Expert Selection for Wordlist-Based DGA Detection

Reynier Leyva La O, Carlos A. Catania, and Rodrigo Gonzalez

***Abstract*—Domain Generation Algorithms (DGAs) have evolved beyond traditional pseudorandom patterns, with wordlist-based variants generating linguistically coherent domains that evade conventional detection methods. While previous research has primarily focused on generalist detection approaches across multiple DGA types, systematic expert model selection specifically targeting wordlist-based variants remains largely unexplored. This work addresses expert model selection for wordlist-based DGA detection, where expert models refer to specialized architectures trained exclusively on specific DGA categories. We conduct systematic evaluation of seven candidate models across transformer, convolutional neural network (CNN), and traditional machine learning approaches. Models were trained on a balanced dataset of 160,000 domains spanning eight wordlist-based DGA families and evaluated using a rigorous two-phase protocol that measures both performance on training families and generalization to previously unseen variants. Our comparative analysis identifies fine-tuned ModernBERT as the optimal expert model, achieving 86.7% F1-score on known families while maintaining 80.9% performance on unknown families with 26ms inference time on NVIDIA Tesla T4 GPUs, enabling processing of approximately 38,000 domains per second. The study validates that domain-specific expert training significantly outperforms generalist approaches trained on diverse DGA families, with F1-score improvements of 9.4% on familiar variants and 30.2% on unseen families. This performance gain indicates that focused expertise develops transferable linguistic patterns rather than memorization of specific family characteristics.**

***Index Terms*—Domain Generation Algorithms, Expert Selection, Cybersecurity, Real-time Detection, Model Comparison**

I. Introduction

DOMAIN Generation Algorithms (DGAs) have evolved from simple pseudorandom string generators into sophisticated linguistic mimics that challenge traditional cybersecurity defenses. Modern malware families leverage DGAs to establish resilient Command and Control (C&C) channels by generating thousands of candidate domains daily, effectively bypassing blacklist-based detection systems even when individual domains are discovered and blocked [1].

Among DGA variants, wordlist-based algorithms present the most challenging detection problem. While conventional DGAs generate easily recognizable character sequences like `qwerty123.com`, wordlist variants produce semantically plausible domains such as `secure-banking-portal.com` that closely resemble legitimate web services. This semantic plausibility renders entropy-based and character frequency methods ineffective [2], [3].

The urgency for effective wordlist-based DGA detection is exemplified by recent threats in Latin America. In 2024, the `Grandoreiro` banking trojan targeting over 1,500 banks across Brazil, Mexico, Argentina, and other Latin American countries demonstrated sophisticated domain generation capabilities with multiple algorithmic variants. IBM X-Force analysis revealed that the latest `Grandoreiro` iteration employs a reworked DGA containing multiple seeds to calculate different domains for each operational mode, generating up to 14 possible (C&C) domains daily [4]. This advanced DGA implementation, combined with the malware's semantic plausibility in domain generation, underscores the critical need for specialized detection models capable of identifying linguistically coherent yet malicious domains in production environments.

The fundamental challenge for cybersecurity practitioners lies in selecting appropriate detection models for wordlist-based DGAs. Character-level CNNs excel at capturing pseudo-random patterns but lack linguistic sophistication for wordlist detection [5]. Large language models possess the necessary semantic capabilities but introduce prohibitive computational overhead for real-time deployment. Traditional machine learning approaches offer computational efficiency but may miss subtle semantic patterns that characterize modern wordlist DGAs. Hybrid approaches combining multiple detection techniques have shown promise in network security applications [6], though these methods typically focus on general anomaly detection rather than the specific linguistic challenges posed by wordlist-based domain generation.

Production cybersecurity systems require both high detection accuracy and millisecond-level response times when analyzing millions of DNS (Domain Name System) queries daily. This constraint necessitates systematic model evaluation to balance detection accuracy with deployment feasibility.

False positive rates in production environments directly impact operational costs, with enterprise DNS monitoring systems processing millions of queries where each misclassified legitimate domain triggers unnecessary security investigations. Industry reports indicate that high false positive rates ($> 5\%$) can overwhelm security operations centers, necessitating detection models that maintain both high accuracy and acceptable operational thresholds [7]. Recent approaches have explored explainable AI techniques to better identify and reduce false positives in intrusion detection systems [8], though these methods focus on general network anomalies rather than the specific semantic challenges of wordlist-based DGA detection.

In this context, expert model selection refers to the sys-

Reynier Leyva La O is with GridTICs, Facultad Regional Mendoza, Universidad Tecnológica Nacional, Mendoza, Argentina, and also with the National Scientific and Technical Research Council (CONICET), Godoy Cruz 2290, C1425FQB, CABA, Argentina (e-mail: rleyvalao@mendoza-conicet.gob.ar).

Carlos A. Catania is with LABSIN, Facultad de Ingeniería, Universidad Nacional de Cuyo, Mendoza, Argentina, and also with the National Scientific and Technical Research Council (CONICET), Godoy Cruz 2290, C1425FQB, CABA, Argentina (e-mail: harpo@ingenieria.uncuyo.edu.ar).

Rodrigo Gonzalez is with GridTICs, Facultad Regional Mendoza, Universidad Tecnológica Nacional, Mendoza, Argentina.

tematic identification of specialized detection architectures optimized for specific threat categories. Unlike generalist approaches that distribute learning capacity across diverse attack types, expert models concentrate their representational power on the distinctive characteristics of a particular DGA variant. This specialization enables deeper pattern recognition within the target domain while maintaining computational efficiency for operational deployment. The expert selection process must therefore balance specialized performance against generalization capability to ensure robust detection of both known families and emerging variants within the same threat category.

