D-Fence:
A Flexible, Efficient, and Comprehensive Phishing Email Detection System

Jehyun Lee[†], Farren Tang[†], Pingxiao Ye[†], Fahim Abbasi[†], Phil Hay[†], Dinil Mon Divakaran[† *]

[†]Trustwave, [*]NUS

Jehyun.Lee@trustwave.com

Trustwave® NUS National University of Singapore

Abstract

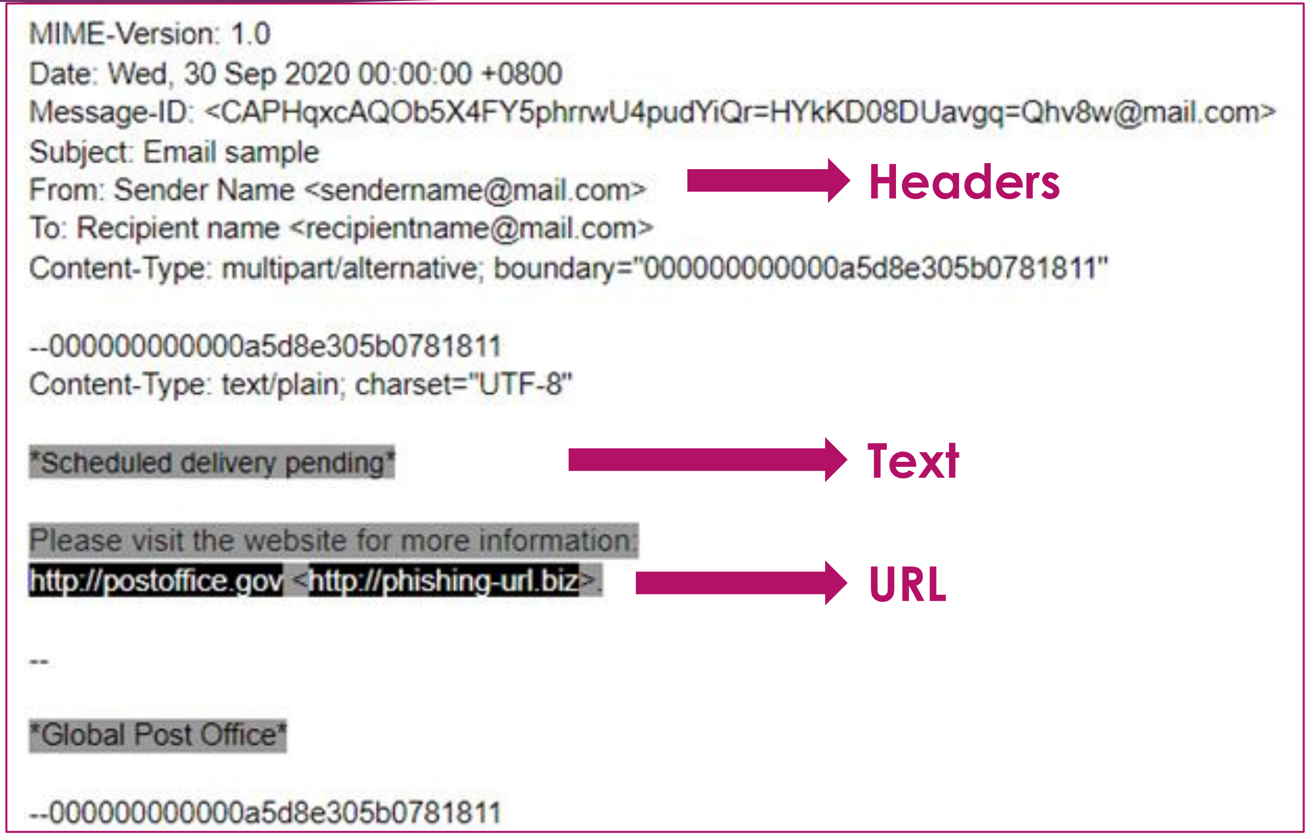
- Phishing Email: Major Security Concern for Organizations
- Previous works
  - Focusing on specific email component: Evadable by changing attack vector
  - Limited single model performance: Limitation of ML models in nature
- **Proposal: Multi-modular phishing email detection system with sophisticated analysis models**
  - **Structure module**: Email headers and HTML structures capturing statistical characteristics.
  - **Text module**: Text classification with pre-trained text vectorization model (BERT)
  - **URL module**: Deep-learning-based URL string modelling and classification
- **0.99+ detection sensitivity (Recall) at a low false-positive rate (1 in 10K)**
  - Evaluated with 68K of recent phishing email samples and 224K of benign samples

Motivation

Shortcomings in Targeting single email component

- ❑ **Email Header Analysis**
  - ❑ [+] Useful in detecting (large-scale) spamming of phishing emails
  - ❑ [-] Easy to evade in spear phishing

- ❑ **Readable text Analysis**
  - ❑ [+] Useful in Message-centric phishing
  - ❑ [-] Evadable by Image-based emails
  - ❑ [-] Bad at short / neutral texts



