Machine learning attack on copy detection patterns: are 1x1 patterns cloneable?

Roman Chaban, Olga Taran, Joakim Tutt, Taras Holotyak, Slavi Bonev and Slava Voloshynovskiy
Department of Computer Science, University of Geneva, Switzerland
{roman.chaban, olga.taran, joakim.tutt, taras.holotyak, slavi.bonev, svolos}@unige.ch

***Abstract*—Nowadays, the modern economy critically requires reliable yet cheap protection solutions against product counterfeiting for the mass market. Copy detection patterns (CDP) are considered as such a solution in several applications. It is assumed that being printed at the maximum achievable limit of a printing resolution of an industrial printer with the smallest symbol size 1×1, the CDP cannot be copied with sufficient accuracy and thus are unclonable. In this paper, we challenge this hypothesis and consider a copy attack against the CDP based on machine learning. The experimental results based on samples produced on two industrial printers demonstrate that simple detection metrics used in the CDP authentication cannot reliably distinguish the original CDP from their fakes under certain printing conditions. Thus, the paper calls for a need of careful reconsideration of CDP cloneability and search for new authentication techniques and CDP optimization facing the current attack.**

***Index Terms*—Copy detection patterns, machine learning fakes, supervised authentication, one-class classification.**

I. Introduction

The problem of the security of physical objects and their protection against counterfeiting is among the highly demanded features for the modern society and the economy. The counterfeited products affect numerous life aspects such as health-critical products (medicine, food), luxury products (objects of art, watches), identification documents and banknotes.

Besides the big variety of different anti-counterfeit technologies, all of them have pros and cons and no technique can guarantee perfect protection. For example, holograms are known to be difficult to clone. At the same time, they are not directly suitable for machine authentication and require the users' education and are still quite expensive in production. This makes their usage limited for massive markets. The RFID [1] and chip-based solutions are still expensive, require expensive infrastructure and extra equipment for the verification. One of the popular nowadays technologies is copy detection patterns (CDP) [2] [3] that belongs to the group of technologies based on hand-crafted randomness. Moreover, CDP are considered as an efficient yet cheap and user-friendly solution due to their low cost of production. Additionally, the recent achievements of mobile phone technologies allow the products' verification directly by the end customers.

The advantages of CDP also include their easy integration into product design, the wide area of applications and low computational complexity for both enrollment and authentication. Besides the mentioned benefits that make the CDP to be a valuable protection technology, it has been shown in [4], [5] that the CDP might be cloned under certain conditions. However, in the prior works, the authors investigated the cloneability aspects of the CDP printed on the desktop printers in contrast to industrial digital offset printers. In this respect, the current work aims at investigating the cloneability aspect of the CDP printed on the industrial printers under conditions close to those used in the industrial large-scale settings.

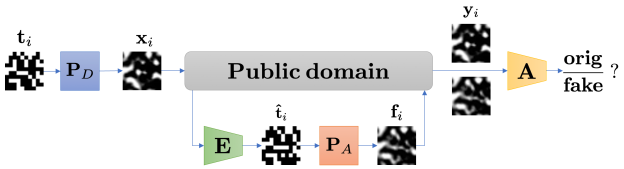

Fig. 1: The cycle of the CDP: $\mathbf{t}_i \in \{0,1\}^{n \times m}$ is a digital template, P_D is the original industrial printer. E is an attacker's estimation network for digital template estimation and $\hat{\mathbf{t}}_i$ is the estimated digital template. $\mathbf{x}_i$ and $\mathbf{f}_i$ are original and fake CDP printed by P_D and P_A, respectively. A is the authentication module, which makes a decision whether CDP is original or fake.

The main contributions of the current paper are:

- Creation of a public dataset of CDP printed on the industrial printers HP Indigo 5500 and HP Indigo 7600 at the resolution 812.8 dpi with the symbol size 1×1.
- Investigation of the impact of the different codes' density 30%, 35%, 40%, 45% and 50% on the cloneability aspects of the CDP.
- Investigation of different authentication methods of the CDP under the produced machine-learning fakes.

II. Problem formulation

As it is shown in Fig. 1, the life cycle of CDP starts from the creation of binary digital templates $\mathbf{t}_i \in \{0,1\}^{n \times m}$, where $n \times m$ denote the size of the code, by the manufacturer and printing them on the industrial printer P_D. Printed packages with CDP are distributed in the public domain. An attacker accesses an original package with the CDP, digitizes it and creates a digital fake using a trained estimation model E. After that, a fake package with the estimated CDP is printed. In this paper, we consider a possibility that the attacker has an access to a high-quality industrial printing machine. The

S. Voloshynovskiy is a corresponding author.
This research was partially funded by the Swiss National Science Foundation SNF No. 200021_182063.

