Online Advertisement Detection by Lightweight URL Analysis

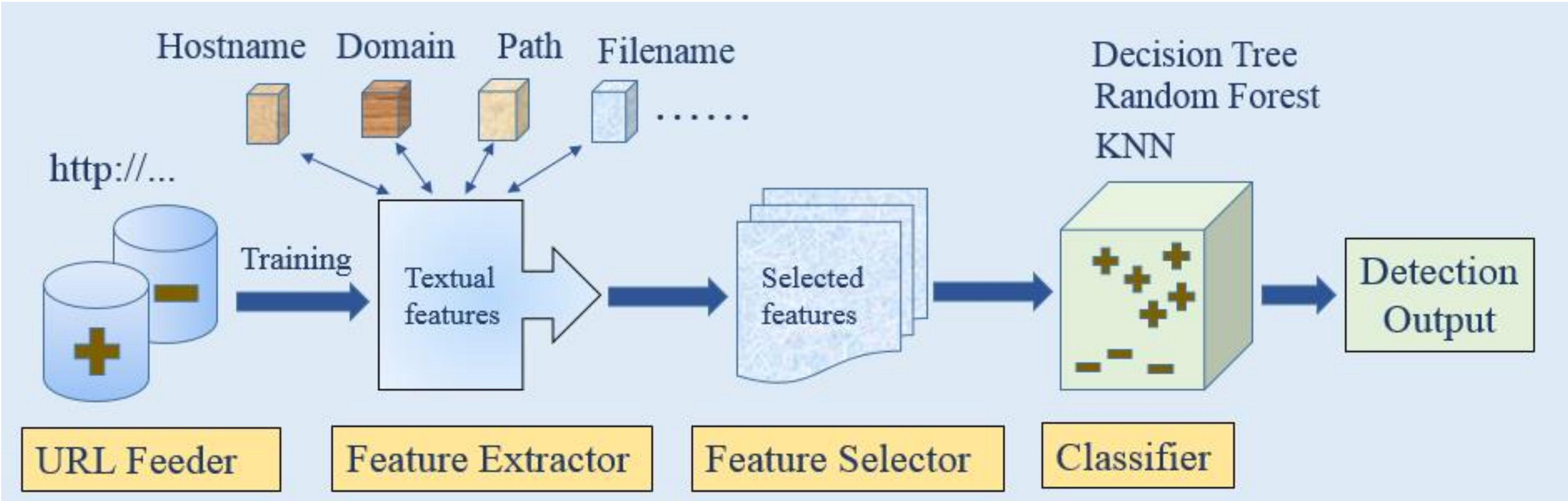
Xichen Zhang, Arash Habibi Lashkari, Ali A. Ghorbani

Canadian Institute for Cybersecurity (CIC), University of new Brunswick (UNB)

ABSTRACT

Due to the fast development of online advertising, malicious advertisements have become one of the major issues to distribute scamming information, click fraud and malware. Current approaches are involved with filtering lists for advertisement detection and blocking, which are not scalable and need manual maintenance. This study presents a lightweight online advertising classification system using lexical-based features as an alternative solution. In order to imitate real-world cases, three different scenarios are generated depending on three different URL sources. Then a set of URL lexical-based features are selected from previous researches for the purpose of training and testing the proposed model. Results show that by using lexical-based features, advertising detection accuracy is about 97% in certain scenario.

