Vilniaus universitetas

Doktorantas:
Paulius Vaitkevičius

Vadovas:
Dr. Virginijus Marcinkevičius

IV metų II pusmečio ataskaita
2025 m. rugsėjo 24 d.

Mašininiu mokymusi grįstų atvirųjų šaltinių žvalgybos informacijos išskyrimo ir analizės metodai

Doktorantūros laikotarpis: 2018 - 2025

Vilniaus universitetas

TURINYS

1. Studijų plano vykdymas
2. Trumpas per pusmetį gautų mokslinių rezultatų pristatymas
3. Problemos apibrėžimas, tyrimo objektas, tikslai ir planuojami gauti rezultatai
4. Kito pusmečio darbo planas

STUDIJŲ PLANO VYKDYMAS

Vilniaus universitetas

Visų studijų planas, vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose		Publikacijos		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė
I (2018/2019)	1	1		1			
II (2019/2020)	1	3		1		1	Publikuota
III (2020/2021)	2		1		1		Įteikta - nepriimta
IV (2021/2025)			1	1	1	1	Įteikta į kitą žurnalą
Iš viso:	**4**	**4**	**2**	**3**	**2**	**2**	

Vilniaus universitetas

Visų mokslinių tyrimų ir disertacijos rengimo etapai

1. Mokslinių tyrimų disertacijos tema apžvalga ir analizė
2. Mokslinio tyrimo vykdymas:
    1. Tyrimo metodikos sudarymas
    2. Teorinis tyrimas
    3. Empirinis tyrimas
    4. Gautų duomenų analizė, apibendrinimas, išvadų parengimas
3. Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas **(90%)**
4. Daktaro disertacijos parengimas ir svarstymas padalinyje
5. Daktaro disertacijos gynimas

Vilniaus universitetas

Ataskaitinio pusmečio planas ir rezultatai

Egzaminai		Dalyvavimas konferencijose		Publikacijos	
Planas	**Įvykdyta**	**Planas**	**Įvykdyta**	**Planas**	**Įvykdyta**
		Tyrimo rezultatų pristatymas tarptautinėje mokslinėje konferencijoje.	„International Conference on Science & Technology, 23-24 September 2025, Budapest“ (online).	Empirinio tyrimo rezultatų publikavimas (recenzuojamame leidinyje, CA WoS su Impact Factor).	Įteikta publikacija žurnalui „Advances in Distributed Computing and Artificial Intelligence Journal“

- Pakoreguota ir pakartotinai įteikta publikacija.
- Rezultatai pristatyti tarptautinėje konferencijoje.
- Disertacija 90%.

PROBLEMOS APIBRĖŽIMAS, TYRIMO OBJEKTAS, TIKSLAI

Vilniaus universitetas

Tyrimo tikslas

Sukurti apsimetinėjimo atakoms atsparų metodą, grįstą giliaisiais neuroniniais tinklais ir natūralios kalbos apdorojimo algoritmais, kuris leistų efektyviai ir patikimai atpažinti **duomenų išviliojimo internete** (angl. „Phishing“) tinklapius.

Vilniaus universitetas

Tyrimo objektas

1. Mašininio mokymo ir giliojo mašininio mokymo algoritmai, skirti atpažinti duomenų išviliojimo internete tinklapius.
2. Atsparūs priešiškoms atakoms algoritmai (angl. „Adversarial Machine Learning“).

Vilniaus universitetas

Tyrimo uždaviniai

1. Atlikti literatūros analizę, išanalizuoti state-of-the-art algoritmus duomenų išviliojimo internete tinklapių atpažinimui.
2. Atkartoti *state-of-the-art* algoritmų rezultatus.
3. Sukurti duomenų rinkinius eksperimentų vykdymui.
4. **Pasiūlyti naują efektyvesnį duomenų išviliojimo internete tinklapių atpažinimo metodą.**
5. Atlikti eksperimentinius tyrimus, palyginant pasiūlytą metodą su *state-of-the-art* algoritmais.

