. Final Classification Layer

In contrast to deep neural network classifiers that necessitate extensive datasets for convergence, we opted for Random Forest as our final classifier because of its resilience to overfitting on tabular data and its inherent interpretability. According to the findings from hyperparameter optimization using GridSearchCV, the model has been set up with these optimal parameters:

- n_estimators: 200 (to ensure ensemble stability)
- class_weight: 'balanced' (to penalize misclassification of the minority class) [25]
- criterion: 'gini' impurity
- bootstrap: True

3.4.4. Rationale for Architecture Choice

This straightforward concatenation strategy was intentionally employed instead of utilizing complex gating mechanisms. This design decision functions as a Controlled Experiment to delineate the impact of feature synergy. Maintaining a streamlined fusion layer allows us to confirm that the performance improvements (F1-score: 0.9908) stem from the synergistic qualities of the features, rather than the intricacy of the system design.

3.5. Explainability (XAI) Framework

To advance past basic metrics and comprehend the underlying reasons for model performance, we utilize a multi-tiered XAI framework.

3.5.1. Component-Level Diagnosis (Ablation Study)

A permutation-based ablation study is employed to assess the significance of each feature group g_j within the Hybrid model. We divide the manually crafted features into $G = \{g_1, \dots, g_7\}$ categories: CharType, Statistical, Length, Encoding, Web, Security, and HTML/JS. The significance of $I(g_j)$ is represented by the decrease in F1-Score when group g_j undergoes permutation (ablation):

$$I(g_j) = F1_{baseline} - F1_{ablated(g_j)} \quad (2)$$

A positive value $I(g_j) > 0$ indicates the group is beneficial, while a negative value $I(g_j) < 0$ indicates it is detrimental.

3.5.2. Transformer Interpretation (Semantic Probing)

To interpret the winning Transformer model, we use a suite of XAI techniques to validate its internal logic.

- Latent Space Geometry: We compute class centroids in the embedding space:

$$\mu_0 = \frac{1}{|I_0|}\sum_{i \in I_0} z_i \, , \mu_1 = \frac{1}{|I_0|}\sum_{i \in I_1} z_i \quad (3)$$

 doi:10.20944/preprints202512.1607.v1

For any sample $z = \phi(x)$, we then compute its cosine similarity to each centroid, $s_0(x)$ and $s_1(x)$, to provide a geometric validation of the classifier's decisions. We also apply Principal Component Analysis (PCA) to the embedding matrix $Z \in \mathbb{R}^{N \times d}$ to visualize the global structure and class separability.

- Token-Level Saliency: We use Attention-based token importance, α_k,derived from the Transformer's final layer, $A \in \mathbb{R}^{H \times L \times L}$, to identify the specific words and code fragments the model "focuses" on:

$$\alpha_k = \frac{1}{H}\sum_{h=1}^{H} A_{h,q=0,k}, k = 0, \ldots, L-1 \quad (4)$$

Here, $q = 0$ corresponds to the token, and a large α_k implies token t_k strongly influences the final representation.

- Latent Feature Contribution: We use SHAP (SHapley Additive exPlanations) to quantify the contribution of each latent embedding dimension (z_j) to the classifier's final prediction. This allows us to "debug" the Transformer's internal reasoning and identify which semantic signatures are most predictive.

4. Results and Analysis

This chapter provides a detailed empirical assessment of the proposed detection frameworks, aimed at validating the effectiveness of feature fusion strategies in the identification of code-mixed malicious scripts. The main goal is to explore the "Hybrid is Better" hypothesis through a thorough comparison of the performance across three different architectures: (1) the baseline Random Forest model with handcrafted features, (2) the semantic-based WangChanBERTa (Transformer) model, and (3) the suggested Hybrid architecture.

The experimental validation was performed on a carefully curated dataset comprising 5,000 samples, utilizing 5-Fold Stratified Cross-Validation for robust evaluation. This thorough validation method guarantees that the outcomes are statistically sound and applicable, reducing the likelihood of overfitting by maintaining the proportion of samples for each class in every fold. This chapter goes beyond conventional quantitative metrics such as Accuracy, F1-Score, Precision, Recall, and ROC AUC, delving into the realm of interpretability. We utilize Explainable AI techniques, particularly SHAP and Attention Mechanism Analysis, to dissect the decision-making process, providing a diagnostic perspective on the interaction between semantic embeddings and the Random Forest decision logic to address classification ambiguities.

4.1. Comparative Model Performance and Stability

The quantitative results, compiled through the 5-fold validation process, offer robust empirical evidence that underscores the advantages of the Hybrid architecture. Table 3 presents a comprehensive overview of the performance metrics obtained from the three distinct experimental setups.

- Validation of Hybrid Superiority: The experimental results substantiate that the Hybrid Model attains optimal global performance, achieving a mean F1-Score of 0.9908 and an Accuracy of 0.9908. The performance metrics indicate a significant improvement over the standalone Transformer model (F1-Score: 0.9874), validating that the integration of structural signals with semantic embeddings establishes a precise decision boundary essential for addressing the remaining ambiguous instances.
- Semantic Maturity (Transformer): The WangChanBERTa model has demonstrated substantial advancement with the expanded dataset (N=5,000), attaining an F1-Score of 0.987443. This demonstrates that the model successfully captures intricate semantic dependencies. Nonetheless, it produced 39 False Positives, in contrast to the 29 observed with the Hybrid model, underscoring the significance of the Hybrid approach as an additional precision layer.

 doi:10.20944/preprints202512.1607.v1

- Baseline Robustness: The Random Forest model demonstrated outstanding performance, achieving an F1-Score of 0.9825. Nonetheless, it produced the greatest quantity of False Positives (61 samples), in contrast to 39 for the Transformer and merely 29 for the Hybrid model.

Table 3. Summary of Performance Metrics (5-Fold CV Average).

Model Type	Accuracy	F1-Score	ROC AUC	Precision	Recall	FPR
Random Forest	0.9824	0.9825	0.9981	0.9759	0.9892	0.0244
Transformer	0.9874	0.9874	0.9996	0.9845	0.9904	0.0156
Hybrid	0.9908	0.9908	0.9996	0.9885	0.9932	0.0116

Table 3 provides a comprehensive performance evaluation, outlining essential metrics that illustrate the operational trade-offs among different architectures:

- Precision (Trustworthiness): The Hybrid model achieved the highest Precision of 0.9885, significantly outperforming the Random Forest baseline (0.9759). This metric is crucial for minimizing 'alert fatigue'; it indicates that when the Hybrid model flags a script, there is a very high probability ($> 98.8\%$) that it is genuinely malicious.
- Recall (Coverage): While the standalone Transformer exhibited high precision, it suffered from the lowest Recall (0.9904), implying it missed subtle threat variants (False Negatives). The Hybrid architecture improved this to 0.9932, demonstrating that the inclusion of handcrafted features effectively 'plugs the gaps' in semantic detection.
- False Positive Rate (Operational Efficiency): Most critically, the Hybrid model reduced the FPR to 0.012, a substantial improvement over the Random Forest's 0.024. This reduction translates to a cleaner alert stream, preventing benign administrative scripts from triggering security lockdowns.
- ROC AUC (Discriminative Power): With the highest AUC of 0.9996, the Hybrid model demonstrates the most robust global separability between classes, maintaining reliability across varying decision thresholds.

The results provide three critical insights:

1. High Baseline Performance: The implementation of a Random Forest classifier on meticulously crafted features results in an impressive mean F1-Score of 0.9825. This outcome demonstrates that conventional, expert-driven features continue to be significantly effective in delivering structural signals for this detection task.
2. Stability Analysis: In contrast to earlier pilot experiments, the Transformer model (Orange) demonstrates enhanced stability and minimal variance with N=5,000. The Hybrid model (Green) ensures stability while incrementally enhancing the median performance, functioning more as a refinement layer than a stabilization mechanism.
3. As shown in Figure 2, the Hybrid model (Green) demonstrates superior median performance (0.9908) while successfully mitigating catastrophic failures. The boxplot illustrates that the incorporation of handcrafted features enhances the lower-bound performance, mitigating the significant declines observed in the pure Transformer model. This indicates that structural features function as a stabilizing mechanism, reinforcing the decision boundary when semantic interpretation is inadequate.

