AUTHENTICATION OF COPY DETECTION PATTERNS UNDER MACHINE LEARNING ATTACKS: A SUPERVISED APPROACH

B. Pulfer, R. Chaban, Y. Belousov, J. Tutt, O. Taran, T. Holotyak, S. Voloshynovskiy

University of Geneva
Department of Computer Science
Switzerland

ABSTRACT

Copy detection patterns (CDP) are an attractive technology that allows manufacturers to defend their products against counterfeiting. The main assumption behind the protection mechanism of CDP is that these codes printed with the smallest symbol size (1x1) on an industrial printer cannot be copied or cloned with sufficient accuracy due to data processing inequality. However, previous works have shown that Machine Learning (ML) based attacks can produce high-quality fakes, resulting in decreased accuracy of authentication based on traditional feature-based authentication systems. While Deep Learning (DL) can be used as a part of the authentication system, to the best of our knowledge, none of the previous works has studied the performance of a DL-based authentication system against ML-based attacks on CDP with 1x1 symbol size. In this work, we study such a performance assuming a supervised learning (SL) setting.

Index Terms— Copy detection patterns, supervised authentication, deep learning, machine learning fakes.

1. INTRODUCTION

Counterfeited products negatively affect world economy. They are present in multiple industries, varying from pharmaceutical and luxury products to identification documents, banknotes and even food and agricultural products.

Among all available technologies to defend against counterfeiting, copy detection patterns (CDP), based on hand-crafted randomness, represent a convenient, efficient and user-friendly solution [1, 2]. The advancement of mobile phone cameras' resolutions empowers the CDP technology even further, allowing the verification of authenticity of a product by the end customers. Furthermore, CDP are easily integrable into product design, are applicable in many different areas and feature low computational complexity for enrollment and authentication.

Previous works have shown that CDP with a symbol size of 3x3 and 5x5 printed on desktop printers at 600 dpi might be
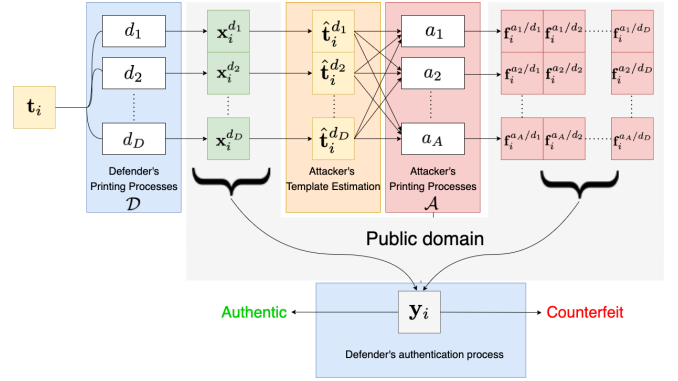
Fig. 1: Schematic representation of the problem. The defender creates a digital code $\mathbf{t}_i$, which is printed using the set of original printing processes $\mathcal{D}$. The printed original CDP are made public. An attacker estimates the digital template $\mathbf{t}_i$ given an original code $\mathbf{x}_i^d$ as $\hat{\mathbf{t}}_i^d$, and prints the estimations through its printing processes $\mathcal{A}$. The defender has to design an authentication system that, given original or fake code $\mathbf{y}_i$, predicts its authenticity with respect to $\mathbf{t}_i$.

cloned under certain conditions [3, 4, 5]. More recent works, which used industrial printers such as HP Indigo capable of producing symbols of size 1x1 at the resolution of 812.8 dpi, have shown that attacks based on ML-methods make the authentication based on hand-crafted features and SVM-based classifiers harder for the defender [6, 7].

However, up to our best knowledge, none of the previous works has combined CDP printed on industrial printers, and attacked using ML-based attacks, with Deep Learning (DL) based authentication strategies. To fill this gap, in this study we empirically evaluate the capabilities of a DL-based authentication system to detect ML-based counterfeits.

We assume a supervised learning (SL) setting, where the defender possesses of at least one type of non-authentic codes. In our case, this assumption translates in the defender observing fakes produced by the attacker using one (potentially more) particular printing process(es). We refer to "distribution shift" of fakes for cases where the fake codes used dur-

ing the training and test are printed on different printers. We address the following research questions (RQ):

RQ1: Is it possible to reliably detect ML-based fakes using a two-class classifier? We measure the performance of the supervised classifier w.r.t. a test set of original and fake codes printed on the same printer as the training data.