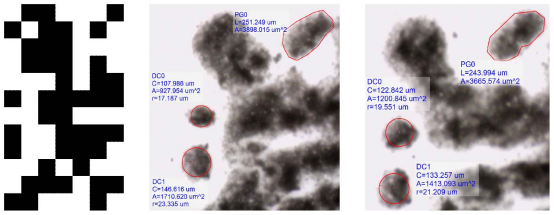

Fig. 2: Images of CDP captured with digital microscope[1]. Digital template has symbol size 1×1. The subscript denotes the index of CDP, and the superscript denotes the printer used for printing, for the fake $\mathbf{f}$ the second number in superscript denotes the printer used for digital template estimation.

printed fake package is released to the public domain at any point in time or logistics. At this moment, the authentication algorithm A has to decide whether the received product is original or counterfeited one based on a probe $\mathbf{y}_i$.

In the present work, we assume that the attacker has the same set of knowledge, equipment and algorithms as the defender to implement his/her attack strategies. This includes both access to printing and digitizing equipment. It is a common assumption [3] that a natural obstacle on a way to produce a perfect fake for the attacker is a loss of information in the process of printing due to various factors that are generally referred to as a *dot gain*. To train the CDP estimator E, the attacker can create his own training set of digital templates and printed CDP for the model training.

Besides the same access to the equipment, the defender has a freedom in the selection of the authentication algorithm. In this paper, we do not consider the defense by designing a special copy-resistant CDP that looks to be a very attractive defense strategy. Therefore, the defense is only based on a passive authentication. In this work, we investigate both supervised and unsupervised authentication approaches. The supervised authentication might produce more reliable results but needs both originals and fakes for the training of the classifier. On the other hand, a one-class unsupervised authentication requires only a set of originals.

III. Dataset overview

In this paper, we present a new public CDP dataset to investigate CDP cloneability[2]. Even though this topic is considered to be quite popular nowadays there is still a lack of datasets produced on industrial printers. This is mainly due to a fact that designing, manufacturing and acquisition of such dataset with further post-processing takes a lot of time and is costly.

The produced dataset consists of 720 randomly generated CDP $\mathbf{t}_i \in \{0,1\}^{228\times 228}$. Moreover, to investigate the printer's degradation, we created CDP with different densities. The term density referes to a probability of a black pixel in the CDP, e.g. if $density = 30\%$, then the probability of

[1]Model: Dino Lite AM7515MT8A (RA), magnification: approximately 689X with made in advance calibration.

[2]The dataset is available https://github.com/sip-group/snf-it-dis/tree/master/datasets/indigo1x1base.

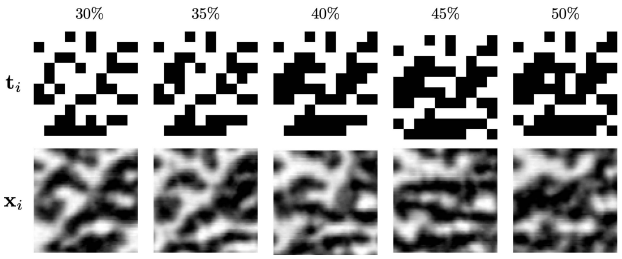

Fig. 3: The examples of CDP with different densities. $\mathbf{t}_i$ represents original digital template and $\mathbf{x}_i$ denotes the corresponding printed samples.

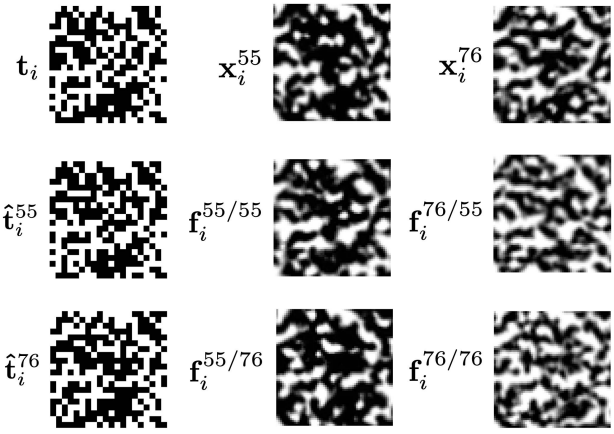

Fig. 4: The examples of both digital templates and printed CDP. The digital templates are estimated from printed codes $\mathbf{x}_i^{55}$ and $\mathbf{x}_i^{76}$ and are denoted as $\hat{\mathbf{t}}_i^{55}$ and $\hat{\mathbf{t}}_i^{76}$ respectively. The produced fakes are $\mathbf{f}_i^{55/55}$, $\mathbf{f}_i^{55/76}$, $\mathbf{f}_i^{76/55}$ and $\mathbf{f}_i^{76/76}$, where the first number is a printer on which fake was printed and second number is a printer by which estimation was done.

black pixel $P_{bp} = P[\mathbf{t}_{(m,n)} = 0] = 0.3$. We investigated $30\%, 35\%, 40\%, 45\%, 50\%$ densities. The difference between densities of the same index CDP is shown in Fig. 3.

