ACM CODASPY 2022
This is a self-archived pre-print version of this article.
The final publication is available at ACM via
https://doi.org/10.1145/3508398.3519363.

Poisoning Attacks against Feature-Based Image Classification*

Robin Mayerhofer
mayerhofer1998@gmail.com
Vienna University of Technology
Vienna, Austria

Rudolf Mayer
rmayer@sba-research.org
SBA Research gGmbH & Vienna University of Technology
Vienna, Austria

ABSTRACT

Adversarial machine learning and the robustness of machine learning is gaining attention, especially in image classification. Attacks based on data poisoning, with the aim to lower the integrity or availability of a model, showed high success rates, while barely reducing the classifiers accuracy – particularly against Deep Learning approaches such as Convolutional Neural Networks (CNNs). While Deep Learning has become the most prominent technique for many pattern recognition tasks, feature-extraction based systems still have their applications – and there is surprisingly little research dedicated to the vulnerability of those approaches.

We address this gap and show preliminary results in evaluating poisoning attacks against feature-extraction based systems, and compare them to CNNs, on a traffic sign classification dataset. Our findings show that feature-extraction based ML systems require higher poisoning percentages to achieve similar backdoor success, and also need a consistent (static) backdoor position to work.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Supervised learning**;

KEYWORDS

Adversarial machine learning, Poisoning attacks, Feature-Based Image Classification

ACM Reference Format:
Robin Mayerhofer and Rudolf Mayer. 2022. Poisoning Attacks against Feature-Based Image Classification. In *Proceedings of the Twelveth ACM Conference on Data and Application Security and Privacy (CODASPY '22), April 24–27, 2022, Baltimore, MD, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3508398.3519363

1 INTRODUCTION AND RELATED WORK

Robustness and security in Machine Learning (ML) is critical due to the rapid growth and increased deployment of Machine Learning applications in real-life. Attacks can often be categorised to address the *confidentiality*, *integrity*, or *availability*, and performed either during the training or prediction phase of the process[2]. Research on attacks and defences primarily focused on deep learning. For example, [3] has shown that only small percentages of poisoned data are needed to embed *backdoors* successfully into Convolutional Neural Networks (CNNs), and thus attack its integrity or availability. Poisoned data contains a specific pattern added to inputs, such as a pixel combination superimposed on an image – a certain type of sunglasses on a face, or stickers on traffic signs. These are used in combination with purposefully labelling the samples wrongly, so that the model learns a wrong association. This backdoor can then be exploited to manipulate the prediction of ML based systems, e.g. to be recognised as a specific person when wearing the sunglasses. This is an imminent threat, due to the growing popularity of ML-based systems, e.g. in self-driving vehicles, or face recognition.

One assumption as to why poisoning attacks on CNNs are successful is that CNNs tend to overfit, and thus easily memorise the pattern and its intended association to a specific class. While it has been shown that the shape and colour of the patterns play a role in the success rate of the attack (e.g. [6]), in general, these attacks are very effective against CNNs.

The vulnerability of shallow learning was investigated in [1], where samples are manipulated to change the decision boundary of Support Vector Machines (SVMs). However, rather little attention has been put on the vulnerability of feature-based image classification, coupled with shallow learning approaches such as Random Forests or SVMs. It thus remains open whether this approach is similarly vulnerable as CNNs, or whether feature extraction already reduces the prominence of the backdoor pattern, and the shallow classifiers subsequently do not learn the backdoor association.

In this paper, we thus empirically compare the success of backdoors in feature extraction based machine learning pipelines to deep-learning ones, on traffic sign classification. Our experiments show that a consistent (static) position of the backdoor trigger is vital for these approaches, contrarily to CNNs (e.g. [3]).