GAUTŲ MOKSLINIŲ REZULTATŲ PRISTATYMAS

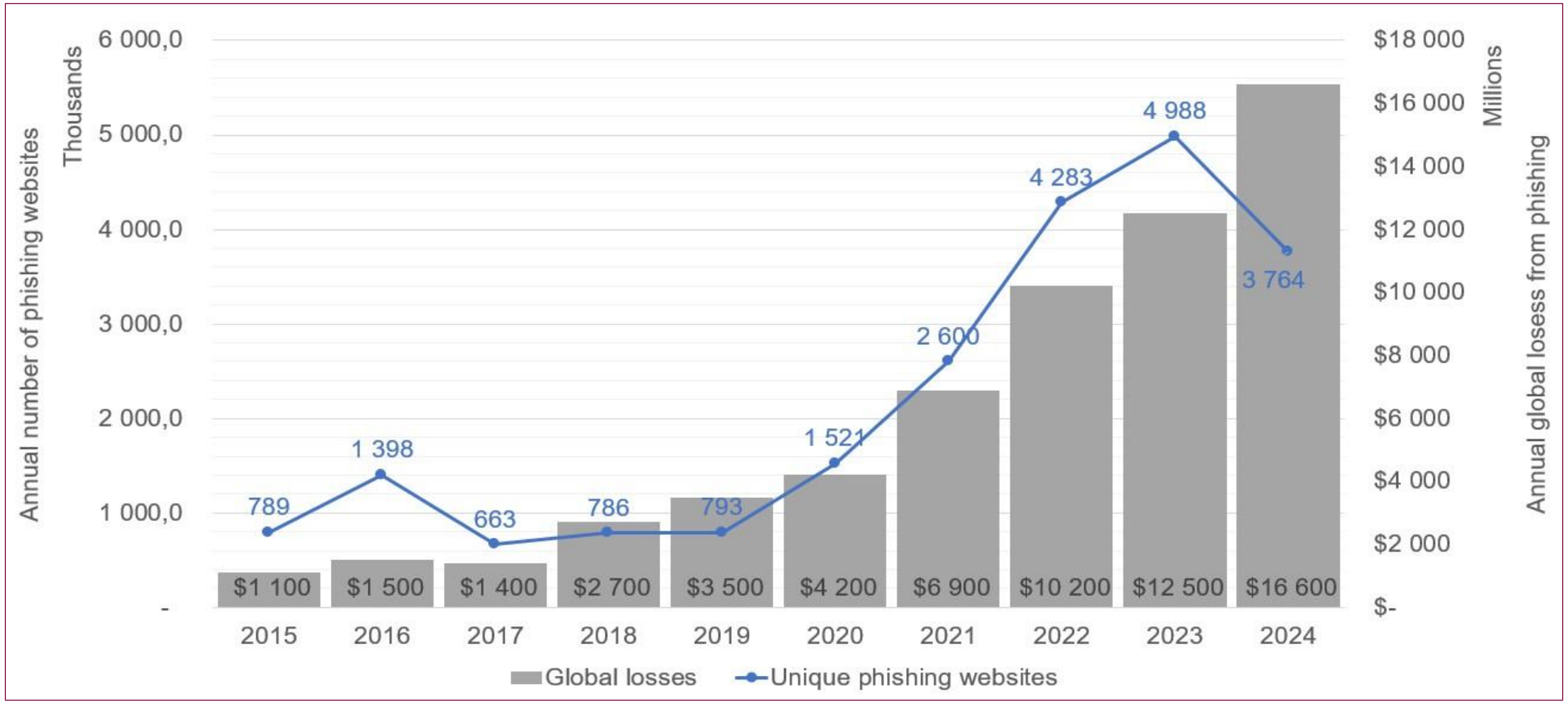
The Phishing Problem: A Growing Threat

- **Phishing is a cyber-attack aimed at stealing sensitive user data like passwords and credit card details.**
- It's a gateway for more severe attacks like ransomware and intrusions.
- Phishing is the #1 cybercrime reported to the FBI's Internet Crime Complaint Center

