Computer Communications 233 (2025) 108052

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Just a little human intelligence feedback! Unsupervised learning assisted supervised learning data poisoning based backdoor removal ☆

Ting Luo [a,1], Huaibing Peng [a,1], Anmin Fu [a], Wei Yang [a], Lihui Pang [b,*], Said F. Al-Sarawi [c], Derek Abbott [c], Yansong Gao [d]

[a] School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing, JiangSu, China
[b] Sino-German College of Intelligent Manufacturing, Shenzhen Technology University, China
[c] School of Electrical and Electronic Engineering, The University of Adelaide, Australia
[d] Department of Computer Science and Software Engineering, The University of Western Australia, Australia

ARTICLE INFO

Keywords:
Backdoor attack
Human intelligence
Unsupervised clustering

ABSTRACT

Backdoor attacks on deep learning (DL) models are recognized as one of the most alarming security threats, particularly in security-critical applications. A primary source of backdoor introduction is data outsourcing such as when data is aggregated from third parties or end Internet of Things (IoT) devices, which are susceptible to various attacks. Significant efforts have been made to counteract backdoor attacks through defensive measures. However, the majority of them are ineffective to either evolving trigger types or backdoor types. This study proposes a poisoned data detection method, termed as LABOR (unsupervised **L**earning **A**ssisted supervised learning data poisoning based **B**ackd **Or** **R**emoval), by incorporating a little human intelligence feedback. LABOR is specifically devised to counter backdoor induced by dirty-label data poisoning on the most common classification tasks. The key insight is that regardless of the underlying trigger types (e.g., patch or imperceptible triggers) and intended backdoor types (e.g., universal or partial backdoor), the poisoned samples still preserve the semantic features of their original classes. By clustering these poisoned samples based on their original categories through unsupervised learning, with category identification assisted by human intelligence, LABOR can detect and remove poisoned samples by identifying discrepancies between cluster categories and classification model predictions. Extensive experiments on eight benchmark datasets, including an intrusion detection dataset relevant to IoT device protection, validate LABOR's effectiveness in combating dirty-label poisoning-based backdoor attacks. LABOR's robustness is further demonstrated across various trigger and backdoor types, as well as diverse data modalities, including image, audio and text.

1. Introduction

Deep learning (DL) has been widely adopted in numerous applications, including object detection, face recognition, fraud detection, self-driving systems, and shard scheduling problem [1], due to its exceptional performance [2]. This success has extended to artificial intelligence generated content (AIGC), leveraging large models such as text-to-image diffusion models [3] and large language models (LLMs) [4]. The increasingly complex model architectures are a key factor driving this impressive performance. However, as the number of model parameters scales up to billions, the benefits of solely enhancing model complexity diminish. It is now recognized that data quality plays an increasingly crucial role in determining model performance [5].

However, acquiring high-quality data is challenging due to the labor-intensive nature and domain knowledge requirements involved. This often necessitates data crowdsourcing or outsourcing through third-party platforms, such as Amazon Mechanical Turk[2] and Scale AI.[3] Despite its efficiency in collecting and annotating data, such data

☆ This work was supported by Research Center for Materials for Advanced MEMS Sensor Chip (No. 2022GCZX005).
* Corresponding author.
E-mail addresses: onelastick@njust.edu.cn (T. Luo), paloze@njust.edu.cn (H. Peng), fuam@njust.edu.cn (A. Fu), generalyzy@njust.edu.cn (W. Yang), sunshine.plh@hotmail.com (L. Pang), said.alsarawi@adelaide.edu.au (S.F. Al-Sarawi), derek.abbott@adelaide.edu.au (D. Abbott), garrison.gao@uwa.edu.au (Y. Gao).
[1] Contributed equally
[2] https://www.mturk.com/.
[3] https://www.scaleai.ca/.

https://doi.org/10.1016/j.comcom.2025.108052
Received 6 September 2024; Received in revised form 2 January 2025; Accepted 7 January 2025
Available online 18 January 2025

outsourcing is vulnerable to the security threat of data poisoning [6–9]. One primary objective of data poisoning attacks is to implant a backdoor into the DL model trained on the compromised data. Data poisoning-based backdoor attacks have been demonstrated to be perilous in various applications, including object detection [8], malware detection [10], and even in the rapidly emerging field of LLMs [11]. Also note that sensing data from Internet of Thing (IoT) devices is vulnerable to data poisoning because these devices often operate in untrusted environments with limited security, making it easier for attackers to inject false or malicious data that can corrupt the integrity of the system's decision-making processes.