This work addresses expert model selection for wordlist-based DGA detection through a comprehensive empirical evaluation across seven candidate architectures. We systematically compare transformer-based models, convolutional neural networks, and traditional machine learning approaches across eight wordlist-based DGA families using a rigorous two-phase evaluation protocol that measures both specialized performance and generalization to three previously unseen variants.

Our analysis identifies ModernBERT as the optimal expert model after fine-tuning on wordlist-based DGA data, achieving 86.7% F1-score on known families while maintaining 80.9% on unseen variants with 26ms inference latency. Controlled comparison validates that domain-specific training significantly outperforms generalist approaches, demonstrating 9.4% improvement on familiar variants and 30.2% enhancement on unseen families.

The study makes three contributions. First, we establish a systematic evaluation framework specifically designed for wordlist-based DGA detection with rigorous generalization testing. Second, we provide comprehensive empirical analysis focused exclusively on wordlist-based DGA families, demonstrating effective detection without performance degradation. Third, we validate that domain-specific expert training significantly outperforms generalist approaches when confronting wordlist-based DGAs. This evaluation scope, encompassing eleven distinct wordlist-based families with systematic statistical validation, addresses the documented underrepresentation of this threat category in existing detection literature.

II. Related Work

A. Wordlist-Based DGA Characteristics

Wordlist-based DGAs employ predefined dictionaries of common words, brand terms, and linguistic tokens, combining them through deterministic or probabilistic rules to generate semantically plausible domain names. Families such as `Suppobox` and `Matsnu` generate domains like `secure-login-check.com` that closely resemble legitimate services [1]. Despite representing only 4 of 92 cataloged botnet families in DGArchive [9], these variants occupy distributional regions close to legitimate traffic, creating disproportionate detection difficulties [7].

Advanced implementations use Hidden Markov Models (HMM) and Probabilistic Context-Free Grammars (PCFG) for generation that systematically evade traditional detection metrics [10]. Traditional entropy-based and character frequency methods fail against these semantically plausible domains, as they lack obvious randomness signatures [11].

B. Detection Approaches and Limitations

Current detection approaches span diverse architectural paradigms with distinct trade-offs. Random Forest classifiers with n-gram features provide computational efficiency but struggle with linguistic variants due to hand-crafted feature limitations [12], [13]. Convolutional neural networks excel at local character patterns [11], while recurrent architectures model sequential dependencies [14], [15].

Transformer-based approaches leverage pre-trained language models to capture semantic relationships, showing improved performance on wordlist variants [16], [17]. However, inference times range from sub-millisecond for traditional models to over 1000ms for large language models, limiting real-time deployment [7].

Embedding techniques have shown promise, with Embeddings from Language Models (ELMo) and Bidirectional Encoder Representations from Transformers (BERT)-based approaches demonstrating effectiveness across different scenarios [18], [19]. Mixed embedding approaches combining n-grams and word representations prove effective in few-shot learning scenarios [20].

C. Evaluation Challenges in Literature

Meta-analysis of 38 DGA detection studies reveals systematic evaluation biases. Most studies underrepresent wordlist-based families, often including minimal samples to avoid performance deterioration since these domains substantially increase false positive rates [7]. This selective evaluation creates incomplete understanding of detection capabilities against sophisticated variants.

While ensemble and few-shot learning approaches have been explored [21], [22], no systematic comparison specifically targets wordlist-based DGA detection across diverse architectural paradigms with rigorous generalization testing. This evaluation gap, combined with the operational necessity for balanced accuracy-efficiency trade-offs in production cybersecurity systems, motivates the development of a comprehensive expert selection framework that addresses both specialized performance on known threat families and robust generalization to emerging variants.

III. Problem Formulation

This work addresses expert model selection for wordlist-based DGA detection through systematic evaluation of candidate architectures. Given a set of models $E = \{E_1, E_2, \ldots, E_n\}$ and a dataset D containing benign domains and samples from multiple wordlist-based DGA families, we identify the optimal expert model E^* that maximizes detection performance while maintaining operational efficiency.

A. Evaluation Framework

We evaluate candidate models using standard classification metrics combined with operational constraints [23], [24]. Core performance measures include precision, recall, F1-score, and false positive rate, complemented by inference time measurements for operational feasibility assessment.

B. Expert Selection Methodology

The optimal expert model E^* maximizes a composite scoring function that balances multiple performance dimensions:

$$E^* = \arg\max_{E_i \in E} f(F1_i, P_i, R_i, FPR_i, T_i) \tag{1}$$

where $F1_i$, P_i, R_i, FPR_i, and T_i represent F1-score, precision, recall, false positive rate, and inference time for model E_i.

The composite scoring function integrates performance and efficiency metrics:

$$f(\cdot) = \alpha_1 \cdot F1 + \alpha_2 \cdot (1 - FPR) + \alpha_3 \cdot P + \alpha_4 \cdot R + \alpha_5 \cdot \left(1 - \frac{T}{T_{\max}}\right) \tag{2}$$

where coefficients $\alpha_1, \ldots, \alpha_5$ reflect relative metric importance with constraint $\sum_{i=1}^{5} \alpha_i = 1$.

This formulation enables practitioners to customize expert selection according to deployment priorities. Real-time systems may emphasize low latency ($\alpha_5 > 0.4$), while investigative applications may prioritize recall ($\alpha_4 > 0.4$). For this comparative study, we conduct two scoring analyses: first, we employ equal weighting ($\alpha_i = 0.2$) to provide unbiased assessment across all performance dimensions, ensuring that no single metric dominates the expert selection process and enabling fair comparison across architecturally diverse models. We also evaluate a case where FPR reduction takes priority, setting $\alpha_1 = \alpha_3 = \alpha_4 = \alpha_5 = 0.1$ and $\alpha_2 = 0.6$, which represents environments where the cost of investigating false alarms becomes a critical operational concern.