DALL-E in style of Lithuanian painter and composer M. K. Čiurlionis

Vilnius University

The Phishing Problem: A Growing Threat



(source: authors, based on Anti-Phishing Working Group and Federal Bureau of Investigation, 2024

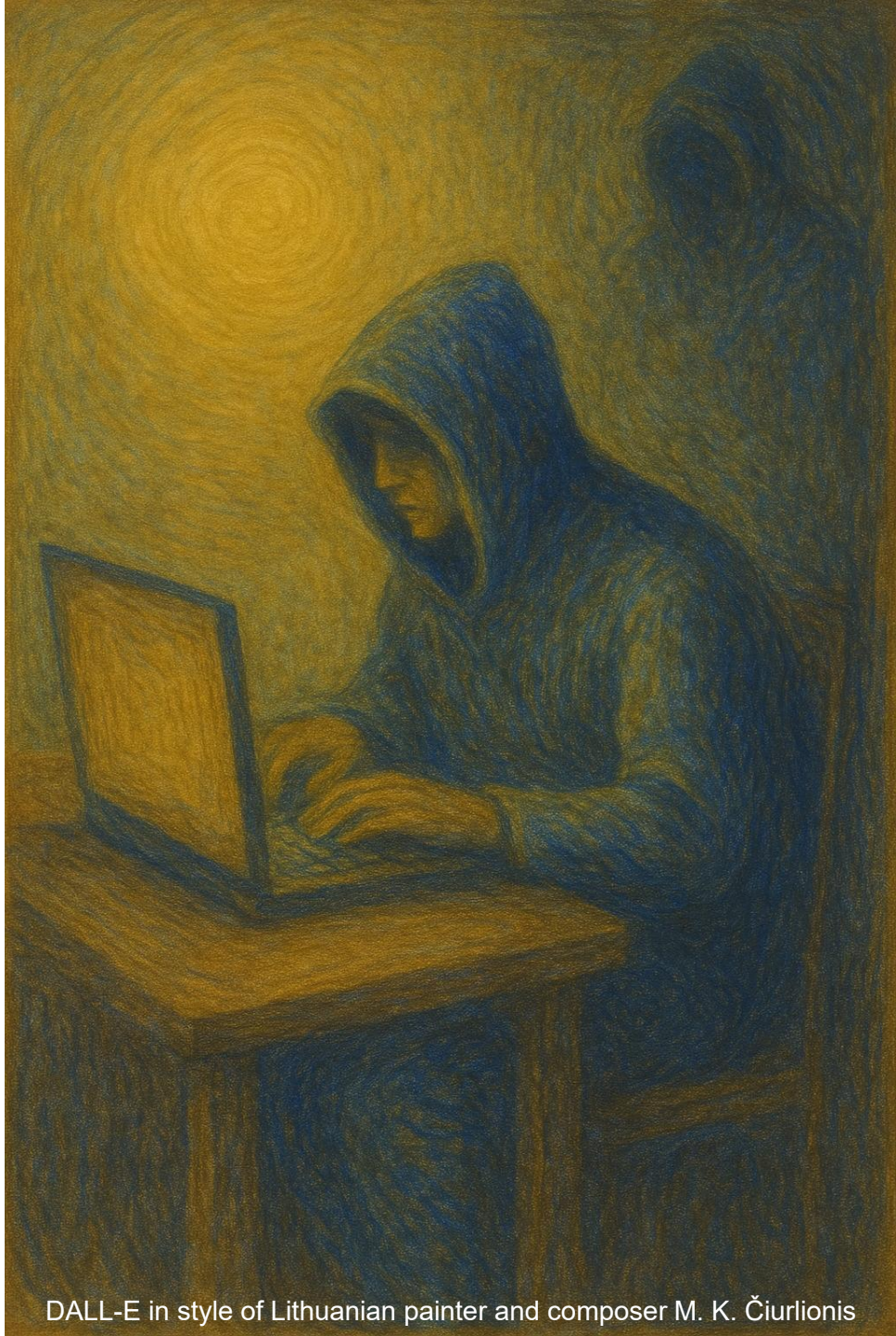
The Vulnerability of Modern Detectors

- SOTA classifiers declare extremely high accuracy, up to 99.98% on standard test datasets.
- When tested against adversarial URLs (specifically crafted to evade detection) their **performance collapses to 30-55% lower** than claimed.
- Studies show evasion attacks can achieve success rates of over **67%** against deep learning models and even up to 100% in some cases.
- **The Core Problem:** Models trained on standard datasets don't learn the patterns of evasive attacks, leaving them vulnerable in real-world scenarios.