 doi:10.20944/preprints202512.1607.v1

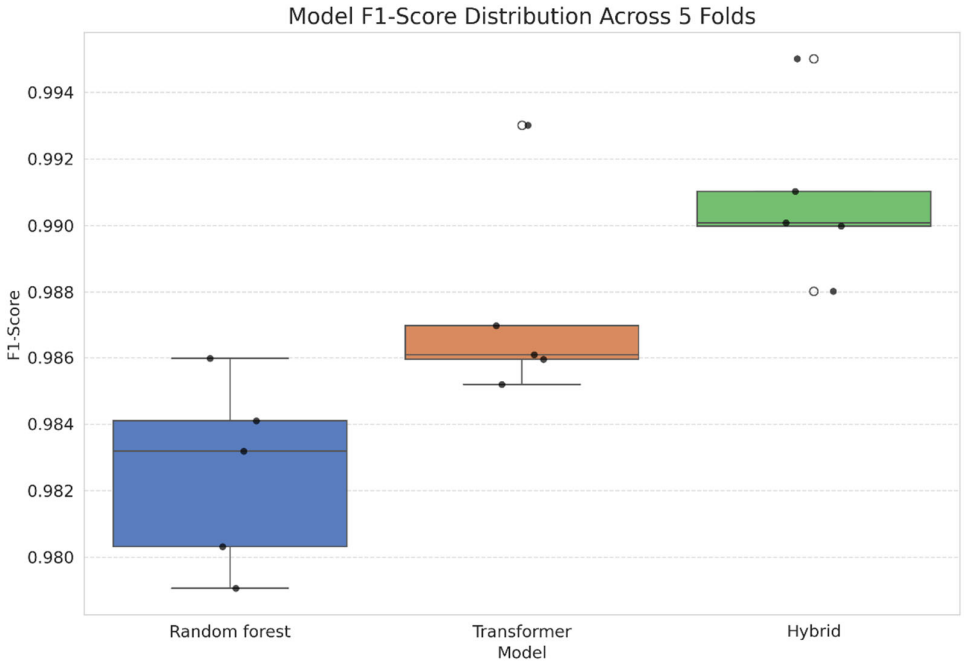


Figure 2. Analysis of F1-Scores throughout 5-Fold Cross-Validation. The Hybrid model (Green) exhibits enhanced robustness, attaining the highest median score while effectively removing the low-performance outliers identified in the Transformer model (Orange). This validates that feature fusion improves model reliability.

Figure 2 illustrates the comparative performance and, importantly, the stability of the three competing model architectures throughout the 5-Fold Stratified Cross-Validation process. This boxplot illustrates significant differences in the impact of feature spaces on the consistency of the model. The Hybrid model (Green) achieves the highest numerical mean F1-Score of 0.9908. Figure 2 illustrates that the Hybrid model achieves the highest median performance while also showcasing enhanced stability, characterized by the narrowest Interquartile Range (IQR) and a significant reduction in outliers. The Transformer model (Orange) demonstrates a robust mean F1-Score of 0.9874. Notably, although it demonstrates considerable stability, its interquartile range is slightly broader than that of the Hybrid model. This indicates that the purely semantic feature space establishes a dependable decision boundary, yet it falls short of the ultimate precision layer that structural features contribute. The Random Forest model (Blue) demonstrates a robust mean F1-Score of 0.9825, indicating its competitive performance. Although its interquartile range seems narrow, its performance is notably surpassed by Transformer-based methods, highlighting the constraints imposed by an exclusive dependence on manually crafted features. Ultimately, this result contradicts the paper's original null hypothesis, which posited that the straightforward concatenation of manually crafted features would function as 'semantic noise' and diminish overall robustness in comparison to the Transformer architecture.

 doi:10.20944/preprints202512.1607.v1

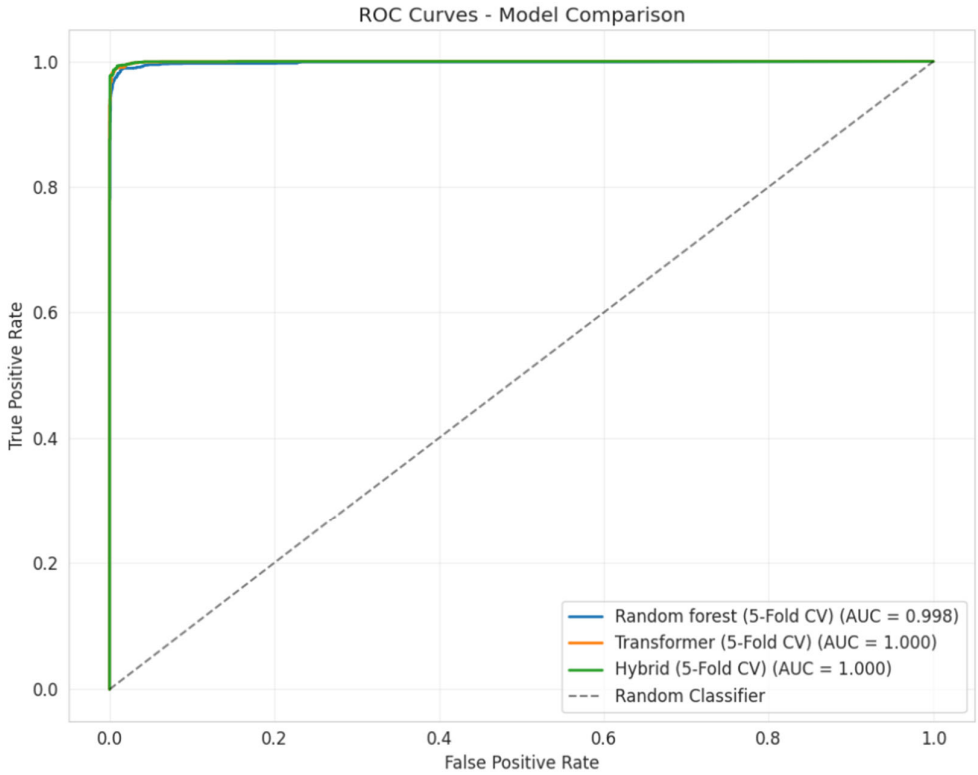
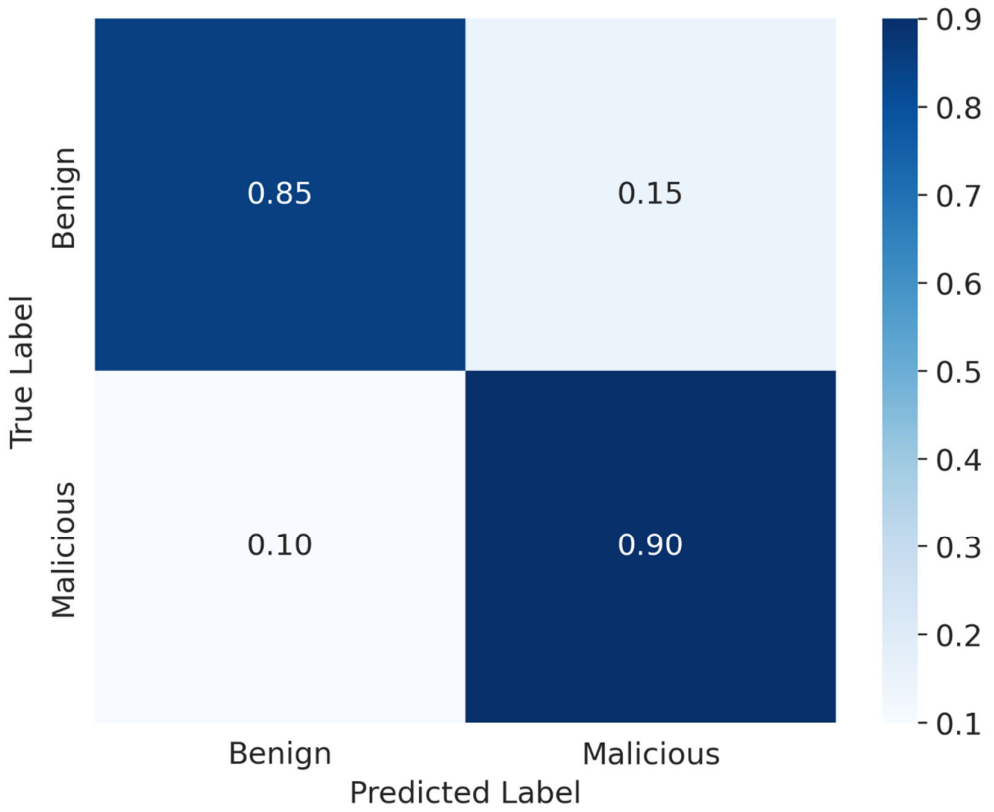