The CDP were printed on two industrial printers HP Indigo 5500 (HPI55) and HP Indigo 7600 (HPI76) with **812.8** dpi resolution on Invercote G paper. The expected symbol size should be around $30\mu m$ in diameter. The chosen resolution is considered to be the best trade-off between print quality and distortions for the chosen digital offset printers. The overall technical conditions of both printers were not inspected and these factors may lead to certain deviations in printing.

The printed CDP are scanned with Epson Perfection V850 Pro using a particular set of settings[3]. The CDP acquisition is performed for the template estimation and CDP authentication. For the CDP estimation considered as a part of fake production, it is important to preserve as much information as possible in the scanned codes and 6400 ppi is the upper optical limit of this scanner model. The acquisition for the authentication stage simulates a hypothetical mobile phone camera resolution which is limited by 2400 ppi. Thus, we used 6400 ppi for the fake production on the side of the attacker

[3]OS: MacOS 11.3, used software: Epson Scan 2 v6.4.94, unsharp mask: High, brightness = 35, bit depth = 16, color mode = gray.

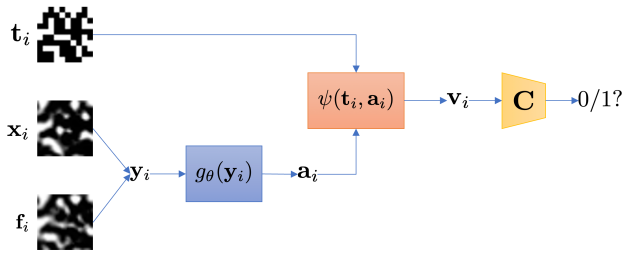

Fig. 5: The authentication scheme: $\mathbf{t}_i$ is the digital template. Probe $\mathbf{y}_i$ represents a probe corresponding to $\mathbf{x}_i$ or $\mathbf{f}_i$. $g_\theta(\mathbf{y}_i)$ is the CDP preprocessing function, which consists of synchronization, dynamic range normalization and Otsu's binarization for the Hamming and Jaccard distances and yields a processed image $\mathbf{a}_i$. $\psi(\mathbf{t}_i, \mathbf{a}_i)$ is function or set of functions measuring the differences between the input CDP and digital template. C is the classifier, which can be both one-class and two-class SVM.

and 2400 ppi was used for the authentication to simulate hypothetical mobile phone camera resolution. The resulting digital image of CDP would have about 8×8 pixels per each printed symbol with the scanning resolution 6400 ppi and 3×3 for 2400 ppi.

Each CDP is placed inside a special design pattern with synchromarkers, which are used for aligning scanned images to digital templates with pixel-wise precision. The visualization of the resulting aligned CDP is shown in Fig. 4.

IV. Attack and Authentication Strategies

A. Attacking strategy

The generation of machine learning (ML) fakes is based on an idea of digital template estimation from the printed counterparts. At the second stage, the estimated digital templates are printed on fake packages. In the considered setup, we assumed that besides the publicly available printed codes $\{\mathbf{x}_i\}_{i=1}^M$, the attacker has an access to the corresponding original digital templates $\{\mathbf{t}_i\}_{i=1}^M$. There are various ways to obtain these training pairs. As one possible scenario, one can assume that the attacker can print the digital templates $\{\mathbf{t}_i\}_{i=1}^M$ on the same printer as the defender and scan them as $\{\mathbf{x}_i\}_{i=1}^M$. Such a setup allows to fully explore the power of training on the side of the attacker.