DALL-E in style of Lithuanian painter and composer M. K. Čiurlionis

Research Questions

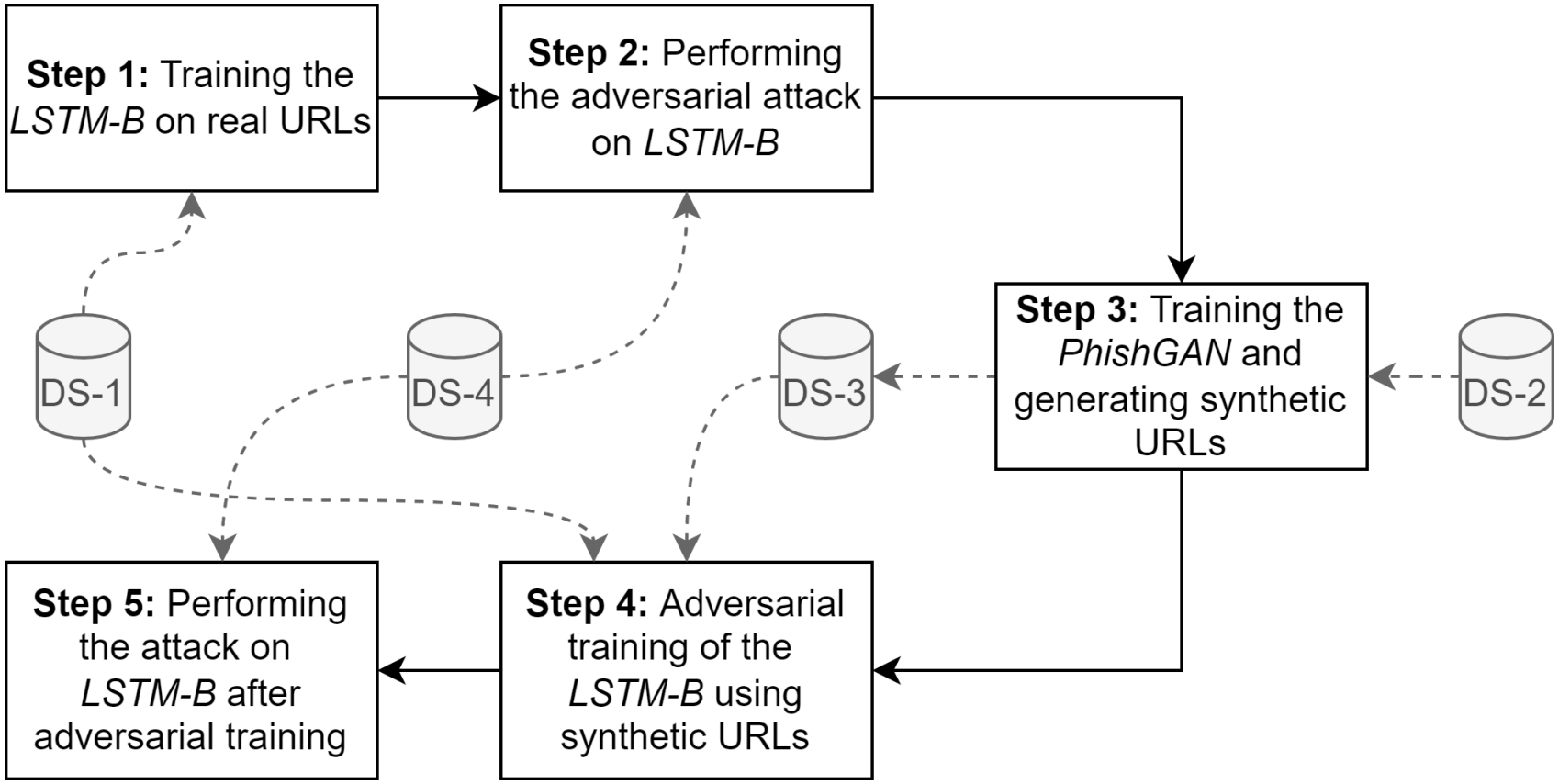
- Can we build a GAN to effectively generate synthetic phishing URLs for training data augmentation?
- What is the optimal architecture for a resilient, deep learning-based phishing classifier?
- Can our model, after adversarial training, achieve an attack success rate lower than the 79.85% seen in prior work on retrained SOTA models?