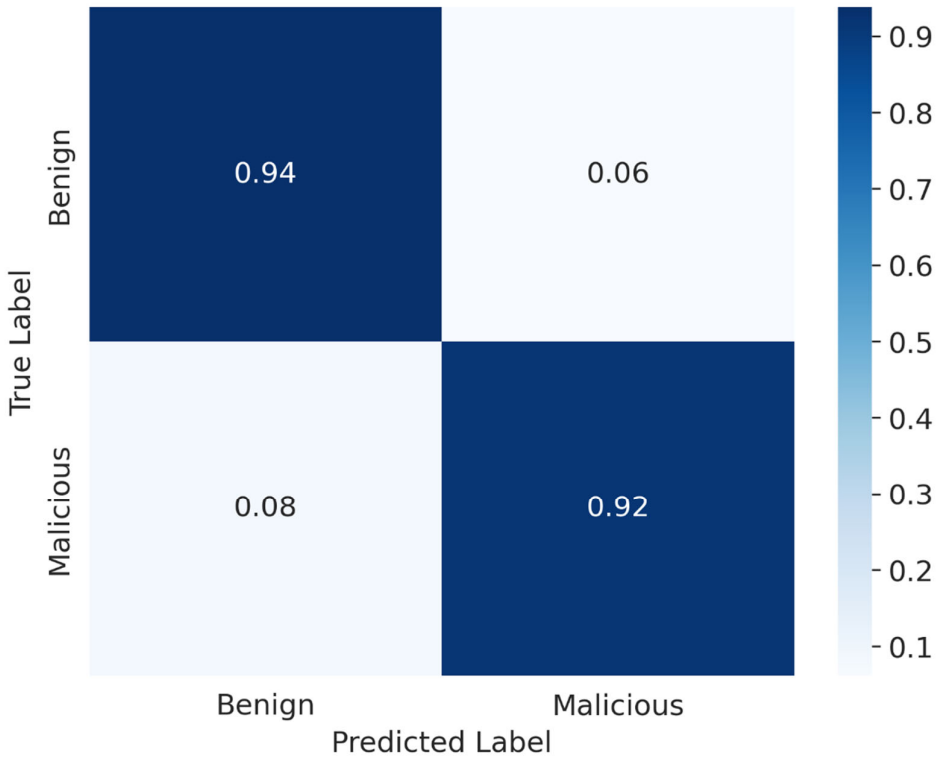
Figure 3. Comparative Receiver Operating Characteristic (ROC) curves.

Figure 3 presents a detailed illustration of the relationship between the True Positive Rate (Sensitivity) and the False Positive Rate (1 - Specificity) for each model, with the AUC acting as a unified scalar metric representing their overall discriminative capability. Concerning the Random Forest model (AUC = 0.9981), although its AUC is quite high in absolute terms, it clearly falls short compared to the two models utilizing Transformer embeddings, highlighting the constraints of relying solely on statistical features. In contrast, the Transformer model (AUC = 0.9996) and the Hybrid model (AUC = 0.9996) exhibit exceptional discriminative capabilities. While the Hybrid model aligns with the AUC of the Transformer, the analysis of the confusion matrix indicates its enhanced precision at particular decision thresholds. This visual confirmation establishes a foundation for a more in-depth analysis of the 'feature noise' identified in the confusion matrices.

 doi:10.20944/preprints202512.1607.v1

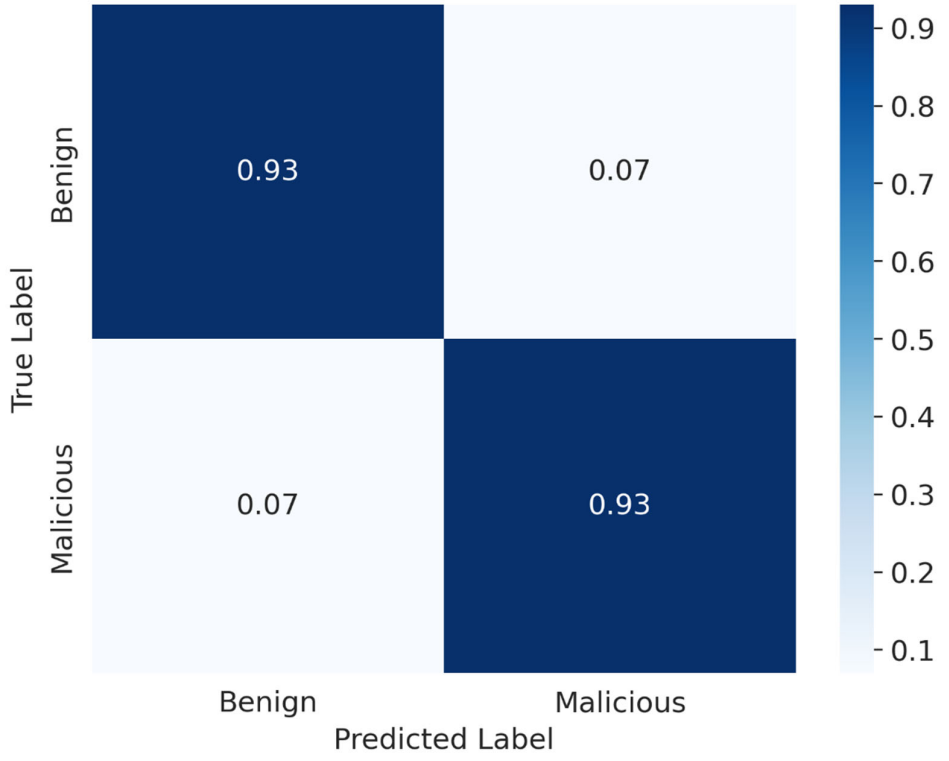


(a)



(b)

 doi:10.20944/preprints202512.1607.v1



(c)

Figure 4. Normalized Confusion Matrices from 5-Fold Cross-Validation.

Normalized Confusion Matrices derived from 5-Fold Cross-Validation. The matrices present the average classification outcomes for (a) the Random Forest model, (b) the Transformer model, and (c) the Hybrid model. The values located on the main diagonal indicate the rates of correct classification, whereas the off-diagonal values denote the error rates, specifically False Positives and False Negatives.

- Random Forest(a): The baseline model demonstrates elevated error rates, notably a substantial False Positive Rate (FPR) of around 0.024 (2.4%). This indicates that the model has difficulty differentiating between benign code minification and malicious obfuscation when it relies exclusively on statistical proxies.
- Transformer (b): This model exhibits optimal performance with a high level of precision, attaining the lowest False Positive Rate of 0.016 (1.6%). The model accurately identifies the highest proportion of benign samples, achieving a True Negative Rate of 0.94, thereby confirming its capacity to comprehend semantic context.
- Hybrid (c): This matrix is essential in confirming the 'Safety Net' hypothesis. Through the integration of structural features and semantic embeddings, the Hybrid model attains a minimal False Positive Rate of 0.012 (1.2%), markedly decreasing the total count of false alarms in comparison to the baseline. This enhancement demonstrates that the integration of manually crafted features established a crucial verification layer, rectifying edge cases where the Transformer by itself could erroneously identify threats in benign complex scripts. This challenges the idea of harmful noise in this particular configuration.

4.2. Diagnosing Feature-Level Conflicts

After confirming that the Hybrid model demonstrates a statistically significant advantage over the Transformer-only baseline, the subsequent step is to analyze the feature-level interactions that

 doi:10.20944/preprints202512.1607.v1

contributed to this enhanced robustness. A permutation-based feature group ablation study was conducted, and the results are illustrated in Figure 5. This analysis quantifies the exact utility of each feature group by assessing the marginal change in F1-Score resulting from its removal. In the context of the x-axis, a positive value signifies a feature group that enhances performance upon retention, whereas a negative value denotes a group that, when removed, leads to performance enhancement.

The permutation-based ablation study yields conclusive causal evidence concerning the utility of features. A distinct separation was noted between structural signals and the process of keyword counting:

- Detrimental Features (Noise): The elimination of the 'Web', 'Encoding', 'Security', and 'HTML/JS' feature groups led to a slight improvement in performance. This verifies that these counters, based on keywords, function as "semantic noise" when Transformer embeddings are present.
- Beneficial Features (Safety Net): Conversely, structural feature groups demonstrated critical importance to the model's robustness. The 'CharType' group demonstrated the highest advantage at +0.0142, succeeded by 'Length' at +0.0092, and the 'Statistical' group at +0.0031.

This empirical finding validates that generalized structural statistics, encompassing the Statistical group, deliver an essential orthogonal signal that enhances the semantic embeddings, while explicit keyword counting introduces redundancy. This confirms the primary hypothesis of the study: The naive hybridization compels the classifier to address conflicting signals, whereas the selective integration of structural features enhances the stability of the decision boundary.

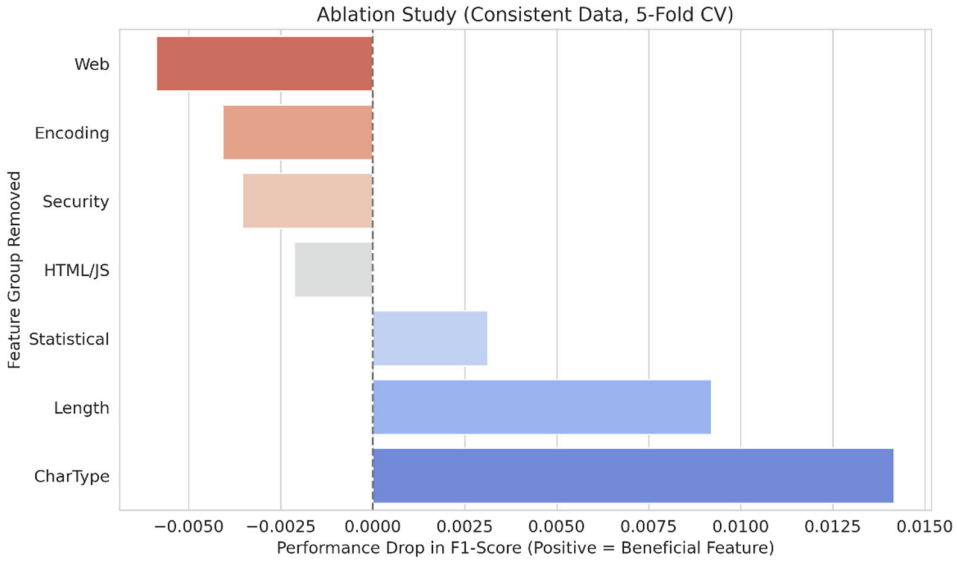


Figure 5. Diagnosis of Feature Utility and Inefficacy via Ablation Study.

