Open Universiteit

Botnet Detection

Detection of DGA-generated Domain Names

Harald Vranken

OUrsi, 10 May 2022

Introduction

- Harald Vranken and Hassan Alizadeh, *Detection of DGA-Generated Domain Names with TF-IDF*, MDPI Electronics 2022, 11, 414, https://doi.org/10.3390/electronics11030414
- Lars Kuipers, *Effectiveness of features in DGA detection*, Research internship thesis, Radboud University, January 2022

electronics MDPI

Article

Detection of DGA-Generated Domain Names with TF-IDF

Harald Vranken [1,2,*] and Hassan Alizadeh [1]

1 Department of Computer Science, Open Universiteit, P.O. Box 2960, 6401 DL Heerlen, The Netherlands; hassan77.alizadeh@gmail.com
2 Institute for Computing and Information Sciences, Radboud University, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
* Correspondence: harald.vranken@ou.nl

Abstract: Botnets often apply domain name generation algorithms (DGAs) to evade detection by generating large numbers of pseudo-random domain names of which only few are registered by cybercriminals. In this paper, we address how DGA-generated domain names can be detected by means of machine learning and deep learning. We first present an extensive literature review on recent prior work in which machine learning and deep learning have been applied for detecting DGA-generated domain names. We observe that a common methodology is still missing, and the use of different datasets causes that experimental results can hardly be compared. We next propose the use of TF-IDF to measure frequencies of the most relevant n-grams in domain names, and use these as features in learning algorithms. We perform experiments with various machine-learning and deep-learning models using TF-IDF features, of which a deep MLP model yields the best results. For comparison, we also apply an LSTM model with embedding layer to convert domain names from a sequence of characters into a vector representation. The performance of our LSTM and MLP models is rather similar, achieving 0.994 and 0.995 AUC, and average F1-scores of 0.907 and 0.891 respectively.

Keywords: DGA; botnet; TF-IDF; machine learning; deep learning

check for updates

Citation: Vranken, H.; Alizadeh, H. Detection of DGA-Generated Domain Names with TF-IDF. *Electronics* 2022, *11*, 414. https://doi.org/10.3390/electronics11030414

Academic Editors: Constantinos Kolias, Georgios Kambourakis and Weizhi Meng

Received: 1 December 2021
Accepted: 24 January 2022
Published: 29 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Botnets pose a severe threat to the security of systems connected to the Internet and their users. A botnet is composed of a collection of compromised systems ('bots') that receive and respond to commands from a Command and Control (C&C) server. A C&C server acts as rendezvous point between the bots and the botmaster, who controls the botnet. By updating the malware running on the bots, the botmaster can configure the botnet to perform different types of attacks, such as launching DDoS attacks, sending spam, or stealing credentials. This versatility causes that botnets are considered as the Swiss army knife of cybercriminals.

C&C servers and the communication channels between botmaster and bots are critical components of a botnet. By taking down the C&C servers, or by blocking the communication channels, the link between bots and botmaster is broken, which renders the botnet useless. Numerous techniques have been applied to provide stealthy botnet operation and to increase resilience against take-down attempts [1]. An eminent technique to evade detection is the application of domain name generation algorithms (DGAs) in bot malware that generate large numbers of pseudo-random domain names for contacting the C&C server, of which only few are actually registered shortly by the botmaster. Due to the dynamic DGA operation and short-lived domain names, the communication between C&C servers and bots is protected against take-down attempts.

The presence of botnets that use DGAs can be revealed by analysing network traffic. For instance, most of the domain names that are generated by DGAs are not registered and hence DNS lookups for resolving such domain names into IP addresses will result in NXDomain responses from name servers. Hence, by monitoring and analysing NXDomain

Electronics 2022, *11*, 414. https://doi.org/10.3390/electronics11030414 https://www.mdpi.com/journal/electronics

Research internship
Cyber Security

Radboud University

Effectiveness of features in DGA detection

Author:
Lars Kuipers
s1005864

First supervisor/assessor:
Dr. ir. Harald Vranken
harald.vranken@ou.nl

Second assessor:
Dr. Katharina Kohls
kkohls@cs.ru.nl

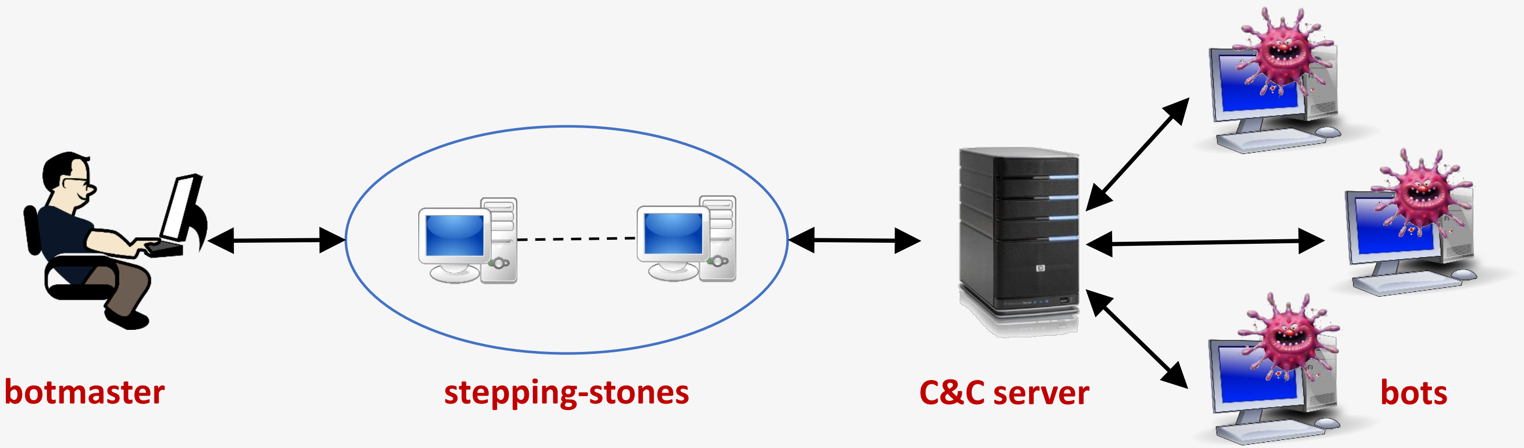
January 29, 2022

Outline

- Botnets
- DGA
- DGA detection with TF-IDF
- Effectiveness of features for DGA detection

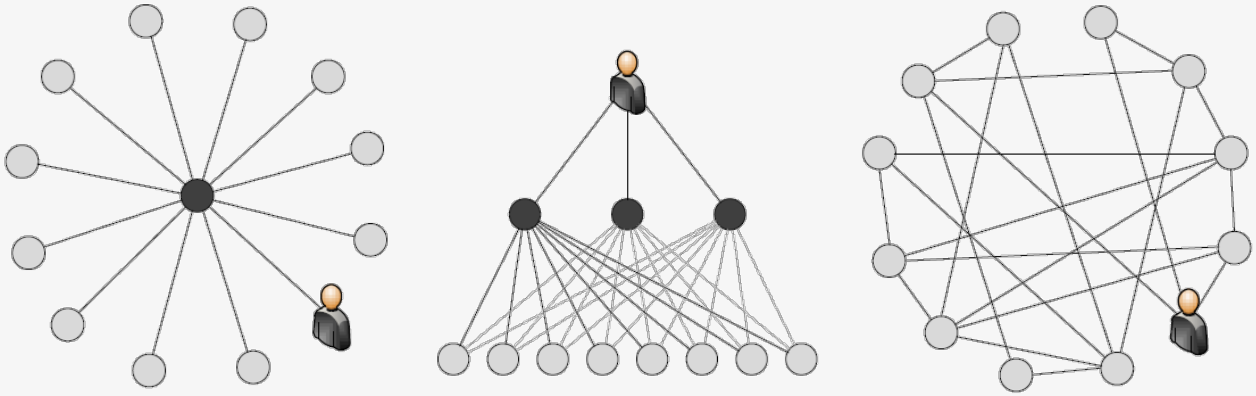
Botnet

- Network of *bots* (computer systems infected with malicious software)
- Bots are controlled remotely by a *botmaster* through *C&C server*
- Botmaster can employ proxy machines (*stepping-stones*) to evade detection
- Botnets are major *cybersecurity threat* (‘Swiss-army knife’ of cyber criminals)



