Semantic-Structural Evidence-Decomposed Gated Fusion for Explainable Spam Email Detection

Saja Fadhil Zakour and Rawaa Ismeal Farhan
Department of Computer Science, College of Education for Pure Sciences, Wasit University, 52001 Al-Kut, Iraq
std.2024205.s.zakur@uowasit.edu.iq, ralrikabi@uowasit.edu.iq

Keywords: Explainable Spam Detection, Gated Fusion, Semantic Evidence, Structural Evidence, Email Security.

Abstract: Modern anti-spam email systems have faced a growing problem in the form of multi-modal campaigns that not only employ compelling textual information but also utilize structural manipulations; moreover, numerous existing high performing models lack explainability. In this research paper, we present a novel explainable approach based on gated fusion of text-based semantic features and metadata-related structural patterns, where the decision process is divided into three types of evidences – semantic-dominant, structural-dominant and mixed mode. For the evaluation of our proposed framework, the Assassin_clean data set has been selected and five different random seeds have been tested. Our results indicate that among the proposed neural models, Variant F performed best, scoring an average F1 score of 0.9712 and an average AUC score of 0.9971; additionally, it outperformed Variant E and demonstrated comparable results with respect to linear SVM. It was also noted that the mode of explanation might differ across the random seeds without deteriorating the predictive performance.

1 INTRODUCTION

The email spam problem persists due to the combination of elements of persuasion, malice, impersonation of brands, and tampering with metadata included in the current generation of email spam campaigns [1], [2]. Current comparative analyses prove that classical machine-learning models are highly effective baseline systems, while deep learning frameworks are utilized in cases where rich semantic context, diverse set of features, and automatic feature learning are essential [3], [4]. Meanwhile, phishing detection and email spam filtering have been evolving in directions of explainability and multimodality since modern analysts need not only the result but the reasoning behind it [5], [6].

While significant advancements have been made, there are still three clear shortcomings. Firstly, many of the current detectors rely solely on the semantic meaning or structural features of the text without combining their impacts on the final decision [4], [7]. Secondly, many models based on deep learning cannot reveal how the prediction is influenced by the context and the structure of the text [8], [9]. Finally, most studies do not conduct multiple runs, and thus, the better model may only be superior due to a lucky random split [1], [10].

The gaps mentioned above are filled by this paper in which an explainable gated fusion of a semantic-structural architecture is proposed, where a model comprises two branches and the decomposition of predictions into evidence modes dominated by semantics, structure, and both of them respectively is performed. There are four contributions of this research: branch-aware gated fusion, evidence decomposition framework, five-seed experiment based on “Assassin_clean” dataset, and threshold-based analysis of explanation variability. The rest of the paper is organized as follows.

2 RELATED WORK AND RESEARCH GAP

The evolution to transformer-supported and multimodal spam detection is recently reviewed [1], [4], [5]. Comparing traditional machine learning approaches, CNNs, recurrent architectures, and language model approaches reveals that no one class of representation outperforms the rest, since it depends on the diversity of features, properties of

datasets, and evaluation settings [2], [3]. Hybrid approaches to spam filtering are thus compelling since they provide flexibility to model short lexical cues and contextual dependencies at once [8], [11].

Second, many approaches leverage multimodal representations or multi-branch integration mechanisms. The integration of DistilBERT with metadata and gradient optimization models [7], as well as a behavior-conscious framework for phishing using transformers [10], has shown that structural clues can help complement input semantics. To further analyze branch contributions, Uddin et al. [12] presented a transformer-based interpretable model using a large-scale linguistic model approach to provide interpretable clues, while Sharma et al. [13] presented a comprehensive review of interpretable artificial intelligence (XAI) for deciphering black-box mechanisms in cybersecurity. Building on these foundations, the ExplainableDetector framework proposed by Uddin et al. [14] was introduced to explore transformer-based phishing detection with interpretable clues. Additionally, Ibrahim et al. [15] evaluated the phishing email detection capabilities of the BERT and RoBERTa models specifically within similar sub-environments. For example, the advanced multimedia interpretable AI system developed by Vulfin et al. [16] achieves high accuracy (F1 metric of 0.989), but it relies on complex HTML, URL, and visual analysis. This massive infrastructure underscores the need for a lightweight, text-focused alternative capable of clearly categorizing its interpretations within these specific linguistic dimensions.