The problem of training an estimator of digital templates from the printed counterparts given the training data $\{(\mathbf{t}_i, \mathbf{x}_i)\}_{i=1}^M$ generated from a joint distribution $p(\mathbf{t}, \mathbf{x})$ is formulated as a training of a parameterized network[4] $p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})$ that is an approximation of $p(\mathbf{t}|\mathbf{x})$ originating from the chain rule decomposition $p(\mathbf{t}, \mathbf{x}) = p_{\mathcal{D}}(\mathbf{x})p(\mathbf{t}|\mathbf{x})$, where $p_{\mathcal{D}}(\mathbf{x})$ corresponds to the empirical data distributions of the original printed codes. The training of the estimation network $p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})$ is performed based on the maximisation of the mutual information $I_{\boldsymbol{\phi}}(\mathbf{T}; \mathbf{X})$ between $\mathbf{x}$ and $\mathbf{t}$ via $p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})$:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\, I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T}) = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\phi}), \tag{1}$$

[4]The code is available at https://github.com/romaroman/cdp-ml-fakes.

where $\mathcal{L}(\boldsymbol{\phi}) = -I_{\boldsymbol{\phi}}(\mathbf{T}; \mathbf{X})$.

It was shown in [6] that the mutual information can be lower bounded as $I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T}) \geq I^L_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T})$, where:

$$\begin{aligned} I^L_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T}) \triangleq & \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})}\left[\log p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})\right]\right]}_{\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}} \\ & - \underbrace{D_{\mathrm{KL}}\left(p_t(\mathbf{t}) \| p_{\boldsymbol{\phi}}(\mathbf{t})\right)}_{\mathcal{D}_{\mathrm{t}}} \triangleq -\mathcal{L}^L(\boldsymbol{\phi}), \end{aligned} \tag{2}$$

where $D_{\mathrm{KL}}\left(p_t(\mathbf{t}) \| p_{\boldsymbol{\phi}}(\mathbf{t})\right) = \mathbb{E}_{p_t(t)}\left[\log \frac{p_t(\mathbf{t})}{p_{\boldsymbol{\phi}}(\mathbf{t})}\right]$ is a Kullback–Leibler divergence between the true $p_t(\mathbf{t})$ and the posterior $p_{\boldsymbol{\phi}}(\mathbf{t})$.

The final minimization problem reduces to:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\, \mathcal{L}^L(\boldsymbol{\phi}) = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} -(\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}} - \mathcal{D}_{\mathrm{t}}). \tag{3}$$

Remark: The term $\mathcal{D}_{\mathrm{t}}$ can be implemented based on the density ratio estimation [7]. The term $\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}$ can be defined explicitly using Gaussian priors as: $p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x}) \propto \exp(-\lambda \|\mathbf{t} - g_{\boldsymbol{\phi}}(\mathbf{x})\|_2)$ with a scale parameter λ, which leads to ℓ_2-norm and $g_{\boldsymbol{\phi}}(\mathbf{x})$ denotes the parameterized mapper.

The model for digital template estimation E as shown in Fig. 1 is built on a well-known U-Net architecture [8] and we refer to it as a template estimation model. The estimator model can be both stochastic and deterministic. The stochastic estimator assumes that the independent additive Gaussian noise with the $std = 0.001$ is added to the input before the estimation while the deterministic one does not inject any noise. The results of the obtained estimated and printed CDP are shown in Fig. 4.

B. Authentication strategy

The general scheme of the authentication is shown in Fig. 5. The authentication has to establish whether the probe $\mathbf{y}_i$ representing the original CDP $\mathbf{x}_i$ or fake $\mathbf{f}_i$ is authentic with respect to the template $\mathbf{t}_i$ or not.

The probe $\mathbf{y}_i$ is pre-processed to estimate the template $\mathbf{t}_i$ via the mapping $\mathbf{a}_i = g_\theta(\mathbf{y}_i)$. The parameters of mapping θ can be learnable at the training stage or estimated from data at the moment of the authentication, e.g. like Otsu's thresholding. A similarity metric $\psi(\mathbf{t}_i, \mathbf{a}_i)$ is computed between the template $\mathbf{t}_i$ and the estimate $\mathbf{a}_i$.

The similarity score might include several metrics. In our study, we consider normalized Hamming distance for binary images (HAMMING), structural similarity index (SSIM) [9], Jaccard index (JACCARD) and normalized cross-correlation (CORR). The output vector $\mathbf{v}_i \in \mathbb{R}^4$ is a concatenation of the four above similarity scores. The support vector machine (SVM) classifier C is trained on the vector $\mathbf{v}_i$ to produce a decision about the probe $\mathbf{y}_i$.

V. Results and Discussion

A. Attack results

Besides the described template estimation method, we consider two base-line template estimation alternatives based on the LDA algorithm [10] and binarization based on Otsu's

TABLE I: The average probability of error P_{error} in % of estimation of digital templates from the scanned codes relatively to original template based on normalized Hamming distance. A_M is the attacking method, P_D is the original printer.