RQ2: Is the supervised classifier robust to a distribution shift of fakes? We measure the performance of our system against fake CDP, which were printed on a different printer w.r.t. CDP used for training. We are interested in learning if the system is robust to such a distribution shift. Also, we investigate which type of fake codes should the defender use during training such that the authentication system yields the best performance under a distribution shift.

2. PROBLEM FORMULATION

To defend its products from counterfeiting, a manufacturer (or defender) creates N digital templates $\mathbf{t}_i \in \{0,1\}^{w \times h}, i \in \mathcal{I} = \{1, ..., N\}$ of size $w \times h$ that are then printed through one or more defender's printing processes $d \in \mathcal{D}$ ($|\mathcal{D}| = D$), to obtain original CDP $\mathbf{x}_i^d \in \mathbb{R}^{w \times h}$. The original code $\mathbf{x}_i^d$ will be available to the customer upon purchase, which in turn can verify the authenticity of the product by acquiring it.

An attacker accesses the original printed CDP in the public domain and estimates, for an observed original $\mathbf{x}_i^d$, the digital template $\mathbf{t}_i$ as $\hat{\mathbf{t}}_i^d$ through a ML-based estimation model [4, 5, 6, 7]. Having the estimation $\hat{\mathbf{t}}_i^d$, the attacker is then able to produce a fake code $\mathbf{f}_i^{a/d} \in \mathbb{R}^{w \times h}$ using one of the attacker's printing processes $a \in \mathcal{A}, |\mathcal{A}| = A$. The superscript "$a/d$" denotes the fact that the digital template $\hat{\mathbf{t}}_i^d$ is estimated from the CDP $\mathbf{x}_i^d$ printed on printer d by the defender and reproduced on the printer a by the attacker. The resulting fake is then put into the public domain as shown in Figure 1. The attacker can thus generate $D \times A$ fake CDP for a single template. We denote $\mathbf{x}^d = \{\mathbf{x}_i^d | i \in \mathcal{I}\}$ and $\mathbf{f}^{a/d} = \{\mathbf{f}_i^{a/d} | i \in \mathcal{I}\}$.

The goal of the defender is to design an authentication system which, given a probe $\mathbf{y}_i \in \{\mathbf{x}_i^d, \mathbf{f}_i^{a/d} | d \in \mathcal{D} \wedge a \in \mathcal{A}\}$, can correctly determine whether the probe is an original one, i.e., $\mathbf{y}_i \in \{\mathbf{x}_i^d | d \in \mathcal{D}\}$ or fake, i.e., $\mathbf{y}_i \in \{\mathbf{f}_i^{a/d} | d \in \mathcal{D} \wedge a \in \mathcal{A}\}$. The authentication system should minimize the probability of accepting fake codes, P_{fa}, while also minimizing the probability of missing original codes, P_{miss}. Notice that, because of the phenomenon of *dot gain*, which is related to the interaction between the printing ink and the substrate, it is assumed that it is generally not possible for the attacker to generate fakes that are identical to the originals [1, 2].

In this study, we assume that the attacker has an access to the same printing processes as the defender, that is: $\mathcal{D} \subseteq \mathcal{A}$. This represents the worst case for the defender, as the difference in printing artifacts is minimal. We will validate this assumption in section 5. Furthermore, while it would be possible for the manufacturer to use the set of enrolled printed
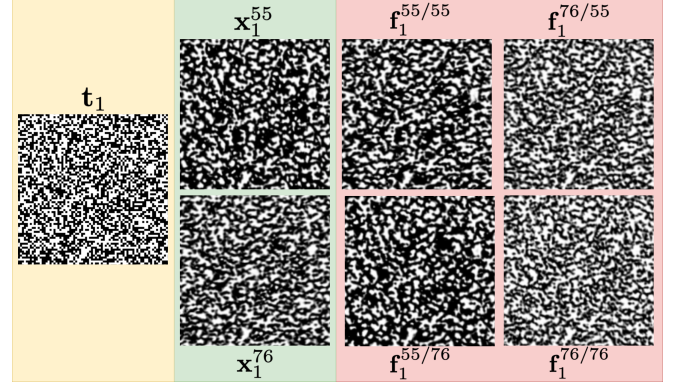