C. Evaluation Protocol

Our evaluation employs a two-phase protocol measuring both specialized performance and generalization capability. The first phase evaluates models on known wordlist-based DGA families using previously unseen domains from training families. The second phase assesses generalization to completely unknown DGA families, reflecting operational scenarios where new threats emerge continuously.

Statistical robustness is achieved through randomized batch sampling, with 30 batches per family containing 100 domains each (50 benign, 50 malicious). This methodology ensures reliable performance estimates while maintaining computational tractability for comprehensive model comparison.

IV. Methodology

We designed a systematic evaluation framework to identify optimal expert models for wordlist-based DGA detection. Our approach encompasses dataset construction, model selection across diverse architectures, and rigorous evaluation under realistic deployment constraints.

A. Dataset Construction

We constructed a balanced dataset reflecting operational network conditions and wordlist-based DGA diversity. The training dataset comprised 160,000 domains split equally between DGA samples (80,000) and legitimate domains (80,000). Table I provides detailed breakdown of family-specific sample allocation across training and evaluation phases.

For evaluation, each family present in training contributes 1,500 DGA samples combined with 1,500 fresh legitimate domains (distinct from training data) to create balanced 3,000 sample family evaluations. The generalization testing employs 1,500 samples from each unknown family paired with 1,500 different legitimate domains (separate from both training and in-family test sets) to maintain evaluation independence and prevent data leakage across experimental phases.

TABLE I
Sample distribution across wordlist-based DGA families and legitimate domains for training and testing

Category	Family	Training	In-Family Test	Generalization Test
Training Families	charbot	10,000	1,500	—
	deception	10,000	1,500	—
	gozi	10,000	1,500	—
	manuelita	10,000	1,500	—
	matsnu	10,000	1,500	—
	nymaim	10,000	1,500	—
	rovnix	10,000	1,500	—
	suppobox	10,000	1,500	—
Generalization Families	bigviktor	—	—	1,500
	ngioweb	—	—	1,500
	pizd	—	—	1,500
Legitimate	**All scenarios**	**80,000**	**1,500**	**1,500**

DGA samples were sourced from established repositories including DGArchive [9], 360 Netlab [25], and UMUDga [26]. Repository data were supplemented with additional domains generated using family-specific implementations such as the charbot generator [27] to ensure adequate sample sizes across all families. Legitimate domains came from the Tranco top sites list [28], providing a curated baseline free from commercial manipulation bias.

B. Model Selection Strategy

Our evaluation encompasses seven models representing three architectural paradigms. This diversity enables comprehensive assessment of the accuracy-efficiency trade-offs central to expert selection:

Large Language Models include Gemma 3 4B [29], [30] and LLaMA 3.2 3B [31], representing state-of-the-art semantic understanding capabilities with substantial computational requirements. These models were selected based on recent studies demonstrating superior performance on wordlist-based DGA classification tasks [16], [17], with size configurations chosen to balance semantic capabilities with computational feasibility for expert deployment scenarios.

Specialized Transformers comprise ModernBERT [32] and DomBertUrl [33], offering domain-specific pretraining advantages with moderate resource demands. ModernBERT represents a contemporary BERT variant specifically optimized for classification tasks with enhanced performance characteristics, while DomBertUrl was selected for its demonstrated effectiveness in domain classification metrics reported in recent literature.

Traditional Models encompass CNN [5], Random Forest [34], and LA_Bin07 [22], providing computational efficiency baselines with established detection capabilities. These architectures were included based on their proven classification performance in their respective evaluations, enabling comparison between modern transformer approaches and established methodologies for wordlist-based DGA detection.

C. Training Configuration

All experiments were conducted under consistent conditions using NVIDIA Tesla T4 GPUs within Google Colab [35] environments to ensure reproducible results. All source code and datasets used in this study are publicly available in the MoE-word-list-DGA GitHub repository [36].

Figure 1 illustrates the dataset construction process, where eight wordlist-based DGA families each contribute 10,000 samples combined with 80,000 legitimate domains to create the balanced 160,000-domain training dataset.

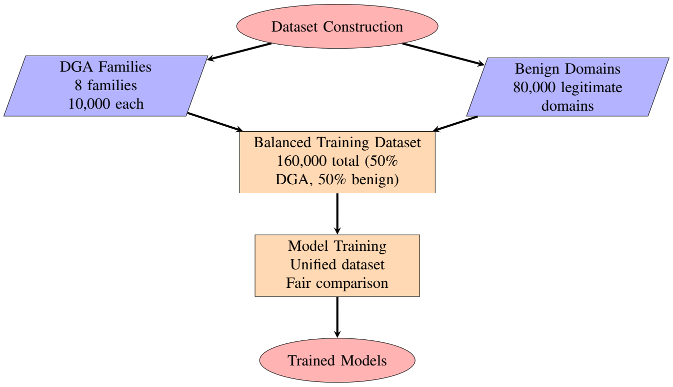

Fig. 1. Training configuration showing balanced 160K domain dataset construction from eight wordlist-based DGA families and legitimate domain samples.

D. Two-Phase Evaluation Protocol

We implemented a rigorous two-stage evaluation designed to assess both specialized performance and generalization capability, moving beyond traditional single-phase validation approaches.

The first phase evaluates model performance on known wordlist-based DGA families using fresh domains from the same families encountered during training. This in-family assessment measures how effectively models learn family-specific linguistic patterns while avoiding overfitting to training samples.

The second phase assesses generalization capability on completely unknown DGA families (bigviktor, ngioweb, and pizd) never encountered during training. This generalization evaluation reflects operational scenarios where new threat variants emerge continuously, testing model robustness beyond familiar attack patterns.