A_M	Density / P_D	30%	35%	40%	45%	50%
Otsu	HPI55	12.54 (±0.0)	14.95 (±0.0)	16.89 (±0.0)	18.18 (±0.0)	20.01 (±0.0)
	HPI76	12.88 (±0.0)	14.97 (±0.0)	16.5 (±0.0)	17.61 (±0.0)	18.13 (±0.0)
LDA	HPI55	6.97 (±0.02)	9.34 (±0.03)	11.9 (±0.01)	13.84 (±0.01)	16.34 (±0.01)
	HPI76	7.39 (±0.03)	9.57 (±0.03)	11.76 (±0.03)	13.73 (±0.02)	15.24 (±0.03)
$-\mathcal{D}_{\mathbf{t}\hat{\mathbf{t}}} + \mathcal{D}_{\mathbf{t}}{}^{deter}$	HPI76	**1.68 (±0.1)**	**2.58 (±0.1)**	**3.67 (±0.04)**	**5.12 (±0.15)**	**6.17 (±0.03)**
$-\mathcal{D}_{\mathbf{t}\hat{\mathbf{t}}} + \mathcal{D}_{\mathbf{t}}{}^{stoch}$	HPI55	1.68 (±0.02)	2.65 (±0.07)	4.04 (±0.04)	5.25 (±0.08)	7.57 (±0.03)

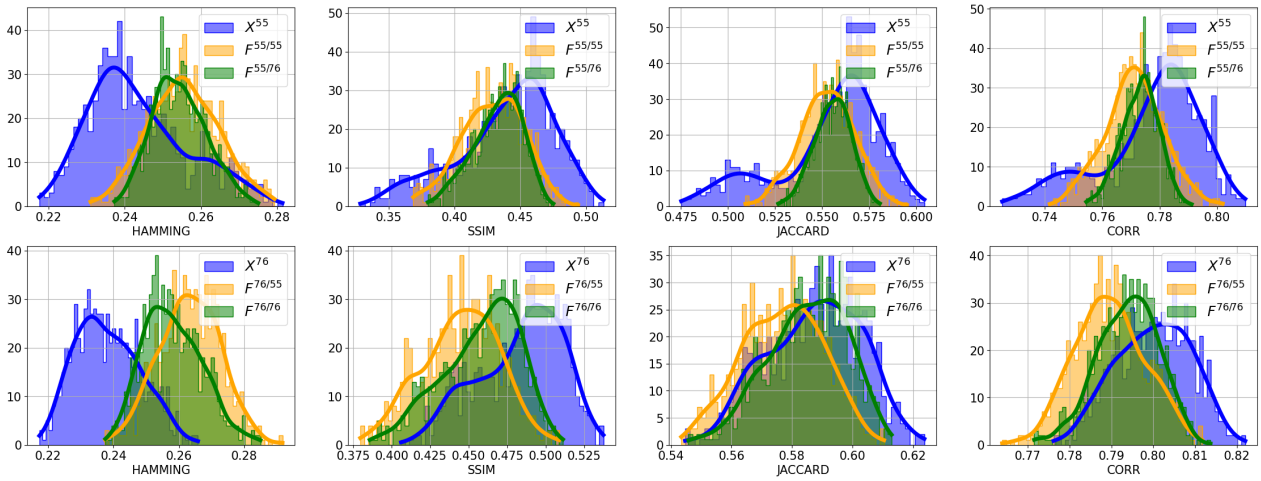

Fig. 6: The distributions of metrics and respective approximated kernel density estimation (solid line).

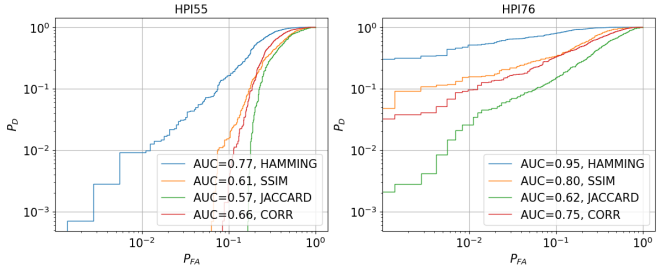

Fig. 7: ROC curves with the corresponding area under the curve (AUC) scores for all metrics. The curves for JACCARD, SSIM and CORR are calculated for $1 - \mathbf{s}$, where $\mathbf{s}$ is an original score.

adaptive thresholding [11] for benchmarking purposes. The accuracy of template estimation is measured as a normalized Hamming distance between the estimated template $\hat{\mathbf{t}}_i$ and its original counterpart $\mathbf{t}_i$. The template estimation results are presented in Table I as a corresponding P_{error} which denotes the Hamming distance between binarized estimated CDP $\hat{\mathbf{t}}_i$ and original digital template $\mathbf{t}_i$.