DALL-E in style of Lithuanian painter and composer M. K. Čiurlionis

Vilnius University

Overall Methodology

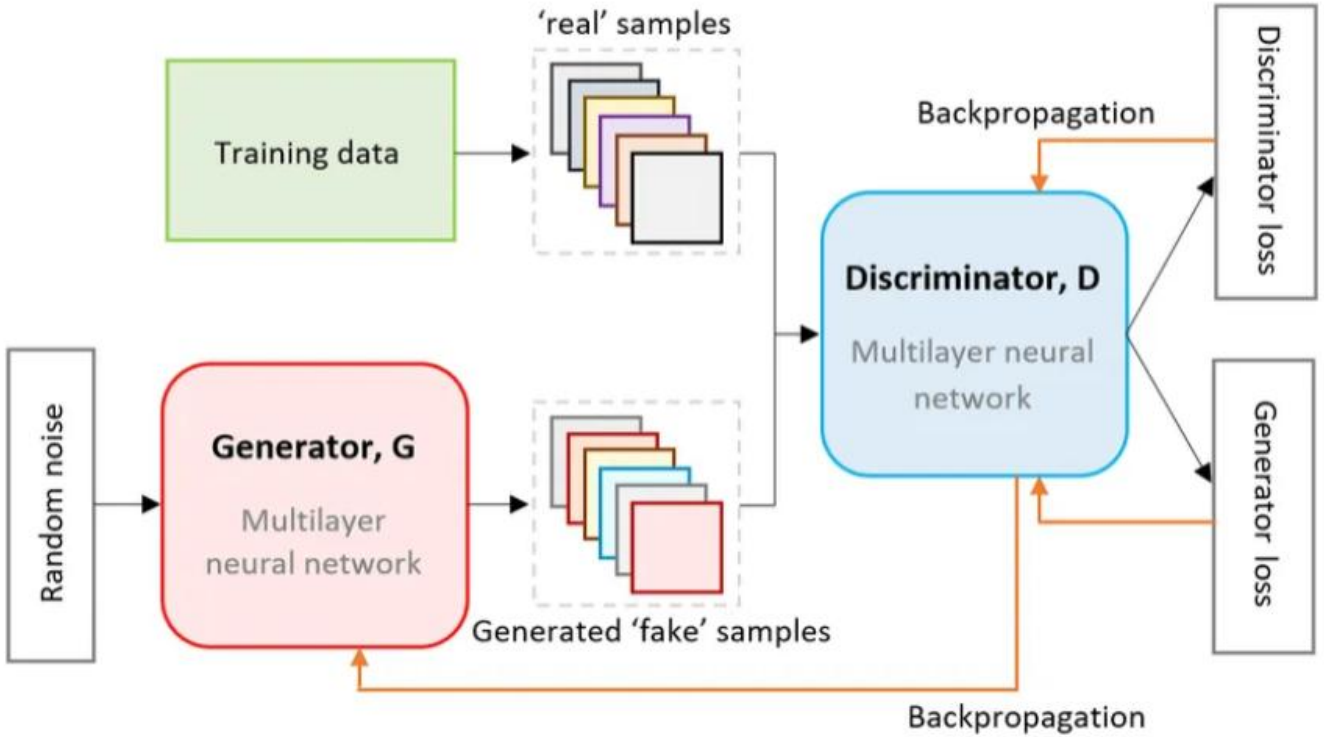


LSTM-B – our baseline phishing detection model
PhishGAN – proposed method to generate synthetic phishing URLs