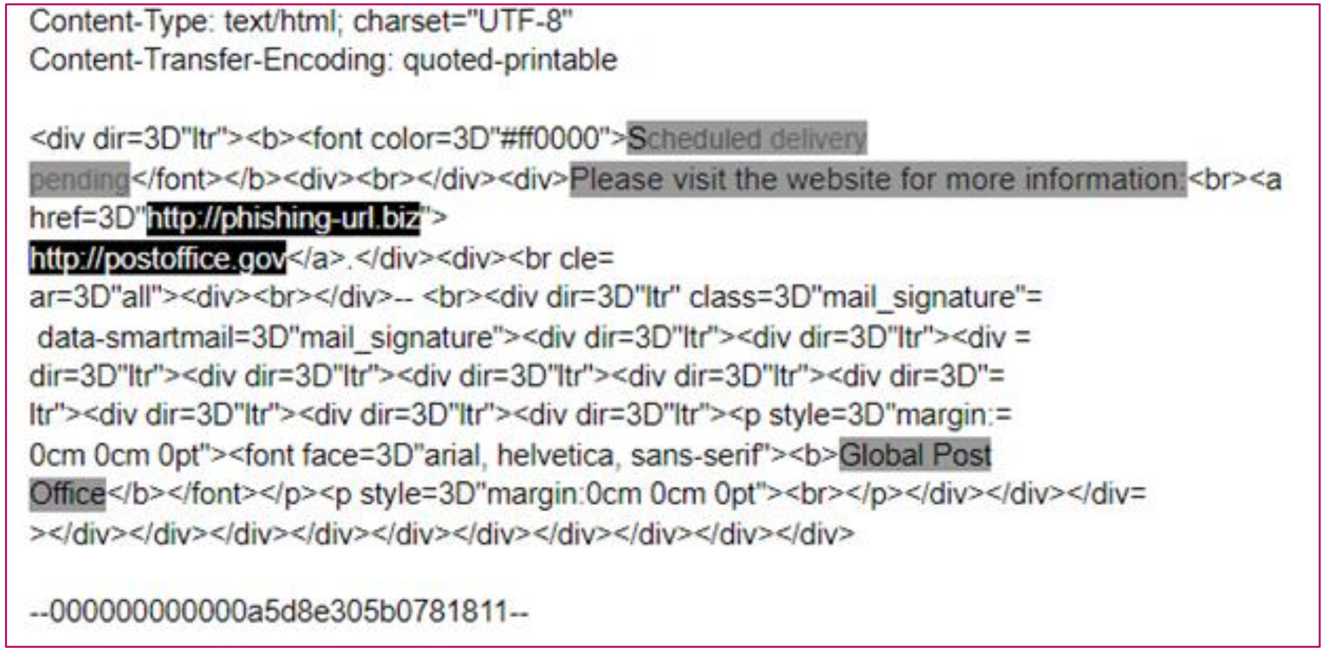
```
MIME-Version: 1.0
Date: Wed, 30 Sep 2020 00:00:00 +0800
Message-ID: <CAPHqxcAQOb5X4FY5phrrwU4pudYiQr=HYkKD08DUavgq=Qhv8w@mail.com>
Subject: Email sample
From: Sender Name <sendername@mail.com>
To: Recipient name <recipientname@mail.com>
Content-Type: multipart/alternative; boundary="000000000000a5d8e305b0781811"

--000000000000a5d8e305b0781811
Content-Type: text/plain; charset="UTF-8"

*Scheduled delivery pending*

Please visit the website for more information:
http://postoffice.gov <http://phishing-url.biz>.

--

*Global Post Office*

--000000000000a5d8e305b0781811
```

Email sample with various Email Components (Header and Plain text Section)

Motivation (cont')

Shortcomings in Targeting single email component

- ❑ **HTML structure Analysis**
  - ❑ [+] Source of phishing techniques
    - ❑ e.g., Scripts, Hidden hyperlinks
  - ❑ [-] Do not cover Message-centric phishing
- ❑ **Embedded URL Analysis**
  - ❑ [+] Wide phishing coverage
    - ❑ Most of the phishing email has a URL
  - ❑ [-] Short living contents



```
Content-Type: text/html; charset="UTF-8"
Content-Transfer-Encoding: quoted-printable

<div dir=3D"ltr"><b><font color=3D"#ff0000">Scheduled delivery
pending</font></b><div><br></div><div>Please visit the website for more information:<br><a
href=3D"http://phishing-url.biz">
http://postoffice.gov</a>.</div><div><br cle=
ar=3D"all"><div><br></div>-- <br><div dir=3D"ltr" class=3D"mail_signature"=
 data-smartmail=3D"mail_signature"><div dir=3D"ltr"><div dir=3D"ltr"><div =
dir=3D"ltr"><div dir=3D"ltr"><div dir=3D"ltr"><div dir=3D"ltr"><div dir=3D"=
ltr"><div dir=3D"ltr"><div dir=3D"ltr"><div dir=3D"ltr"><p style=3D"margin:=
0cm 0cm 0pt"><font face=3D"arial, helvetica, sans-serif"><b>Global Post
Office</b></font></p><p style=3D"margin:0cm 0cm 0pt"><br></p></div></div></div=
></div></div></div></div></div></div></div></div></div>

--000000000000a5d8e305b0781811--
```

Email sample with Various Email Component (HTML Section)

D-Fence: Overview

Three Independent Analysis Modules

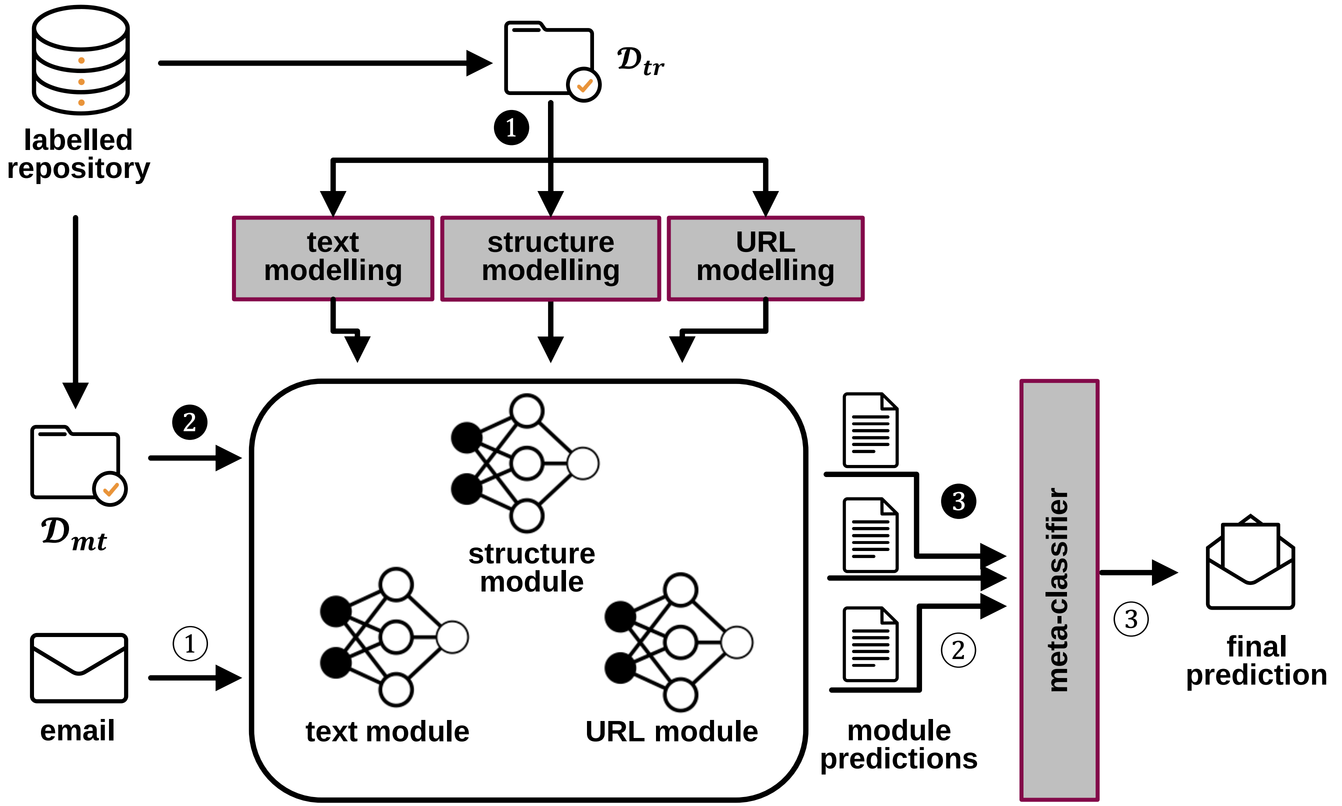
- Wide component coverage
- Extensible

No External Information Sources

- Stand-alone solution
- No up-to-date repository required
- No external communications