The comparison between the three estimation techniques indicates that the proposed ML estimation attack produces the highest accuracy for all code densities in comparison to both the Otsu and LDA estimates. Furthermore, the deterministic version of the proposed attack slightly outperforms its stochastic counterpart.

It is important to emphasize the role of code density

TABLE II: The results of the one-class SVM training using pairs of metrics M_1, M_2. P_D indicates the original CDP printer subset and P_A the fake CDP printer.

Metrics	P_D	P_A	P_{miss}	P_{fa}
CORR, JACCARD	HPI55	HPI55	**5.08 (±1.78)**	**42.70 (±12.71)**
		HPI76	31.94 (±7.50)	82.53 (±3.97)
	HPI76	HPI55	42.72 (±6.43)	11.90 (±6.46)
		HPI76	**5.29 (±1.39)**	**50.85 (±3.85)**
HAMMING, SSIM	HPI55	HPI55	**5.13 (±1.74)**	**23.53 (±15.46)**
		HPI76	93.35 (±5.01)	0.00 (±0.00)
	HPI76	HPI55	99.12 (±0.35)	61.05 (±6.43)
		HPI76	**5.05 (±1.52)**	**6.88 (±4.41)**

on attack estimation accuracy. The codes with the highest density of 50%, i.e., the codes with the highest entropy, are characterized by higher estimation error in comparison to the codes with the lower density.

Therefore, to address the most challenging task in the investigation of the cloneability of CDP, we proceed with the codes of density 50%, because we consider that if the authentication system cannot distinguish the fakes with a larger amount of errors, then it will be incapable to distinguish ones with a smaller amount of errors with respect to the original digital template.

At the second stage of the attack, the binary estimates of codes of density 50% were printed on the same printers HPI55 and HPI76 with the symbol size 1×1. The original scans of codes are denoted as $\mathbf{x}^{55}$ and $\mathbf{x}^{76}$ for HPI55 and HPI76,

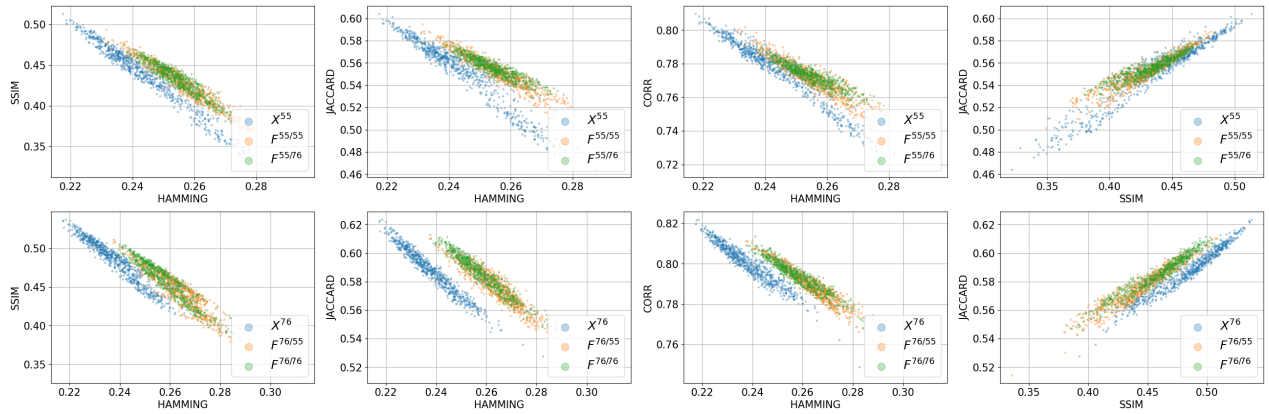

Fig. 8: The visualization of pairwise metrics, where each dot represents a particular sample.

TABLE III: The results of the one-class SVM training using all metrics. P_D indicates the printer used for the originals and P_A is the attacker's printer. Both P_{miss} and P_{fa} are calculated over the test subset $\mathbf{x}^{55}, \mathbf{f}^{55/55}$, $\mathbf{f}^{55/76}$ for P_A HPI55 and $\mathbf{x}^{76}, \mathbf{f}^{76/55}$, $\mathbf{f}^{76/76}$ for P_A HPI76.

P_D	P_A	P_{miss}	P_{fa}
HPI55	HPI55	**10.91 (±2.28)**	**26.39 (±8.90)**
	HPI76	99.76 (±0.15)	0.00 (±0.00)
HPI76	HPI55	100.00 (±0.00)	1.28 (±0.40)
	HPI76	**9.92 (±1.86)**	**0.00 (±0.00)**

respectively. The fakes of the printed $\mathbf{x}^{55}$ and $\mathbf{x}^{76}$ codes are then reprinted on the same printers and are denoted as $\mathbf{f}^{55/55}$, $\mathbf{f}^{55/76}$, $\mathbf{f}^{76/55}$, $\mathbf{f}^{76/76}$. One can observe some minor differences between printed codes produced by two printers in Fig. 2. At this stage, we were not able to establish a source of these differences due to the limited access to the tuning of the printers and not knowing the history of their prior usage.