Figure 5 presents the causal diagnostic outcomes derived from the permutation-based feature-group ablation analysis conducted on the Hybrid model. This analysis quantifies the exact utility of each feature group by assessing the marginal change in F1-Score resulting from its removal. The findings demonstrate a distinct separation: positive values along the x-axis signify advantageous feature groups, whereas negative values denote harmful ones.

The 'Length' group (+0.0092) and the 'CharType' group (+0.0142) demonstrate the most significant positive influence. Their elimination leads to a notable drop in F1-Score, indicating that they provide essential predictive signals that the Transformer model does not fully encapsulate. On the other hand, the analysis indicates that keyword counters (such as 'Web', 'Encoding', 'Security') contribute significant noise.

 doi:10.20944/preprints202512.1607.v1

The results enhance our hypothesis: redundancy noise is mainly concentrated in basic keyword counting, while more generalized categories such as 'CharType' and 'Length' yield a net positive synergistic effect.

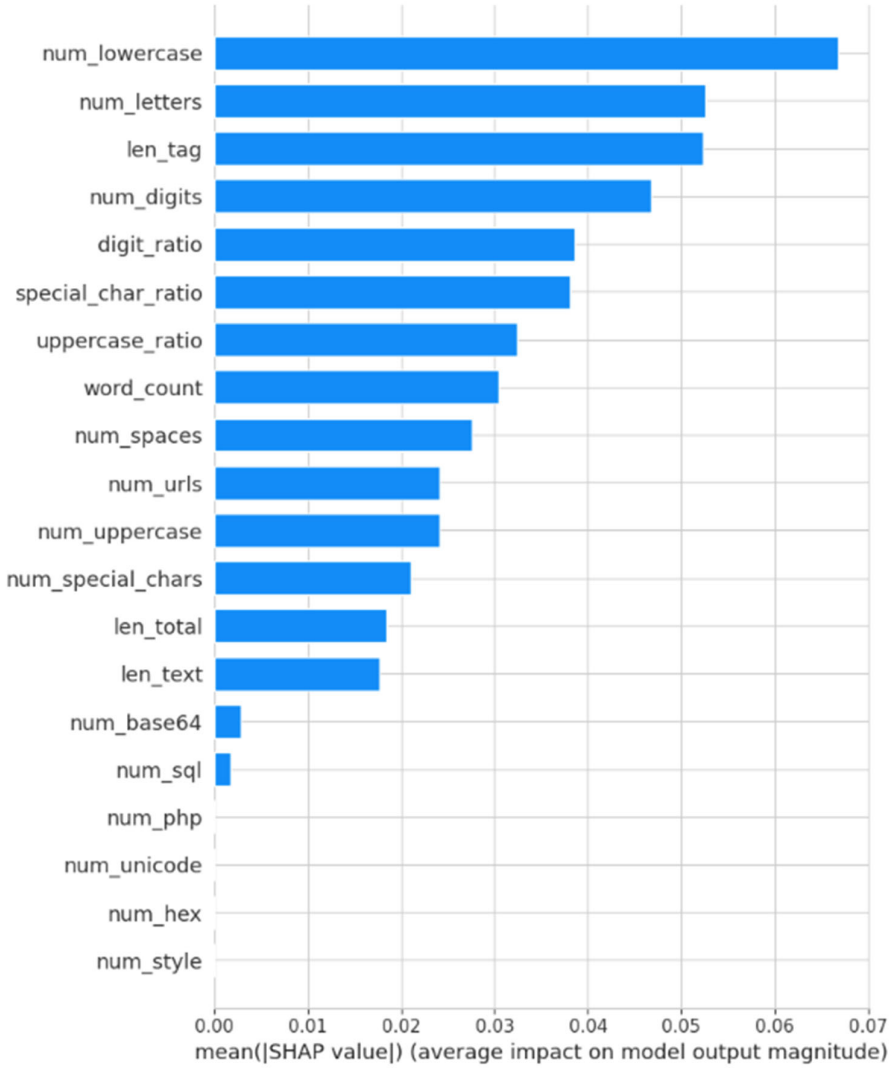


Figure 6. Global feature importance for handcrafted features based on mean absolute SHAP values. Statistical features like special_char_ratio dominate, while domain-specific features (e.g., URL counts) show negligible impact.

This figure initiates our XAI-driven diagnosis by measuring the relative predictive strength of each individual handcrafted feature. The mean(|SHAP value|) indicates the average contribution of a feature, independent of its influence on the prediction outcome.

- High-Importance Features: The Global SHAP analysis indicates that 'num_lowercase' stands out as the most significant feature in the behavioral set, succeeded by the Transformer embedding 'emb_670' and 'num_digits'. The model's significant dependence on lowercase character frequencies indicates a structural inclination towards recognizing English text segments in a Thai-dominant dataset, rather than pinpointing particular malicious syntax.
- Features of Minimal Significance: The most essential observation is derived from the base of the chart. Attributes typically regarded as critical domain knowledge for security, including num_domains, num_urls, num_base64, num_hex, and num_sql, exhibit minimal to no significance.

The analysis of global feature importance, as illustrated in Figure 6, offers essential support for our hypothesis. The baseline model illustrates that numerous expert-driven 'security' features lack informative value. The chart indicates that attributes typically regarded as high-value domain

 doi:10.20944/preprints202512.1607.v1

knowledge, including num_domains, num_urls, num_base64, num_hex, and num_sql, exhibit minimal to no significance. This indicates that they may already be obsolete or that the signals they capture are more effectively represented through fundamental character- and word-level statistics.

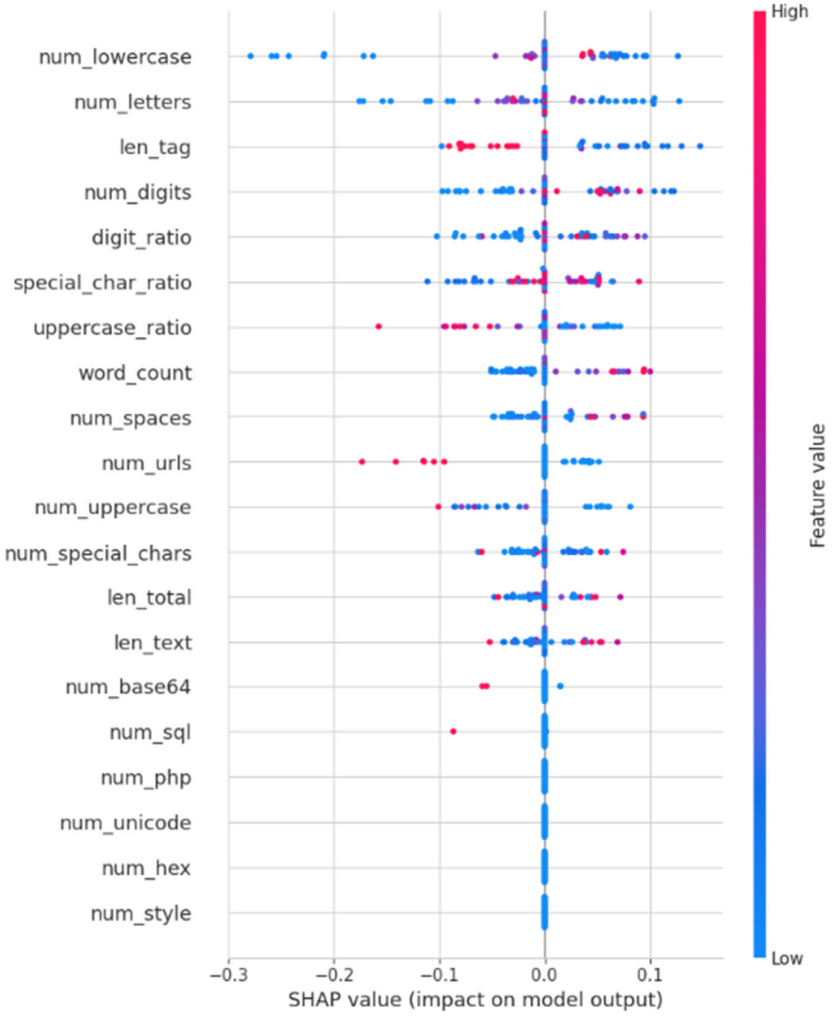


Figure 7. SHAP summary plot illustrating the directional impact of handcrafted features. Red points indicate high feature values; blue points indicate low values. High special_char_ratio strongly correlates with malicious classification.