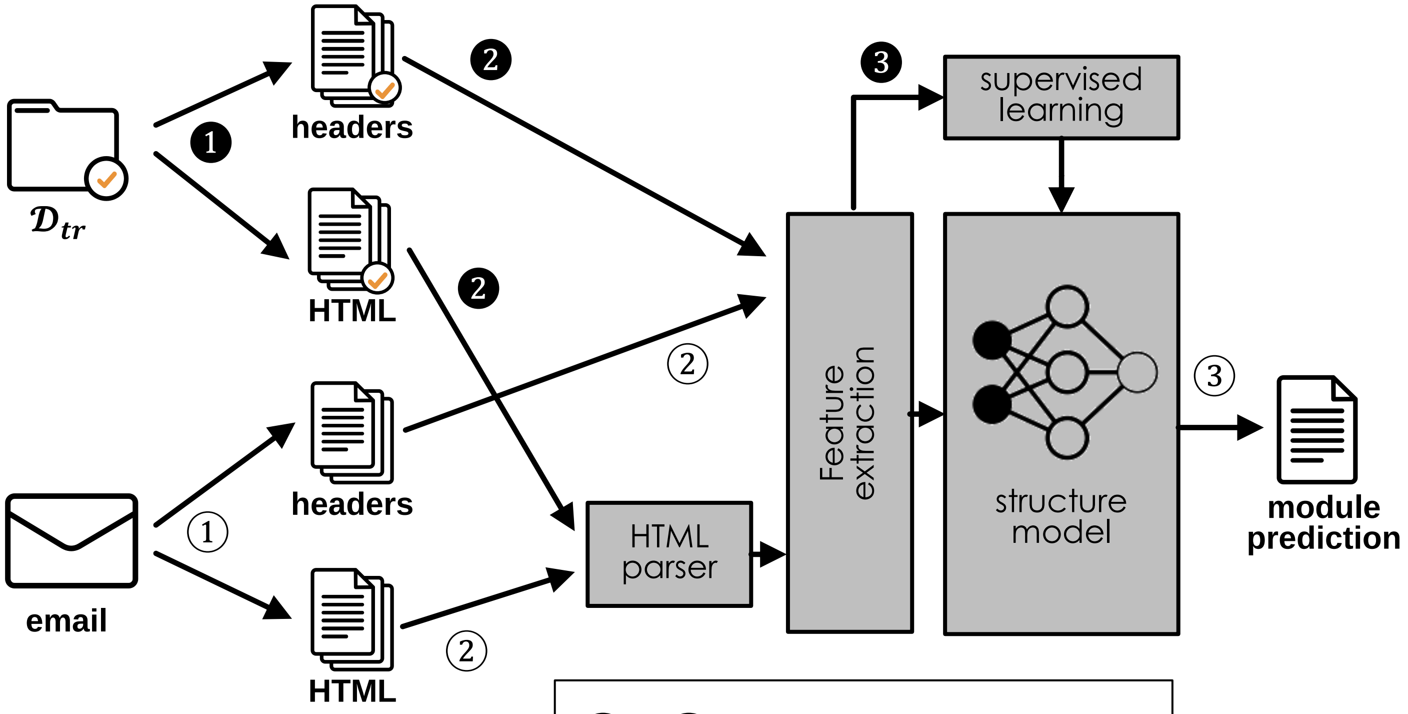
Flexible model configuration / Update

- e.g., Feature modification, model update, module addition., etc.



D-Fence: Structure Module (1/4)

- **Analysis Component**
  - Email Header and HTML section
- **Feature set**
  - 63 Structural features
  - 10 Feature categories
- **Classification**
  - Probability prediction with a supervised learning model

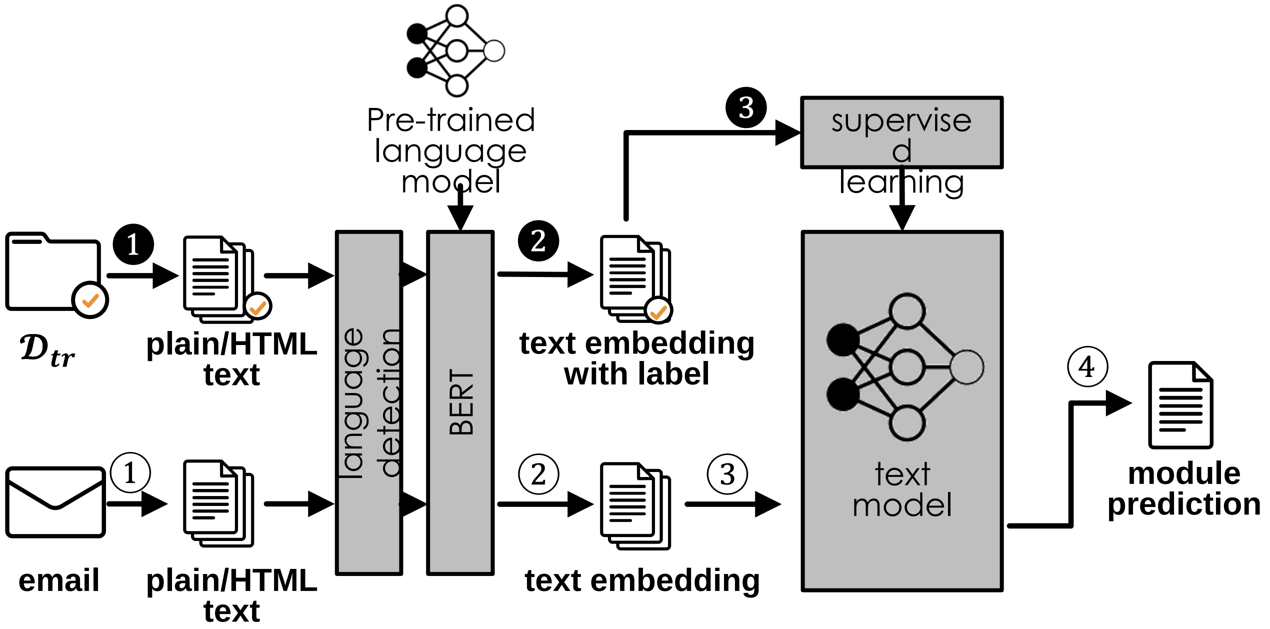


❶~❸: Training process

①~③: Prediction process

D-Fence: Text Module (2/4)