Fig. 2: Example of the first tuple of our dataset. For a single template code (cropped part), two originals are printed on two industrial printers. Fakes are obtained by estimating the digital template starting from the two original printed CDP and then printing the estimations on both printers.

codes $\{\mathbf{x}_i^d | d \in \mathcal{D}\}$ to help authenticate a probe $\mathbf{y}_i$, we assume that the defender does not use physical references and relies only on the set of digital templates $\{\mathbf{t}_i | i \in \mathcal{I}\}$ when authenticating probe $\mathbf{y}_i$. This assumption is due to a typical operational industrial scenario where the defender tries to minimize its costs for the enrollment of physical references for each printed item and Information Technology (IT) infrastructure management. Thus, the authentication is performed based on the digital reference only. Finally, we focus on supervised learning based authentication systems. We thus assume that the defender has an access to at least one (or more) set(s) of fake CDP $\mathbf{f}^{a/d}$, which could have been observed from the public domain, for the authentication system training.

3. DATASET

We use 720 tuples of CDP of size 1x1 pixel from the Indigo 1x1 base dataset presented in [6] and publicly available at [8], where each tuple includes one digital template, two original and four fake printed codes as shown in Figure 2. Each digital template is a binary image of size 684x684 pixels, whereas printed originals and fakes are gray-scale images of the same size obtained after synchronization based on special markers. Templates were printed using industrial printers HP Indigo 5500 (HPI55) and HP Indigo 7600 (HPI76) with resolution 812.8 dpi. Originals are 2400 ppi scans of the obtained printed templates (to a simulate mobile phone camera), while for fakes 6400 ppi scans were used to obtain the ML-estimations that were then printed with the same conditions [6]. We denote the printing processes based on the printer number as $\{55, 76\}$. Tuple i of our dataset is thus presented as $(\mathbf{t}_i, \mathbf{x}_i^{55}, \mathbf{x}_i^{76}, \mathbf{f}_i^{55/55}, \mathbf{f}_i^{55/76}, \mathbf{f}_i^{76/55}, \mathbf{f}_i^{76/76})$, where $\mathbf{t}, \mathbf{x}$ and $\mathbf{f}$ represent digital templates, original and fake printed codes respectively. The superscript for originals indicates the used

(a) $\mathbf{x}^{55}$ and fakes   (b) $\mathbf{x}^{76}$ and fakes
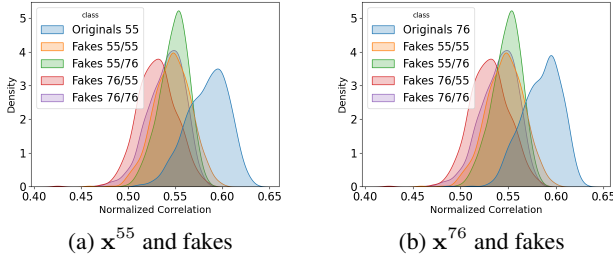
Fig. 3: KDE plot of normalized correlation between the digital templates and printed original and fake codes.
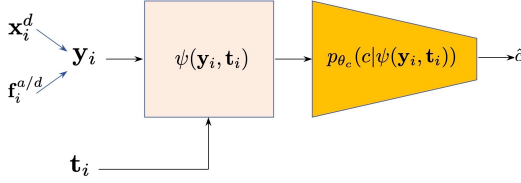
Fig. 4: Considered system of supervised authentication based on reference digital template $\mathbf{t}_i$. The sample under investigation $\mathbf{y}_i$, which could be an original $\mathbf{x}_i^d$ or fake $\mathbf{f}_i^{a/d}$, is aggregated with the corresponding template $\mathbf{t}_i$ through aggregating function $\psi(.)$. The deep classifier p_{θ_c}, parametrized by learnable parameters θ_c, outputs binary prediction $p_{\theta_c}(c|\psi(\mathbf{y}_i, \mathbf{t}_i)) = \hat{c}$ indicating the degree of acceptance of $\mathbf{y}_i$ as an original code.

printer. For fakes, the superscript a/d indicates that the estimation of the template was originated from original $\mathbf{x}_i^d$ and then printed with printer a. In Figure 3, we show the Kernel Density Estimation (KDE) plot of normalized correlation with the template codes for both original and all fake. The plots show that while originals (shown in blue) correlate more with templates, there is a decent overlap with fakes, making the authentication task challenging. Therefore the linear separation of originals and fakes in the considered metric is not feasible.