Therefore, it is imperative to winnow poisoned data points in the training dataset, preferably as a once-off operation, so that the audited training dataset can be safely later for training. In this context, the training phase poisoned data detection studies [12–18] have been proposed to counter data poisoning by identifying and removing them (detailed in Section 2.5). These defenses have shown to be effective against different backdoor types and trigger types, however, the majority of them [14–18] rely on a (small) reserved clean dataset. Such a requirement may not be met in practice, rendering them fall short in such situations. Some detection methods [12,13] are only effective for universal backdoor types.

We note that these extant poisoned data detection studies all focus on an automated detection operation, which has achieved great success in defeating data poisoning based backdoor attacks. In contrast, human intelligence feedback has not been explored in detecting poisoned samples. Considering the success of human intelligence feedback in various other fields or applications Section 2.2, this work attempts to answer the following research questions.

(**R1**) Can human intelligence be helpful for identifying poisoned samples? (**R2**) If so, to what extent is it helpful?

We propose LABOR to explore human intelligence in detecting poisoned samples used for supervised classification tasks. The key is to first leverage unsupervised learning to group all data points into groups (the clusters are not labeled) and then to annotate each group with *little* human intelligence. In this context, we conclude that human intelligence is of great importance for detecting poisoned samples (**R1**). We show that LABOR is independent of trigger types as well as backdoor types as long as data poisoning is carried out through typical dirty-label poisoning (**R2**).

We summarize our contributions in threefold:

1. We propose LABOR, for the first time, exploiting (little) human intelligence in the loop of detecting poisoned samples.
2. We constructively utilize unsupervised learning to assist poisoned sample detection on supervised classification tasks, building upon little human intelligence feedback.
3. We extensively validate the efficacy and effectiveness of LABOR on 8 benchmark datasets in detecting dirty-label data poisoning regardless of trigger types, backdoor types, and even data modalities.

The remainder of this work is structured as follows. Section 2 presents the preliminaries and related work. Section 3 defines the threat model for LABOR and elaborates on its implementation. Section 4 provides comprehensive experimental validations of LABOR. Section 5 provides further discussion and additional experiments. Finally, Section 6 concludes the study.

2. Related work

2.1. Supervised learning and unsupervised learning

Supervised learning is a machine learning paradigm where models are trained on labeled datasets, which means that each training example is paired with a ground-truth label. The goal is to learn a mapping from inputs to outputs that generalizes well to unseen data. This paradigm is widely used for tasks such as classification and regression. Algorithms such as support vector machines (SVMs) [19], decision trees [20], and deep neural networks [2], depend on annotated data to determine decision boundaries or regression functions.

Unsupervised learning deals with data that do not have annotated labels. The objective is to discover the inherent structure within a set of data points. It is useful for tasks like clustering, association and dimensionality reduction. Algorithms such as k-means clustering [21], principal component analysis (PCA) [22], and hierarchical clustering [23], do not rely on annotated data, but instead identify patterns and relationships within the data itself. These techniques are particularly valuable when exploring new datasets or when labeling data is prohibitively expensive.

This work mainly uses clustering. It is one of the primary techniques in unsupervised learning, aiming to group a set of objects so that objects in the same group (or cluster) are more similar to each other than to those in other groups. For clustering algorithms, this work mainly uses MoCo [24] and k-means [21] clustering to identify potentially contaminated samples. MoCo is a contrastive learning-based clustering algorithm. It leverages a contrastive learning mechanism to map similar data samples into a close representation space, thus achieving data clustering. MoCo employs a dynamic 'queue' mechanism and a momentum encoder to maintain consistent representations, enabling it to capture similarities between samples without labels. This makes MoCo well-suited for extracting structured information from unlabeled data. k-means is a classical unsupervised clustering algorithm that partitions data into a predefined number of clusters (k) by minimizing the distance of each data point to its assigned cluster center. Through iterative adjustments of the cluster centers, k-means continues until all data points are stably assigned. It is widely used for preliminary clustering tasks in data analysis. Common clustering algorithms include k-means [21], DBSCAN [25], and Gaussian Mixture Models (GMM) [26]. These methods are widely used in various fields, such as market research [27], bioinformatics [28], and image segmentation [29]. The choice of clustering algorithm depends on the nature of the data and the specific requirements of the task. For example, k-means is effective for well-separated clusters, while DBSCAN is robust to noise and outliers [30]. Now clustering also proposes more emerging clustering approaches such as IK-USPEC [31], an ultra-scalable spectral clustering algorithm that integrates the Isolation Kernel to improve handling of datasets with heterogeneous densities.

