Understanding the Process of Data Labeling in Cybersecurity

Tobias Braun
University of Liechtenstein
tobias.braun@uni.li

Irdin Pekaric
University of Liechtenstein
irdin.pekaric@uni.li

Giovanni Apruzzese
University of Liechtenstein
giovanni.apruzzese@uni.li

ABSTRACT

Many domains now leverage the benefits of Machine Learning (ML), which promises solutions that can autonomously learn to solve complex tasks by training over some data. Unfortunately, in cyberthreat detection, high-quality data is hard to come by. Moreover, for some specific applications of ML, such data must be labeled by human operators. Many works "assume" that labeling is tough/challenging/costly in cyberthreat detection, thereby proposing solutions to address such a hurdle. Yet, we found no work that specifically addresses the process of labeling *from the viewpoint of ML security practitioners*. This is a problem: to this date, it is still mostly unknown how labeling is done in practice—thereby preventing one from pinpointing "what is needed" in the real world.

In this paper, we take the first step to build a bridge between academic research and security practice in the context of data labeling. First, we reach out to five subject matter experts and carry out open interviews to identify pain points in their labeling routines. Then, by using our findings as a scaffold, we conduct a user study with 13 practitioners from large security companies, and ask detailed questions on subjects such as active learning, costs of labeling, and revision of labels. Finally, we perform proof-of-concept experiments addressing labeling-related aspects in cyberthreat detection that are sometimes overlooked in research. Altogether, our contributions and recommendations serve as a stepping stone to future endeavors aimed at improving the quality and robustness of ML-driven security systems. We release our resources.

CCS CONCEPTS

• **Security and privacy** → **Intrusion/anomaly detection**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Labeling, ML, Practitioners, User Study, Cyberthreat Detection

1 SUPPLEMENTARY FIGURES

This supplementary document extends our paper by providing the figures reporting the performance metrics of our evaluation (which we could not include in the main paper due to page limitations). For consistency, the numbering of the figures starts from the one of the main paper. Additional details can be found in our repository [1].

Training Size Impact (Baseline)

We report in Figs. 5 the complete results of the experiments discussed in §5.1 of the main paper (wherein only Accuracy was reported—corresponding to the current Fig. 5d).

Human Error Impact (Mislabeling / Poisoning)

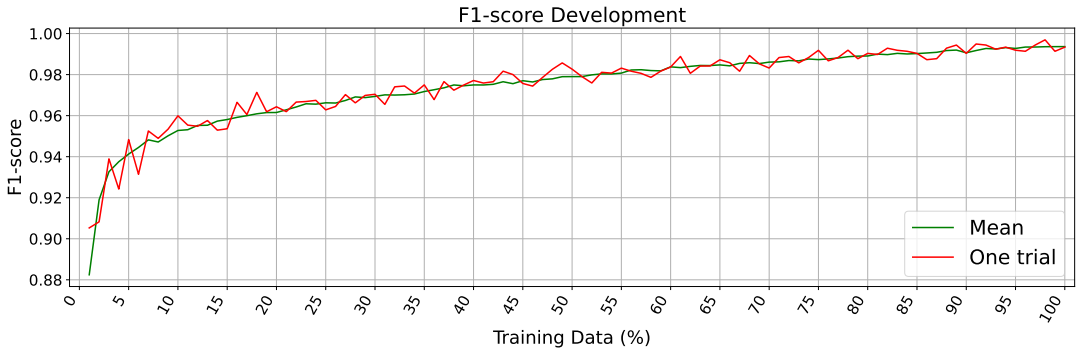
We report in Figs. 6 the complete results of the experiments discussed in §5.2 of the main paper (wherein only F1-score – corresponding to the current Fig. 6a – and absolute false positive were reported).