2 DATASET AND EXPERIMENT SETUP

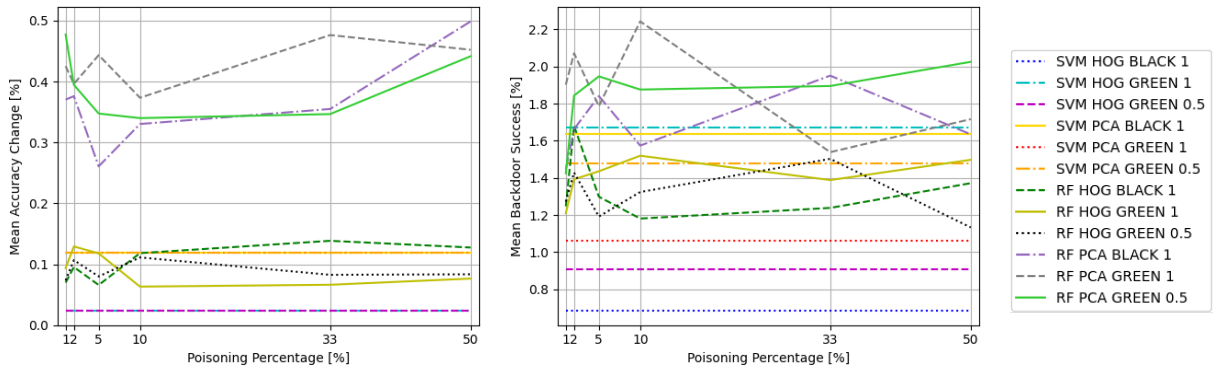
For our experiments, we utilise the *German Traffic Sign Recognition Benchmark* (GTSRB) [7], which consists of ~50,000 photos of 43 traffic signs types. The images are distributed very unevenly among the classes, from ~200 to more than 2,000 samples per class. The signs differ in size and aspect ratio. In line with [7], which also provides HoG features, we thus resize the images to the same 40×40 dimension before applying any feature extraction. Due to the fact that we also need HoG features of the poisoned images, we did not directly use the computed HoG features but computed them ourselves. The best found HoG configuration used a window size of (20, 20), block size of (10, 10), cell size of (5, 5) and signed gradients; we use nine bins, from 0° to 160° with a step size of 20°.

The backdoor trigger is a sticker of fixed size and colour that is placed on traffic signs; we use 0.5% and 1% of the image size

*SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna; COMET is managed by FFG.
This work is supported by BMK and FFG under Grant No. 873979 (project PRIMAL).

CODASPY '22, April 24–27, 2022, Baltimore, MD, USA

ACM ISBN 978-1-4503-9220-4/22/04.
https://doi.org/10.1145/3508398.3519363



(a) Mean accuracy change **(b) Backdoor success rate**

Figure 1: Mean of metrics of all classifier/feature extraction/pattern type combinations (positioned with small variations)

as trigger size for a green colour backdoor, and also investigate a 1% sized black pattern, as in [4]. We use poisoning percentages of 1/2/5/10/33 and 50%.

Following the terminology of [3], we implemented the one-to-one (OTO) backdoor, which focuses on backdooring one class (the "source" class) to another (the "target" class)[1]. We first analysed the semantic meaning of the traffic signs, and selected target classes that lead to adverse effects if the misclassification actually happens. Therefore, we evaluated among others the following one-to-one backdoors: (i) from "speed limit" to "the highest speed limit" (120) (ii) from "speed limit" to "end of the speed limit" (iii) from "curve left" to "curve right" (iv) from "stop" to various other ones (speed limit 20/50/80, deer crossing, traffic light, right of way).

3 RESULTS

For the baseline traffic sign classification task, we achieve results comparable to the state-of-the-art: [8] reports an accuracy of 96.14% with HoG Features and Random Forests; we obtain a slightly higher 96.18% with SVM, and a slightly lower 95.94% with Random Forests.

For the poisoning attack, we report four metrics: (i) overall accuracy of the classifier (ii) recall of the source class (iii) precision in the target class (iv) success rate of the backdoor (i.e. percentage of poisoned samples classified to their intended target class).

Figure 1 summarises the results, by providing averages from all OTO backdoor combinations when placing the backdoor-trigger randomly in an area around the centre of the image, with a 5% potential shift in the exact position. All three backdoor triggers result in barely any impact regarding our measured metrics, as shown in Figure 1. This happens for both tested feature extraction methods (HoG and PCA), as well as for both classifiers (SVMs and Random Forests). In detail, we can see that the mean accuracy (left plot) does not decrease, as it would be expected when training on partially poisoned data. This is also confirmed when inspecting in detail the impact on the source and target class, which would be affected the most. For the source class recall, the decrease does not exceed 1.5%, and similarly, the effect on the target class precision is dropping just by 1.7% (not depicted). Most relevant for the attacker, the backdoor success (right plot in Figure 1) does not even reach 2.3%. This is in contrast to reference results when attacking a CNN, which reaches near-perfect attack success, e.g. [4, 5].