2.2. Human intelligence

Incorporation of human intelligence feedback has multiple advantages. Firstly, human experts can discover and fix the loopholes and errors in the model in a timely manner by reviewing and analyzing the model output results [32]. Secondly, human intelligent feedback can help the model better understand and handle abnormal situations, improving the robustness and adaptability of the model [33]. Finally, human intelligent feedback can continuously improve the performance and reliability of the model through continuous monitoring and optimization of the model [34].

Human intelligence feedback is widely used in several fields. For example, in medical diagnosis, human experts can ensure the accuracy and reliability of the diagnosis by reviewing the diagnostic results of the model [35]. In automated driving, human experts can improve the safety and reliability of the automated driving system by supervising the decision-making process of the model [36]. In addition, human intelligence feedback plays an equally important role in areas such as financial analyses and smart homes [32]. It has recently been used to align behaviors of large language models [37].

2.3. Backdoor attacks

The backdoor attack causes the infected backdoor to misbehave according to the attacker's willingness when input contains the secret trigger [6,38]. However, the backdoored model functions normally as its clean model counterpart in the absence of the trigger. The backdoor can be implanted through data poisoning [8,39] e.g., under data outsourcing or/and model training regularization, e.g., under model outsourcing [40].

•**Backdoor Type.** The most well-studied backdoor type is the universal backdoor or source-class-agnostic backdoor. Regardless of the sample's source classes, any sample containing the trigger will trigger the universal backdoor [41]. The partial backdoor or source-class-specific backdoor is only activated when the sample is from a source class and contains the trigger at the same time. The partial backdoor will not be fired even if the sample contains the trigger but is from a non-source class [42,43]. There are other backdoor-type variants upon the universal backdoor or partial backdoor.

The All-to-All (A2 A) attack is a variant of a partial backdoor, where a sample from the, e.g., i_{th} class will be misclassified into the target $(i+1)_{\text{th}}$ class [44,45]. This means that the backdoor effect depends on the source class. The Multiple Trigger Multiple Backdoor (MTMB) attack is a variant of a universal backdoor. An attacker implants multiple backdoors in the model, each backdoor is associated with a trigger. Any sample carrying a trigger will hijack the infected model and misclassify it into the trigger corresponding backdoor (e.g., targeted label) [46].

•**Trigger Type.** A trigger is embedded in the input and can activate the backdoor in an infected model once the trigger-carrying input is processed by the model. For image-based modalities, the trigger might be a patch [44], which can be placed at a fixed location or a dynamic one [47]. Triggers can be either visible or invisible [41]. While most existing studies focus on digital triggers, natural objects [8] and phenomena can also serve as triggers [48]. It is important to note that a new type of trigger does not necessarily result in a new backdoor type. More specifically, backdoor type and trigger type are orthogonal concepts—different trigger designs can be employed to achieve the same backdoor effects, such as universal or partial backdoor effects.

2.4. IoT under backdoor attacks

The Internet of Things (IoT) refers to a network of interconnected devices that collect and exchange data, playing a pivotal role in various applications, from smart homes to industrial automation. As IoT devices proliferate, they become attractive targets for cyber attacks, including backdoor attacks. These devices often operate in untrusted environments, making them particularly vulnerable to data poisoning attacks that can implant malicious backdoors within their machine-learning models.

The importance of securing IoT systems against backdoor attacks cannot be overstated. These attacks can compromise the integrity of IoT applications, leading to unauthorized access and manipulation of critical data. Recent studies [49] have highlighted the inherent vulnerabilities of IoT devices, emphasizing the need for robust defenses to mitigate these risks. For example, it has been shown [50] that many IoT devices lack sufficient security measures, making them easy targets for attackers to exploit weaknesses in their machine-learning models.