The examples of the original codes and their fakes are shown in Fig. 2 and Fig. 4. Fig. 4 presents the results scanned by the scanner at the resolution 2400 ppi, while Fig. 2 shows the magnified prints acquired by Dino microscope. As one can conclude based on the visual inspection of images that the produced fakes very closely resemble the original CDP. Moreover, the dot size of fakes and originals is almost identical. However, it is obvious, that even there are some minor noticeable differences between the originals and fakes under the microscope examination, it is very important to verify, whether such differences can be reliably authenticated under the mobile phone or even scanner acquisition. The next section validates this question.

B. Authentication results

In this section, we investigate the performance of the authentication system represented in Fig. 5. The authentication is analyzed for the original CDP of density 50% printed on both printers and the fakes considered in the previous section. The performance validation includes three stages. First, we analyze the discrimination in each of the considered metrics $\psi(\mathbf{t}_i, \mathbf{a}_i)$ by plotting the corresponding histograms and receiver operating characteristic (ROC) curves. Second, we investigate the classification under the absence of information about the fakes based on a one-class classifier. Third, we consider the classification in the fully informed setting based on a supervised classifier.

The discrimination in different metrics is shown in Fig. 6. We plot the distribution of resulting scores between the reference templates $\mathbf{t}_i$ and original prints $\mathbf{x}^{55}$ versus the scores between $\mathbf{t}_i$ and fakes $\mathbf{f}^{55/55}$ and $\mathbf{f}^{55/76}$ and the same is done for HPI76. In all considered metrics, the fakes produced on the same printer as the original CDP are closer to the originals than the cross-printed ones[5].

The ROC curves produced for two printers HPI55 and HPI76 and the most similar fakes $\mathbf{f}^{55/55}$ and $\mathbf{f}^{76/76}$ are shown in Fig. 7. Although none of the considered metrics allows to reliably detect fakes, the Hamming distance between $\mathbf{t}_i$ and binarized $\mathbf{y}_i$ based on the Otsu's thresholding (HAMMING) produces the best results for both printers.

Since none of the considered metrics is capable to achieve satisfactory authentication accuracy of considered CDP under the proposed ML attack, we investigate a simultaneous usage of several scores. Fig. 8 shows a pair-wise separability of originals and fakes. The most promising separability plots are observed for the pair of *HAMMING* and *CORR*. For the printer HPI76 the set of original $\mathbf{x}^{76}$ is visually separable from sets of $\mathbf{f}^{76/55}$ and $\mathbf{f}^{76/76}$ fakes. However, such a phenomenon is not observed for the HPI55 printer.

Furthermore, before the training of SVM models, it is important to decide whether to continue with a whole set of metrics or with the best separable ones. To confirm this hypothesis the one-class SVM was trained[6] on the subset of 2 selected metrics, training subset consists only of original samples printed on a single printer (HPI55 or HPI76). The results are presented in Table II with P_{miss} denoting the

[5]It should be noted that in the case of the distribution of the original codes (blue histogram) printed on the HPI55 printer one can observe a small hump that deviates from the main peak. It is related to the instability in the printing quality. This is a very important factor that might seriously impact the CDP authentication and should be properly addressed in the future investigations.

[6]Training set size = 144 samples, number of runs with different seeds = 20, $\gamma = 0.3$, kernel = RBF, for one-class SVM $\nu = 0.01$.

TABLE IV: The results of the supervised two-class SVM training using all metrics. P_D indicates the printer of the training subset and P_A the printer of the test subset.

Trained on	P_D	P_A	P_{miss}	P_{fa}
x^{55}, $f^{55/55}$	HPI55	HPI55	**2.87 (±0.63)**	**2.25 (±0.71)**
		HPI76	47.73 (±5.60)	3.86 (±2.41)
x^{55}, $f^{55/76}$	HPI55	HPI55	**2.16 (±0.58)**	**5.24 (±0.80)**
		HPI76	27.54 (±2.86)	22.59 (±5.14)
x^{76}, $f^{76/55}$	HPI76	HPI55	9.99 (±10.68)	79.81 (±14.86)
		HPI76	**0.42 (±0.36)**	**0.35 (±0.23)**
x^{76}, $f^{76/76}$	HPI76	HPI55	5.12 (±9.50)	90.39 (±14.43)
		HPI76	**0.20 (±0.34)**	**0.22 (±0.23)**