Botnet structure

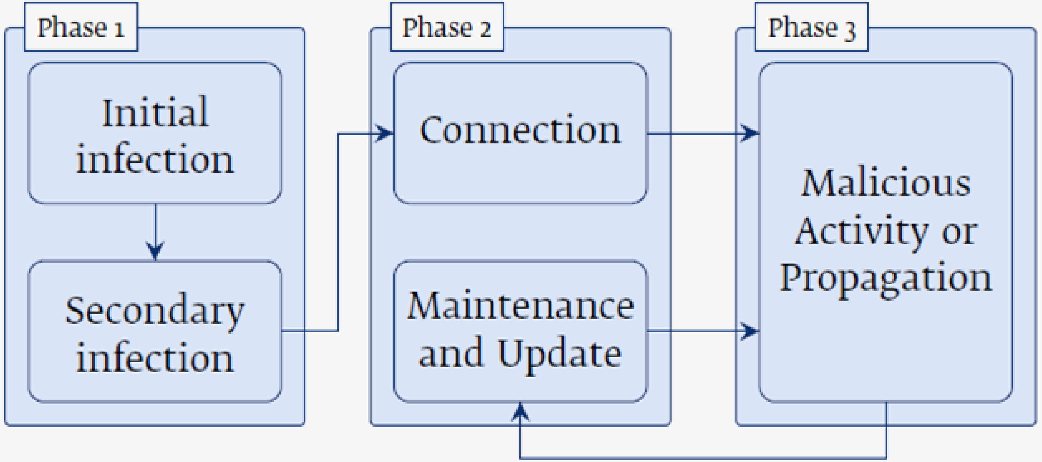
- C&C channels
  - push or pull
  - IRC, HTTP, DNS, …

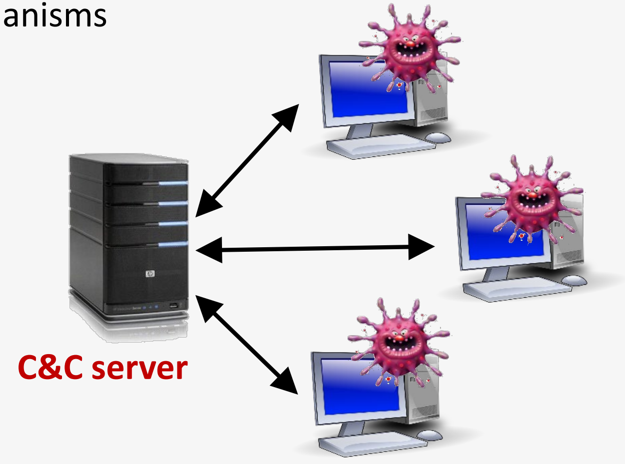


(a) Centralized (b) Semi-Distributed (c) Peer-to-Peer

Bot lifecycle

- *Infection*: bot is infected with malware (initial infection) and downloads bot binary (secondary infection)
- *Rallying*: bot contacts C&C server and announces its presence
  - establishes *C&C channel* through which bot receives updates and commands
- *Passive*: bot waits for commands (and bot binary may be updated)
- *Active*: bot carries out malicious activity
  - optionally spreads infection to other hosts using *propagation* mechanisms





C&C channels

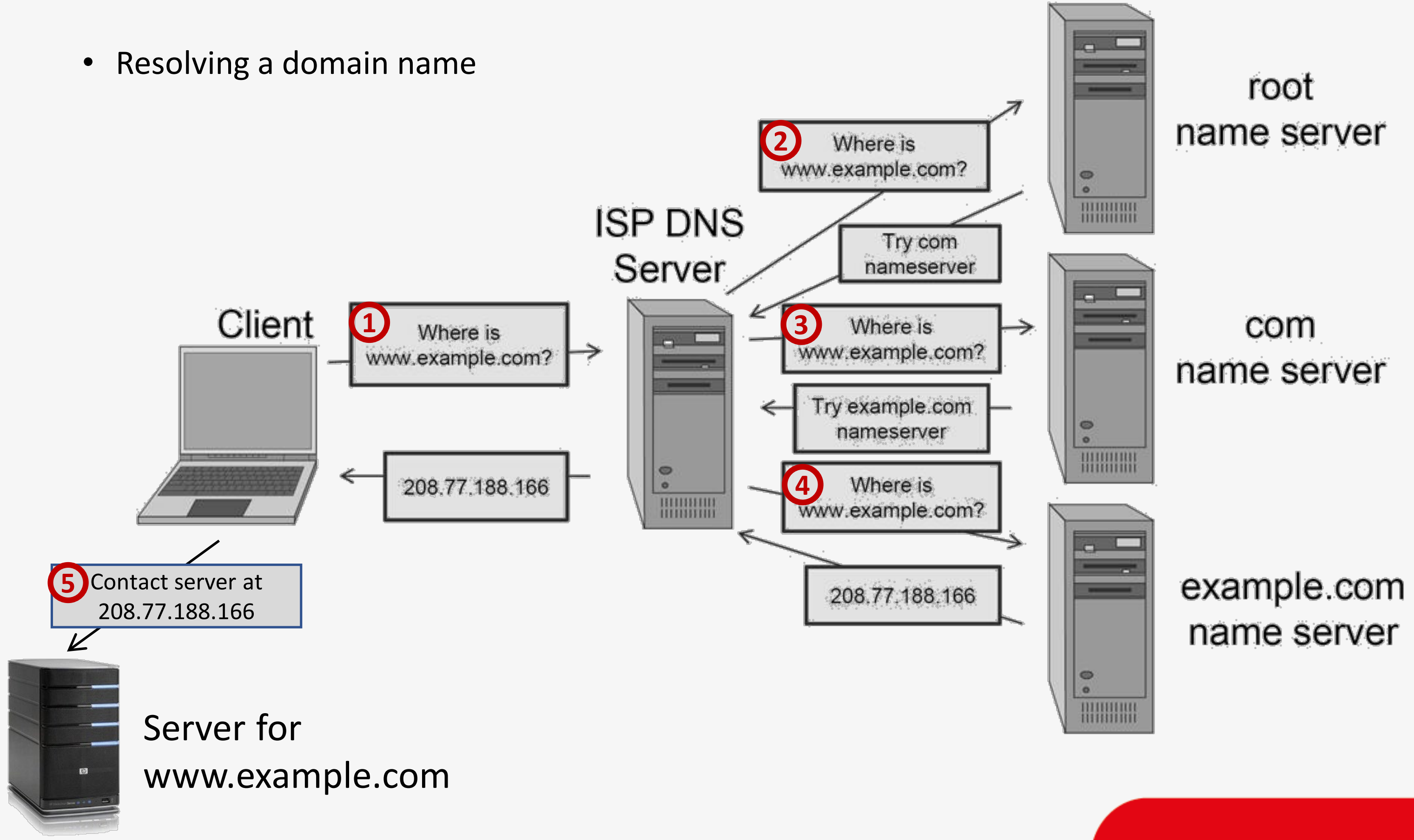
- Bot has to know *domain name* or *IP address* of C&C server
- Reverse engineering of bot binary may reveal domain name or IP address of C&C server

- Bot knows *domain name* of C&C server
  - static: hardcoded in bot binary
  - dynamic: generated using DGA (Domain name Generation Algorithm)
  - requires DNS lookup to resolve domain name into IP address

- Bot knows *IP address* of C&C server
  - static: hardcoded in bot binary
  - dynamic: seeding by providing initial list of peers (P2P botnet)
  - eliminates DNS lookup (stealthy)