4. METHODOLOGY

The proposed supervised model takes as an input an aggregation $\psi(\mathbf{y}_i, \mathbf{t}_i)$ of the code under investigation $\mathbf{y}_i$ with the respective template $\mathbf{t}_i$. In the general case, the aggregation might be performed in different ways, for example, channel-wise concatenation, subtraction, and so on. In our experiments, $\psi(.)$ is the channel-wise concatenation, such that $\psi(\mathbf{y}_i, \mathbf{t}_i) \in \mathbb{R}^{2 \times 684 \times 684}$. The classifier $p_{\theta_c}(c|\psi(\mathbf{y}_i, \mathbf{t}_i))$, parametrized by learnable parameters θ_c, is trained to output $c = 1$ when $\mathbf{y}_i$ is an original CDP, and $c = 0$ when it is a counterfeit. The investigated system architecture is schematically shown in Figure 4. Let us denote $\phi_{\theta_c}(\mathbf{y}, \mathbf{t}) = p_{\theta_c}(c|\psi(\mathbf{y}, \mathbf{t}))$ for simplicity. Our model is trained to minimize the following loss function:

$$\mathcal{L}(\theta_c) = \frac{1}{N} \sum_{i=1}^{N} \left[\mathcal{L}_{miss}(i) + \mathcal{L}_{fa}(i) \right], \qquad (1)$$

where:

$$\mathcal{L}_{miss}(i) = \frac{1}{D} \sum_{d \in D} d_{BCE}(\phi_{\theta_c}(\mathbf{x}_i^d, \mathbf{t}_i), 1), \qquad (2)$$

$$\mathcal{L}_{fa}(i) = \frac{1}{AD} \sum_{a \in \mathcal{A}, d \in \mathcal{D}} d_{BCE}(\phi_{\theta_c}(\mathbf{f}_i^{a/d}, \mathbf{t}_i), 0), \qquad (3)$$

and the binary cross-entropy loss is defined as:

$$d_{BCE}(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}). \qquad (4)$$

4.1. Setups

We conduct experiments on setups which represent interesting study cases for the defender. **Setup 1:** The defender produces originals and observes fakes from a single printer. **Setup 2:** The defender prints on one printer but observes fakes from multiple printers. **Setup 3:** The defender prints multiple originals and observers multiple types of fakes.

The first setup could represent, for example, a situation where a defender prints originals with a unique printing process and produces his own fakes with some kind of manipulation (e.g., scanning original and re-printing). The second setup is as the first one, except that the defender has a way to create a multitude of manipulations to its original codes. Finally, in the third setup the defender also possesses multiple printing processes (e.g., multiple printers).

4.2. Training

For each setup, we train a standard ResNet18 backbone [9] with an added 2-layer MLP head. Each model was trained with a 40/10/50% train-validation-test split for 1'000 epochs using early stopping and a learning rate of 0.005 [1].

For training data, we use the following random augmentations: horizontal flip, vertical flip, 90° rotations, and gamma correction with $\gamma \in [0.4, 1.3]$. Each augmentation is chosen randomly, but then applied equally to all CDP in the tuple. For example, a clockwise rotation of 90° is applied to the template $\mathbf{t}_i$ as well as to all $\mathbf{x}_i^d$ and $\mathbf{f}_i^{a/d}$ used in the specific setup. No augmentation is carried out for validation and test data.

5. RESULTS

In Table 1 we present the average P_{miss}, P_{fa} and AUC score for all settings over 5 runs with different seeds. The first four

---

[1]Source code is publicly available at `https://gitlab.unige.ch/sip-group/cdp_supervised_ad`.