- **Analysis Component**
  - Texts from ***text/plain*** and ***text/html*** sections
- **Text Vectorization**
  - Sentences to numeric vectors
  - **BERT: B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- **Classification**
  - Probability prediction with a supervised learning model

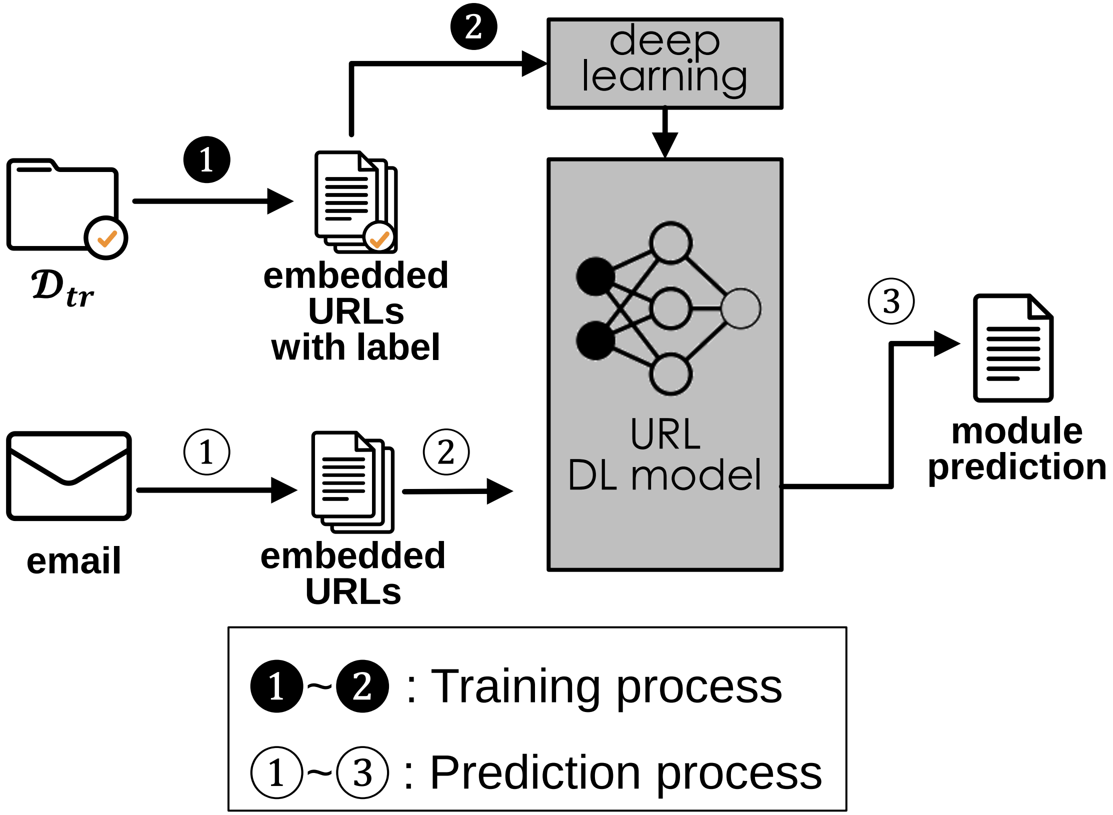


❶~❸ : Training process

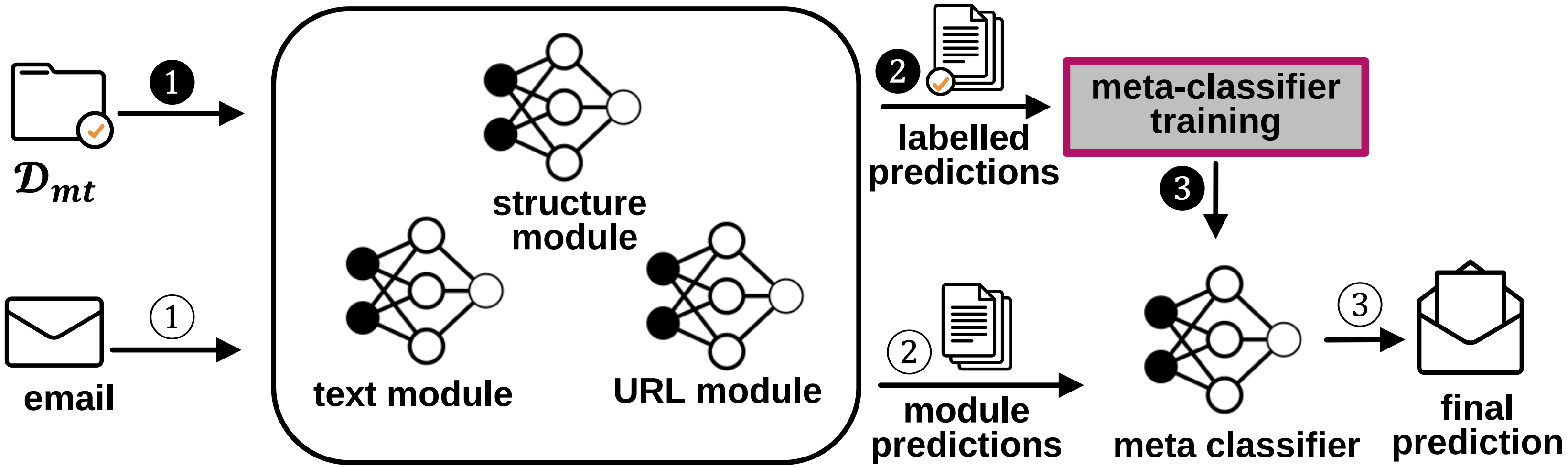
①~④ : Prediction process

D-Fence: URL Module (3/4)

- ❑ **Analysis Component**
  - ❑ URL strings in ***text/plain*** and ***text/html*** sections
- ❑ **Feature set**
  - ❑ Encoded characters in a URL string
- ❑ **Modelling and Classification**
  - ❑ CNN-LSTM
  - ❑ Multiple URLs in an email: multiple predictions
  - ❑ Classification of an email:
    Maximum prediction of all embedded URLs