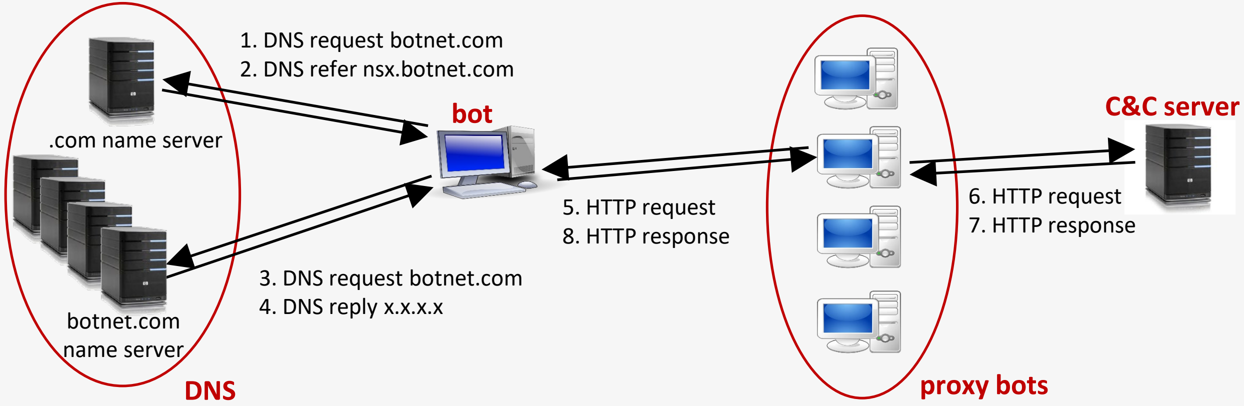
DNS

- Resolving a domain name



Evasion tactics of botnets

- *IP flux*
  - frequently change IP address to evade blacklisting and blocking of IP addresses
  - real-time update of DNS facilitated by Dynamic DNS (DDNS) services
- *Fast flux:* IP addresses refer to proxy bots, that relay communication to C&C server
- *Double flux:* also IP address of name server changes frequently



Evasion tactics of botnets

- *Domain flux*
  - frequently change domain name for contacting C&C server
  - helps evade URL-based detection
  - achieved by
    - domain wildcarding (DNS service)
    - *DGA (domain name generation algorithm)*

DGA

- Bot applies DGA to periodically generate a (large) number of domain names
  - only one/few are registered by botmaster
  - bot uses DNS to resolve domain names one by one
    - unregistered domain names result in Non-Existent Domain (NXDomain) responses from name servers
    - successfully resolved domain name refers to proxy bot or C&C server

- *Re-engineering* DGA by analysis of botnet binary to predict what domain names a bot will try
  - unfeasible to register all those domains by law enforcement or check which ones are malicious
  - prohibited if DGA uses dynamic seed

DGA

- DGA generates large number of pseudo-random domain names from a *seed*
  - seed is shared secret between botmaster and bots
- *Static/deterministic seed*
  - eg. seed derived from current date (Torpig), GMT (Conficker)
  - eg. Conficker.C generated 50,000 domain names of which bots daily tried up to 500
    - law enforcement would have to pre-register and check 50,000 domain names
    - if botmaster registers only 1 domain name, bot has 1% chance per day to contact C&C server, hence bot will contact C&C server once every 100 days on average
- *Dynamic seed*
  - eg. foreign exchange reference rates published daily by European Central Bank (Bedep), trending topics on Twitter (Torpig)
  - domain names cannot be precomputed in advance (small time window, also for botmasters)

DGA types

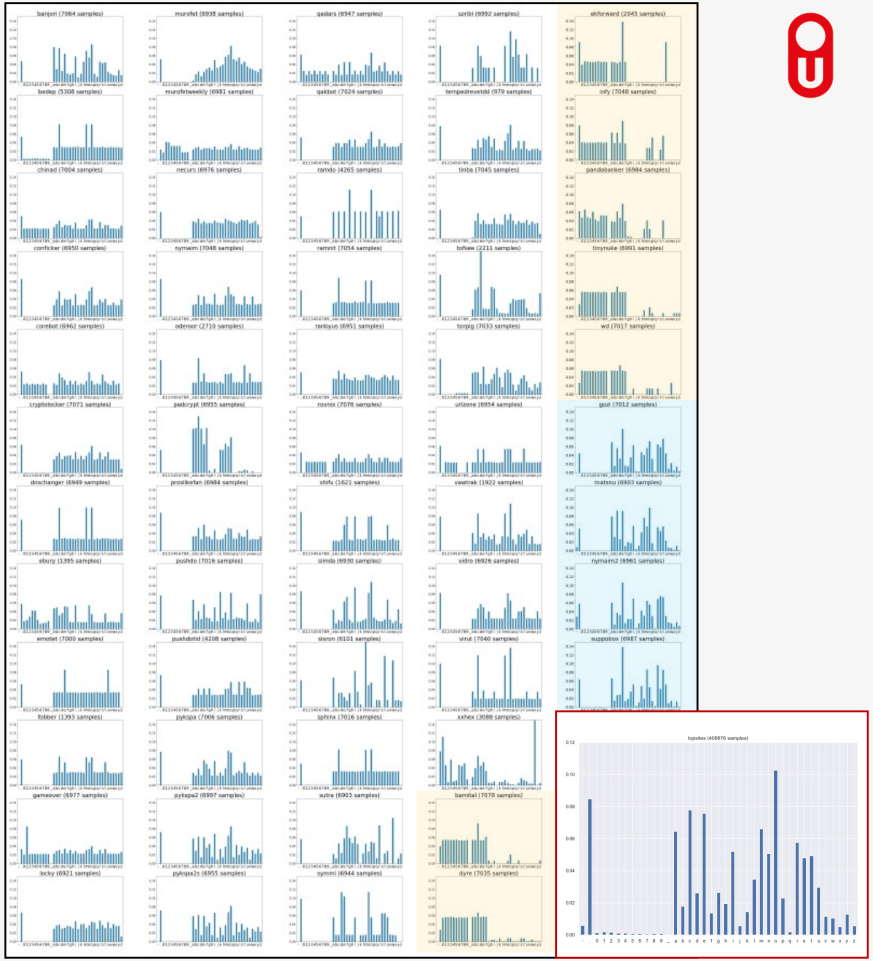
- *Arithmetic-based*: generate random *sequences of ASCII characters* vhljakiutpq7.com
  - domain names contain random letters and digits
- *Hash-based*: apply *hashing algorithms* such as MD5 and SHA256 52efedef74d4.com
  - domain names contain hexadecimal numbers
- *Wordlist-based*: concatenate *sequences of words* from dictionaries formsworkfreeall.com
  - domain names are less random, but contain no digits
- *Permutation-based*: permutate given domain name redotntexplore.com
  - domain names look similar to regular domain names

Plohmann, D.; Yakdan, K.; Klatt, M.; Bader, J.; Gerhards-Padilla, E. A Comprehensive Measurement Study of Domain Generating Malware. 25th USENIX Security Symposium (USENIX Security 16); USENIX Association: Austin, TX, 2016; pp. 263–278.