To illustrate the growing concern surrounding IoT vulnerabilities, recent work has focused on developing methodologies to detect and prevent data poisoning in these devices. Such studies emphasize the necessity of incorporating advanced defense mechanisms, like LABOR, to enhance the resilience of IoT systems against sophisticated attacks.

2.5. Training phase backdoor detection

In the context of data outsourcing, backdoors are often introduced through data poisoning. To counteract this, two main strategies are employed: prevention and detection. Generally, prevention methods aim to train a clean model on a poisoned training dataset [51,52]; however, they struggle to accurately identify poisoned data points. Consequently, these methods are unsuitable when the same dataset is used to train different models.

In contrast, detection methods focus on making the poisoned training dataset reusable by accurately identifying and removing these poisoned data points. Such detection methods are crucial, as they allow data curators to thoroughly cleanse the data before it is made available for use. Significant efforts have been made in this area of research.

The Spectral Signature method [12] detects poisoned samples by analyzing the covariance spectral characteristics of the model's potential representations. Similarly, the Spectre [14] effectively identifies and removes poisoned samples by amplifying their spectral characteristics using clean samples. However, both of these methods require knowledge of the poisoning rate to set a threshold for removing poisoned samples based on anomaly scores, which may not be feasible in practice.

Other approaches include AC [13], which applies 2-means clustering to analyze the activation of hidden layers and detect categories affected by poisoning. The SCAn method [15] decomposes image representations into identity and change components, detecting the presence of multiple identity vectors within categories to identify poisoned samples. While effective for partial backdoors, SCAn is sensitive to dynamic trigger designs. These methods typically rely on the first-moment differences between benign and triggered samples in potential representations. Beatrix [16] approaches trigger sample detection as an out-of-distribution detection problem by performing higher-order statistical analysis in the Gram feature space. The CT method [53] trains a model on a weighted combination of clean and poisoned data with random labels, preventing the model from fitting the clean portion and thus identifying labeled, consistent poisoned samples. The ASSET method [18] first minimizes the loss on clean data and then maximizes the same loss on the entire dataset to amplify the difference between poisoned and clean samples, facilitating the identification of poisoned samples. The TellTale [54] leverages discernible trajectory between poisoned and benign samples to identify poisoned ones, in which differentiation is enhanced through spectrum transformation.

We note that all of these detection methods [13,15,16,18,53,54] primarily focus on automated processes, with none incorporating human intelligence feedback into the loop. It remains unclear whether human intelligence could be beneficial in detecting poisoned data points, and how to effectively integrate human expertise into the detection process has yet to be explored.

3. LABOR

This section begins with the definition of the threat model. Subsequently, we present an overview and then detail the specific steps of its implementation.

3.1. Threat model

Attacker. This work focuses on the data outsourcing scenario, where the training dataset is sourced from third parties. In this context, the attacker can poison a portion of the dataset before sending it to the data curator; however, the attacker has no control over the training process. The dataset is intended for supervised classification tasks, and we assume the attacker employs dirty-label poisoning. This type of poisoning is particularly concerning because it can achieve a high attack success rate with a low poisoning rate. Although there is an

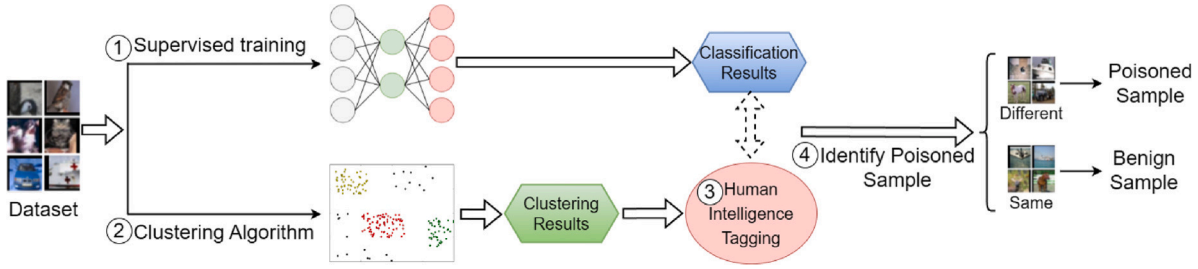

Fig. 1. The overview of LABOR.