3 METHODOLOGY

The experiments were carried out using the Assassin_clean dataset, where there are 5,787 emails labeled after excluding 39 samples without a label. The final class distribution had 3,896 class 0 emails and 1,891 class 1 emails, with an imbalance ratio of 2.06:1. In each experiment, word-level tokenization was applied (vocabulary capped at the 10,000 most frequent tokens; the exact tokenizer implementation should be specified in a future revision for full reproducibility) and a partitioning strategy of 4,166/463/1,158 for training/validation/testing, respectively. Five random seeds (42, 52, 62, 72, and 82) were tested to evaluate the robustness of the models.

Prior to Table 1, the experimental setup is provided, considering that the validity of any claim about spam detection heavily relies on the datasets.

Table 1: Dataset and evaluation protocol.

Item	Setting
Dataset	Assassin_clean
Usable samples	5,787
Dropped missing labels	39
Train/Validation/Test	4,166 / 463 / 1,158
Vocabulary size	10,000
Seeds	42, 52, 62, 72, 82
Main metrics	Accuracy, Precision, Recall, F1-score, AUC

Variants A-E correspond to progressively simpler ablations of the full architecture (e.g., single-branch semantic-only or structure-only models, and versions without the learnable gate or without evidence decomposition), used to isolate the contribution of each architectural component; full architectural specification of each variant is provided in Table 2. In version F, the complete proposed architecture, there are two branches, one dedicated to extracting semantic features using contextual cues from fragmented text content via a CNN-LSTM network architecture [17]. Within this branch, the LSTM component is based on the underlying architecture proposed by Hochreiter and Schmidhuber [18], which uses specialized gate mechanisms to capture long-range sequential dependencies and preserve semantic context without suffering gradient fading. and another that encodes features relevant to metadata. These results are then combined via a learnable gate and classified as either spam or ham. Equations (1)-(5) detail the formulation for the most part.

3.1 Model Formulation

Equation (1) defines the gated fusion weight that balances semantic and structural information:

$$\alpha = sigmoid(W_g\,[h_{sem}\,;\,h_{str}] + b_g). \quad (1)$$

In (1), h_{sem} is the semantic latent representation, h_{str} is the structural latent representation, W_g and b_g are learnable gate parameters, and α is the gate output constrained to [0,1]. A larger α indicates stronger semantic influence [8], [19].

Equation (2) defines the fused latent vector:

$$z = \alpha \odot h_{sem} + (1 - \alpha) \odot h_{str}. \quad (2)$$

In (2), z is the fused representation used for downstream classification and $\odot$ denotes element-wise multiplication. The term $(1 - \alpha)$ allocates the complementary share to the structural branch.

Equation (3) converts the fused vector into the spam posterior:

$$\hat{y} = sigmoid(W_c\, z + b_c). \quad (3)$$

In (3), $\hat{y}$ is the predicted probability of the spam class, while W_c and b_c are classifier parameters.

Equation (4) gives the binary cross-entropy objective:

$$L_{BCE} = -\left(\frac{1}{N}\right) \Sigma_{\{i=1\}}^{\{N\}} [y_i\, log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]. \quad (4)$$

In (4), N is the batch size, y_i is the true label of sample i, and $\hat{y}_i$ is the predicted probability. Binary cross-entropy is the standard loss for binary neural classification tasks [20].

Equation (5) reports the F1-score used in evaluation:

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}. \quad (5)$$

In (5), $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$, where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Because spam datasets are typically imbalanced, F1-score is more informative than accuracy alone [1], [4].

3.2 Explainable Evidence Decomposition

Once the inference process is completed, Variant F documents the contribution made by each branch to the outcome and then classifies each prediction according to three possible types of evidence: semantic-dominant, structure-dominant, and mixed. This technique will help us determine whether the decision-making process is influenced primarily by contextually-based linguistic expressions, structural indicators, or both.