DGArchive

DGA family	DGA type	Count	Length	Sample 1	Sample 2
banjori	A	10,000	11 - 30	eihspartbulkyf.com	ochqfordlinnetavox.com
bedep	A	7,458	16 - 22	vhljakiutpq7.com	csejdvmqgmqj.com
chinad	A	10,000	19 - 21	3vainry4stex8arf.cn	vfuupsix5ki5omg0.cn
conficker	A	10,000	8 - 16	qzvwnnije.biz	dovcujbpg.biz
corebot	A	10,000	15 - 32	kr105hivgrqvo8e8ijqh1bc.ws	i472uvy6qjyvgh18mhw4k85.ws
cryptolocker	A	10,000	15 - 21	leojfthetfvk.com	thtatcpfomflk.com
dnschanger	A	10,000	14 - 14	xxxfuhkjzu.com	viwnolcsqf.com
ebury	A	2,000	17 - 18	r2g1v3mau7h4k.info	k1i5q3w5r1x4i.net
emotet	A	10,000	19 - 19	iqpucsfnnijdnbii.eu	olahnvuhbiitauve.eu
fobber	A	2,000	14 - 21	phtatognxg.com	vzuopketsrtaqttgk.net
gameover	A	10,000	18 - 37	iz6tx9jwre387brksimxpkcp.net	d2u8ds1aif9oryzft8f1u052m5.org
locky	A	10,000	8 - 23	viuoabuc.fr	rkwaoicjullpc.click
murofet	A	10,000	13 - 21	prkwwoswewwkfzuy.com	udumozptkqqpo.info
murofetweekly	A	10,000	35 - 51	jyi35d10gwgqlrmrhupudxdqoyc69n40d20dq.ru	buiuj26gvhxk57pvmrk17d50bwfzlxa17hrls.ru
necurs	A	10,000	10 - 28	yaatqhjjgicemhoeiu.nf	inlclnelid.ug
nymaim	A	10,000	8 - 16	xhhtaldw.net	uckvk.net
oderoor	A	3,833	10 - 16	uyftputndw.cc	mdnaizofvm.cc
padcrypt	A	10,000	19 - 24	fkaokkbfaalfbdeb.info	menccfmdkcmaemfk.de
proslikefan	A	10,000	9 - 17	zrimegy.in	vnmwww.co
pushdo	A	10,000	11 - 16	katcetutyx.kz	lakeotux.kz
pushdotid	A	6,000	13 - 14	gxmdgfmjcx.com	opgrexsbif.net
pykspa	A	10,000	10 - 17	rldbwwarp.net	myhmexr.net
pykspa2	A	10,000	10 - 19	iugzosiugkeq.net	wkuglwiugkeq.biz
pykspa2s	A	9,957	10 - 19	pkpycifox.com	wudmdgeoya.biz
qadars	A	10,000	16 - 16	ysmoq4esi0q0.org	gt6b8tirkh2r.net
qakbot	A	10,000	12 - 30	xvvluuabuftqilmnynimpipb.info	tugfpmprjspprbwxdzi.biz
ramdo	A	6,000	20 - 20	skuqesksmewsckwg.org	iqgieiyuigamowca.org
ramnit	A	10,000	11 - 25	ixrghbaytyaksgug.com	bwqkmskfwpvljd.com
ranbyus	A	10,000	17 - 21	ndgpkwlmftaryloae.cc	gttfhnegjtmegkhrt.cc
rovnix	A	10,000	21 - 22	jaitc336ybcds71ykg.cn	oar7juqajea1wnyopo.cn
shifu	A	2,331	10 - 12	vhqrdfg.info	xxuissv.info
simda	A	10,000	8 - 14	rynezev.info	qebol.eu
sisron	A	8,800	16 - 17	mjcwmziwmtqa.net	mjmwotiwmtga.net
sphinx	A	10,000	20 - 20	libuybegcrlrfyof.com	oixwkitoiqseltry.com
sutra	A	9,882	19 - 29	gweqifjejtoaemgw.info	hpwazeehjwpfwgaj.ru
symmi	A	10,000	17 - 24	oqmievkeedloovm.ddns.net	esitkoelmei.ddns.net
szribi	A	10,000	12 - 12	ddpuuddd.com	grawspwe.com
tempedrevetdd	A	1,380	12 - 14	gbuxwrwx.org	crwhchuda.org
tinba	A	10,000	10 - 23	bcjwxxumttmh.net	rwtopxoocwtt.cc
tofsee	A	3,140	10 - 11	drndrng.biz	drodroi.biz
torpig	A	10,000	11 - 13	bfcmulj.net	bhksvgrpa.com
urlzone	A	10,000	8 - 19	ehw5jdkwkv.com	rc5iycl4suf.com
vawtrak	A	2,700	10 - 15	dmzqvyn.top	misohnatl.com
vidro	A	10,000	11 - 23	prjbemepgzkp.com	rakrfxs.com
virut	A	10,000	10 - 10	yzraho.com	ehuquf.com
xxhex	A	4,400	12 - 13	xxa5c1b019.sg	xx3603da38.sg
bamital	H	10,000	36 - 38	43f3d094f08dd1a2df2869352e2a9712.cz.cc	f0b79a9253cf7c58f0e1f54426f45bf4.cz.cc
dyre	H	10,000	37 - 37	rdf36ed41339f9abd57a5a1c9f2143f513.ws	u28c43d53bb3ecafbdfd29fa34a47dae09.to
ekforward	H	2,919	8 - 11	80a118c7.eu	9356c774.eu
infy	H	10,000	12 - 14	1e60c5f5.space	a56bc6c6.top
pandabanker	H	10,000	16 - 17	52efedef74d4.com	0b16dca48547.com
tinynuke	H	10,000	36 - 36	ec893776679264b90cfff916cc5f0eaf.com	84b4a55d8ac046a9816dda8b866893b7.top
wd	H	10,000	36 - 38	wd679ab775d15bbee733b8545f20452504.win	a0e433f4c96c6b8f3ece607d791d6546.pro
gozi	W	10,000	15 - 29	formsworkfreeall.com	allowdisalloallow.me
matsnu	W	10,000	16 - 28	bitpersuadebutton.com	structuresurvey.com
nymaim2	W	10,000	11 - 33	sculpturenegative.net	shuttlefatty.it
suppobox	W	10,000	11 - 30	senseinto.ru	threeslept.net

Character distribution



Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

Prior work on detection with ML/DL

- Detecting DGA-generated domain names with *machine learning*
  - context-free features from domain name: length, entropy, ratios (letters, digits, vowels), pronounceability

Reference	Year	Model	Dataset (Benign/Malicious)	Number of Features	
				Context-Free	Context-Aware
Chiba et al. [14]	2018	RF	Alexa/hpHosts	-	55
Schüppen et al. [15]	2018	**RF**, SVM	Private/DGArchive (72 DGAs)	21	-
Ashiq et al. [16]	2019	FFNN (2-4 hidden layers)	From [17]	8	-
He et al. [18]	2019	Adaboost, DT, kNN, **RF**	Alexa/various sources	21	153
Li et al. [19]	2019	Adaboost, C4.5, **kNN**, NB	.cn name server/Rustock DGA	1	31
Liu et al. [20]	2019	SVM	Alexa/DGArchive (87 DGAs)	-	18
Selvi et al. [21]	2019	RF	Alexa/26 DGAs	18	-
Yang et al. [22]	2019	DT, ET, NB, SVM, **ensemble (NB,ET,LR)**	Cisco Umbrella/Netlab, synthetic	24	-
Akhila et al. [23]	2020	DT, **GBT**, LR, RF, SVM	Alexa/Bambenek	10	-
Alaeiyan et al. [24]	2020	**RF**, RNN, SVM	Alexa/MasterDGA	18	-
Almashhadani et al. [25]	2020	BT, DT, **kNN**, NB, SVM	Alexa/DGArchive (20 DGAs)	16	-
Anand et al. [26]	2020	**C5.0**, CART, GBM, kNN, RF, SVM	Alexa/Netlab (19 DGAs)	45	-
Hwang et al. [27]	2020	LightGBM	KISA/KISA (20 DGAs)	110	-
Liang et al. [28]	2020	**RF**, SVM, XGBoost	Alexa/various blacklists	5	5
Mao et al. [29]	2020	NB, LSTM, **MLP**, RF, SVM, XGBoost	Alexa/Netlab (40 DGAs)	5	-
Palaniappan et al. [30]	2020	LR	Alexa/various blacklists	4	13
Sivaguru et al. [31]	2020	RF	Alexa, private/DGArchive	26	9
Wu et al. [32]	2020	**MLP**, NB	Alexa/Netlab	4	-
Zhang et al. [33]	2020	DT, LR, NB, RF, SVM, **XGBoost**, Voting	Alexa/UMUDGA (37 DGAs)	18	-
Zago et al. [13]	2020	Adaboost, DT, kNN, **NN**, RF, SVM	Majestic/various sources (16 DGAs)	40	-
Cucchiarelli et al. [34]	2021	**MLP**, RF, SVM	Alexa/Netlab (25 DGAs)	$4n + 5$ (n DGAs)	-
Patsakis et al. [35]	2021	RF	Alexa, unipi/DGArchive, synthetic (13 DGAs)	32	-

Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

Prior work on detection with ML/DL

- Detecting DGA-generated domain names with *deep learning*
  - word embedding of domain names