Statistical robustness is achieved through randomized batch sampling, with 30 batches per family containing 100 domains each (50 benign, 50 malicious). Results are reported with means and standard deviations following established sampling theory principles [37].

E. Specialization versus Generalization Analysis

To validate the benefits of expert specialization, we compare our wordlist-focused approach against a generalist model trained on the comprehensive 54 DGA families used in the training dataset described in Table 2 of La O et al. [17]. This expanded training set includes our eight wordlist families plus 46 additional families spanning arithmetic, hash-based, and hybrid generation strategies.

Both models use identical architectures and hyperparameters, differing only in training data composition. This controlled comparison directly addresses whether specialized training on wordlist characteristics outperforms broad exposure to diverse DGA types.

V. Experimental Results

A. Individual Model Performance

We evaluate seven candidate models across wordlist-based DGA families using the previously described two-phase protocol. Table II presents comprehensive performance metrics for known and unknown families, revealing significant performance variations across architectural paradigms.

The results demonstrate distinct architectural trade-offs and generalization patterns. Large language models exhibit extreme precision-recall imbalances: LLaMA achieves 92.4% precision but only 41.9% recall on known families, while Gemma maintains exceptional precision (95.4-95.7%) across both scenarios but struggles with recall (66.5% known, 60.3% unknown). This pattern indicates conservative classification behavior that misses substantial portions of malicious domains.

Specialized transformers show contrasting generalization capabilities. ModernBERT demonstrates consistent balanced performance with minimal precision-recall variance (89.7%/86.6% known vs 89.0%/75.5% unknown), while DomBertUrl exhibits superior generalization stability, actually improving its F1-score from 72.4% on known families to 84.6% on unknown families. This represents a remarkable 12.2 percentage point increase suggesting effective transfer learning.

Traditional approaches reveal fundamental limitations for semantic detection tasks. FANCI shows catastrophic performance degradation on unknown families, dropping from 70.5% to 39.1% F1-score, while CNN maintains moderate consistency (78.9% vs 65.5%) but fails to capture linguistic nuances.

Inference time analysis establishes clear operational boundaries: CNN achieves sub-millisecond inference but limited semantic understanding, specialized transformers operate in the practical 13-35ms range with superior accuracy, while large language models require 656-1413ms. This represents two orders of magnitude slower performance, making them impractical for real-time DNS monitoring systems processing millions of queries daily.

B. Expert Selection Analysis

Based on our composite scoring methodology, we evaluate overall expert qualification by integrating performance across

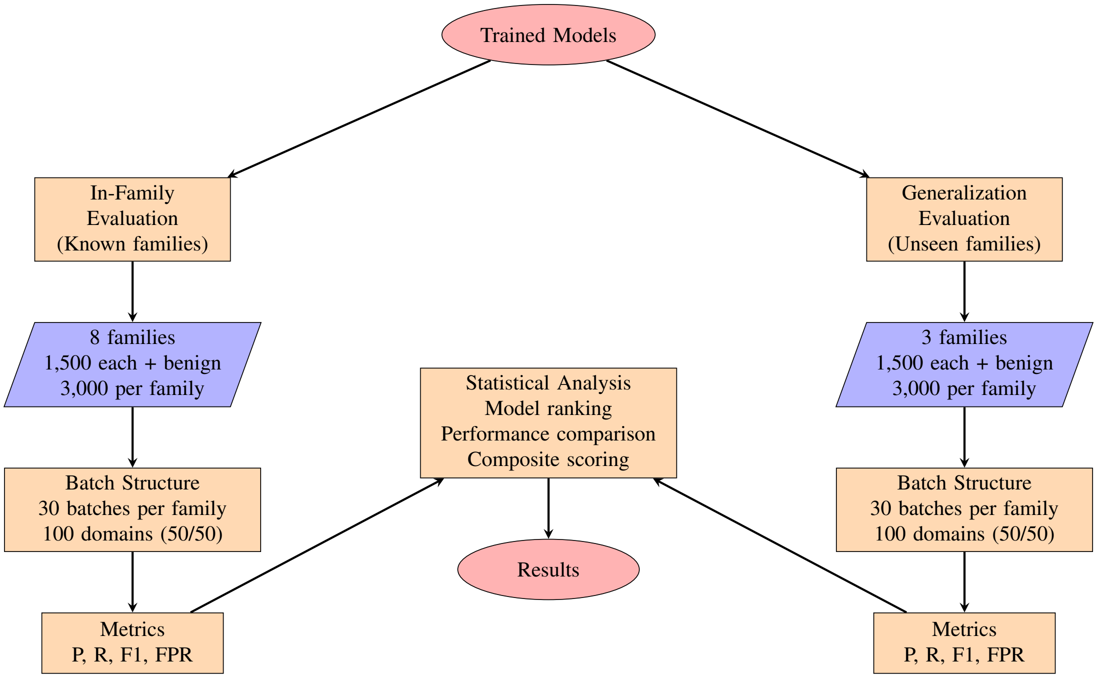

Fig. 2. Two-phase evaluation methodology emphasizing statistical robustness through randomized batch sampling across known and unknown family distributions.