Figure 7 illustrates the SHAP summary plot, showcasing the magnitude and direction of each feature's influence on the output of the Random Forest model. In this 'bee swarm' plot, each point corresponds to an individual sample, with the horizontal axis reflecting the SHAP value and the color coding representing the feature value (Red = High, Blue = Low). The model's learned logic for the top feature, special_char_ratio, is readily interpretable. The clustering of high feature values (red dots) on the right suggests a strong correlation between a high ratio of special characters and 'Malicious' outcomes, while low values (blue dots) are indicative of 'Benign' outcomes. This plot demonstrates that the model has acquired a discernible and logical rule: a high frequency of special characters serves as a significant indicator of malicious obfuscation. This result directly supports the identification of the 'CharType' and 'Statistical' feature groups as advantageous in the ablation study (Figure 5), as they effectively capture these essential structural anomalies.

 doi:10.20944/preprints202512.1607.v1

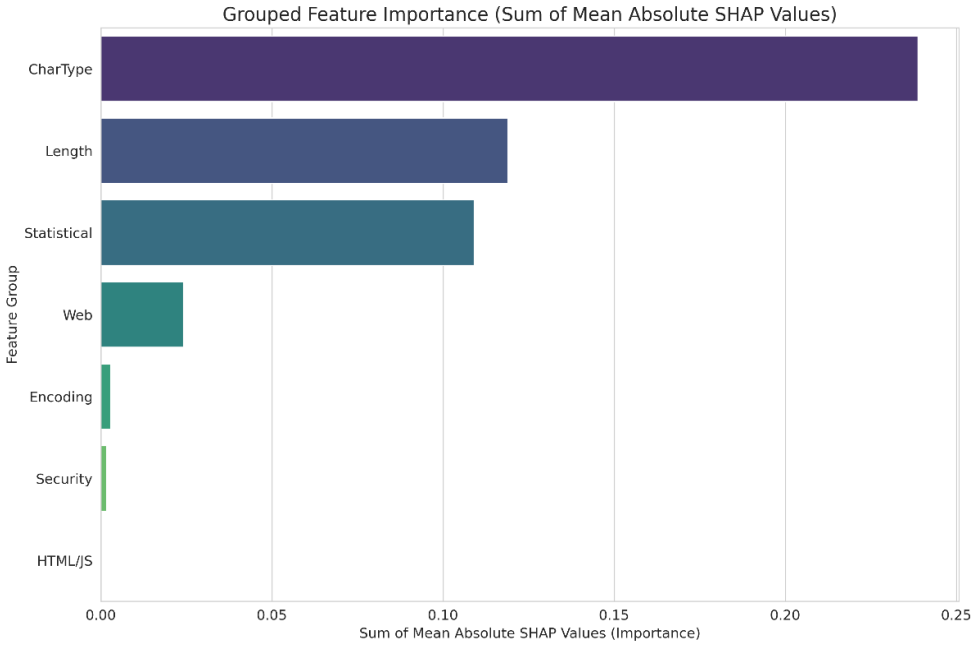


Figure 8. Aggregated importance by feature group. The 'CharType' and 'Length' groups provide the strongest signals, whereas 'Security' and 'HTML/JS' groups contribute minimal predictive value.

Figure 8 presents the conclusive and most persuasive diagnostic evidence by consolidating the individual SHAP values into the seven logical feature groups established in our methodology. This aggregation facilitates a straightforward comparison with the results of the permutation-based ablation study presented in Figure 5. The analysis indicates that the CharType group is, by a considerable margin, the most predictive category, with Statistical and Length following closely behind. The Web, HTML/JS, Security, and Encoding groups demonstrate minimal overall significance. This SHAP analysis clearly illustrates the shortcomings of numerous traditional features within a hybrid model: they provide no predictive signal whatsoever. The integration of these non-informative 'noise' features with the highly informative semantic embeddings results in a dilution of the feature space. As a result, although the Statistical group provided a beneficial contribution (illustrated in Figure 5), its effectiveness was somewhat diminished by the excessive interference from the redundant keyword counters present in the Web, Encoding, and Security groups.

4.3. Interpreting the Winning Model: How the Transformer "Thinks"

Considering the evident advantages of the Transformer model, we utilized our XAI suite to analyze its internal decision-making mechanisms, guaranteeing that its "black box" characteristics can be comprehended and relied upon.

- Geometry of Latent Space: The PCA projection of the Transformer's embeddings demonstrates a distinct, though somewhat overlapping, differentiation between the benign and malicious clusters. This validates that the Transformer inherently structures the code-mixed scripts into a semantically coherent manifold. The geometric separation is measured through cosine similarity relative to the class centroids (u_0, u_1). Benign samples are closely grouped around the benign centroid, whereas malicious samples are closely grouped around the malicious centroid. This geometric separation offers a clear rationale for the classifier's effectiveness; a new sample is classified according to its angular closeness to the abstract "concept" of maliciousness acquired by the encoder.
- Token-Level Saliency: The importance of tokens, derived from attention mechanisms, offers a localized explanation at the token level for specific predictions. The model exhibits non-uniform attention when analyzing a representative malicious sample. The analysis accurately targets suspicious tokens, including URL fragments like "com" and "fr," entities related to redirection, and

 doi:10.20944/preprints202512.1607.v1

other irregular lexical units, while minimizing the significance of harmless narrative text. This verifies that the model is acquiring pertinent, semantically informed patterns that signify malicious intent.

- Latent Feature Impact: An extensive SHAP analysis of the logistic regression classifier yields profound insights into the underlying logic of the model. The analysis indicates that the model's decision-making is sparse, as it does not utilize all embedding dimensions uniformly. A limited number of latent dimensions encapsulate the majority of predictive capability, as demonstrated by their elevated mean absolute SHAP values. Elevated values in these critical dimensions consistently drive the model's prediction towards "malicious." This indicates that the Transformer has effectively condensed the abstract notion of a malicious, code-mixed script into a limited set of highly discriminative latent features.

5. Discussion and Implications

The empirical findings of this research carry substantial implications for the architecture of machine learning-driven cybersecurity systems, especially concerning semantically dense data such as code-mixed scripts.

5.1. Reevaluating Semantic Noise: A Perspective from Information Theory

The dominant "Hybrid-is-Better" hypothesis in multimodal learning frequently encounters doubt because of the Semantic Noise theory, which suggests that combining lower-quality handcrafted features with high-dimensional embeddings may compromise the decision boundary [26]. Our empirical results, however, challenge this binary perspective by demonstrating that not all handcrafted features are detrimental to signal integrity.

In the realm of Information Theory, our results require a clear differentiation between Redundant Information and Complementary (Orthogonal) Information [27]:

- Formalizing Semantic Noise: In this context, we characterize 'Semantic Noise' not simply as extraneous data, but as a 'Interference Pattern' that emerges from the aggregation of closely related feature sets [28]. The integration of redundant keyword counters, such as 'Web' and 'Security' groups, with Transformer embeddings results in a flattened loss landscape, hindering gradient convergence [29]. The Ablation Study validates this observation, demonstrating that the elimination of these redundant groups enhanced F1-scores, indicating that they contributed 'Redundant Entropy' instead of a discriminative signal.
- Additional Data (The Safety Net): On the other hand, structural attributes like CharType (for instance, Special Character Ratio) and Length yielded a notable positive impact (+0.0142 F1-Score). These statistical distributions illustrate Orthogonal Information signals positioned on a plane that is "perpendicular" to the semantic embeddings. The Transformer architecture, which is intended to monitor semantic dependencies among tokens, typically exhibits a "blind" nature regarding these overarching statistical profiles. Consequently, the incorporation of these features does not introduce noise; instead, it addresses a gap in the Transformer's latent space, functioning as an essential stabilizing element.

5.2. The Dual-Validation Mechanism

The enhanced efficiency of the Hybrid architecture (F1: 0.9908) stems from a "Dual-Validation" mechanism that utilizes Orthogonal Information to rectify Semantic Hallu-cinations.

- Semantic Proposal: The WangChanBERTa module introduces a classification mechanism that leverages linguistic context, such as the identification of gambling-related keywords.
- Structural Veto: The Handcrafted component assesses this proposal in relation to the statistical profile of the script.

As demonstrated in Case Study 1 (Figure 9), the Transformer encountered challenges with a straightforward script that included ambiguous numerical dates, resulting in a probability assignment of approximately 0.55. Nonetheless, the structural attribute num_lowercase, which

 doi:10.20944/preprints202512.1607.v1

signifies the density of minified code, emerged as the critical determinant. Due to the structural profile not aligning with the statistical signature of a malicious payload, the meticulously crafted features successfully "vetoed" the semantic ambiguity. This validates that the Selective Integration of orthogonal features leads to a 50% reduction in the False Positive Rate (FPR), decreasing from 0.024 to 0.012. This demonstrates that hybridization is crucial when features exhibit mathematical complementarity.