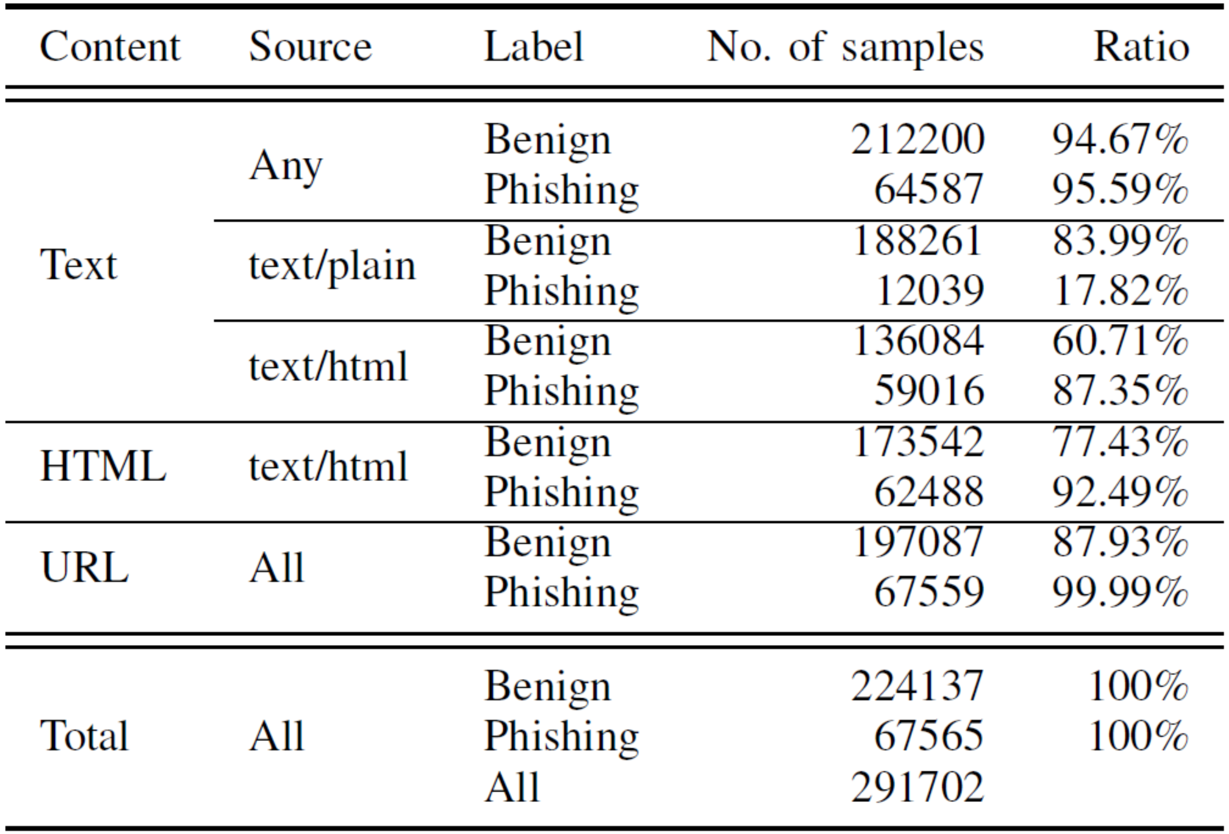
D-Fence: Meta-classifier (4/4)



- Learning prediction confidence and correlation of the individual module’s prediction
- **Training**: Prediction values from individual modules for Meta-classifier training set $\mathcal{D}_{mt}$
- **Prediction**: Three module prediction values into one final prediction value

Evaluation: Enterprise Email Dataset (EES 2020)

- **Email samples from enterprises**
  - Benign emails reviewed by users as Benign
  - Phishing emails detected by multiple solutions
  - Collected in 2018 ~ 2020
- **292K unique samples**
  - Benign: 224K, Phishing: 68K

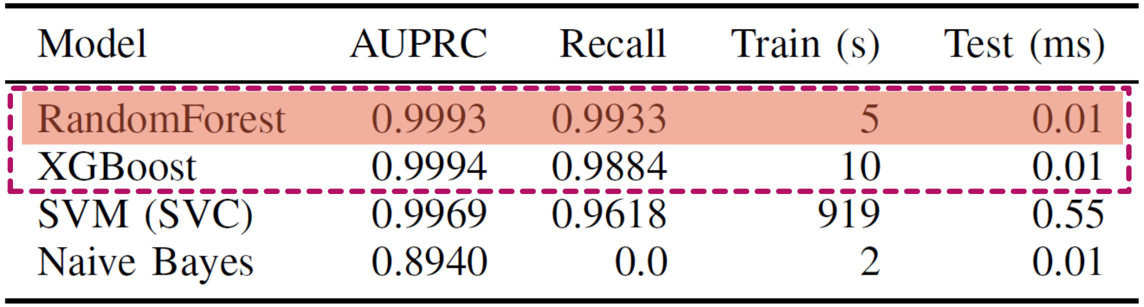


Content	Source	Label	No. of samples	Ratio
Text	Any	Benign	212200	94.67%
		Phishing	64587	95.59%
	text/plain	Benign	188261	83.99%
		Phishing	12039	17.82%
	text/html	Benign	136084	60.71%
		Phishing	59016	87.35%
HTML	text/html	Benign	173542	77.43%
		Phishing	62488	92.49%
URL	All	Benign	197087	87.93%
		Phishing	67559	99.99%
Total	All	Benign	224137	100%
		Phishing	67565	100%
		All	291702	

Evaluation: Model Selection

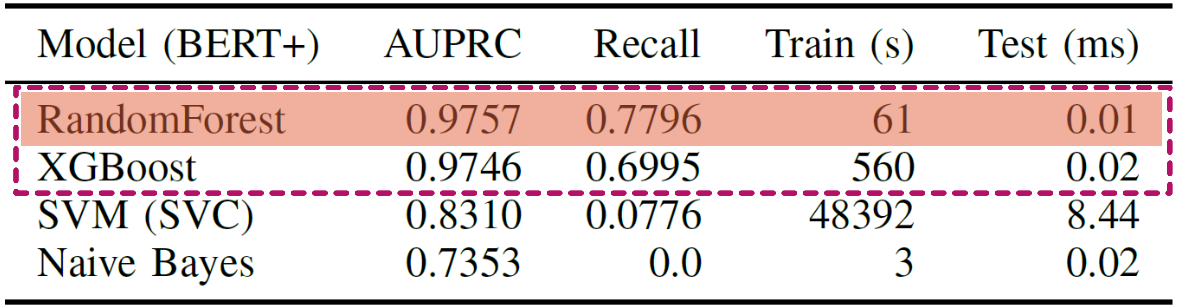
AUPRC, and Recall at Fixed False-positive rate **0.001 (10^{-3}).** Tested with EES 2020 dataset