The poor results for all three backdoor triggers seem to be due to the (semi-)random position around the centre. In combination with the feature extraction methods and chosen classifiers, this seems to prevent a backdoor from being embedded into the resulting model. For CNNs, it has been observed that the pattern position does not exert a significant influence (e.g. [3]).

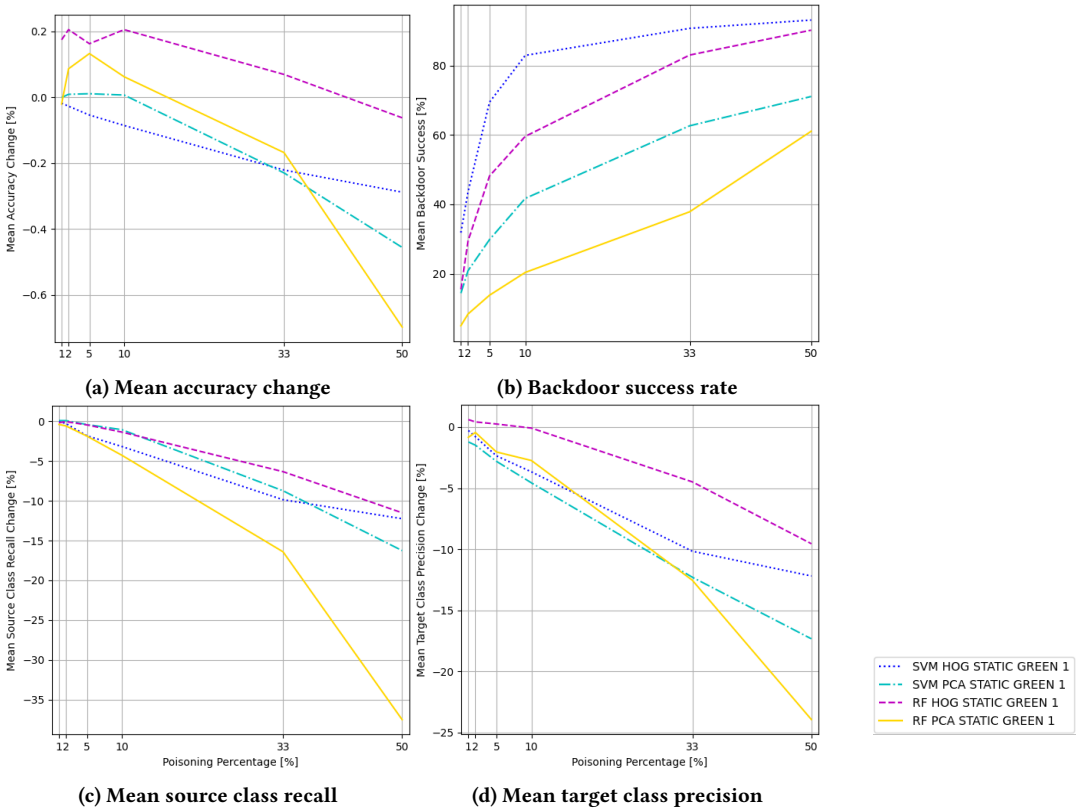
We thus also perform an attack with a static trigger position. Figure 2 shows again the results, averaged over the different OTO combinations. They confirm the assumption that the exact position of the backdoor is relevant for our classification pipelines. Here, the expected trend in backdoor success manifests - the success is growing with the increase of poisoning percentage. Further, poisoning has also a noticeable impact on overall accuracy, and especially on the source and target class performance, where the source class recall and the target class precision drop more than 10% at the highest poisoning percentage.