inconsistency between the label and its content, the trigger can be imperceptible, and the poisoning rate can be minimal. As a result, if the poisoned samples are visually detected, the attacker can dismiss them as noisy data points. Given that human inspection of each sample is prohibitively expensive, this strategy is especially effective. We also consider the implications of clean-label poisoning as an alternative attack vector, where the attacker injects samples with imperceptible triggers but maintains consistent labels. Although clean-label poisoning often presents a lower attack success rate compared to dirty-label poisoning, its subtler nature can evade human inspections and some defenses. But it can be effectively defeated through SOTA defenses such as ASSET [18]. We propose complementarily employing LABOR and SOTA defenses to mitigate a diverse range of attack vectors including clean-label poisoning, which details are deferred to Section 5.7.

Defender. As for the defender, who is the data curator, the defender aims to detect poisoned samples once-off, so that the cleansed training dataset can be reused later on. The defender has full access to the training dataset and controls the training process. However, unlike extant poisoned data detection methods [14–18] that rely on a (small) reserved clean dataset, we relax the assumption that the defender has no such reserved clean dataset. In addition, the attacker has no prior knowledge of the trigger type and backdoor type that the attacker exploits.

3.2. Detection overview

The LABOR method constructively integrates unsupervised learning to enhance the detection of poisoned data points in supervised learning tasks. It achieves this by leveraging minimal human intelligence feedback to effectively annotate the clusters identified through unsupervised learning. The overview of LABOR is presented in Fig. 1. LABOR consists of the following steps:

Supervised Training. Initially, a dataset containing various types of images or other modal data is prepared. This dataset is used to train and test the model. Step ① trains the classification model through supervised learning without any specific regularization. In case the training dataset is poisoned, the trained model is backdoored and will misclassify the poisoned points into the targeted label.

Unsupervised Training. In step ②, the same training dataset is input into a clustering algorithm, a widely used unsupervised learning method. Since the number of categories is known to the data curator, the number of clusters is set to match the number of categories. The clustering algorithm organizes the data points in an unsupervised manner, revealing the natural distribution and structure within the dataset. The resulting clusters are then used to compare and evaluate the inference outcomes of the supervised learning model for each data sample.

Human Intelligence Feedback. The clustered data points from step ② do not initially have labels. The LABOR③ incorporates human intelligence by manually tagging these clusters. This tagging process requires minimal effort, as the data curator only needs to randomly

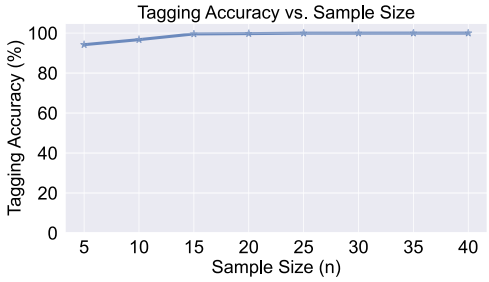

Fig. 2. Tagging accuracy as a relationship of the manually inspected number of samples in a cluster. Clustering accuracy is 80%.

inspect a few data points within each cluster to assign its ground-truth label accurately. To formulate, suppose the clustering accuracy is p, the curator randomly selects n samples, the tag of i_{th} label is determined if the majority of samples belong to the i_{th} label. In this process, the number of samples n chosen for incorporating human intelligence feedback varies based on the cluster's characteristics, with large or diverse clusters requiring 15–20 samples, while small or homogeneous clusters need only 5–10 samples for efficient tagging. The probability of accurate tagging P_{tagging} is expressed as:

$$P_{\text{tagging}} = \sum_{k=\lceil \frac{n+1}{2} \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