4 RESULTS AND DISCUSSION

The results prior to Table 2 represent the overall analysis performed to test if the explainable model suggested outperformed the best neural networks. The values presented in Table 2 show the mean F1 score, standard deviation, and mean AUC of the analyzed models on Assassin_clean.

Table 2: Cross-seed aggregate performance on Assassin_clean.

Model	Mean F1	F1 std	Mean AUC
Variant A	0.9462	0.006	0.9935
Variant B	0.9499	0.0053	0.9927
Variant C	0.9442	0.0057	0.9924
Variant D	0.9517	0.0068	0.9909
Variant E	0.968	0.0074	0.9976
Variant F	0.9712	0.0064	0.9971
Naive Bayes	0.9621	0.0018	0.9975
Linear SVM	0.9862	0.0029	0.9995
Logistic Regression	0.962	0.003	0.9961

As shown in Table 2, Variant F achieved the highest average F1-score among all neural variants, reaching 0.9712 with a standard deviation of 0.0064, while Variant E achieved 0.9680. Thus, the explainable model not only remained competitive but also surpassed the strongest non-explainable neural alternative on average. Although linear SVM still produced the overall best mean F1-score at 0.9862, Variant F offered a clear interpretability advantage while preserving high predictive quality.

To further examine robustness, Figure 1 presents the mean F1-score and standard deviation of the strongest models. The figure is placed here because stability across seeds is central to the claim that the proposed model is not a single-run artifact.

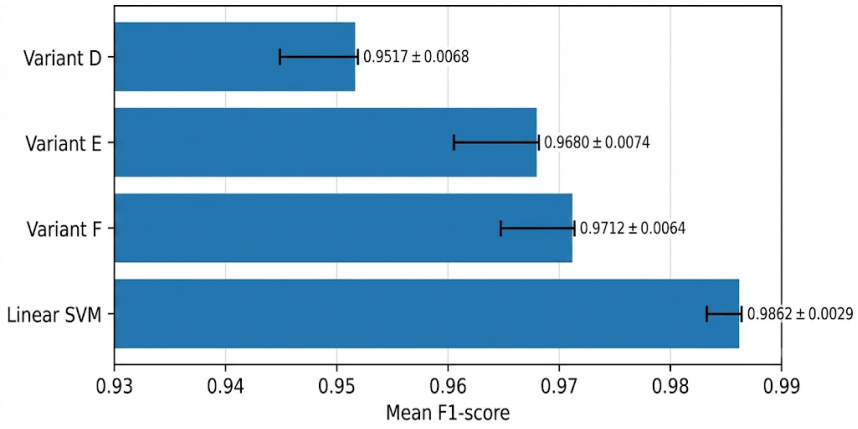

Figure 1: Cross-seed mean F1-score with standard-deviation error bars for the strongest models on Assassin_clean.

The results depicted in Figure 1 show that Variant F is always superior to other seeds and usually outperforms Variant E on F1-score. Moreover, the difference between Variant F and other neural network baselines is quite large. It proves that the contribution of our technique is greater than superficial complexity.

4.1 Seed-Wise Comparison

This can be seen from Table 3, which makes a comparison of the best models on a seed-by-seed basis. Table 3 is presented prior to the interpretation figures, since it proves that the proposed model performed better under several different runs, particularly for seeds 62 and 82.

The values in Table 3 show that Variant F surpassed Variant E in three of the five seeds and remained close to it in the remaining runs. The strongest single-run result for Variant F was obtained at seed 62 with F1 = 0.9788, closely followed by seed 82 with F1 = 0.9787. This pattern supports the claim that the explainable architecture is not merely averaging out weak runs; instead, it can achieve genuinely strong operating points while remaining stable overall.

Table 3: Seed-wise F1-score comparison of the strongest models.