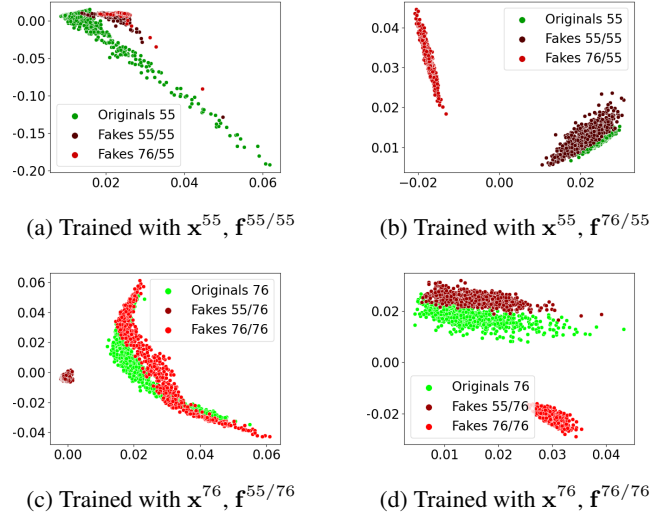| $\mathcal{D}_{train}$	$\mathcal{A}_{train}$	$\mathcal{D}_{test}$	$\mathcal{A}_{test}$	P_{miss}	P_{fa}	AUC
Setup 1						
$\mathbf{x}^{55}$	$\mathbf{f}^{55/55}$	$\mathbf{x}^{55}$	$\mathbf{f}^{55/55}$	0.01	0.00	**1.00**
			$\mathbf{f}^{76/55}$	0.01	0.00	**1.00**
$\mathbf{x}^{55}$	$\mathbf{f}^{76/55}$	$\mathbf{x}^{55}$	$\mathbf{f}^{55/55}$	0.00	1.00	0.70
			$\mathbf{f}^{76/55}$	0.00	0.00	**1.00**
$\mathbf{x}^{76}$	$\mathbf{f}^{55/76}$	$\mathbf{x}^{76}$	$\mathbf{f}^{55/76}$	0.01	0.00	**1.00**
			$\mathbf{f}^{76/76}$	0.01	1.00	0.26
$\mathbf{x}^{76}$	$\mathbf{f}^{76/76}$	$\mathbf{x}^{76}$	$\mathbf{f}^{55/76}$	0.00	1.00	0.25
			$\mathbf{f}^{76/76}$	0.00	0.00	**1.00**
Setup 2						
$\mathbf{x}^{55}$	$\mathbf{f}^{55/55}$, $\mathbf{f}^{76/55}$	$\mathbf{x}^{55}$	$\mathbf{f}^{55/55}$	0.00	0.00	**1.00**
			$\mathbf{f}^{76/55}$	0.00	0.00	**1.00**
$\mathbf{x}^{76}$	$\mathbf{f}^{55/76}$, $\mathbf{f}^{76/76}$	$\mathbf{x}^{76}$	$\mathbf{f}^{55/76}$	0.00	0.00	**1.00**
			$\mathbf{f}^{76/76}$	0.00	0.00	**1.00**
Setup 3						
$\mathbf{x}^{55}$, $\mathbf{x}^{76}$	$\mathbf{f}^{55/55}$, $\mathbf{f}^{76/55}$, $\mathbf{f}^{55/76}$, $\mathbf{f}^{76/76}$	$\mathbf{x}^{55}$	$\mathbf{f}^{55/55}$	0.00	0.00	**1.00**
			$\mathbf{f}^{76/55}$	0.00	0.00	**1.00**
		$\mathbf{x}^{76}$	$\mathbf{f}^{55/76}$	0.00	0.00	**1.00**
			$\mathbf{f}^{76/76}$	0.00	0.00	**1.00**

Table 1: Mean P_{miss}, P_{fa} and AUC over 5 runs on test data for the considered setups. P_{miss} and P_{fa} are found setting the acceptance threshold $\tau = 0.5$.