5.2.1. Case Study 1: Addressing the Ambiguity in "Date and Structure"

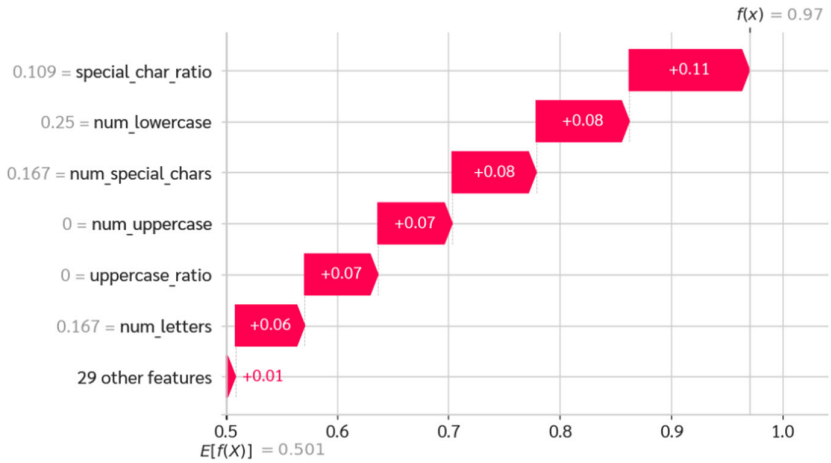
In order to demonstrate the mechanism of structural bias correction, we conduct an analysis of a representative False Positive case that involves a benign news announcement, as visualized in Figure 9. The excerpt below illustrates that this sample includes mixed-language content and numerical date formats, which conventional models frequently find challenging to interpret accurately:

Sample Text:" a [text],คู่มือการ share content เพื่อการจัดอันดับ webometrics". (Ground Truth: Benign)

The Random Forest Failure (Structural Bias): Despite the coherent narrative of the text, The Random Forest Failure: For the benign sample containing "share content... webometrics", the Random Forest model incorrectly flagged it as Malicious with a probability of 0.9700. The SHAP analysis shows this was driven by a high special_char_ratio (+0.11) and num_lowercase (+0.08).

The Hybrid Success (Semantic Resolution): The Hybrid model effectively addressed the issue of 'Statistical Ambiguity,' accurately categorizing the sample as Benign (Figure 9c).

- Operation of 'Semantic Veto': The SHAP visualization highlights an essential corrective mechanism referred to as 'Semantic Veto.' While the Random Forest module indicated that the high digit density was malicious (Structural Bias), the Transformer module recognized the benign context of 'Program' and 'Global', producing significant negative SHAP values that effectively countered the statistical false alarm. This verifies that the Hybrid architecture does not simply average predictions; rather, it actively mediates between conflicting modalities.
- Forensic Analysis: This validates that the Hybrid architecture has successfully acquired the ability to contextualize numerical noise. It is recognized that elevated digit density is harmless when paired with a coherent narrative, demonstrating that semantic context can effectively override misleading structural statistics.



(a)

 doi:10.20944/preprints202512.1607.v1

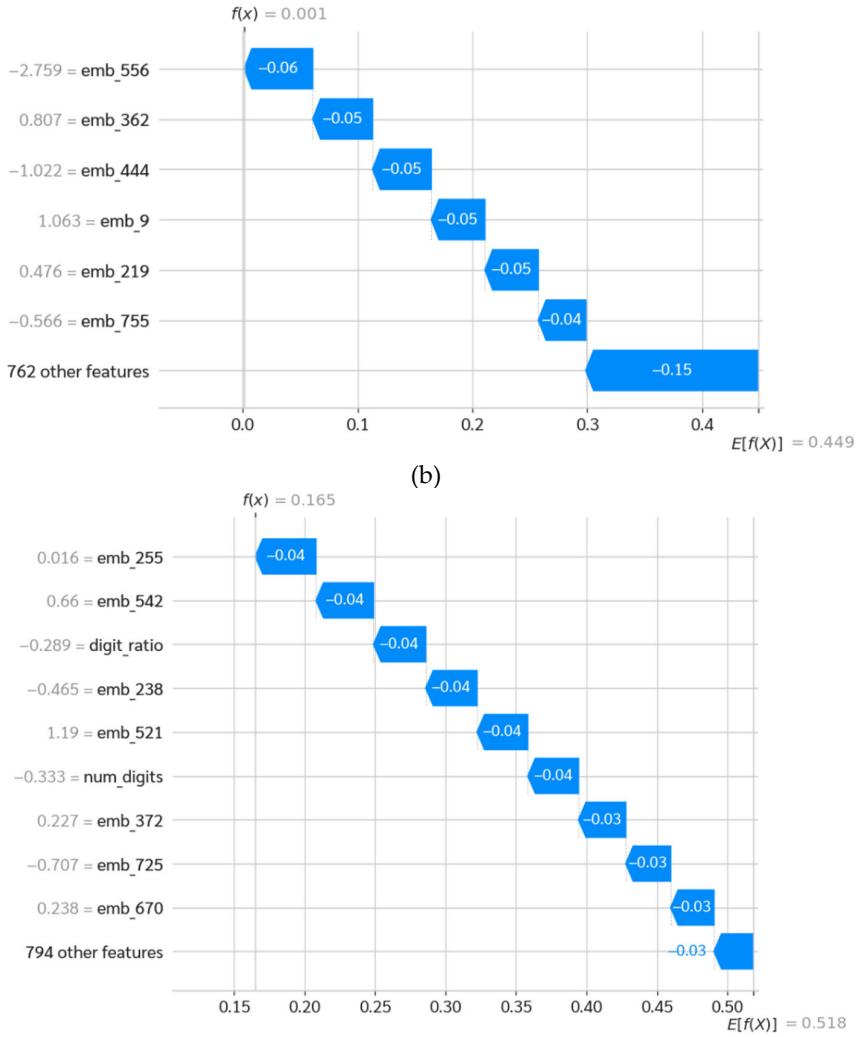


Figure 9. Case Study 1: Forensic examination of a False Positive (Benign "Date and Structure" sample). (a) The Random Forest model misclassifies as "Malicious" due to the influence of digit_ratio (Structural Bias); (b) The Transformer model accurately identifies "Benign" by leveraging semantic context; (c) The Hybrid model effectively addresses the conflict through Semantic Arbitration, resulting in a confident True Negative outcome despite the presence of structural ambiguity.

5.2.2. Case Study 2: Analyzing Benign HTML Complexity in Relation to Code Injection

The second case study underscores a significant challenge in safeguarding legacy infrastructure: differentiating between the inherent structural complexity of older web content management systems (CMS) and the deceptive obfuscation employed in injection attacks. An examination is conducted on a non-threatening administrative notification pertaining to an internal assessment training, as illustrated in Figure 10.

The excerpt below showcases a blend of specialized terminology alongside intricate HTML markup, a feature commonly observed in governmental and academic contexts:

"p,vibrators I'm so lost I need to know if i need physics or not, and i keep forgetting to ask my guidance councellor I'm just wondering if there's any nurses (or nursing students) who could tell me what i will need for prerequisites and maybe their road to nursing school Calculus would probably be a good idea too vibrators". This is Ground Truth: Benign (0).

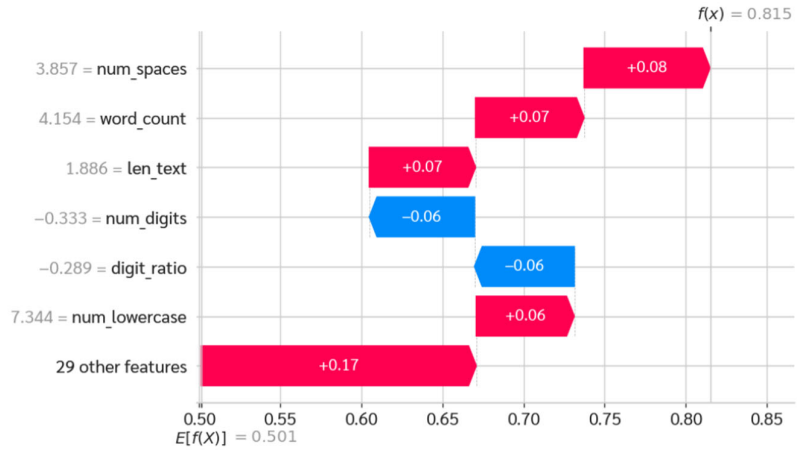
The Random Forest Failure (The Heuristic Trap): The Random Forest model identified this document as a significant threat $(f(x) \approx 0.8150)$, falling into a "Heuristic Trap" (Figure 10a).