TABLE II
PERFORMANCE COMPARISON OF EXPERT MODELS FOR WORDLIST DGA DETECTION ACROSS KNOWN AND UNKNOWN FAMILIES

Model	**Precision (%)**	**Recall (%)**	**F1 (%)**	**FPR (%)**	**Time (ms)**	**Score**
Known families (n=8)						
LLaMA 3.2 3B	92.4±5.5	41.9±8.8	54.7±8.8	2.9±2.2	656±95	0.679
Gemma 3 4B	**95.4±3.6**	66.5±5.7	75.2±4.8	**2.5±2.2**	1413±1309	0.669
ModernBERT	89.7±4.1	**86.6±3.1**	**86.7±3.0**	9.0±3.8	26±3	**0.904**
DomBertUrl	81.2±6.4	69.0±6.9	72.4±5.8	12.8±5.0	13±2	0.818
CNN	80.9±5.7	80.0±4.1	78.9±4.0	15.3±5.5	<**1**	0.849
FANCI	70.3±4.8	72.7±5.4	70.5±4.9	27.6±5.5	310±23	0.728
LA_Bin07	84.6±5.9	82.3±3.3	81.7±3.8	12.0±5.9	80±23	0.862
Unknown families (n=3)						
LLaMA 3.2 3B	60.5±4.4	68.8±4.9	63.4±4.2	39.8±5.8	693±270	0.607
Gemma 3 4B	**95.7±4.4**	60.3±5.9	70.8±5.0	**2.2±2.1**	1390±1146	0.652
ModernBERT	89.0±4.4	75.5±5.6	80.9±4.5	9.1±4.1	35±19	0.867
DomBertUrl	87.7±4.2	**82.3±4.5**	**84.6±3.5**	11.5±4.3	13±2	**0.884**
CNN	76.9±6.9	60.2±4.9	65.5±5.3	15.9±5.4	<**1**	0.773
FANCI	51.8±7.6	32.0±6.5	39.1±6.7	27.6±5.5	284±17	0.550
LA_Bin07	73.0±9.1	45.7±5.3	53.7±5.7	14.1±5.6	80±22	0.705

both evaluation scenarios. Table III presents the combined scoring analysis using equal weighting ($\alpha_i = 0.2$) that guides our expert selection decision.

The composite scoring analysis identifies ModernBERT as the optimal expert with a combined score of 0.886, demonstrating superior balance across detection accuracy, generalization capability, and operational constraints. While DomBertUrl excels in generalization scenarios (0.884 vs 0.867), ModernBERT's substantial advantage on known families (0.904 vs 0.818) combined with practical inference latency establishes it as the preferred expert for operational deployment.

TABLE III
EXPERT SELECTION SCORING ANALYSIS ACROSS KNOWN AND UNKNOWN DGA FAMILIES

Model	**Known Score**	**Unknown Score**	**Combined Score**	**Rank**
ModernBERT	**0.904**	0.867	**0.886**	**1**
DomBertUrl	0.818	**0.884**	0.851	2
CNN	0.849	0.773	0.811	3
LA_Bin07	0.862	0.705	0.784	4

However, given the operational costs associated with high false positive rates discussed earlier, where enterprise DNS monitoring systems processing millions of queries can be overwhelmed by misclassified legitimate domains, we examine how expert selection changes when FPR minimization becomes the primary concern. As a practical case study, we recalculate composite scores using $\alpha_1 = \alpha_3 = \alpha_4 = \alpha_5 = 0.1$ and $\alpha_2 = 0.6$, reflecting production environments where false positive costs significantly outweigh other performance considerations.

TABLE IV
EXPERT SELECTION UNDER FPR-PRIORITIZED SCORING ($\alpha_2 = 0.6$)

Model	Known Score	Unknown Score	Combined Score	Rank
ModernBERT	**0.907**	**0.888**	**0.898**	**1**
DomBertUrl	0.845	0.885	0.865	2
CNN	0.848	0.807	0.827	3
LA_Bin07	0.871	0.782	0.827	4

Table IV reveals that ModernBERT maintains its position as the optimal expert even under FPR-prioritized scoring, with the ranking order remaining unchanged across all models. This stability validates ModernBERT's robustness across different operational priorities and confirms its suitability for production deployment where false positive minimization is paramount. The consistent expert selection across scoring configurations demonstrates that ModernBERT achieves superior FPR performance without compromising other detection capabilities.

C. Family-Level Performance Analysis

Table V presents F1-scores across individual DGA families, revealing significant variance in detection difficulty. The manuelita family presents consistent challenges with F1-scores below 45% across all models, generating domains with sophisticated linguistic patterns that closely resemble legitimate structures. In contrast, matsnu and suppobox achieve detection rates above 90% for transformer models, indicating reliance on identifiable semantic patterns.

ModernBERT demonstrates the most consistent performance across families, achieving the highest scores on 6 out of 8 known families and maintaining competitive performance on unknown families.

D. Specialization vs. Generalization Validation

We compare our wordlist-focused expert against a generalist model trained on 54 DGA families to validate the benefits of domain-specific specialization. Both models use identical ModernBERT architecture and hyperparameters, differing only in training data composition. Table VI shows the F1-score comparison between specialized and generalist approaches.

Domain-specific training consistently outperforms the generalist approach. On known families, the specialist achieves 9.4% relative improvement, while on unknown families the advantage increases to 30.2%. This substantial generalization gain demonstrates that focused training enables the model to capture fundamental linguistic patterns underlying wordlist generation rather than memorizing family-specific characteristics.

Figure 3 illustrates the F1-score distribution differences between approaches. The specialized model demonstrates markedly higher consistency on known families, with F1-scores concentrated near the upper bound and minimal dispersion. In generalization scenarios, the specialist maintains compact F1-score distribution around 0.79, while the generalist shows broader variance and lower central tendency, highlighting diminished reliability on unfamiliar threats.