Datasets

- DS-1: 300k URLs (150k phishing, 150k legit). Sources: PhishTank, OpenPhish, PhishStorm, Common Crawl, Alexa.
  Train/test baseline LSTM-B model.
- DS-2: 350k phishing URLs (unique, no overlap with DS-1/DS-4).
  Train PhishGAN.
- DS-3: 150k adversarial phishing URLs synthesized by PhishGAN.
  Adversarial training of LSTM-B model.
- DS-4: 25k handcrafted adversarial phishing URLs (Sabir et al. 2020).
  Used for adversarial attacks on LSTM-B model.

GAN

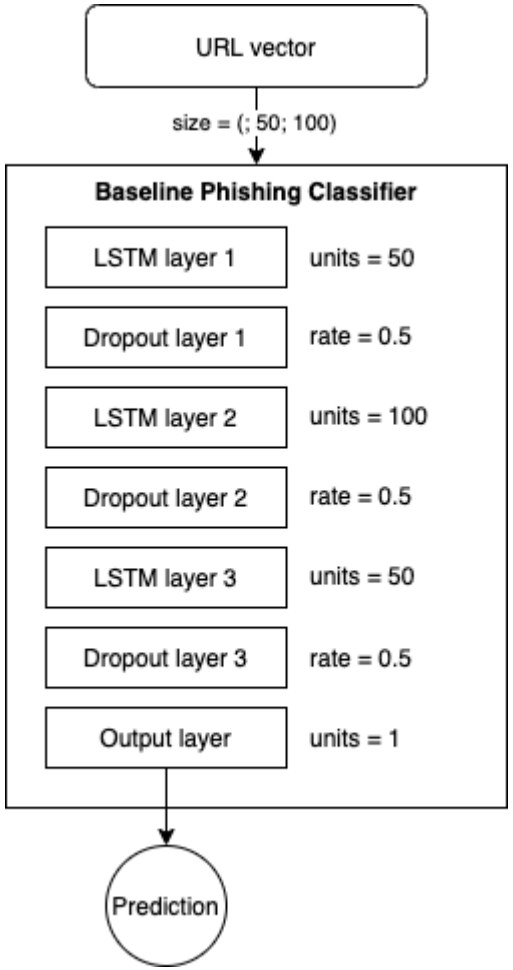


- GANs were proposed by Goodfellow et al. (2014) for estimating generative models via an adversarial process
- A generative model G and a discriminative model D are trained simultaneously in a *minimax* game