Reference	Year	Model	Dataset (Benign/Malicious)
Woodbridge et al. [36]	2016	LSTM	Alexa/Bambenek
Lison and Mavroeidis [37]	2017	RNN	Alexa/DGArchive (63 DGAs), Bambenek (11 DGAs)
Koh and Rhodes [38]	2018	LSTM	OpenDNS/Bader, Abakumov
Tran et al. [39]	2018	LSTM.MI	Alexa/Bambenek (37 DGAs)
Vinayakumar et al. [40]	2018	**LSTM**, GRU, IRNN, RNN, CNN, **hybrid (CNN-LSTM)**	Alexa, OpenDNS/Bambenek, Bader (17 DGAs)
Xu et al. [41]	2018	CNN-based	Alexa/DGArchive (16 DGAs)
Yu et al. [42]	2018	LSTM, BiLSTM, stacked CNN, parallel CNN, hybrid (CNN-LSTM)	Alexa/Bambenek
Akarsh et al. [43]	2019	LSTM	OpenDNS, Alexa/20 public DGAs
Qiao et al. [44]	2019	LSTM	Alexa/Bambenek
Liu et al. [45]	2020	Hybrid (BiLSTM-CNN)	Alexa/Netlab (50 DGAs), Bambenek (30 DGAs)
Ren et al. [46]	2020	CNN, LSTM, CNN-BiLSTM, **ATT-CNN-BiLSTM**, SVM	Alexa/Bambenek, Netlab (19 DGAs)
Sivaguru et al. [31]	2020	hybrid (RF-LSTM.MI)	Alexa, private/DGArchive
Vij et al. [47]	2020	LSTM	Alexa/11 DGAs
Cucchiarelli et al. [34]	2021	BiLSTM, LSTM.MI, hybrid (CNN-BiLSTM)	Alexa/Netlab (25 DGAs)
Highnam et al. [48]	2021	hybrid (CNN-LSTM-ANN)	Alexa/DGArchive (3 DGAs)
Namgung et al. [49]	2021	CNN, LSTM, BiLSTM, **hybrid (CNN-BiLSTM)**	Alexa/Bambenek
Yilmaz et al. [50]	2021	LSTM	Majestic/DGArchive (68 DGAs)

Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

DGA detection with TF-IDF as features

- *TF-IDF*
  - originates from information retrieval and automated text analysis
  - composed of multiplying *term frequency* (TF) and *inverse document frequency* (IDF)

- Set of *terms* $T = \{t_1, \ldots, t_K\}$ in set of *documents* $D = \{d_1, \ldots, d_N\}$
- TF_{t_i, d_j} indicates how often term t_i occurs in document d_j
  - usually normalized by document length or most frequent term count in document
  - TF is larger if term occurs more often
- IDF_{t_i} indicates the number of documents (n_i) in set D that contain term t_i
  - usually defined as $\log(N/n_i)$
  - IDF is larger if term occurs in fewer documents
- TF-IDF discriminates *key terms* that appear often but in a smaller number of documents

TF-IDF example

D = { "the house had a tiny little mouse",
"the cat saw the mouse",
"the mouse ran away from the house",
"the cat finally ate the mouse",
"the end of the mouse story“
}

T = {‘mouse’, ‘the’, ‘cat’, ‘house’, ‘had’, ‘tiny’, ‘little’, ‘saw’, ‘ran’, ‘away, ’from’, ‘finally’, ‘ate’, ‘end’, ‘of’, ‘story’}

IDF = {1.000, 1.000, 1.693, 1.693, 2.099, 2.099, 2.099, 2.099, 2.099, 2.099, 2.099, 2.099, 2.099, 2.099, 2.099}

TF-IDF = { 0.235, 0.235, 0, 0.398, 0.494, 0.494, 0.494, 0, 0, 0, 0, 0, 0, 0, 0,
…
…
…
…
}

DGA detection with TF-IDF

- Hassan's idea
  - apply TF-IDF as measure for how relevant *n-grams* are in *domain names*
  - use TF-IDF scores as features in ML

vhljakiutpq7.com
csejdvpqgmqj.com

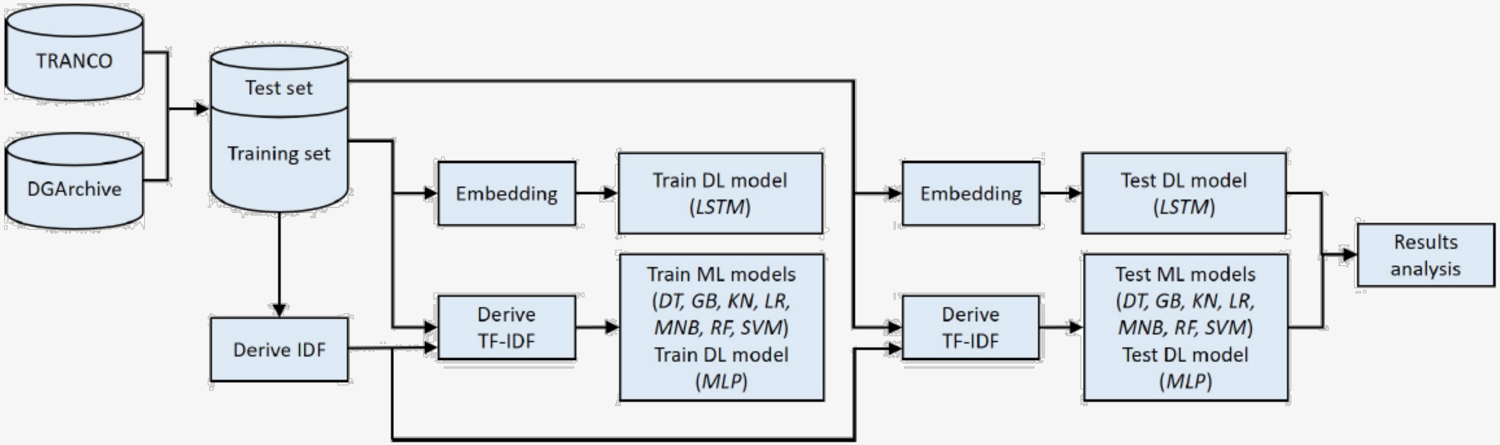
- Created *dataset* with 1,076,754 domain names
  - 583,954 benign domain names; 492,800 malicious domain names from 57 DGA families
  - 70% in training dataset, 30% test dataset

- Determined *top 5,000 of n-grams* (for n=1,2,3) that occur most often in training dataset, and derive IDF
- Transform dataset from set of domain names into a set of vectors with dimension 5,000
  - each vector represents TF-IDF of top 5,000 n-grams in domain name

Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

Research questions and method

- How accurate can *ML/DL models* classify DGA-generated domain names when using *TF-IDF as features*?
  - Considered 7 ML models (DT, GB, KN, LR, MNB, RF, SVM) and 1 DL model (MLP) that give best results as reported in related literature
  - All models are multi-class classifiers with 58 outputs (57 DGA families and non-DGA)
- How good is accuracy when compared to state-of-the-art *DL model* (LSTM) with *word embedding*?



Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

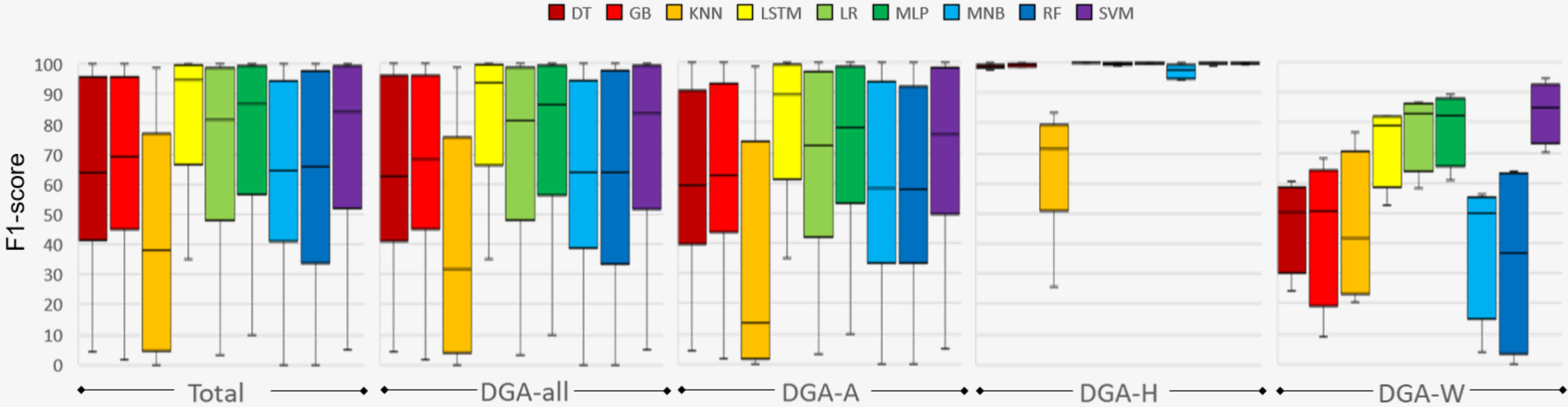
Metrics

- Classification results
  - true positive (TP): *correct* classification of *DGA* domain name
  - false positive (FP): *incorrect* classification of *non-DGA* domain name
  - true negative (TN): *correct* classification of *non-DGA* domain name
  - false negative (FN): *incorrect* classification of *DGA* domain name

- *Precision* (fraction of all positive classifications that are classified correctly): TP / (TP + FP))
- *Recall* (fraction of all DGA domain names that are classified correctly): TP / (TP + FN)
- *F1-score* (harmonic mean of precision and recall): 2 / (precision^{-1} + recall^{-1})

Experimental results

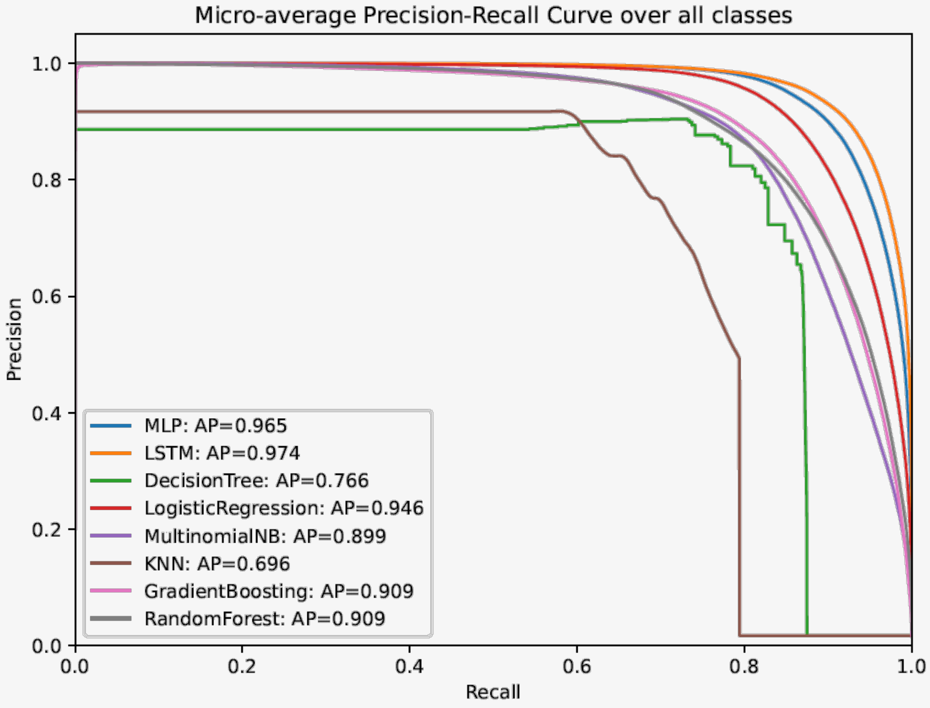
- Best results overall are obtained with *LSTM (90.69% weighted average F1-score)*, closely followed by *MLP (89.08%)* and *SVM (88.08%)*
  - for DGA-W families and non-DGA, best results with MLP, SVM, and LR
  - DGA-H families are very easy to detect; DGA-W families are more difficult to detect
- Models with *highest average F1-score* also have *smallest standard deviation/spread* in F1-score



Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

Experimental results

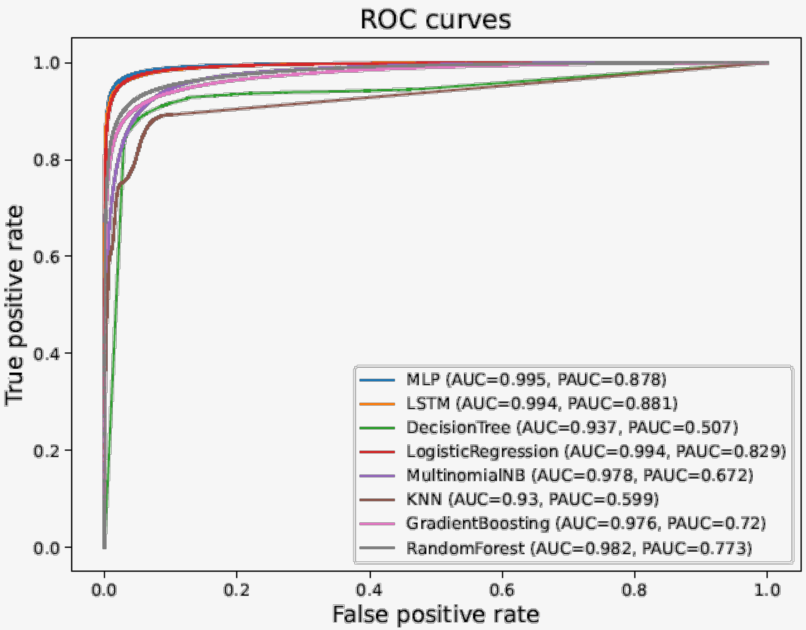
- Precision-recall curves for weighted-average of all classes: *LSTM* performs best, closely followed by *MLP*

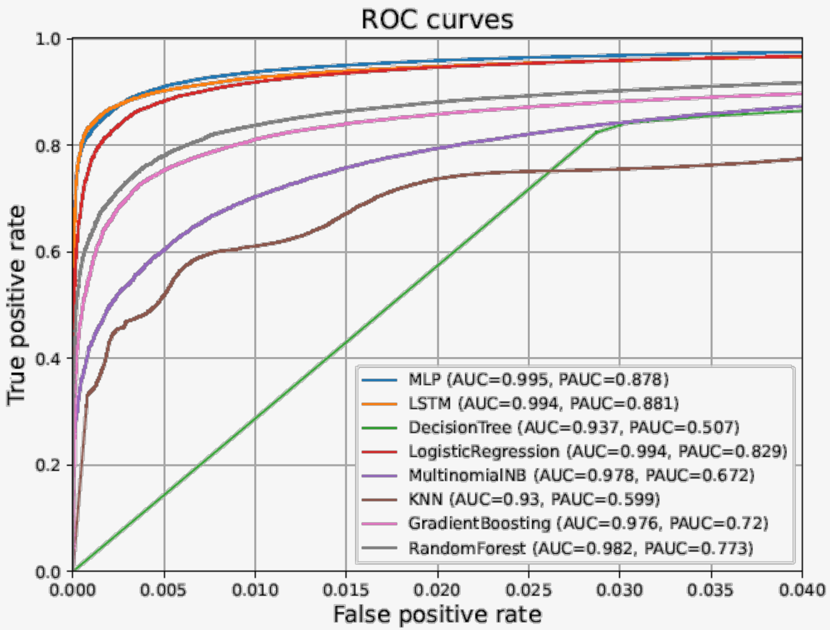


Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

Experimental results

- ROC-curves for binary classification (DGA vs. non-DGA): *MLP* performs best, closely followed by *LSTM*





Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF, MDPI Electronics 2022, 11, 414