Structural Module



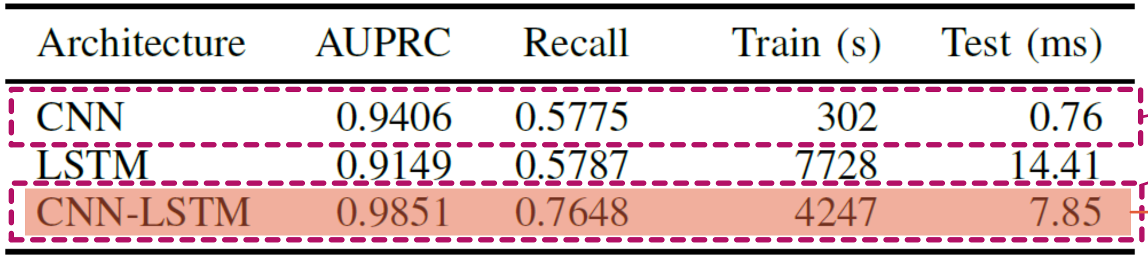
Model	AUPRC	Recall	Train (s)	Test (ms)
RandomForest	0.9993	0.9933	5	0.01
XGBoost	0.9994	0.9884	10	0.01
SVM (SVC)	0.9969	0.9618	919	0.55
Naive Bayes	0.8940	0.0	2	0.01

Text Module



Model (BERT+)	AUPRC	Recall	Train (s)	Test (ms)
RandomForest	0.9757	0.7796	61	0.01
XGBoost	0.9746	0.6995	560	0.02
SVM (SVC)	0.8310	0.0776	48392	8.44
Naive Bayes	0.7353	0.0	3	0.02

URL Module



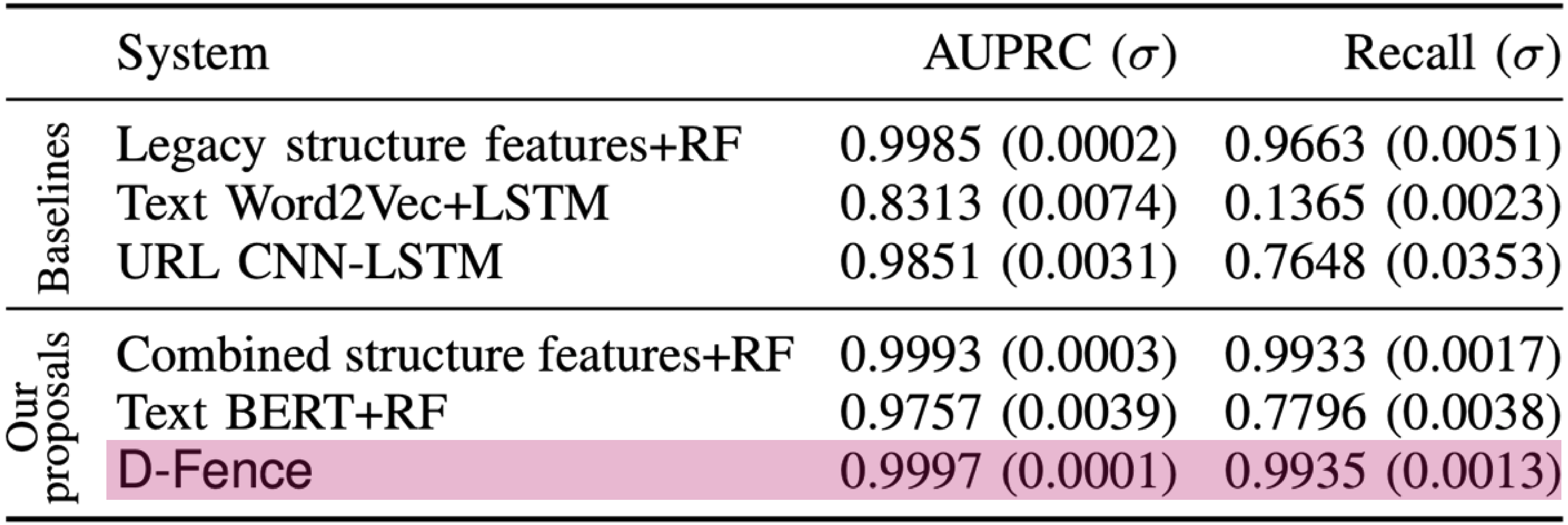
Architecture	AUPRC	Recall	Train (s)	Test (ms)
CNN	0.9406	0.5775	302	0.76
LSTM	0.9149	0.5787	7728	14.41
CNN-LSTM	0.9851	0.7648	4247	7.85

Models for Cost-efficient Configuration Analysis

Models for Best-Accuracy Evaluation

Evaluation: Comparison with Baselines

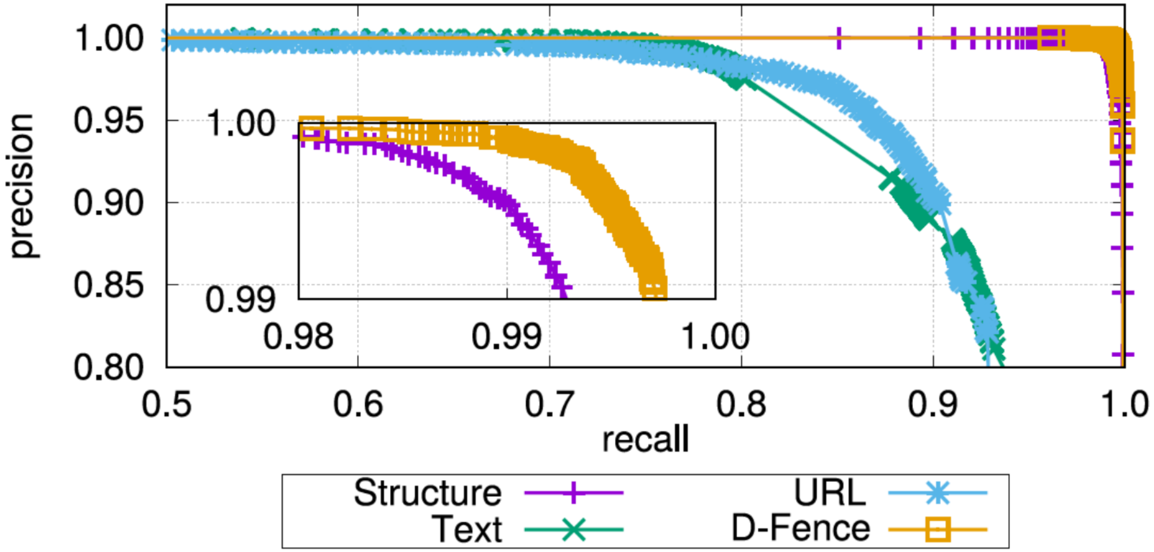
10-Cross-fold validation (90:10 splits). Recall at 10^{-3} FPR



	System	AUPRC (σ)	Recall (σ)
Baselines	Legacy structure features+RF	0.9985 (0.0002)	0.9663 (0.0051)
	Text Word2Vec+LSTM	0.8313 (0.0074)	0.1365 (0.0023)
	URL CNN-LSTM	0.9851 (0.0031)	0.7648 (0.0353)
Our proposals	Combined structure features+RF	0.9993 (0.0003)	0.9933 (0.0017)
	Text BERT+RF	0.9757 (0.0039)	0.7796 (0.0038)
	D-Fence	0.9997 (0.0001)	0.9935 (0.0013)

Evaluation: Recall at 10^{-4} FPR

EES 2020 Dataset. Best Accuracy Configuration.



