

## ABSTRACT

Domain Generation Algorithms (DGAs) automatically produce large numbers of pseudo-random domain names for command-and-control (C2) communication, thereby posing a substantial challenge to network security mechanisms. While machine learning-based detectors have demonstrated high accuracy for DGAs represented in the training data, only 26% of prior studies explicitly evaluate cross-family generalization, and model performance frequently deteriorates when confronted with previously unseen DGA families. To this end, we develop a Random Forest classifier employing a split-ensemble training scheme, comprising 24 stratified sub-ensembles aggregated via majority voting. The model relies on 12 domain-level features, encompassing entropy-based, structural, linguistic, and sequential characteristics.

The experimental evaluation is conducted on 120 DGA families, of which 65 families are strictly held out from the training phase to emulate a realistic zero-day scenario. The proposed split-ensemble model attains a Matthews Correlation Coefficient (MCC) of 0.965 (95% CI: 0.963–0.967) on these zero-day families, with all 65 held-out families surpassing the generalization threshold of  $MCC = 0.70$ . Entropy-related features account for 60.2% of the aggregated feature importance and exhibit stable relative rankings across evaluation settings (Spearman’s  $\rho = 1.0$ ). The split-ensemble training strategy mitigates performance degradation under distribution shift by 73% relative to a single-model baseline ( $\Delta MCC = 0.095$ ,  $p < 0.001$ ).

The resulting false positive rate of 0.17% lies within an acceptable range for operational deployment. Overall, the findings indicate that diversity in the training process rather than increased architectural complexity is the primary factor governing generalization to zero-day DGA families. Consequently, a lightweight, interpretable ensemble model can achieve competitive zero-day detection performance, offering a resource-efficient and operationally transparent alternative to deep learning-based methods, and is thus well suited for integration into Security Operations Center workflows.

**Keyword:** Domain Generation Algorithm, zero-day detection, machine learning, Random Forest ensemble, cybersecurity, malware detection