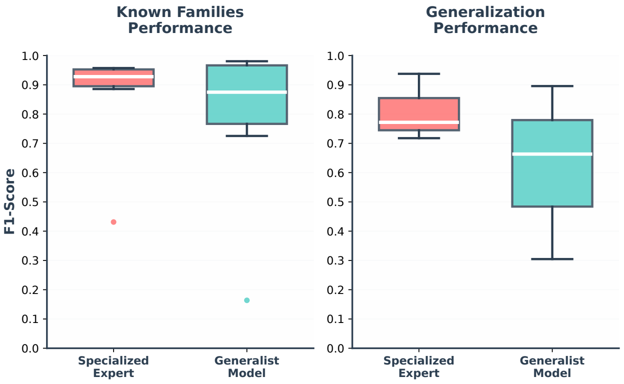

Fig. 3. F1-score distributions comparing specialized wordlist expert versus generalist model on known and unknown DGA families showing superior consistency of the specialized approach

These results validate the expert specialization paradigm for cybersecurity applications. The specialist achieves a combined composite score of 0.886 (averaging performance across known and unknown family scenarios) compared to the generalist's 0.840, demonstrating superior overall expert qualification.

VI. Discussion

A. Expert Model Selection and Performance Trade-offs

ModernBERT emerges as the optimal expert model, demonstrating superior balance between detection accuracy and operational efficiency. The transformer architecture proves most suitable for wordlist-based DGA detection, effectively capturing the semantic patterns that distinguish sophisticated malicious domains from legitimate services.

Traditional machine learning approaches show significant limitations for wordlist-based DGA detection, though with varying degrees of effectiveness. FANCI demonstrates the poorest performance, failing catastrophically due to its reliance on hand-crafted features that cannot capture semantic relationships essential for distinguishing linguistically coherent malicious domains from legitimate ones.

CNN and LA_Bin07 achieve moderate success, maintaining reasonable detection capabilities while offering computational efficiency advantages. However, these approaches still fall short of transformer-based models when confronting the semantic complexity inherent in wordlist variants that closely mimic natural language patterns.

TABLE V
F1-SCORE PERFORMANCE MATRIX ACROSS WORDLIST DGA FAMILIES

Family	**LLaMA**	**Gemma**	**ModernBERT**	**DomBertUrl**	**LA_Bin07**	**CNN**	**FANCI**
			Known families				
charbot	53.2±6.9	68.3±5.4	**88.6±3.1**	66.0±5.3	83.0±4.6	79.3±3.9	64.2±6.0
deception	82.8±4.8	**97.9±1.7**	95.7±1.8	92.8±2.7	94.4±2.6	92.5±2.4	87.9±2.2
gozi	58.6±5.9	81.1±5.3	**90.4±2.9**	63.6±12.1	84.1±5.6	79.6±7.1	74.0±6.5
manuelita	34.5±11.2	29.7±6.4	**43.1±7.8**	24.5±8.4	23.8±6.4	22.6±6.0	29.3±5.6
matsnu	83.4±8.4	90.7±2.5	**95.2±2.1**	89.5±3.3	93.0±2.8	90.1±2.6	82.9±3.4
nymaim	33.6±8.7	55.1±7.7	**89.8±2.7**	62.5±6.9	87.7±3.4	81.9±4.6	67.1±4.4
rovnix	40.7±7.0	**96.4±1.7**	95.5±2.0	92.4±2.2	93.6±2.5	92.4±2.6	85.4±2.8
suppobox	51.4±17.3	82.1±8.2	**95.2±2.0**	87.8±5.9	94.2±2.7	92.9±2.4	73.5±6.0
Average	54.7±9.5	75.2±5.4	**86.7±3.6**	72.4±6.6	81.7±4.1	78.9±4.3	70.5±4.9
			Unknown families				
bigviktor	71.7±4.0	40.2±6.5	77.2±4.8	**79.0±4.2**	35.6±4.8	47.3±6.1	48.3±5.9
ngioweb	81.5±2.5	**88.6±3.5**	71.8±6.4	82.6±3.9	42.0±9.4	57.7±7.2	45.5±6.4
pizd	36.9±6.1	83.5±4.9	**93.8±2.1**	92.4±2.5	83.4±3.0	91.4±2.5	23.5±7.7
Average	63.4±4.4	70.8±5.1	80.9±4.8	**84.6±3.6**	53.7±6.3	65.5±5.6	39.1±6.7
Total Average	57.1±8.4	74.0±5.3	**85.1±3.9**	75.7±5.9	74.1±4.8	75.2±4.7	62.0±5.4

TABLE VI
F1-SCORE COMPARISON: SPECIALIZED WORDLIST EXPERT VS. GENERALIST MODEL TRAINED ON 54 DGA FAMILIES

Evaluation Scenario	**Specialist F1-Score**	**Generalist F1-Score**	**Improvement**
Known families	**86.7**%	79.2%	+9.4%
Unknown families	**80.9**%	62.1%	+30.2%

Large language models, while possessing the necessary semantic understanding, introduce prohibitive computational overhead that renders them impractical for real-time cybersecurity applications. Their multi-second inference requirements create an insurmountable barrier for enterprise DNS monitoring systems that must process queries at scale.

B. Validation of Expert Specialization

The comparative analysis between specialized and generalist approaches validates the benefits of domain-specific training. The specialist model achieves 9.4% F1-score improvement on known families and 30.2% F1-score enhancement on unknown variants compared to the generalist trained on 54 diverse DGA families.

This performance differential reveals a fundamental principle: focused training on specific threat categories develops more effective pattern recognition than broad exposure to diverse attack types. The substantial improvement on unknown families demonstrates that specialization enhances transferability rather than limiting it through overfitting to familiar patterns.

The specialist model's superior generalization capability suggests that concentrated exposure to wordlist characteristics enables the learning of fundamental linguistic structures underlying this DGA category. In contrast, the generalist approach appears to struggle with the semantic complexity inherent in wordlist-based domains when training resources are distributed across multiple threat types.

These results support the deployment of specialized detection models rather than single comprehensive systems for cybersecurity applications where distinct attack categories exhibit fundamentally different characteristics.