where $\lceil \frac{n+1}{2} \rceil$ represents the minimum number of correctly tagged samples required to satisfy the majority rule (i.e., more than half of the samples must be correctly tagged). $\binom{n}{k}$ is the binomial coefficient, representing the number of combinations of choosing k correctly tagged samples out of n samples. p^k is the probability that exactly k samples are correctly tagged, while $(1-p)^{n-k}$ is the probability that the remaining $n-k$ samples are incorrectly tagged. As one example, in Fig. 2, we plot tagging accuracy as a relationship of n on the dataset of CIFAR-10 when the clustering accuracy p is 80%. We can see that by manually examining no more than 15 samples in a cluster, the tagging accuracy is almost 100% (particularly, 99.6%). To further optimize this process, priority is given to clusters with a larger share of the overall dataset to reduce potential labeling errors' impact on LABOR's overall performance. For high-priority clusters, the feedback is weighted up to 70%, ensuring accurate tagging of major data distributions. Additionally, clusters, where the majority of feedback samples (over 80%) share a common label, will have their labels automatically updated across the cluster, improving both efficiency and consistency. Human feedback, through initial tagging and iterative adjustment, enhances LABOR's effectiveness in backdoor detection by reinforcing model precision and adaptability. Meanwhile, an overload of humanly inspecting less than 20 samples per clustering is acceptable for enhancing security, compared to thousands of samples in a given cluster (e.g., for the CIFAR10 dataset).

Identify Poisoned Sample. In step④, each sample is inferred by the classification model trained from the supervised learning and the

clustering model trained from the unsupervised learning at the same time. Each provides an inference label for this sample as the clusters have been tagged to their ground-truth. If both labels mismatch, then such a sample is regarded as a poisoned sample, otherwise, it is a benign sample. Here we argue that backdoor attacks can cause supervised learning to misclassify the poisoned samples and cause the corresponding samples to deviate from their original labels. However, poisoned samples are ineffective in backdooring clustering because the trigger has to be associated with a target label that is different from the ground-truth of the poisoned sample. Clustering learns without the guidance of the label but learns inherent features of classes. In addition, it is noted that the trigger usually does not obscure the main feature of the poisoned sample. Therefore, comparing the consistency of the inference results of the models between the supervised learned and the unsupervised learned can filter out poisoned samples.

4. Evaluation

This section first describes experimental setups. We then evaluate LABOR under seven benchmark datasets across image, audio, and text modalities.

4.1. Experimental setup

Our operating system used for this experiment is Windows 10, the processors are NVIDIA 3070ti and NVIDIA 3090, RAM is 8G, and the programming language is Python.

4.1.1. Dataset

CIFAR10 [55] is a natural color image dataset for object recognition. It consists of 60,000 $32 \times 32 \times 3$ RGB images with 10 classes. There are 50,000 training images and 10,000 test images in total.

MNIST [56] contains 10 types of handwritten digits from 0 to 9 with an image size of $28 \times 28 \times 1$. The number of training/testing images is 60,000/10,000.

SpeechCommand(SC) [57] is an audio dataset of spoken words designed to help train and evaluate keyword spotting systems. The SC contains many one-second `.wav` audio files, and each file has a single spoken English word. These words are from a small set of commands and are spoken by a variety of different speakers. Our test uses 10 classes from 'zero' to 'nine'. The number of training/testing audio is 11,360/9467.

ConsumerComplaint(CC) [58] is a dataset for consumers' complaints about financial products and services into different categories. CC originally had 18 classes. However, some classes are closely related to the other class, such as 'Credit reporting', 'Credit reporting, Credit repair services, or Other personal consumer reports'. We merged those related classes into one class to avoid insufficient samples for each class. In addition, we removed classes of 'Other finance service' or 'Consumer loan', as their samples are too few. Therefore, our test has 10 classes. The number of training and testing text samples is 100,773 and 10,000, respectively.

IMDB [59] is a dataset that has 50K movie reviews for natural language processing or text analytics, which is a binary sentiment classification task. It provides a set of 25,000 highly polar movie reviews for training and 25,000 for testing, where the total number of positive and negative movie reviews is both 25,000.

COVID_CT [60] is an open source medical dataset, which contains 349 COVID-19 CT images from 216 patients and 463 non-COVID-19 CTs. According to the senior radiologist, models with such performance are good enough for clinical usage.

CIC-IDS2017 [61] is a dataset that contains benign and the most up-to-date common attacks, which resembles the true real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files). The total number of data is 2,846,497 of which 227,3097 are benign samples.

Table 1
Clustering accuracy through unsupervised learning. Confidence intervals are shown in parentheses.