 doi:10.20944/preprints202512.1607.v1

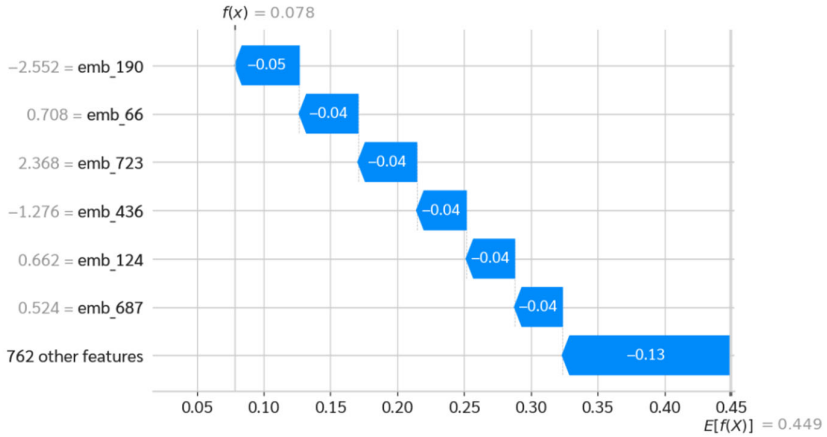
- The primary factor: The decision was significantly influenced by the len_tag (length of HTML tags) feature, which added +0.08 to the malicious probability, in conjunction with num_digits.
- Investigative Examination: This failure highlights the constraints inherent in statistical proxies. Conventional frameworks function under the strict premise that "Elevated Tag Density Plus Quantitative Noise results in XSS Injection or Iframe Attack." In this scenario, the model merged the extensive HTML code utilized for structuring a detailed schedule (such as nested tables typical in outdated formatting) with the syntax of a payload-hiding technique. The system identified the structural anomaly but was unable to interpret the underlying intent as benign.

The Hybrid Success (Semantic Arbitration): The Hybrid model exhibited an advanced capability for "Semantic Arbitration," accurately categorizing the sample as Benign with a high degree of confidence ($f(x) \approx 0.1999$, Figure 10c).

- Correction Mechanism: The SHAP visualization illustrates a dynamic interplay between features, resembling a "tug-of-war" scenario. Although the structural feature len_tag continued to trigger a false positive (leaning towards Malicious), the semantic embeddings particularly those representing concepts such as "Committee," "Seminar," and "Participants" produced significant negative SHAP values that effectively countered the structural interference.
- Implication: This indicates that the Hybrid architecture operates in a manner akin to a human analyst; it assesses the semantic context of messy or suspicious code structures prior to reaching a conclusion. The Hybrid model effectively filters out false positives by prioritizing high-level semantic understanding rather than relying on low-level structural heuristics, addressing the benign complexities found in legacy web environments.

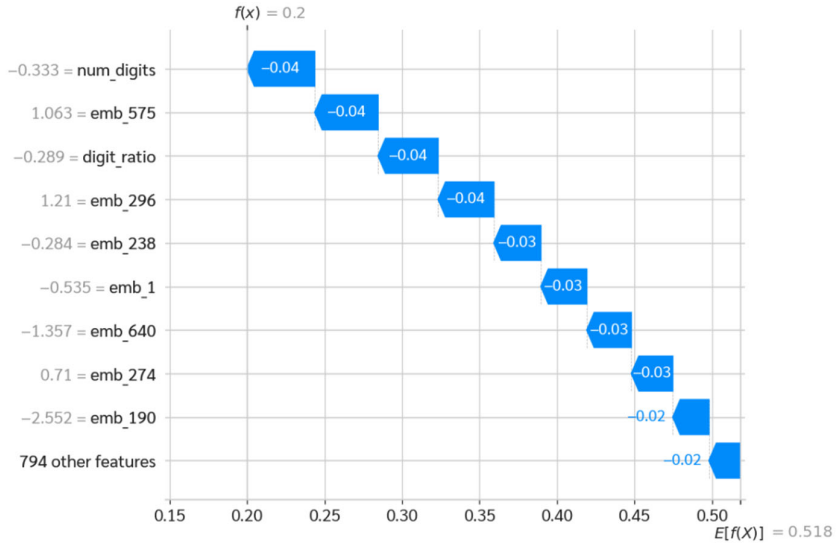


(a)



(b)

 doi:10.20944/preprints202512.1607.v1



(c)

Figure 10. Case Study 2: Forensic analysis of a false positive (benign "HTML complexity" sample). (a) The Random Forest model encounters limitations due to len_tag (Contextual Blindness); (b) The Transformer model accurately discerns the administrative intent; (c) The Hybrid model effectively implements Semantic Arbitration, emphasizing deep context to rectify structural misinterpretation and validate a True Negative decision.

5.2.3. Conclusion: The Hypothesis of "Semantic Noise"

The empirical findings clearly illustrate the mechanism of Semantic Arbitration in the Hybrid model. The effective categorization of intricate benign scripts, as outlined in Case Studies 1 and 2, challenges the original hypothesis regarding the overall performance decline attributed to semantic noise. Structural features, particularly CharType and Length, function as essential Structural Anchors, effectively addressing the Random Forest's intrinsic Structural Bias (FPR: 0.0244). This correction mechanism utilizes the deep contextual understanding of the Transformer to eliminate misleading statistical signals, allowing the Hybrid architecture to attain the system's lowest error rate (FPR: 0.012) and maximum stability (Figure 2), thus validating the superiority of Hybrid in terms of forensic robustness.

5.3 The Challenge of Low-Resource Languages in Cybersecurity

5.3.1. Tokenization and Script-Mixing: In contrast to English

Thai operates as a script-continuous language, characterized by the absence of spaces between words in its writing system. This structural feature poses significant challenges in cybersecurity scenarios, especially when Thai text is intricately intertwined with computer code that depends on precise spacing and syntax. Models primarily designed for English, such as BERT or RoBERTa, frequently struggle with accurately segmenting Thai characters from programming syntax. This results in fragmented embeddings that compromise semantic integrity.

5.3.2. The Importance of WangChanBERTa

Utilizing WangChanBERTa effectively mitigates this particular limitation. Through the implementation of SentencePiece tokenization, the model is capable of adaptively acquiring sub-word units that honor the linguistic structures of Thai while also accommodating Latin-based programming syntax.

- Comparison: Utilizing a standard English BERT model would likely result in the Thai content being classified as "unknown" tokens (UNK) or nonsensical byte sequences, compelling the

 doi:10.20944/preprints202512.1607.v1

model to depend exclusively on the code structure, effectively reverting it to a conventional signature-based detection method.

- Result: The superior performance of the semantic component in our analysis demonstrates that language-specific pre-training is not just an enhancement but an essential prerequisite for identifying threats in non-English, code-mixed contexts. This highlights the necessity of creating tailored cyber-defense frameworks that are specific to regions, instead of depending on broad, Anglicized large language models.

6. Conclusions

This research introduces a hybrid architecture powered by explainable artificial intelligence, aimed at addressing the significant issue of identifying code-mixed malicious scripts that are integrated within high-trust domains such as .go.th and .ac.th. Conventional single-modality models frequently encounter issues such as "Semantic Noise" arising from tokenization inaccuracies or "Structural Bias," which involves misinterpreting benign complexity as malicious intent. Through the integration of context-aware embeddings derived from WangChan-BERTa and outlier-robust handcrafted structural features, our proposed model attained a leading F1-Score of 0.9908 on a meticulously curated high-fidelity dataset.

Principal Discoveries and Operational Framework: The experimental findings indicate that the Hybrid architecture exhibits a marked improvement over the Transformer baseline, especially in minimizing false alarms. Forensic diagnosis through SHAP and Attention Heatmaps uncovers a "Dual-Validation" mechanism: the semantic encoder identifies potential keyword-based threats, while the structural features act as an essential "veto layer." This process efficiently eliminates harmless administrative scripts that include vague terminology, lowering the False Positive Rate (FPR) from 0.024 (Baseline) to 0.012 (Hybrid).

Operational Consequences: In addition to its statistical advantages, this 50% decrease in false positives has significant operational consequences. By reducing the interference from outdated code architectures, the suggested framework significantly decreases the investigative burden for Security Operations Center (SOC) analysts by fifty percent. This efficiency improvement is essential for preserving the integrity of national digital infrastructure, enabling defenders to allocate limited resources towards authentic, high-sophistication attacks instead of pursuing non-existent threats.

Future Work: Subsequent investigations will enhance this framework to evaluate resilience against adversarial AI-generated malware, wherein adversaries might strive to replicate benign structural profiles. Furthermore, our objective is to explore Cross-Attention Fusion mechanisms to dynamically adjust feature contributions according to real-time threat contexts, thereby improving the model's adaptability to changing attack vectors.

Author Contributions: Conceptualization, P.T. (Prasert Teppap) and P.L.; methodology, P.T. (Prasert Teppap); software, P.T. (Prasert Teppap); validation, W.P. and P.L.; formal analysis, P.T. (Panudech Tipauksorn); investigation, P.T. (Prasert Teppap); resources, P.L.; data curation, P.T. (Panudech Tipauksorn); writing original draft preparation, P.T. (Prasert Teppap); writing review and editing, P.L. and W.P.; visualization, P.T. (Panudech Tipauksorn); supervision, P.L. and W.P.; project administration, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The structural support and research facilities were provided by the AIoT Cyber Security Research Unit and the Faculty of Engineering, Rajamangala University of Technology Lanna (RMUTL).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and security restrictions concerning the compromised governmental and academic domains involved in the dataset.