Active Learning

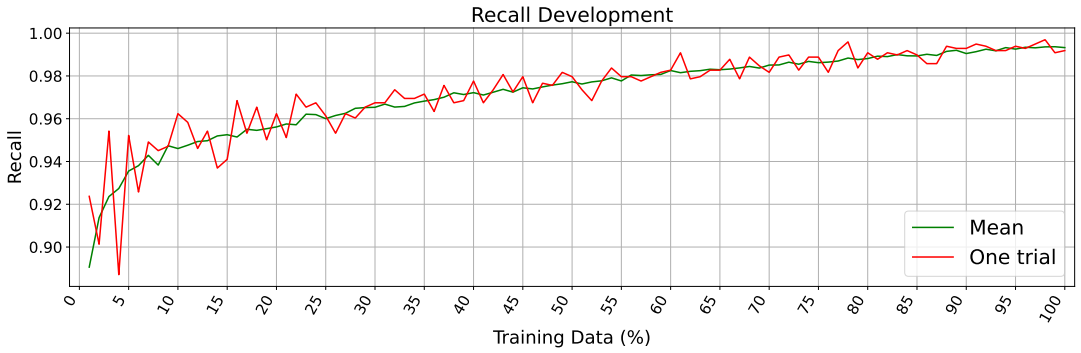
We report in Figs. 7 (for $\mathbb{U}$=100% of $\mathbb{T}$), Figs. 8 (for $\mathbb{U}$=80% of $\mathbb{T}$), and Figs. 9 (for $\mathbb{U}$=50% of $\mathbb{T}$) the complete results of the experiments discussed in §5.3 of the main paper (wherein only Accuracy for 100% of $\mathbb{T}$ was reported—corresponding to the current Fig. 7d).

REFERENCES

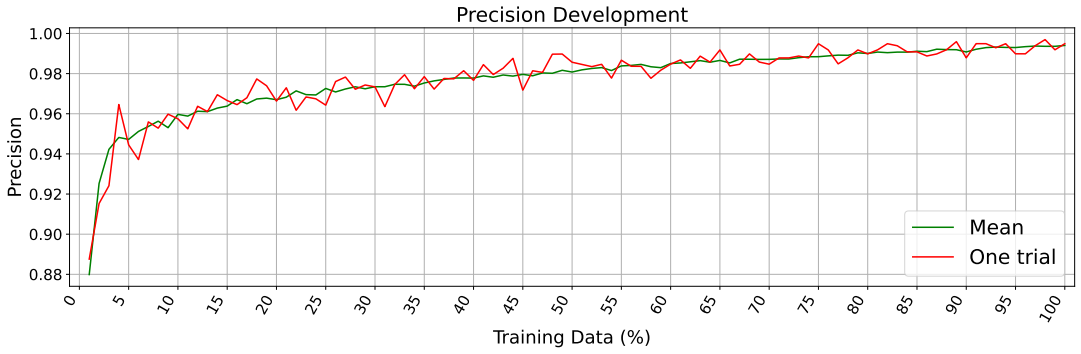
[1] 2024. *Our Repository.* https://github.com/hihey54/sac24_labeling



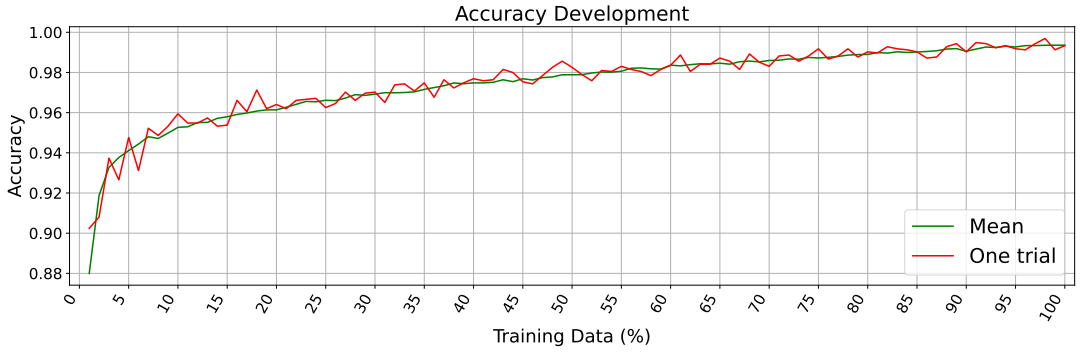
(a) Effect on F1-score (averaged over 30 trials).