In detail, the mean accuracy change (cf. Figure 2a) is larger than for the random positioned pattern – but still in line with expected results. RFs work better than SVMs at poisoning percentages up to 33%, staying below a 0.2% drop. In terms of the feature extraction method, there is not one that is the clearly better one. Mean source class recall (cf. Figure 2c) drops less than 5% for up to 10% poisoning, but exceeds this level above. At 50%, for all tried configurations of classifier types and feature extraction method, the drop is between 11 and 17%, except for RFs using PCA, which drop more than 35%.

In case of the mean target class precision (cf. Figure 2d), we observe that the drops correlate, but are much larger than for the overall accuracy. All combinations, except the RFs using PCA, drop by approximately the same amount as the source class recall, ±3%. The outlier case from the source class recall, RFs using PCA, drops less, namely 24% in mean target class precision compared to more than 35% drop in mean source class recall.

The mean backdoor success (top right of Figure 2) is best for SVMs in terms of classifier type, and best for HoG features in terms of feature representation. At 1% poisoned data, it already exceeds

[1]In contrast, an all-to-all backdoor tries to backdoor all classes at once, but due to computational limitations, this was infeasible with the 43 classes of GTSRB



(a) Mean accuracy change **(b) Backdoor success rate**

(c) Mean source class recall **(d) Mean target class precision**

Figure 2: Mean of metrics of all classifier/feature extraction/pattern type combinations (static pattern position)

30% for SVMs using HoG features. The classifiers using the HoG features exceed a 90% backdoor success at 50% poisoning.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we evaluated the success of poisoning attacks on feature-based image classification systems on an image categorisation task. We found that SVM and Random Forest are also vulnerable to these attacks when using HoG features and PCA as representation. However, triggering the backdoor behaviour is not as easy as for CNNs. Furthermore, we found that the backdoor trigger position is important for our experiments and must be static, i.e. consistently placed on the same position. We believe that this is due to the feature extraction, e.g., for HoG features, the feature vector is influenced at different positions depending on the backdoor trigger position. Future work may evaluate whether backdoors moving across cells of the HoG feature extraction are the issue. To this end, also further evaluation on different datasets needs to be performed. We will moreover investigate whether dataset and feature-extraction dependent triggers, where the attacker tries to optimise the change in the feature vector, i.e., gradient histograms, work better from an attacker's perspective. If our assumption holds, then attacks on HoG feature extraction needs to be more sophisticated than backdoor attacks on CNNs – which in turn could enable new defences. For example, ensembles of classifiers using different HoG feature extraction hyperparameters may further reduce the impact of a non-100% statically placed backdoor trigger.

REFERENCES

[1] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning Attacks against Support Vector Machines. In *Proceedings of the 29th International Conference on Machine Learning (ICML)* (Edinburgh, UK). Omnipress, New York, NY, USA.

[2] Battista Biggio and Fabio Roli. 2018. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. *Pattern Recognition* 84 (2018). https://doi.org/10.1016/j.patcog.2018.07.023

[3] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access* 7 (2019). https://doi.org/10.1109/ACCESS.2019.2909068

[4] Florian Nuding and Rudolf Mayer. 2020. Poisoning Attacks in Federated Learning: An Evaluation on Traffic Sign Classification. In *10th ACM Conference on Data and Application Security and Privacy*. ACM, New Orleans LA USA. https://doi.org/10.1145/3374664.3379534

[5] Florian Nuding and Rudolf Mayer. 2022. Data Poisoning in Sequential and Parallel Federated Learning. In *ACM International Workshop on Security and Privacy Analytics (IWSPA)*. ACM, Baltimore, MD, USA. https://doi.org/10.1145/3510548.3519372

[6] Huma Rehman, Andreas Ekelhart, and Rudolf Mayer. 2019. Backdoor Attacks in Neural Networks – A Systematic Evaluation on Multiple Traffic Sign Datasets. In *Machine Learning and Knowledge Extraction*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-29726-8_18

[7] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *International Joint Conference on Neural Networks*. IEEE, San Jose, CA, USA. https://doi.org/10.1109/IJCNN.2011.6033395

[8] Fatin Zaklouta, Bogdan Stanciulescu, and Omar Hamdoun. 2011. Traffic sign classification using K-d trees and Random Forests. In *International Joint Conference on Neural Networks*. IEEE, San Jose, CA, USA. https://doi.org/10.1109/IJCNN.2011.6033494