C. Architectural Boundaries and Deployment Considerations

The evaluation reveals clear architectural boundaries that guide model selection for production environments. Large language models (LLaMA, Gemma) demonstrate semantic understanding but suffer from prohibitive inference latency (656-1413ms), making them unsuitable for real-time applications. Traditional approache CNN offer computational efficiency but lack the semantic sophistication required for wordlist detection.

Specialized transformers occupy the practical deployment space, offering the optimal balance between detection capability and operational constraints. ModernBERT's 9.0% false positive rate, while higher than large language models (2.5-2.9%), remains operationally acceptable given the substantial improvement in detection capability and processing speed.

The inference time analysis establishes practical thresholds for cybersecurity applications: sub-100ms latency enables real-time DNS monitoring, while multi-second inference times limit models to batch processing scenarios unsuitable for active threat mitigation.

The composite scoring methodology allows practitioners to select the most suitable model for their specific application by adjusting the weighting coefficients ($\alpha_1, \ldots, \alpha_5$) according to operational priorities. For instance, real-time monitoring systems processing millions of DNS queries would benefit from increasing the inference time weight ($\alpha_5 > 0.4$) to prioritize speed, potentially selecting CNN despite lower accuracy. Conversely, security investigation scenarios could emphasize recall ($\alpha_4 > 0.4$) to ensure comprehensive threat detection, favoring models like LLaMA despite slower processing. This approach enables organizations to identify the optimal model that aligns with their deployment constraints and security requirements rather than relying on a one-size-fits-all solution.

D. Current Limitations and Research Directions

The manuelita family's resistance to detection across all evaluated models highlights a critical limitation in current approaches. Rather than indicating fundamental flaws in semantic methods, this challenge suggests the need for deeper understanding of why models fail on specific linguistic patterns. SHAP (SHapley Additive exPlanations) [38] analysis could provide valuable insights into which domain characteristics cause classification errors. For instance, SHAP could reveal whether failures stem from specific character n-grams, syllable patterns, or semantic word combinations that closely mimic legitimate domain structures. Such granular feature attribution would enable targeted improvements: if SHAP identifies that models incorrectly weight certain linguistic tokens as benign, training strategies could emphasize these problematic patterns, or model architectures could incorporate attention mechanisms specifically designed to capture the subtle semantic differences that distinguish sophisticated wordlist-generated domains from legitimate ones.

The temporal dimension of DGA evolution presents ongoing challenges. As threat actors adapt their generation strategies in response to detection systems, models require continuous learning capabilities that maintain performance on established families while adapting to emerging variants. This necessitates research into continual learning approaches that avoid catastrophic forgetting.

Future work should address complete Mixture of Experts (MoE) implementation [39] including dynamic routing mechanisms and expert gating strategies. Following the comparative analysis methodology established in this study for wordlist-based DGAs, similar evaluations should be conducted across other DGA categories (arithmetic, hash-based, hybrid) to identify optimal expert models for each threat type. This comprehensive approach would enable the development of a complete expert ensemble system capable of high-efficiency detection tailored to specific operational requirements.

Evaluation expansion to hybrid DGA categories would establish comprehensive design principles for cybersecurity MoE systems. Additionally, investigation of few-shot learning approaches could improve detection of novel families with limited training data, addressing the challenge of emerging threats that lack substantial training samples.

The integration of additional semantic features, such as domain registration patterns and network-level characteristics, may enhance detection capabilities for linguistically sophisticated families. However, such enhancements must maintain the real-time processing requirements critical for operational deployment.

VII. Conclusions

This work presented a systematic strategy for expert model selection aimed at detecting domains generated by wordlist-based DGAs. Through the evaluation of seven models across different architectural paradigms, *ModernBERT* was identified as the optimal expert, combining high detection performance with inference times suitable for real-time operational environments.

The two-phase evaluation protocol enabled the assessment of both specialization on known families and generalization to previously unseen variants. Results demonstrate that domain-specific training offers significant advantages over generalist approaches, with improvements of 9.4% and 30.2% in F1-score for known and unseen scenarios, respectively. This gap suggests that specialization enhances transferable linguistic pattern recognition rather than limiting it through overfitting to familiar cases.

Nevertheless, persistent challenges remain in families such as *manuelita*, whose linguistic coherence complicates classification even for semantically advanced models. This limitation highlights the need to explore deeper contextual analysis techniques or the integration of complementary features.

Finally, the composite scoring methodology used in this study provides a flexible framework for expert selection, adaptable to various operational priorities such as precision, response time, or false positive rates. Future work will focus on extending this approach toward a full MoE implementation, including dynamic routing and gating mechanisms, and evaluating its effectiveness against other DGA categories.

Acknowledgments

The authors would like to thank the National Scientific and Technical Research Council of Argentina (CONICET) for its institutional support throughout the development of this research.

References

[1] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla, "A comprehensive measurement study of domain generating malware," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 263–278.

[2] L. Yang, J. Zhai, W. Liu, X. Ji, H. Bai, G. Liu, and Y. Dai, "Detecting word-based algorithmically generated domains using semantic analysis," *Symmetry*, vol. 11, no. 2, p. 176, 2019.

[3] X. H. Vu and X. D. Hoang, "A novel machine learning-based approach for detecting word-based dga botnets," *Journal of Theoretical and Applied Information Technology*, 2021.

[4] I. X-Force, "Grandoreiro banking trojan unleashed: X-force observing emerging global campaigns," April 2024, accessed: 2025-06-19. [Online]. Available: https://www.ibm.com/think/x-force/grandoreiro-banking-trojan-unleashed

[5] C. Catania, S. García, and P. Torres, "Deep convolutional neural networks for dga detection," in *Computer Science–CACIC 2018: 24th Argentine Congress, Tandil, Argentina, October 8–12, 2018, Revised Selected Papers 24*. Springer, 2019, pp. 327–340.