(b) Effect on Recall (averaged over 30 trials).

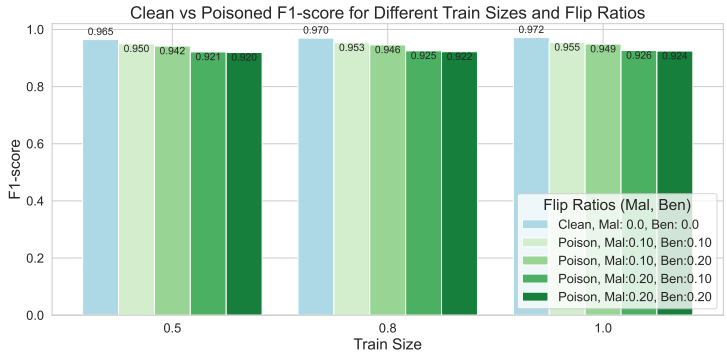


(c) Effect on Precision (averaged over 30 trials).

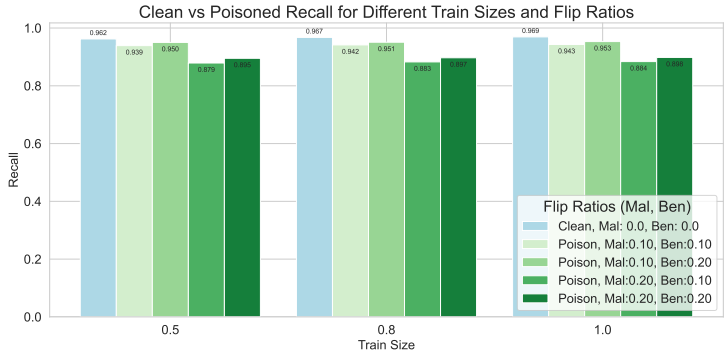


(d) Effect on Accuracy (averaged over 30 trials).

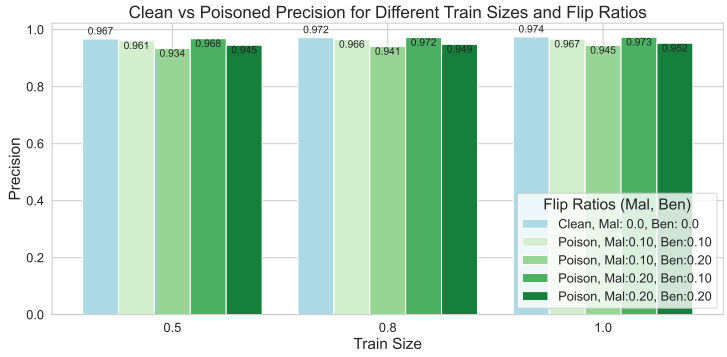
Fig. 5: Performance as a function of the training size. We further compare the average performance (over 30 trials) w.r.t. a single run.



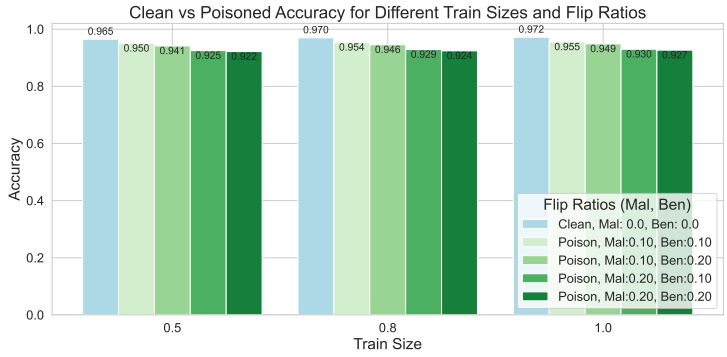
(a) Effect on F1-score (averaged over 30 trials).



(b) Effect on Recall (averaged over 30 trials).