Seed	Linear SVM	Variant E	Variant F
42.0	0.9881	0.9749	0.9635
52.0	0.9826	0.9665	0.9694
62.0	0.9881	0.9671	0.9788
72.0	0.9827	0.9554	0.9658
82.0	0.9894	0.9762	0.9787

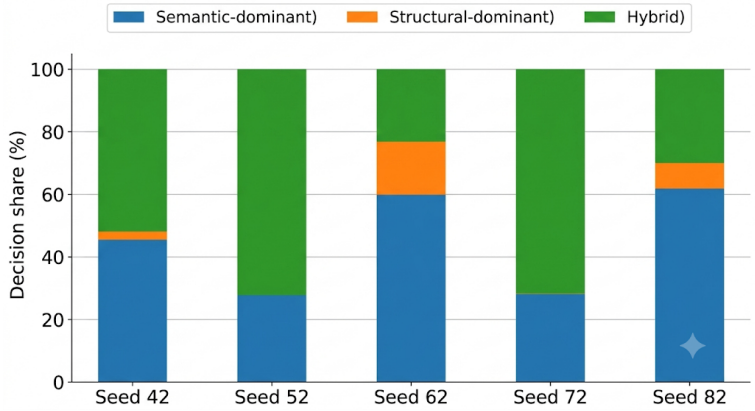

Figure 2: Distribution of semantic-dominant, structural-dominant, and hybrid decisions for Variant F across five seeds.

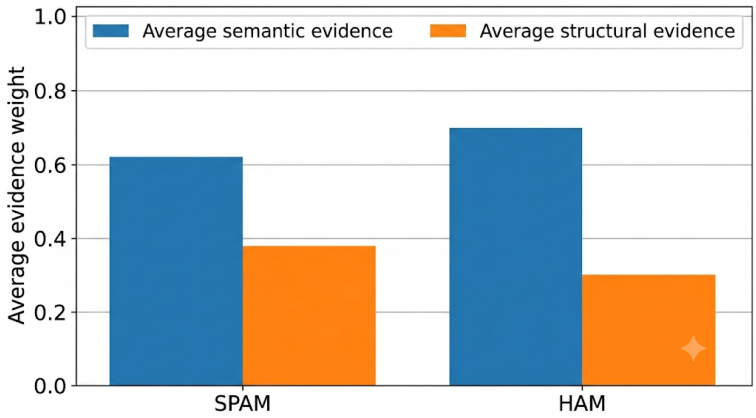

Figure 3: Average semantic and structural evidence by class for seed 62.

4.2 Interpretation Behavior

The subsequent analysis will be conducted for interpretability purposes. The introduction of Figure 2 is made at this point since the distribution of evidence modes provides an explanation for the variations in the proposed model's operation modes.

From Figure 2, it is clear that there was significant diversity in the explanation of branches among the seeds. In particular, seeds 52 and 72 were dominated by hybrid explanations, while seeds 62 and 82 had more semantic-dominated branches. Seed 62 also showed significant structural domination. That explains why seed-based plots of evidence show greater dominance of certain decision regimes. It is worth noting that the difference in plots is not due to failure of the model but rather how the gate adjusted importance among branches.

Figure 3 shows this more concretely at the class level, using the best data-rich run, which was seed 62. This comparison is critical since any contribution made by branches will only be worthwhile if they can distinguish spam from ham messages.

According to Figure 3, it seems that there was relatively more structural information in spam samples from seed 62 than ham samples, whereas the ham messages maintained more semantic information on an average basis. It demonstrates that the proposed classifier is also considering non-lexical features along with lexical features, as there are some irregularities in terms of structure present in suspicious emails. The behavior matches with previous works, which have utilized non-textual features alongside semantic information for malicious email detection [7], [10], [12].

In addition, seed-wise threshold selection affected variations as well. Higher values, like 0.70 in seeds 42 and 52, helped achieve better precision but reduced recall, while lower values, like 0.30 in seed 62, resulted in a better balance between precision and recall. This tendency is common for security classification problems and proves that it is essential to present results not only on average but also with respect to operating characteristics [1], [13].

In conclusion, there are three main points that can be derived from the above discussion. First, semantic-structure fusion through explainability is proven to work effectively with Assassin_clean. Second, explicit proof of evidence decomposition is useful in analysis, and this is something that neural alternatives cannot provide. Third, there is still an obvious gap between the proposed method and linear SVM, implying that the proposed method is not meant to fully substitute existing baselines.