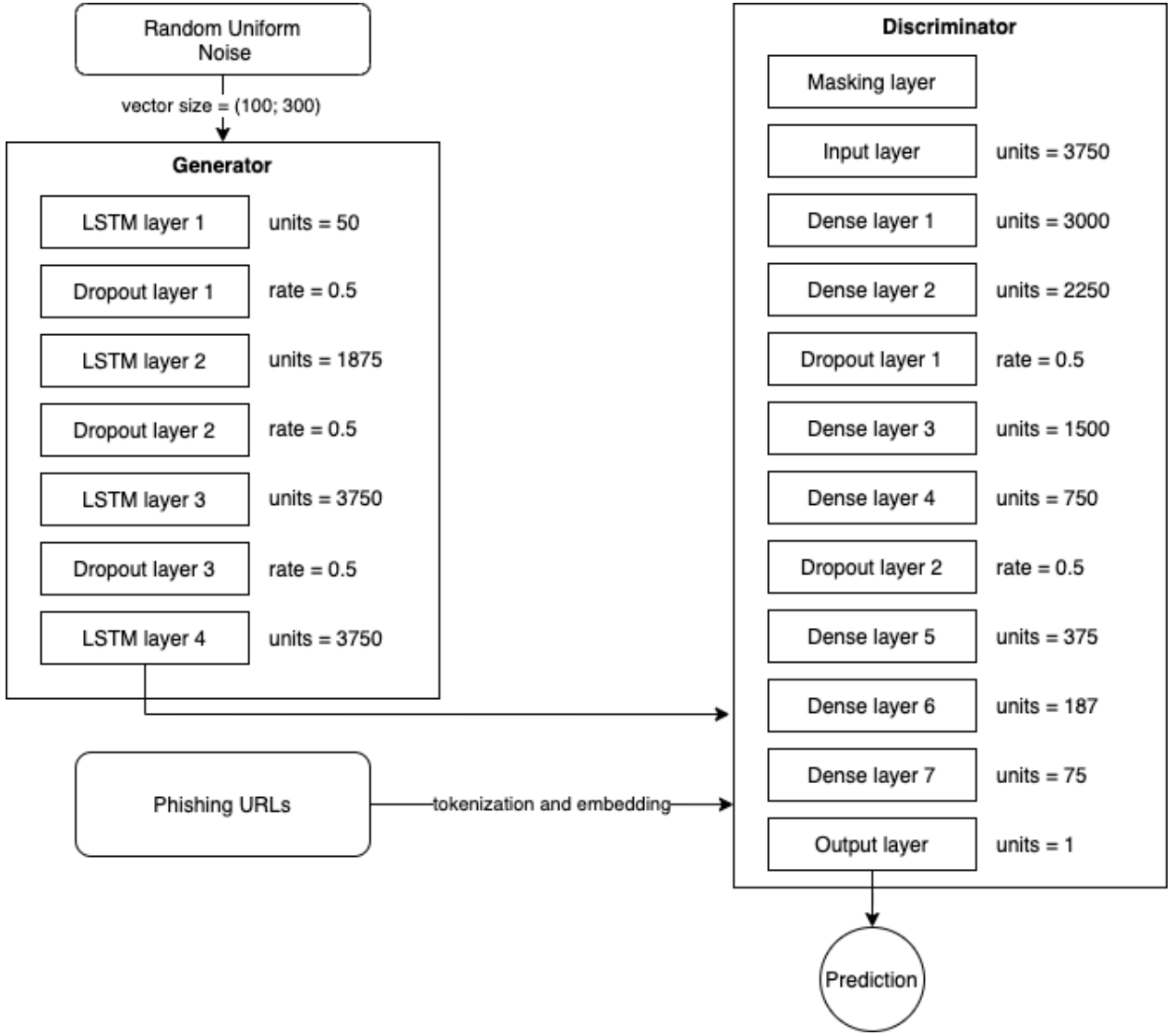
Vilnius University

Model Architectures

Baseline Classifier



PhishGAN



PhishGAN

- Challenge: Generating textual samples (vs. images) is unstable. Problems include mode collapse, gradient vanishing, and gradient boosting

- Stable training of PhishGAN achieved with Wasserstein GAN + Gradient Penalty
- Wasserstein-1 distance (EMD) gives smoother, stable gradients
- Improved with Gradient Penalty, which enforces the Lipschitz constraint (needed for Wasserstein distance) not by weight clipping, but by penalizing gradients