Conclusions

- DL models (LSTM, MLP) clearly yielded *better results* than ML models in multi-class classification
- Results for LSTM with standard embedding are *comparable* with results for MLP with TF-IDF features (F1: 0.907-0.891; AU-PR-C: 0.974-0.965; AU-ROC: 0.994-0.995; TPR: 0.957-0.965; FPR: 0.027-0.025)
- Results *differ per DGA type*
  - DGA-H domain names are easy to classify (up to 99.96% F1-score with LSTM)
  - DGA-W domain names are more difficult to classify (best F1-score of 83.61% with SVM)
- *Not straightforward to compare* our results with prior work
  - Different datasets of benign and malicious domain names,
    from different time periods, and different numbers and types of DGA families
  - Mix of DGA families included in the dataset has large impact
- Observed in prior work: many different (and combinations) of features for ML models are used
  - Large variety, unknown which features are more relevant

Vranken, H. and Alizadeh, H., Detection of DGA-Generated Domain Names with TF-IDF,
MDPI Electronics 2022, 11, 414

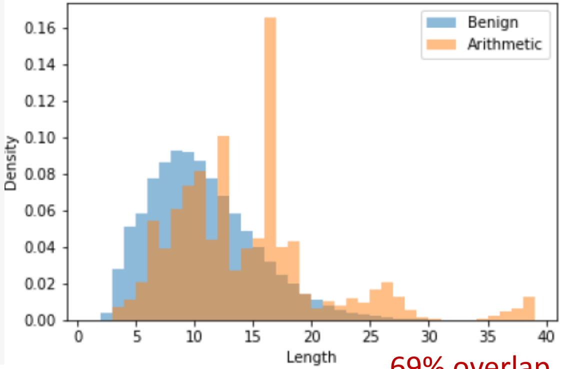
Effectiveness of features

- Research question: What features from domain names are more effective in ML classifiers for DGA detection?

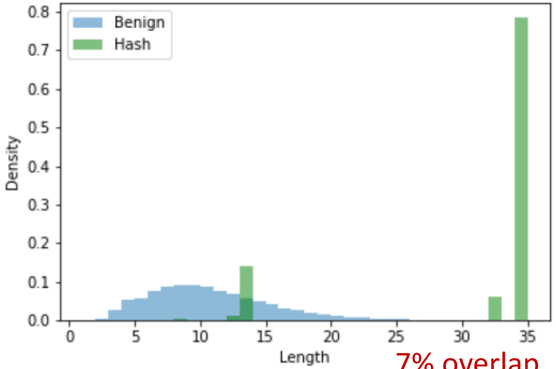
- Research method
  - Considered 80 recent papers, from which 69 features were derived
  - Datasets: retrieved second-level domain name (AAA.BBB.CCC)
    - Benign from TRANCO: 999,913
    - DGA-generated domain names from DGArchive: 2,922,654 DGA-A; 2,616,128 DGA-H; 336,667 DGA-W
    - Computed feature values, frequency distributions and overlap for benign vs. DGA-A/DGA-H/DGA-W

Vranken, H. and Kuipers, L. (to be published)