5 CONCLUSIONS

The proposed model in this work uses semantic-structural information to design an explainable gated-fusion approach for the task of spam classification, which was tested against the dataset Assassin_clean with five different random seeds. Variant F was able to provide the highest F1-scores out of all neural models used, staying close to the highest-performing baseline and providing explicit decision-making components. This provides significant support for the main thesis that explainable branch-level information can be added to an effective spam classifier without impacting its performance.

Moreover, it has been shown that the explanations depend largely on the specific seed due to their dependence on both the chosen data split and the particular choice of threshold. Yet, the predictive power was fairly stable across different seeds, and the explanation patterns provided were consistent and not contradictory. Further research could consider multilingual datasets, temporal issues, more information at the header level, and other ways of calibrating gates.

ACKNOWLEDGMENTS

The author acknowledges the broader master's research context in AI-based cybersecurity detection and the reproducible experimental workflow that enabled the reported analyses.

REFERENCES

[1] E. H. Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin, "Email spam: A comprehensive review of optimization techniques in detection systems," IEEE Access, vol. 12, pp. 1-24, 2024.

[2] K. I. Roumeliotis et al., "Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email classification," Electronics, vol. 15, no. 3, p. 630, 2026.

[3] R. Meléndez, M. Ptaszyński, and F. Masui, "Comparative investigation of traditional machine-learning models and transformers for phishing and spam detection," Applied Sciences, vol. 15, no. 6, p. 3396, 2025.

[4] P. H. Kyaw et al., "A systematic review of deep learning techniques for phishing email detection," Electronics, vol. 14, no. 5, p. 812, 2025.

[5] J. L. Wilk-Jakubowski et al., "Machine learning and neural networks for phishing detection: A systematic review," Sensors, vol. 24, no. 9, p. 2951, 2024.

[6] A. Alhuzali, A. Alloqmani, M. Aljabri, and F. Alharbi, "In-depth analysis of phishing email detection: Evaluating the robustness of modern deep learning approaches," Information, vol. 15, no. 1, p. 45, 2024.

[7] H. Asliyuksek et al., "A comparative evaluation of a multimodal approach for spam email classification using DistilBERT and structured metadata," Computers & Security, vol. 148, Art. no. 104312, 2025.

[8] M. Hosseinzadeh et al., "Improving phishing email detection performance through hybrid deep learning architecture and feature fusion," Expert Systems with Applications, vol. 254, Art. no. 124565, 2024.

[9] C. Patra et al., "Phishing email detection using vector similarity search and transformer-based word embedding," Computers, Materials & Continua, vol. 82, no. 3, pp. 5229-5250, 2025.

[10] M. Murhej et al., "Multimodal framework for phishing attack detection and mitigation through behavior analysis using deep learning," Engineering Applications of Artificial Intelligence, vol. 134, Art. no. 108791, 2024.

[11] S. Rashed and C. Ozcan, "A novel dual-layer deep learning architecture for phishing and spam email detection," Electronics, vol. 14, no. 2, p. 205, 2025.

[12] M. A. Uddin et al., "An explainable transformer-based model for phishing email detection: A large language model approach," IEEE Access, vol. 14, pp. 1-18, 2026.

[13] A. Sharma, S. Rani, and M. Shabaz, "A comprehensive review of explainable AI in cybersecurity: Decoding the black box," Journal of Information Security and Applications, vol. 80, Art. no. 103796, 2024.

[14] M. A. Uddin, M. N. Islam, L. Maglaras, H. Janicke, and I. H. Sarker, "ExplainableDetector: Exploring transformer-based phishing email detection with interpretable evidence," IEEE Access, vol. 13, pp. 1-16, 2025.

[15] M. Ibrahim et al., "Phishing email detection using BERT and RoBERTa," Computation, vol. 14, no. 2, p. 46, 2026.

[16] A. M. Vulfin et al., "A multimodal phishing website detection system using explainable artificial intelligence techniques," Knowledge-Based Systems, vol. 298, Art. no. 112019, 2024.

[17] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, pp. 1746-1751, 2014.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.