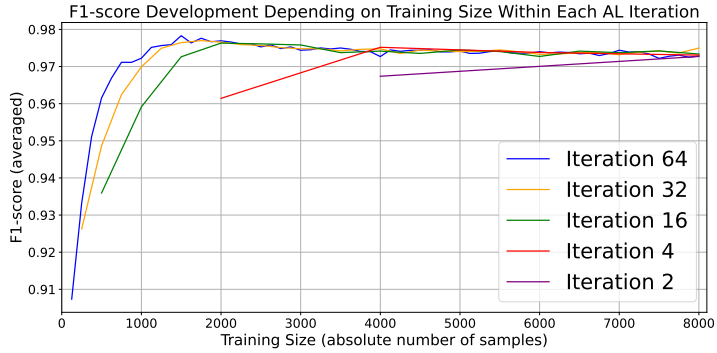


(c) Effect on Precision (averaged over 30 trials).

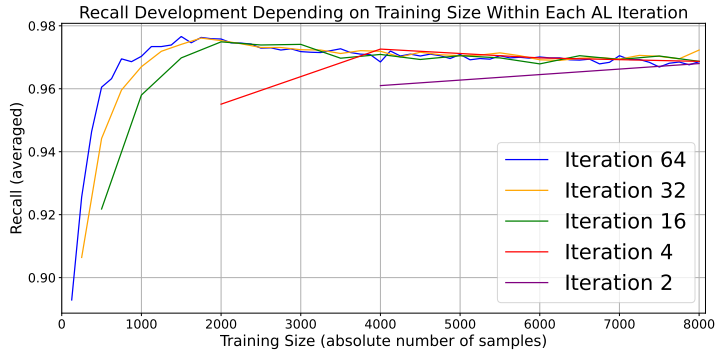


(d) Effect on Accuracy (averaged over 30 trials).

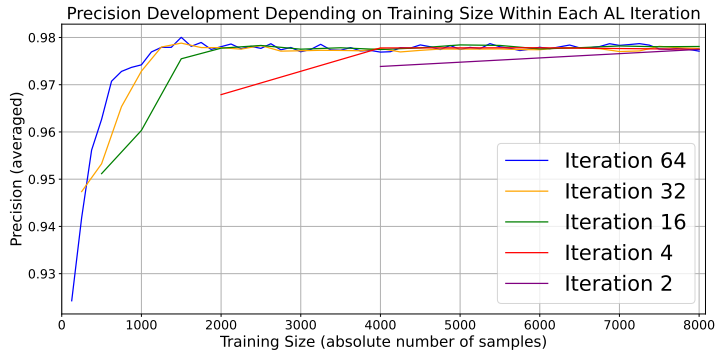
Fig. 6: Impact of mislabeling. We simulate human error by flipping the label of some subsets of the training data to see how much the performance changes.



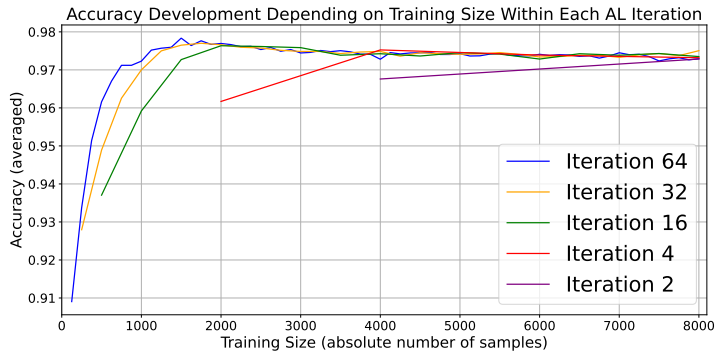
(a) Effect on F1-score (averaged over 100 trials).



(b) Effect on Recall (averaged over 100 trials).

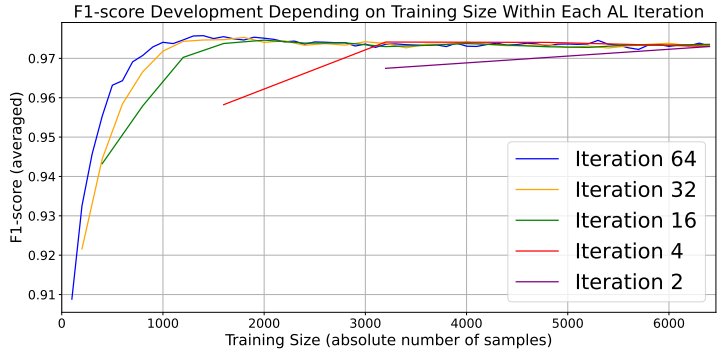


(c) Effect on Precision (averaged over 100 trials).