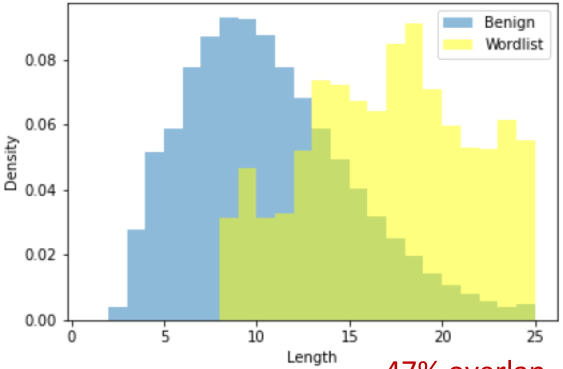
Experimental results



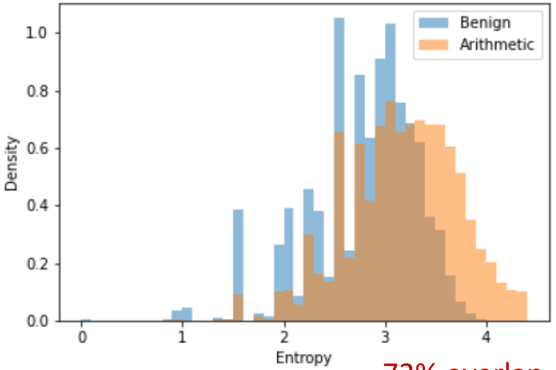
69% overlap



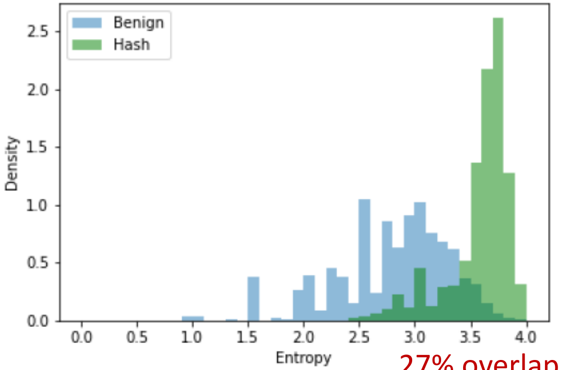
7% overlap



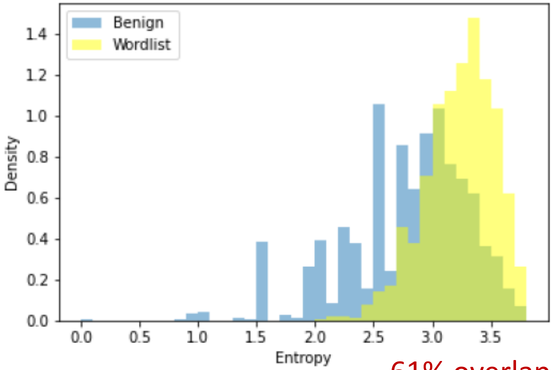
47% overlap



72% overlap

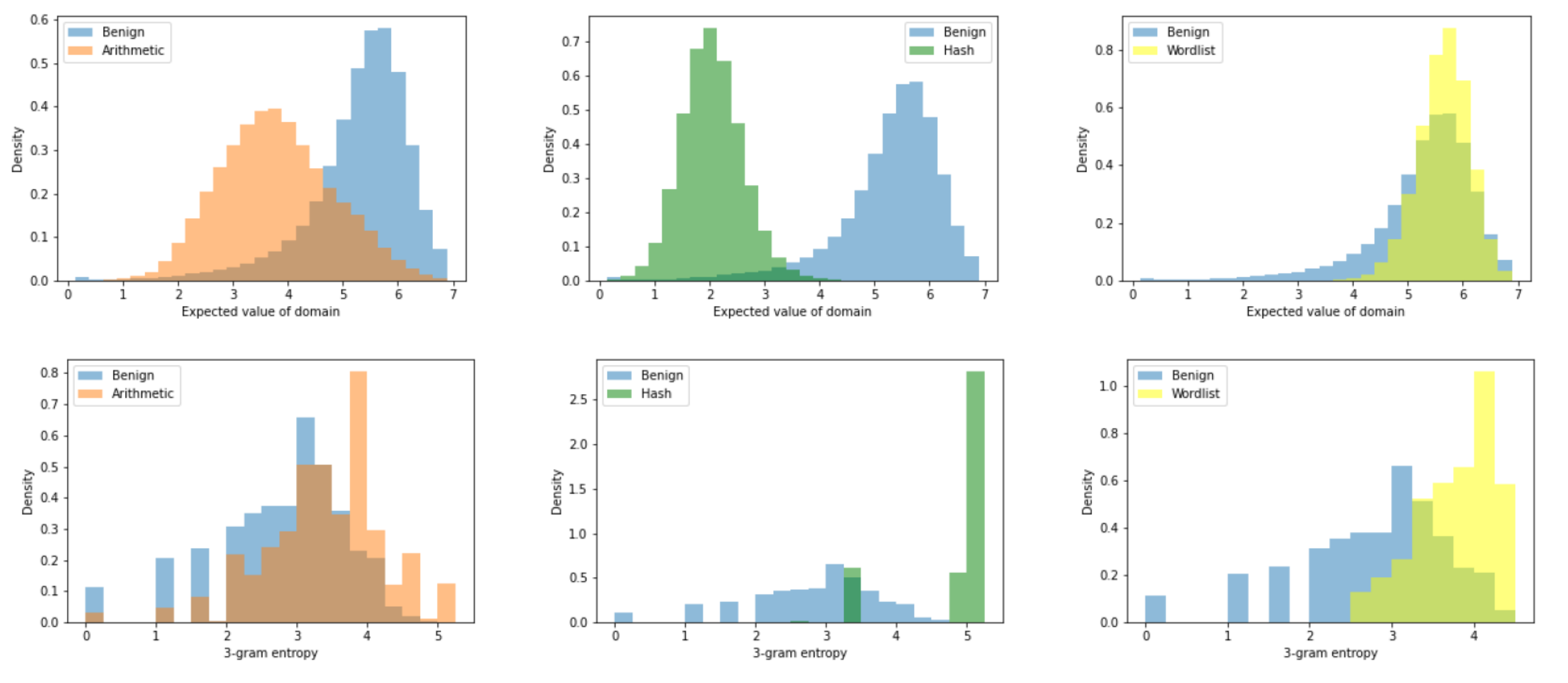


27% overlap

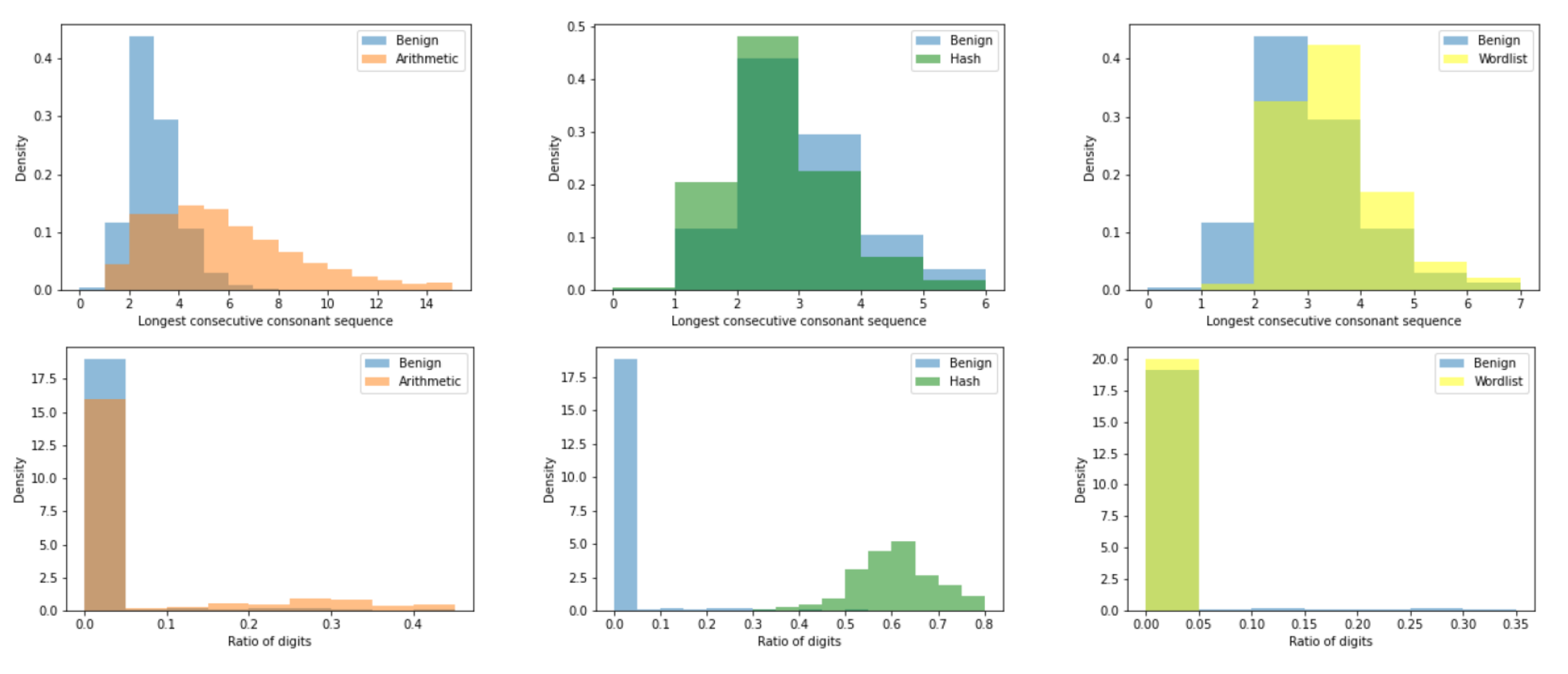


61% overlap

Experimental results



Experimental results



Experimental results

- Overview of effectiveness of features

Feature	Arithmetic		Hash		Wordlist	
length	(X)*	69.28%	X	7.28%	(X)*	46.81%
subdomain length mean			X	19.86%		
entropy	(X)	72.29%	X	26.6%		
#consonants	(X)	63.9%				
#digits	(X)	85.91%	X	0.67%		
unique chars	(X)	72.7%	X	22.34%		
#words over (2)-3 chars	(X)	39.32%	(X)	31.98%		
#num sequences	(X)	82.6%	X	0.61%		
longest consonant sequence	(X)	45.52%				
longest digit sequence			X	3.58%		
longest hex sequence			X	0.04%		
longest prime sequence			X	4.03%		
longest vowelless sequence	(X)	42.58%	X	5.87%		
longest meaningful substring	(X)	39.42%	(X)	29.67%		
digit ratio			X	1.71%		
letter ratio			X	1.9%		
hex ratio			X	0.52%		
prime digit ratio	(X)	86.8%	X	3.89%		
vowel ratio	(X)	48.54%	X	15.82%		
consonant ratio	(X)	61.12%	X	7.89%		
ratio unique chars			X	17.68%	(X)	59.63%
ratio meaningful chars	X	33.42%	X	11.68%		
ratio max seq vowels			X	28.78%		
ratio max seq consonants			X	17.65%		
ratio consecutive digits			X	3.26%		
ratio consecutive consonants	(X)	60.61%	X	28.79%		
ratio repeated characters			X	24.53%		
consonant to vowel ratio	(X)	53.26%				
digit to letter ratio			X	1.46%		
ratio max seq consonants to max seq vowels	(X)	57.85%				
ratio LMS	X	31.76%	X	12.07%		
ratio hex exclusive sub			(X)	36.09%		
ratio entropy			X	15.7%	(X)	49.18%
meaningful length ratio			X	1.51%		
top used letters ratio	X	41.66%	X	7.93%		
least used letters ratio	(X)	44.13%				
four gram score	(X)	42.64%	X	9.57%		
conversion frequency	(X)	84.4%	X	2.99%		
gini index			(X)	34.73%		
classification error			(X)	41.63%		
expected value	X	38.09%	(X)	5.93%		
contains digits			X	5.95%		
first character digit			(X)	88.15%		
is hexadecimal			(X)	60.94%		
2-gram entropy			X	15.9%	(X)	48.87%
3-gram entropy			X	13.23%	(X)	48.79%
1-gram mean of freqs			X	14.83%	(X)	59.45%
2-gram mean of freqs			(X)	29.87%		
3-gram mean of freqs			(X)	92.19%		
1-gram max of freqs			X	22.54%	(X)	57.44%
2-gram max of freqs			(X)	40.47%		
1-gram median of freqs			X	23.59%		
1-gram 25^{th} percentile			(X)	69.23%		
1-gram 75^{th} percentile			X	21.78%	(X)	61.06%
1-gram variance			X	23.11%		
2-gram variance			X	33.01%		
3-gram variane			(X)	92.21%		
1-gram st. deviation			X	24.32%		
2-gram st. deviation			X	39.84%		
3-gram st. deviation			(X)	92.04%		
3-gram circle median	Benign domains stand out from rest in some cases					

* (X): the feature is useful in some specific cases for that DGA type