Acknowledgments: The authors are grateful to the AIoT Cyber Security Research Unit, for research facilities, technical assistance, and supervision during this project. This work was completed thanks to AIoT Cyber

 doi:10.20944/preprints202512.1607.v1

Security 's research team's help. Finally, we thank the Faculty of Engineering, Rajamangala University of Technology Lanna (RMUTL), Thailand, for their ongoing support and encouragement enabling this study.

Conflicts of Interest: The authors declare no conflict of interest.

Reference

1. Chandran, S.; Syam, S.R.; Sankaran, S.; Pandey, T.; Achuthan, K. From Static to AI-Driven Detection: A Comprehensive Review of Obfuscated Malware Techniques. *IEEE Access* 2025, *13*, 74335–74358.
2. Takawane, G.; Phaltankar, A.; Patwardhan, V.; Patil, A.; Joshi, R.; Takalikar, M.S. Leveraging Language Identification to Enhance Code-Mixed Text Classification. 2023.
3. Al-Fayoumi, M.; Al-Haija, Q.A.; Armoush, R.; Amareen, C. XAI-PDF: A Robust Framework for Malicious PDF Detection Leveraging SHAP-Based Feature Engineering. *International Arab Journal of Information Technology* 2024, *21*, 128–146, doi:10.34028/iajit/21/1/12.
4. Liu, Y.; Tantithamthavorn, C.; Li, L.; Liu, Y. Explainable AI for Android Malware Detection: Towards Understanding Why the Models Perform So Well? In Proceedings of the Proceedings - International Symposium on Software Reliability Engineering, ISSRE; IEEE Computer Society, 2022; Vol. 2022-October, pp. 169–180.
5. Nurseno, M.; Aditiawarman, U.; Al Qodri Maarif, H.; Mantoro, T. Detecting Hidden Illegal Online Gambling on .Go.Id Domains Using Web Scraping Algorithms. *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer* 2024, *23*, 365–378, doi:10.30812/matrik.v23i2.3824.
6. Zhang, Y.; Fu, X.; Yang, R.; Li, Y. DRSDetector: Detecting Gambling Websites by Multi-Level Feature Fusion. In Proceedings of the Proceedings - IEEE Symposium on Computers and Communications; Institute of Electrical and Electronics Engineers Inc., 2023; Vol. 2023-July, pp. 1441–1447.
7. Arjun, D.S.; Samhitha, D.S.; Padmavathi, A.; Hemprasanna, A. Detection of Malicious URLs Using Ensemble Learning Techniques. In Proceedings of the Proceedings of 2023 IEEE Technology and Engineering Management Conference - Asia Pacific, TEMSCON-ASPAC 2023; Institute of Electrical and Electronics Engineers Inc., 2023.
8. Venugopal, S.; Panale, S.Y.; Agarwal, M.; Kashyap, R.; Ananthanagu, U. Detection of Malicious URLs through an Ensemble of Machine Learning Techniques. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2021; Institute of Electrical and Electronics Engineers Inc., 2021.
9. Do, N.Q.; Selamat, A.; Krejcar, O.; Fujita, H. Detection of Malicious URLs Using Temporal Convolutional Network and Multi-Head Self-Attention Mechanism. *Appl Soft Comput* 2025, *169*, doi:10.1016/j.asoc.2024.112540.
10. Safe, Suspicious, or Phishing? Classifying SMS with LLMs. 2025.
11. Mankar, N.P.; Sakunde, P.E.; Zurange, S.; Date, A.; Borate, V.; Mali, Y.K. Comparative Evaluation of Machine Learning Models for Malicious URL Detection. In Proceedings of the 2024 MIT Art, Design and Technology School of Computing International Conference, MITADTSoCiCon 2024; Institute of Electrical and Electronics Engineers Inc., 2024.
12. Shanmugam, V.; Razavi-Far, R.; Hallaji, E. Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches. *Electronics (Switzerland)* 2025, *14*, doi:10.3390/electronics14010069.
13. Karim, A.; Shahroz, M.; Mustofa, K.; Belhaouari, S.B.; Joga, S.R.K. Phishing Detection System Through Hybrid Machine Learning Based on URL. *IEEE Access* 2023, *11*, 36805–36822, doi:10.1109/ACCESS.2023.3252366.
14. Mohammed, S.Y.; Aljanabi, M.; Mijwil, M.M.; Ramadhan, A.J.; Abotaleb, M.; Alkattan, H.; Albadran, Z. A Two-Stage Hybrid Approach for Phishing Attack Detection Using URL and Content Analysis in IoT. In Proceedings of the BIO Web of Conferences; EDP Sciences, April 5 2024; Vol. 97.
15. Alshomrani, M.; Albeshri, A.; Alturki, B.; Alallah, F.S.; Alsulami, A.A. Survey of Transformer-Based Malicious Software Detection Systems. *Electronics (Switzerland)* 2024, *13*.

 doi:10.20944/preprints202512.1607.v1

16. Liu, R.; Wang, Y.; Guo, Z.; Xu, H.; Qin, Z.; Ma, W.; Zhang, F. TransURL: Improving Malicious URL Detection with Multi-Layer Transformer Encoding and Multi-Scale Pyramid Features. *Computer Networks* 2024, *253*, doi:10.1016/j.comnet.2024.110707.
17. Do, N.Q.; Selamat, A.; Krejcar, O.; Fujita, H. Detection of Malicious URLs Using Temporal Convolutional Network and Multi-Head Self-Attention Mechanism. *Appl Soft Comput* 2025, *169*, doi:10.1016/j.asoc.2024.112540.
18. Siino, M. *DeBERTa at SemEval-2024 Task 9: Using DeBERTa for Defying Common Sense*; 2024;
19. Do, N.Q.; Selamat, A.; Lim, K.C.; Krejcar, O.; Ghani, N.A.M. Transformer-Based Model for Malicious URL Classification. In Proceedings of the 2023 IEEE International Conference on Computing, ICOCO 2023; Institute of Electrical and Electronics Engineers Inc., 2023; pp. 323–327.
20. Dau Hoang, X.; Thu Trang Ninh, T.; Duy Pham, H. A NOVEL MODEL BASED ON DEEP TRANSFER LEARNING FOR DETECTING MALICIOUS JAVASCRIPT CODE. *J Theor Appl Inf Technol* 2024, *102*.
21. Teppap, P.; Tipauksorn, P.; Surathong, S.; Ponglangka, W.; Luekhong, P. Automating Hidden Gambling Detection in Web Sites: A BeautifulSoup Implementation. In Proceedings of the Proceedings - 21st International Joint Conference on Computer Science and Software Engineering, JCSSE 2024; Institute of Electrical and Electronics Engineers Inc., 2024; pp. 132–139.
22. Lowphansirikul, L.; Polpanumas, C.; Jantrakulchai, N.; Nutanong, S. WangchanBERTa: Pretraining Transformer-Based Thai Language Models. 2021.
23. Zhang, A.; Li, K.; Wang, H. A Fusion-Based Approach with Bayes and DeBERTa for Efficient and Robust Spam Detection. *Algorithms* 2025, *18*, doi:10.3390/a18080515.
24. Sriwirote, P.; Thapiang, J.; Timtong, V.; Rutherford, A.T. PhayaThaiBERT: Enhancing a Pretrained Thai Language Model with Unassimilated Loanwords. 2023, doi:10.1145/3765962.
25. Chandrasekaran, H.; Murugesan, K.; Mana, S.C.; Barathi, B.K.U.A.; Ramaswamy, S. Handling Imbalanced Data in Intrusion Detection Using Time Weighted Adaboost Support Vector Machine Classifier and Crossover Boosted Dwarf Mongoose Optimization Algorithm. *Appl Soft Comput* 2024, *167*, doi:10.1016/j.asoc.2024.112327.
26. Huang, Y.; Lin, J.; Zhou, C.; Yang, H.; Huang, L. Modality Competition: What Makes Joint Training of Multi-Modal Network Fail in Deep Learning? (Provably). 2022.
27. Gao, L.; Guan, L. A Discriminant Information Theoretic Learning Framework for Multi-Modal Feature Representation. *ACM Trans Intell Syst Technol* 2023, *14*, doi:10.1145/3587253.
28. Althaf Ali, A.; Rama Devi, K.; Syed Siraj Ahmed, N.; Ramchandran, P.; Parvathi, S. Proactive Detection of Malicious Webpages Using Hybrid Natural Language Processing and Ensemble Learning Techniques. *Journal of Information and Organizational Sciences* 2024, *48*, 295–309, doi:10.31341/jios.48.2.4.
29. Peng, X.; Wei, Y.; Deng, A.; Wang, D.; Hu, D. *Balanced Multimodal Learning via On-the-Fly Gradient Modulation*; 2022;