(a) Trained with $\mathbf{x}^{55}$, $\mathbf{f}^{55/55}$ (b) Trained with $\mathbf{x}^{55}$, $\mathbf{f}^{76/55}$

(c) Trained with $\mathbf{x}^{76}$, $\mathbf{f}^{55/76}$ (d) Trained with $\mathbf{x}^{76}$, $\mathbf{f}^{76/76}$

Fig. 5: 2D PCA of ResNet backbone for the four possible configurations in setup 1. Only for case (a) the originals and fakes are almost linearly separable.

columns of Table 1 define the particular setting, specifying the originals and fakes used during training and testing of the model. P_{miss} and P_{fa} are found for each model by setting an acceptance threshold $\tau = 0.5$. The AUC score is found varying the values of τ.

We empirically show the ease of a DL-based model at distinguishing fake CDP in setups 2 and 3. Our approach thus yields perfect performance in the absence of distribution shift at testing time.

Also for setup 1 the system performs perfectly against fakes which distribution was observed during training. However, only in one out of four cases the model could generalize to a distribution shift at the test time. This is the case when the system is trained on the originals and fakes printed with the printer HPI55, but then tested with fakes printed on the printer HPI76.

In Figure 5 we show the first two dimensions of PCA obtained from features extracted from the ResNet backbone (last layer) for each model in the setup 1. While in most cases the originals could be distinguished from both sources of fake, only in one case (Figure 5a) this can be done almost linearly. Thus, the obtained results confirm a known shortcoming of DL methods that require the complete knowledge of fake at the training stage. The alternation of fake' statistics at test time from those assumed at training time might have serious consequences as clearly shown in setup 1. When all fakes are known at training time as considered in settings 2 and 3, the DL authentication system demonstrates an excellent performance.

6. CONCLUSION

In this work, we studied the behaviour of the supervised DL-based authentication system against ML-based fake CDP with 1x1 symbol size. While the system behaves perfectly against the types of fake used at training time, it mostly does not generalize well to fake CDP printed on different printers. In summary, we addressed the following research questions:

RQ1: Is it possible to reliably detect ML-based fakes using a two-class classifier? The answer is positive. In all cases, our system yields the perfect performance in the absence of a distribution shift.

RQ2: Is the supervised classifier robust to a distribution shift of fakes? In general, the answer is negative. We show in Table 1 that only in one out of four cases the system generalizes to the fakes printed on different printers w.r.t. fake CDP used for training.

While the proposed system can detect fake CDP which distribution was used at training time, it is generally not robust to a distribution shift at test time. The authentication based on the physical references might be a possible solution to this problem. In this respect, in future work, we will investigate the performance of the authentication model based on the physical references, such that information about the statistics of the CDP can be taken into account when authenticating a probe coming from the public domain.

7. REFERENCES

[1] Justin Picard, "Digital authentication with copy-detection patterns," *Electron. Imaging*, vol. 5310, 06 2004.

[2] Justin Picard, Paul Landry, and Michael Bolay, "Counterfeit detection with qr codes," in *Proceedings of the 21st ACM Symposium on Document Engineering*, New York, NY, USA, 2021, DocEng '21, Association for Computing Machinery.

[3] Olga Taran, *Reliable classification in digital and physical worlds under active adversaries and prior ambiguity*, Ph.D. thesis, University of Geneva, 2021.

[4] Olga Taran, Slavi Bonev, and Slava Voloshynovskiy, "Clonability of anti-counterfeiting printable graphical codes: a machine learning approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019.

[5] Olga Taran, Joakim Tutt, Taras Holotyak, Roman Chaban, Slavi Bonev, and Slava Voloshynovskiy, "Mobile authentication of copy detection patterns: how critical is to know fakes?," 2021.

[6] Roman Chaban, Olga Taran, Joakim Tutt, Taras Holotyak, Slavi Bonev, and Slava Voloshynovskiy, "Machine learning attack on copy detection patterns: are 1x1 patterns cloneable?," 2021.

[7] Rohit Yadav, Iuliia Tkachenko, Alain Trémeau, and Thierry Fournel, "Estimation of Copy-sensitive Codes Using a Neural Approach," in *7th ACM Workshop on Information Hiding and Multimedia Security*, Paris, France, July 2019.

[8] Roman Chaban and Olga Taran, "Indigo 1x1 base," https://github.com/sip-group/ snf-it-dis/tree/master/datasets/ indigo1x1base, 2021.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.