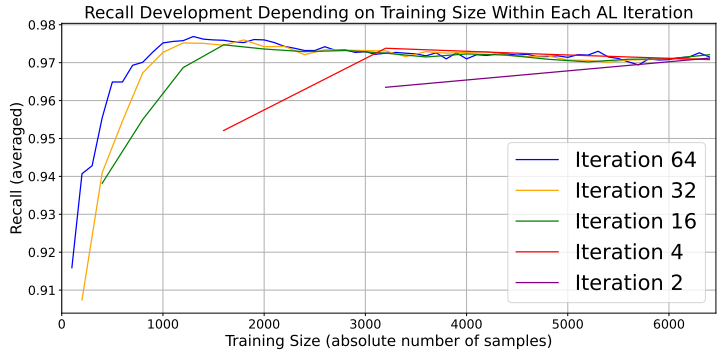


(d) Effect on Accuracy (averaged over 100 trials).

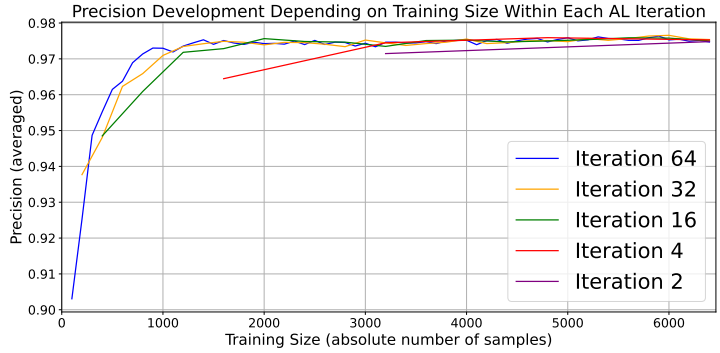
Fig. 7: Impact of Active Learning ($\mathbb{U}$=100% of $\mathbb{T}$). We compare the gains of labeling the suggested samples "all together" w.r.t. doing so over many iterations—each done by updating the model and suggesting new samples. These figures correspond by having an $\mathbb{U}$ of 8k samples (i.e., 100% of our $\mathbb{T}$), while $\mathbb{E}$ is 20% of $\mathbb{D}$.



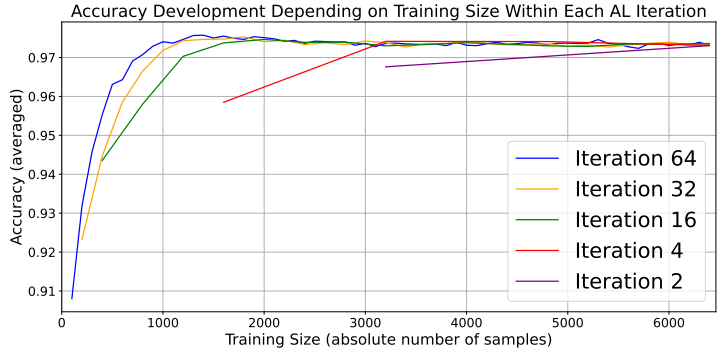
(a) Effect on F1-score (averaged over 100 trials).



(b) Effect on Recall (averaged over 100 trials).

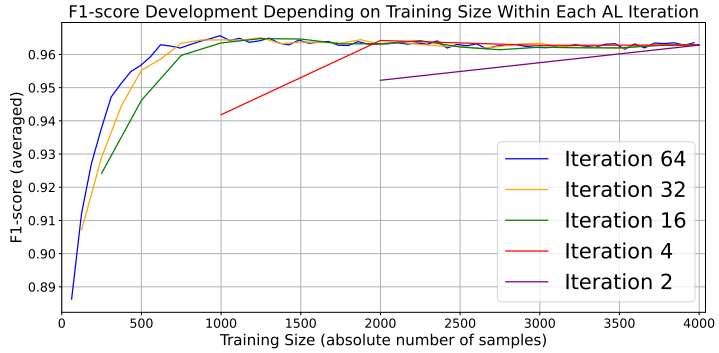


(c) Effect on Precision (averaged over 100 trials).