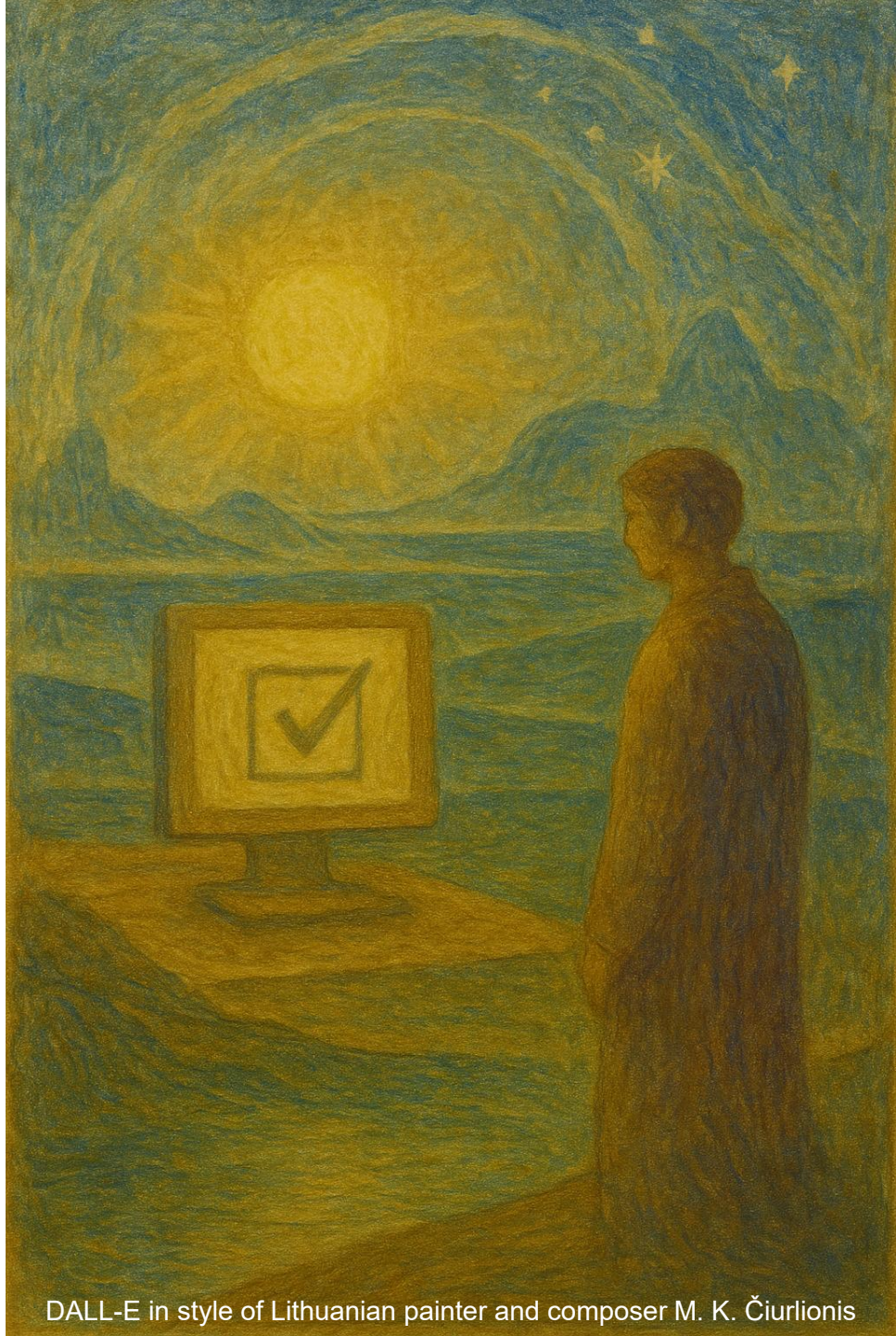
Results: Before Adversarial Training

- Baseline LSTM-B performance: mean accuracy of 95.88%.
- LSTM-B under adversarial attack:
  - mean accuracy of 63.16%
  - ASR of **36.84%**.
- While the accuracy drop is significant, our baseline's ASR of 36.84% is already better than ASR of **67.98%** reported for other SOTA deep learning.
- This suggests LSTMs are inherently more robust.

DALL-E in style of Lithuanian painter and composer M. K. Čiurlionis

Results: After Adversarial Training

- Performance under the same adversarial attack: mean accuracy improved from 63.16% to **68.16%**.
- New attack success rate dropped to 31.84% from 67.98% comparing with SOTA classifiers.



DALL-E in style of Lithuanian painter and composer M. K. Čiurlionis

Conclusions

- We proposed an improved Wasserstein GAN method PhishGAN that successfully generates synthetic URLs for adversarial training.
- Memory-equipped models show greater resilience to character-level URL perturbations than standard DNNs.
- PhishGAN works: augmenting training data with synthetic adversarial examples is an effective strategy to harden classifiers against evasion attacks.
- A 5% improvement is significant in an operational setting, this translates to blocking thousands of additional phishing attempts daily with minimal overhead.

DALL-E in style of Lithuanian painter and composer M. K. Čiurlionis