ARCHITECTURE



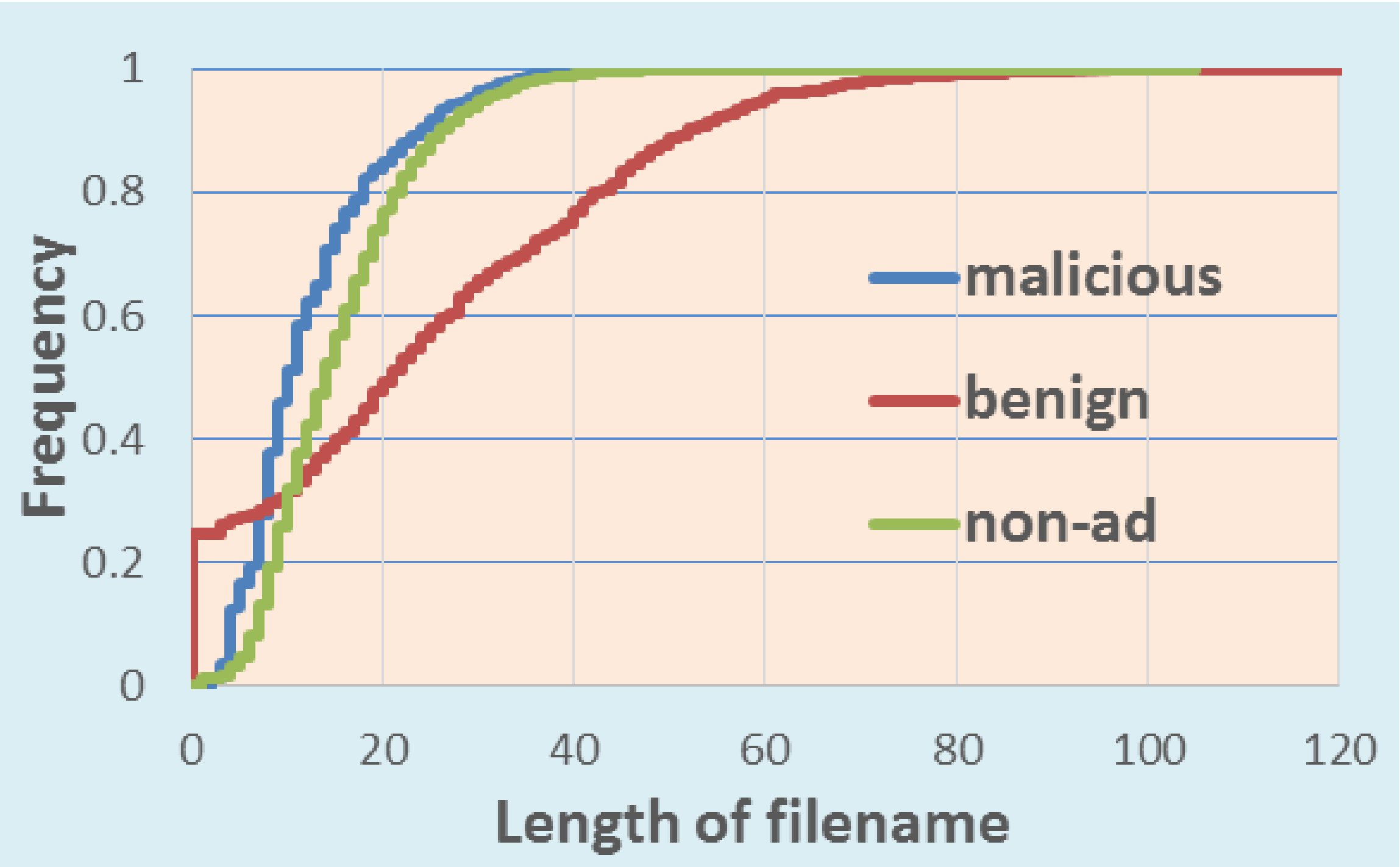
DATASET

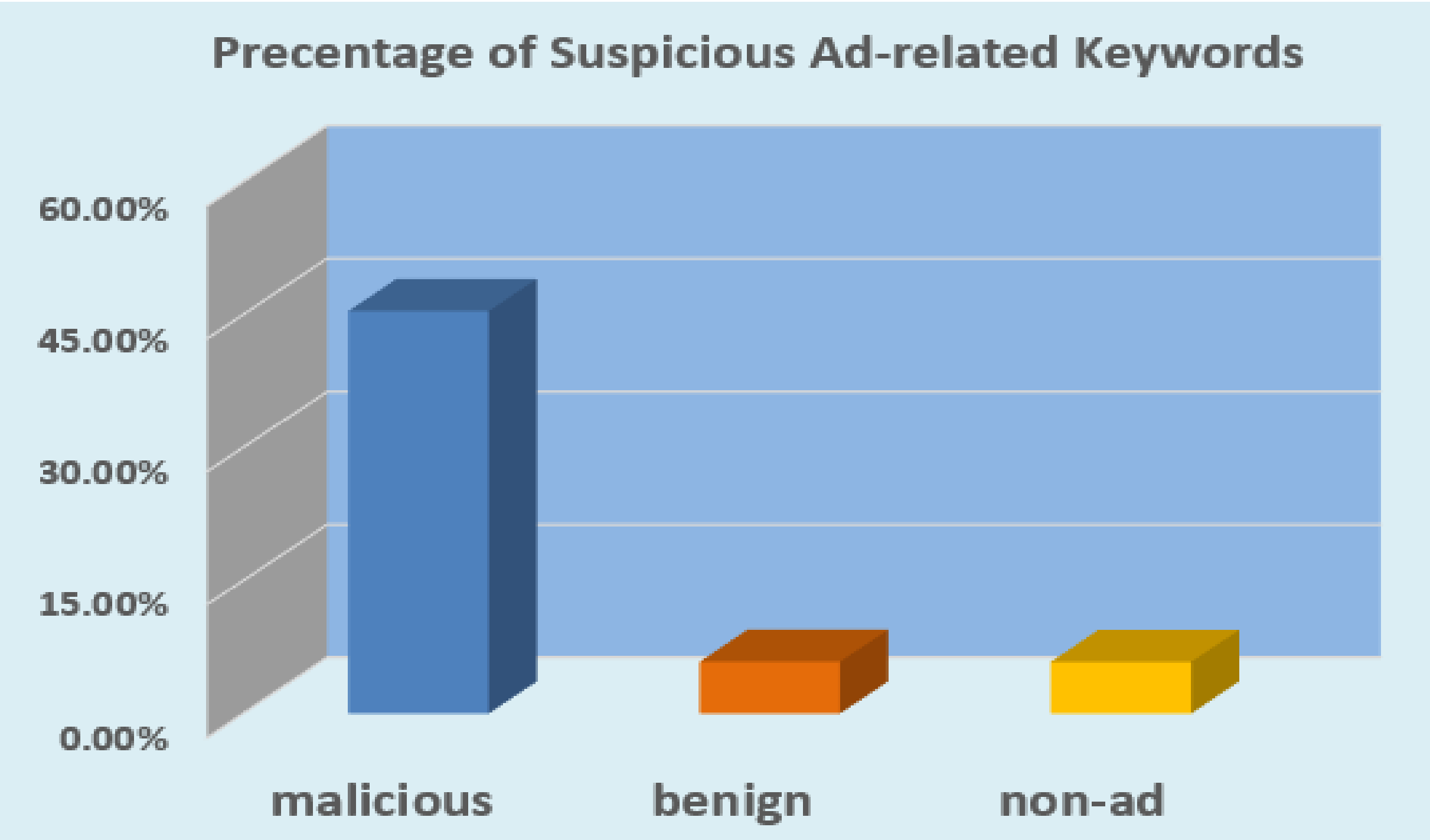
	Malicious-ad	Benign-ad	Non-ad
Scenario A	--	3000	5000
Scenario B	1115	--	5000
Scenario C	1115	3000	5000