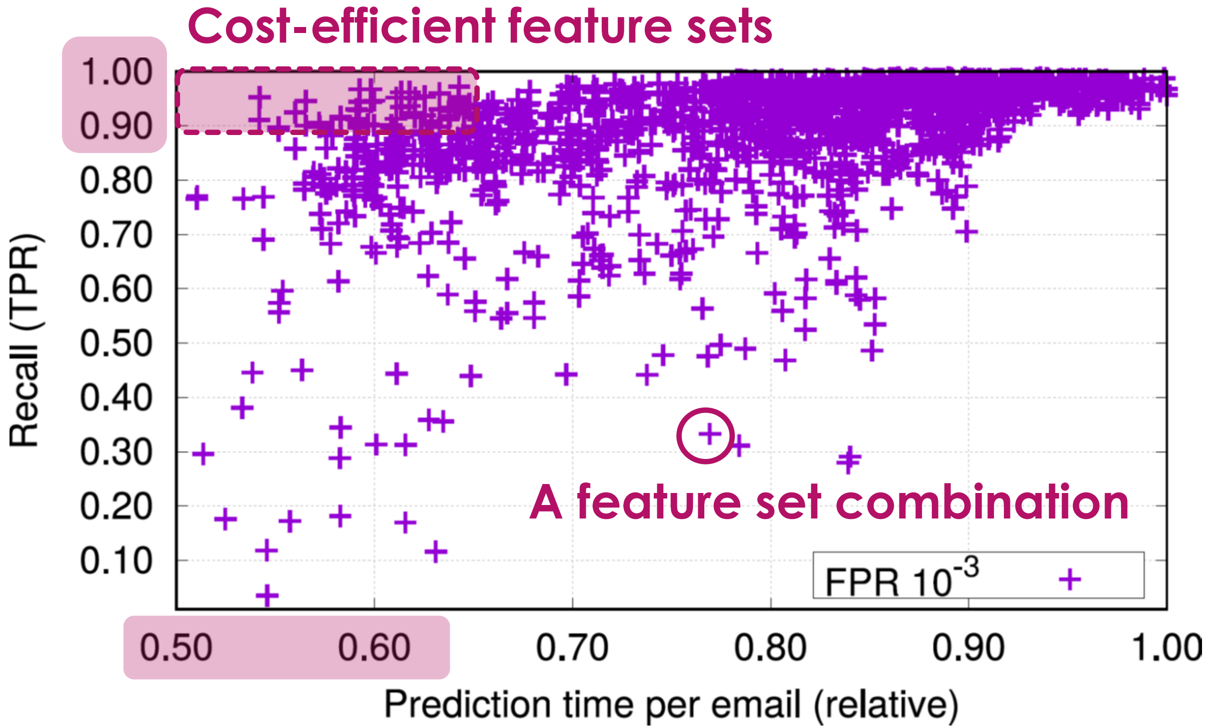
	AUPRC	Recall (10^{-3} FPR)	Recall (10^{-4} FPR)
Structure module	0.9994	0.9878	0.9428
Text module	0.9192	0.6182	0.2710
URL module	0.9492	0.8806	0.7721
D-Fence	**0.9995**	**0.9932**	**0.9844**

4% more detection e.g., 1K more phishing emails in our test set

Cost Reduction: Structural Module

Feature set Reduction

- ❑ **Feature selection by Feature Category**
  - ❑ 10 Feature categories
  - ❑ e.g., Msg-ID features, Link features, .., etc.
- ❑ **Test on 2^{10} Feature set combinations**
  - ❑ A point + on plot indicates one combination
- ❑ **Cost-Efficient Features**
  - ❑ Less feature extraction time but high accuracy