[6] E. W. T. Ferreira, G. A. Carrijo, R. de Oliveira, and N. V. de Souza Araujo, "Intrusion detection system with wavelet and neural artifical network approach for networks computers," *IEEE Latin America Transactions*, vol. 9, no. 5, pp. 832–837, 2011.

[7] B. Cebere, J. Flueren, S. Sebastián, D. Plohmann, and C. Rossow, "Down to earth! guidelines for dga-based malware detection," 2024.

[8] R. da Silveira Lopes, J. C. Duarte, and R. R. Goldschmidt, "False positive identification in intrusion detection using xai," *IEEE Latin America Transactions*, vol. 21, no. 6, pp. 745–751, 2023.

[9] D. Plohmann, "Dgaarchive–a deep dive into domain generating malware," *Dec-2015*, 2015.

[10] Y. Fu, L. Yu, O. Hambolu, I. Ozcelik, B. Husain, J. Sun, K. Sapra, D. Du, C. T. Beasley, and R. R. Brooks, "Stealthy domain generation algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1430–1443, 2017.

[11] B. Yu, J. Pan, J. Hu, A. Nascimento, and M. De Cock, "Character level based detection of dga domain names," in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[13] K. Highnam, D. Puzio, S. Luo, and N. R. Jennings, "Real-time detection of dictionary dga network traffic using deep learning," *SN Computer Science*, vol. 2, no. 2, p. 110, 2021.

[14] J. Woodbridge, H. S. Anderson, A. Ahuja, and D. Grant, "Predicting domain generation algorithms with long short-term memory networks," *arXiv preprint arXiv:1611.00791*, 2016.

[15] R. R. Curtin, A. B. Gardner, S. Grzonkowski, A. Kleymenov, and A. Mosquera, "Detecting dga domains with recurrent neural networks and side information," in *Proceedings of the 14th international conference on availability, reliability and security*, 2019, pp. 1–10.

[16] M. A. Sayed, A. Rahman, C. Kiekintveld, and S. Garcia, "Fine-tuning large language models for dga and dns exfiltration detection," *arXiv preprint arXiv:2410.21723*, 2024.

[17] R. L. La O, C. A. Catania, and T. Parlanti, "Llms for domain generation algorithm detection," *arXiv preprint arXiv:2411.03307*, 2024.

[18] J. J. Koh and B. Rhodes, "Inline detection of domain generation algorithms with context-sensitive word embeddings," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 2966–2971.

[19] W. Huang, Y. Zong, Z. Shi, L. Wang, and P. Liu, "Pepc: A deep parallel convolutional neural network model with pre-trained embeddings for dga detection," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[20] C. Morbidoni, A. Cucchiarelli, and L. Spalazzi, "Mixed-embeddings and deep learning ensemble for dga classification with limited training data," *IEEE Access*, 2025.

[21] D. Tran, H. Mac, V. Tong, H. A. Tran, and L. G. Nguyen, "A lstm based framework for handling multiclass imbalance in dga botnet detection," *Neurocomputing*, vol. 275, pp. 2401–2413, 2018.

[22] T. A. Tuan, H. V. Long, and D. Taniar, "On detecting and classifying dga botnets and their families," *Computers & Security*, vol. 113, p. 102549, 2022.

[23] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[24] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[25] 360NetLab, "360netlab dga dataset," 2023, accessed: 2024-07-04. [Online]. Available: https://data.netlab.360.com/

[26] M. Zago, M. G. Pérez, and G. M. Pérez, "Umudga: A dataset for profiling algorithmically generated domain names in botnet detection," *Data in Brief*, vol. 30, p. 105400, 2020.

[27] J. Peck, C. Nie, R. Sivaguru, C. Grumer, F. Olumofin, B. Yu, A. Nascimento, and M. De Cock, "Charbot: A simple and effective method for evading dga classifiers," *IEEE Access*, vol. 7, pp. 91 759–91 771, 2019.

[28] P. Snyder, C. Taylor, and C. Kanich, "The 2020 tranco list: Improving the alexa ranking," https://tranco-list.eu, 2020, accessed: 2024-07-05.

[29] G. DeepMind, "Gemma documentation," https://ai.google.dev/gemma/docs/core, 2024, accessed: 2025-06-05.

[30] ——. (2025) Gemma 3 4b it. Hugging Face. Accessed: 2025-04-30. [Online]. Available: https://huggingface.co/google/gemma-3-4b-it

[31] M. AI, "Llama 3.2 3b instruct on hugging face," https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024, accessed: 2025-06-05.

[32] Answer.AI, "Modernbert-base on hugging face," https://huggingface.co/answerdotai/ModernBERT-base, 2024, accessed: 2025-06-05.

[33] A. E. Mahdaouy, S. Lamsiyah, M. J. Idrissi, H. Alami, Z. Yartaoui, and I. Berrada, "Domurls_bert: Pre-trained bert-based model for malicious domains and urls detection and classification," *arXiv preprint arXiv:2409.09143*, 2024.

[34] S. Schüppen, D. Teubert, P. Herrmann, and U. Meyer, "{FANCI}: Feature-based automated {NXDomain} classification and intelligence," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1165–1181.

[35] Google, "Google colaboratory: A cloud-based collaborative notebook environment," 2023, available at: https://colab.research.google.com [Accessed: 2024-06-15].

[36] R. L. L. O, "MoE-word-list-dga-detection: Mixture-of-Experts for Wordlist-based DGA Detection," https://github.com/reypapin/MoE-word-list-dga-detection, 2024, accessed: 2025-06-13.

[37] W. G. Cochran, *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons, 1977.

[38] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[39] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.