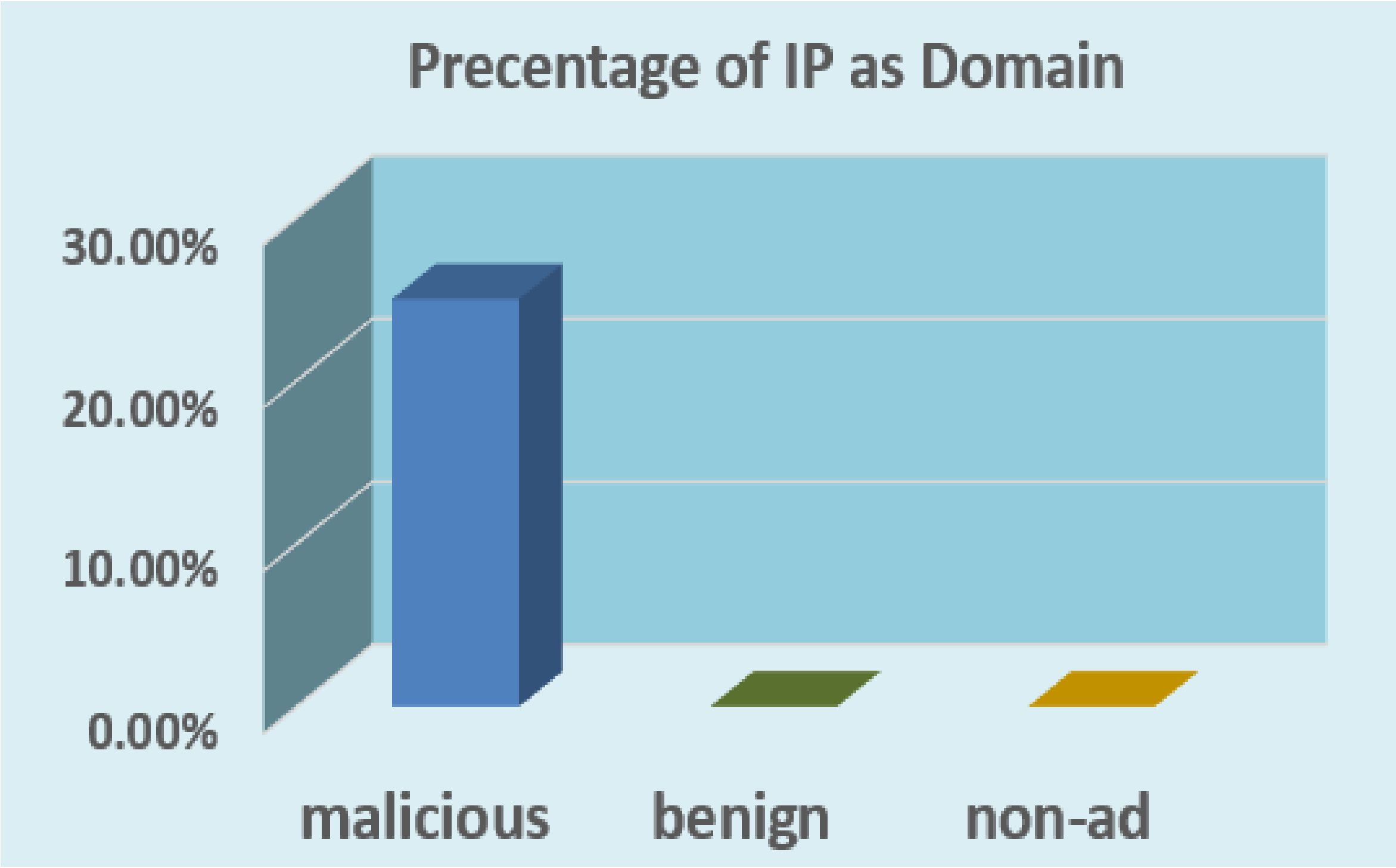
FEATURE SET

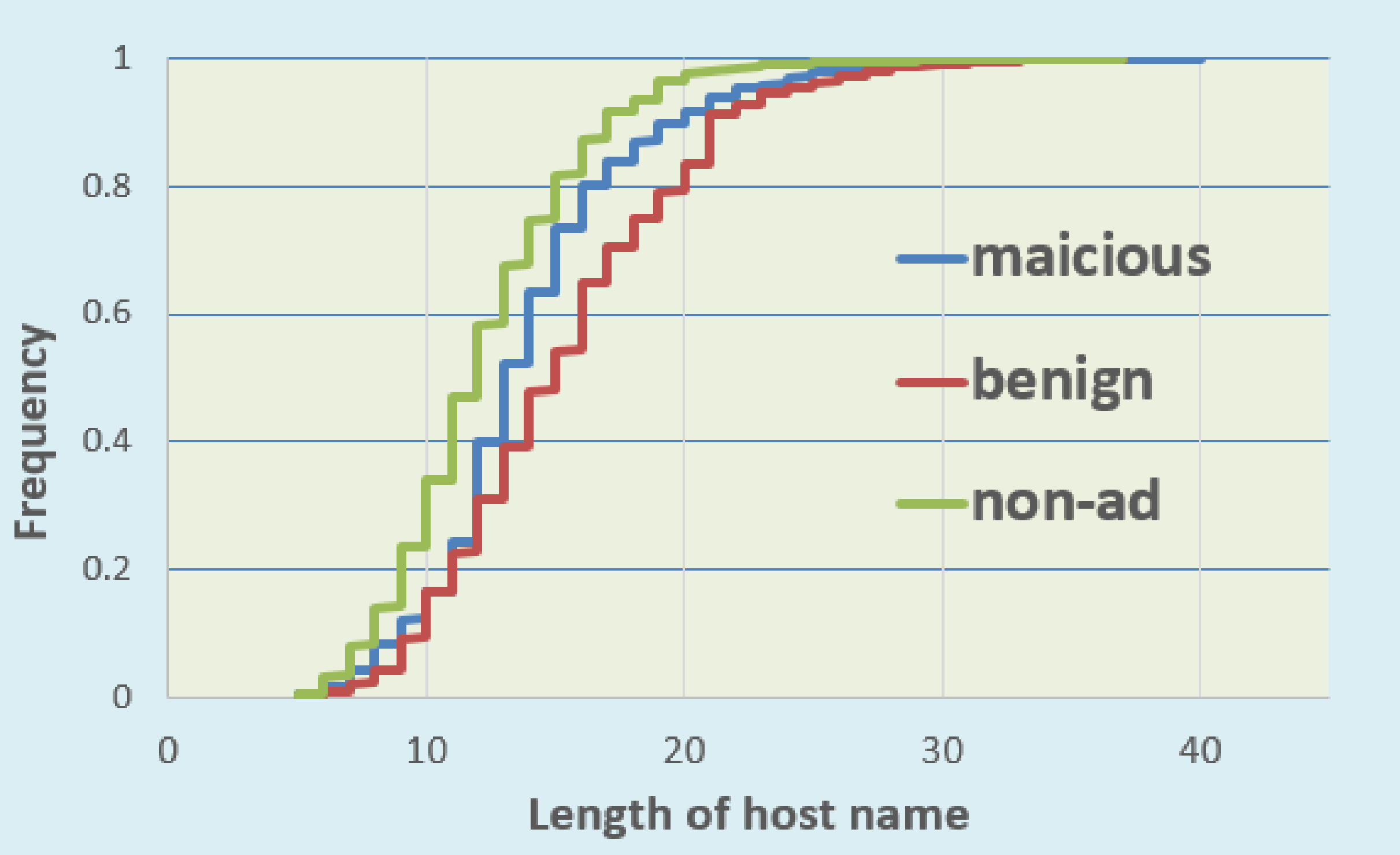
Feature Name	Size
Length of URL / hostname / domain / filename	4
Numeric token / ad-keyword / symbol present	3
Dot / Dash count	2
Longest token / path length	2
Average token / path length	2
Number of symbols / URL component	2
Length ratio	9
Pattern-based features	7
Executable file present	1
Use IP as domain	1
Total	**33**

FEATURE ANALYSIS









Length of filename, Ad-related keyword and **IP as domain** are distinguishing features for ad-detection. **Length of host name** contributes little to the system.

CONTRIBUTION & FUTURE WORK

Scen.	Algo.	Accuracy	Precision	Recall	FPR
A	C4.5	97.6%	97.6%	96.0%	1.5%
B	C4.5	97.5%	97.5%	97.3%	0.6%
C	C4.5	97.4%	97.3%	97.0%	2.2%

- ♦ C4.5 outperforms other algorithms
- ♦ Both full & select feature set provide promising results for online ad-detection
- ♦ More sophisticated algorithms are necessary for improving FNR