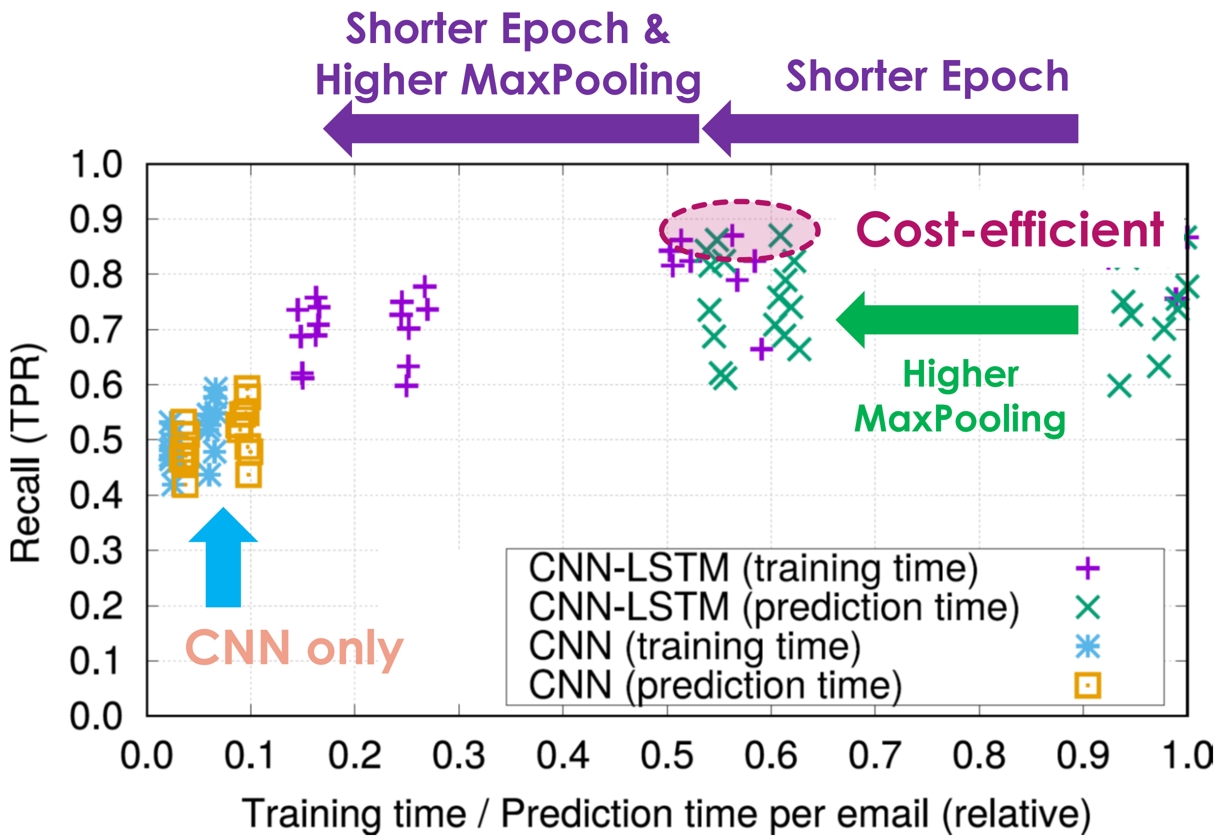


~50% Prediction time reduction from Reduced feature set with keeping 95%+ Recall at FPR 10^{-3}

Cost Reduction: URL Module

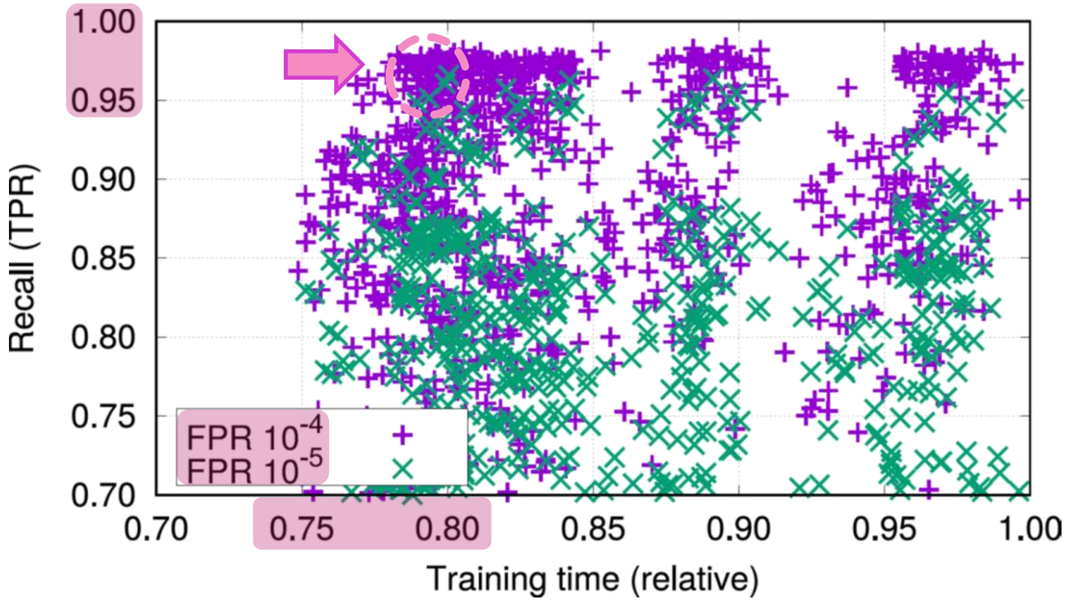
Hyper-parameter tuning: Simpler/Faster Neural network

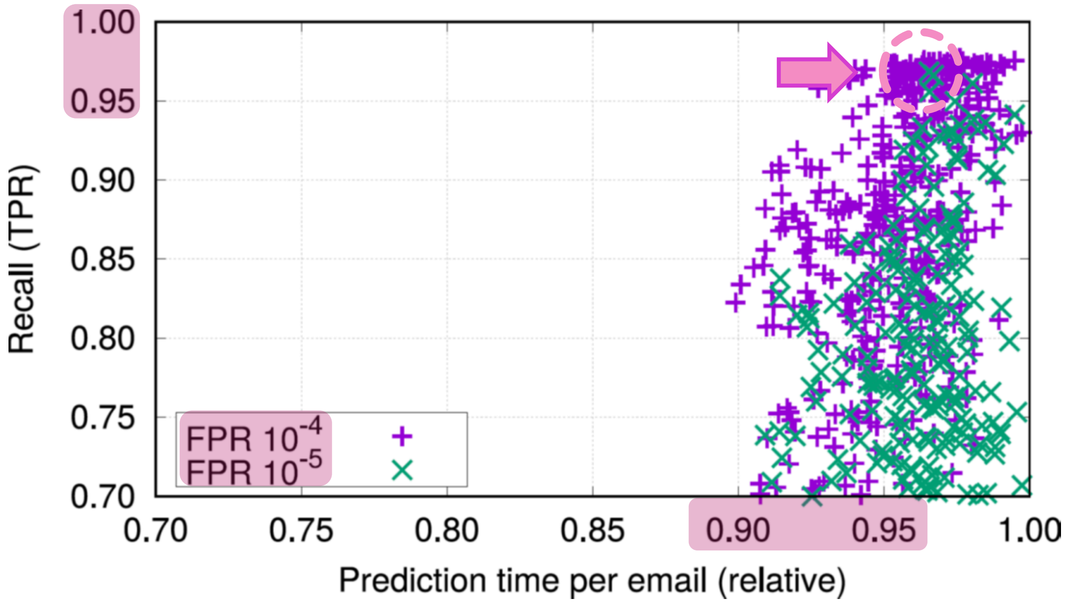
- ❑ **Shorter training Epoch**
  - ❑ Advantage: Shorter training time
  - ❑ Cost: Loss in accuracy
- ❑ **Higher Max Pooling**
  - ❑ Advantage: Shorter training/prediction time
- ❑ **CNN (without LSTM layer)**
  - ❑ Advantage: Faster training/prediction
  - ❑ Cost: Large loss in accuracy



Cost-Efficient Configuration

Combinations of the module configurations





- **Text module fixed as the fastest configuration. (100 words analysis)**
- **A pair of points (purple and green)** : one config combination

~20% of Training time reduction from mainly Deep-learning for URL
~10% of Prediction time reduction from URL and Structure module with 0.95+ Recall at 10^{-5} FPR

Conclusions

D-FENCE: Flexible Multi-modular phishing email detection system

- **Wide component coverage** with comprehensive detection: **little evasion surface**
- **Low False-detection** powered by independent analysis modules supplementing each other
- Evaluated with near **300K** of real-world Enterprise email dataset

Cost-efficient Configuration

- **Synergetic configuration**: Better than combination of the best individual configurations
- **Training time reduction** without harming accuracy

Thank You

Q & A