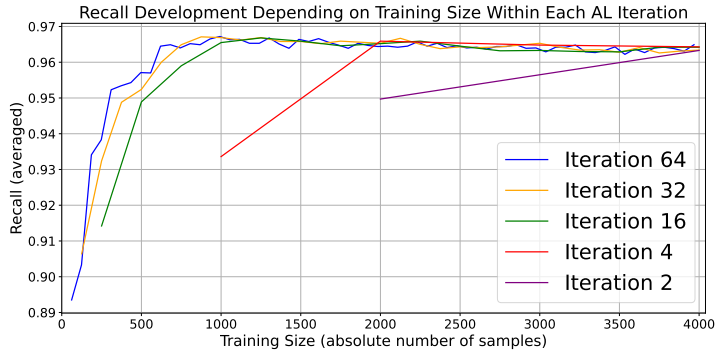


(d) Effect on Accuracy (averaged over 100 trials).

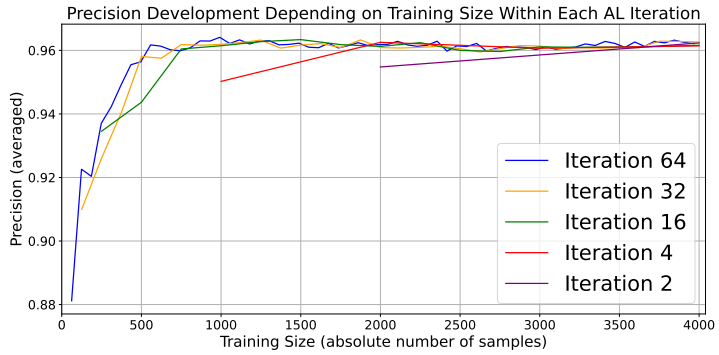
Fig. 8: Impact of Active Learning ($\mathbb{U}$=80% of $\mathbb{T}$). We compare the gains of labeling the suggested samples "all together" w.r.t. doing so over many iterations—each done by updating the model and suggesting new samples. These figures correspond by having an $\mathbb{U}$ of 6.4k samples (i.e., 80% of our $\mathbb{T}$), while $\mathbb{E}$ is 20% of $\mathbb{D}$.



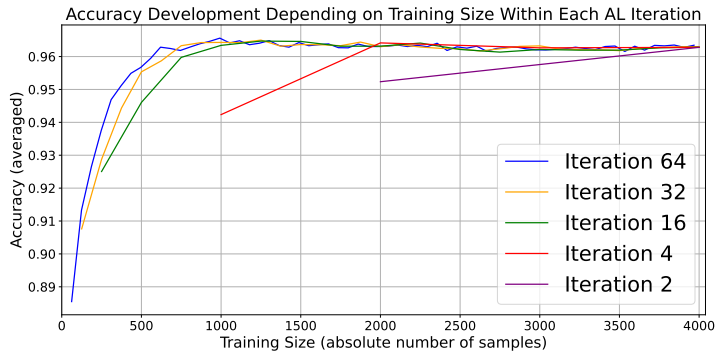
(a) Effect on F1-score (averaged over 100 trials).



(b) Effect on Recall (averaged over 100 trials).



(c) Effect on Precision (averaged over 100 trials).



(d) Effect on Accuracy (averaged over 100 trials).

Fig. 9: Impact of Active Learning ($\mathbb{U}$=50% of $\mathbb{T}$). We compare the gains of labeling the suggested samples "all together" w.r.t. doing so over many iterations—each done by updating the model and suggesting new samples. These figures correspond by having an $\mathbb{U}$ of 4k samples (i.e., 50% of our $\mathbb{T}$), while $\mathbb{E}$ is 20% of $\mathbb{D}$.