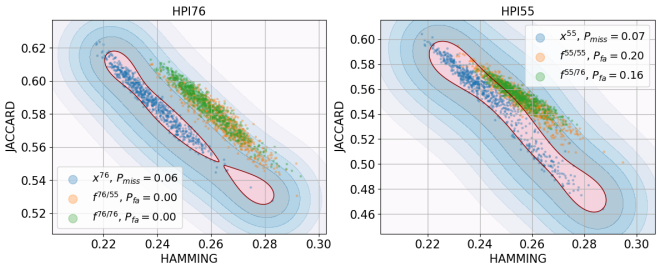

Fig. 9: The visualization of one-class SVM decision regions.

probability of classifying the original CDP as fake and P_{fa} the probability of classifying the fake CDP as original. The obtained results confirm that the one-class SVM trained on the target printer used to print the original CDP produces the best results for the corresponding samples. However, P_{miss} drastically increases, when the one-class SVM is used to the wrong printer.

The obtained results indicate that even if the authentication is performed on the digital scanner with 2400 ppi that is generally superior to the mobile phone acquisition, the reliable authentication of CDP is questionable under the proposed ML attack in the one-class SVM setup. The examples of corresponding decision regions are shown in Fig. 9.

To benefit from a full power of SVM the models were trained once again based on the whole set of metrics with the same hyperparameters and the results are present in Table III. The full set of metrics does not improve overall P_{miss} but achieves $P_{fa} = 0$ for HPI76, while P_{fa} for HPI55 is still high. These results are not satisfactory for practical applications since many originals will be recognized as fakes.

Finally, in the same manner, two-class supervised SVM models were trained as a fully informed system and the results are shown in Table IV. As expected the supervised approach yields much better P_{miss} and P_{fa}. In the best scenario, if SVM is trained on $\mathbf{x}^{76}$ and $\mathbf{f}^{76/76}$, then $P_{miss} = 0.20$ and $P_{fa} = 0.22$. Still, this result is not satisfactory for the large-scale industrial application and demands to have an access to a big variety of fakes. However, due to rapid technological growth one cannot guarantee that the classifier will be aware of all unseen fakes and possible attacks. Nevertheless, the trend of inferior accuracy for HPI55 printer is preserved even for the two-class supervised SVM.

VI. Conclusions

In this work we investigated the possibility to clone CDP with different codes densities produced on two industrial HP Indigo printers.

To perform attacking with the highest possible success chance we assume that the attacker possesses the same equipment and has an access to the original digital templates. This work proves that it is possible to obtain estimations with a relatively low probability of bit error over the whole set of code densities.

The faked CDP were acquired and processed in the same way as originals. The proposed classification system shows that high-quality fakes still preserve some information loss with respect to the originals and can be differentiated. However, the accuracy of authentication should be considerably enhanced for large-scale applications[7].

For future work, we aim at replacing considered SVM classifiers with deep neural networks. Also, it is important to investigate an opportunity to perform authentication without access to the original digital template. Finally, for a complete real-world setup simulation, it is important to acquire present CDP with the mobile phone whereas in this paper it is done on the images acquired by the high-resolution scanner.

References

[1] K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification*. Wiley Publishing, 2003.

[2] S. Voloshynovskiy, T. Holotyak, and P. Bas, "Physical object authentication: Detection-theoretic comparison of natural and artificial randomness," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2029–2033.

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

[4] O. Taran, S. Bonev, and S. 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] R. Yadav, I. Tkachenko, A. Trémeau, and T. Fournel, "Estimation of copy-sensitive codes using a neural approach," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, ser. IH and MMSec'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 77–82. [Online]. Available: https://doi.org/10.1145/3335203.3335718

[6] S. Voloshynovskiy, M. Kondah, S. Rezaeifar, O. Taran, T. Hotolyak, and D. Rezende, "Information bottleneck through variational glasses," in *NeurIPS Workshop on Bayesian Deep Learning*, Vancouver, Canada, December 2019.

[7] I. Goodfellow, "Generative adversarial nets," *arXiv:1406.2661*, 2014.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, May 2015.

[9] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] M. L. Diong, P. Bas, C. Pelle, and W. Sawaya, "Document authentication using 2d codes: Maximizing the decoding performance using statistical inference," in *Communications and Multimedia Security*, B. De Decker and D. W. Chadwick, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 39–54.

[11] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[7]Moreover, the performed experiments raise an important question of the impact of the printing variability on authentication accuracy. We aim at investigating this question in more detail in our future research.