Dataset	Clustering method	Clustering accuracy
CIFAR10	MoCo	82.50% (±0.29%)
MNIST	MoCo	82.50% (±0.26%)
COVID_CT	MoCo	84.30% (±2.15%)
SC	k-means	85.87% (±0.79%)
CC	k-means	78.30% (±0.53%)
IMDB	k-means	86.37% (±0.46%)
CICIDS2017	k-means	89.81% (±0.39%)

4.1.2. Model architecture

Standard model architectures of VGG16 [62] and ResNet18 [63] are used in our image dataset evaluations. We note that in the benchmark datasets of SC, CC and IMDB, a 1D CNN model is provided, which we follow.

4.1.3. Clustering algorithm

For all image datasets, we use MoCo [24] as an unsupervised learning algorithm. In our experiments, we chose to set the initial learning rate to 0.06 for MoCo and used a cosine scheduling strategy to gradually reduce the learning rate. The batch size is 128, and this larger batch supports the generation of richer negative samples to further enhance the comparison effect—recall MoCo is contrastive learning. In addition, the momentum coefficient (0.99) and dictionary size (4096) are used to stabilize the encoder updates and maintain diverse negative samples, respectively. Experiments show that this configuration works well in maintaining the smoothness of model training. The feature dimension is chosen to be 128 to ensure the expressiveness of the contrast embedding space, while the temperature coefficient is set to 0.1 to balance the relative weights of the samples in the contrastive loss. For the rest of the audio and text datasets, we use k-means as an unsupervised clustering algorithm. We adopted the k-means algorithm and configured the key parameters as follows. The number of clusters (k) was experimentally determined to match the number of categories in the dataset. This alignment ensures that the number of clusters is consistent with the dataset's categories, allowing the clustering to better adapt to the diversity of the data and the distribution of the feature space. The initialization method chooses k-means++ to improve the convergence speed and stability of the algorithm; the maximum number of iterations is set to 200 and the tolerance is 0.0001 to ensure the balance between clustering accuracy and efficiency. The distance metric uses Euclidean distance.

Table 1 summarizes the clustering accuracy of each dataset. At the same time we give the corresponding confidence intervals after the data in the table.

4.1.4. Metric

Two main metrics of clean data accuracy (CDA) and attack success rate (ASR) are used to evaluate the backdoor attack and defensive performance.

CDA. The CDA quantifies the probability that a normal input, free of any triggers, is correctly classified by the backdoored model.

ASR. The ASR quantifies the probability that a trigger-carrying input is misclassified by the backdoored model into the target label predefined by the attacker.

The CDA of a backdoored model should match that of its clean counterpart, ensuring that validation accuracy alone does not reveal

Table 2
CDA of clean and backdoored models, and ASR of backdoored models. A white-square patch trigger is used.

Dataset + Model	Backdoor type	Clean	Backdoored	
		CDA	CDA	ASR
CIFAR10 + ResNet18	Universal	93.57%	93.01%	99.67%
MNIST + ResNet18	Universal	99.87%	99.78%	99.96%
COVID_CT + ResNet18	Universal	91.37%	91.20%	87.63%
CIFAR10 + ResNet18	Partial	93.57%	93.32%	97.75%
CIFAR10 + VGG16	Partial	96.57%	95.54%	99.04%

the presence of a backdoor. Conversely, the ASR should be high, ideally close to 100%.

When the poisoned dataset is cleansed using LABOR and the cleansed dataset is employed for training, the CDA of the resulting model should align with that of a model trained on a clean dataset. Simultaneously, the ASR of the model trained on the cleansed dataset should be low, rendering the backdoor ineffective.

4.2. Image modality

4.2.1. Universal backdoor

A white-square patch trigger is placed in the lower right corner of the image. Once the image is stamped with the trigger, its label is altered to the target label, which is the airplane class in the CIFAR10 dataset, digit 0 in MNIST, and the positive label in the COVID_CT dataset. The poisoning rate is 1% for CIFAR-10 and MNIST. As for COVID_CT, Since it only has 463 negative images, we randomly chose 10 images out of 463 to be poisoned and changed their labels to be positive.

The CDA and ASR of the backdoored models infected by Badnet are summarized in Table 2. The CDA of the backdoored model is always on par with that of the clean model, and the ASR is normally high—close to 100%, especially for CIFAR10 and MNIST.

Then we apply LABOR to detect and remove those poisoned samples. There are 5253, 6128, and 29 points removed from the training dataset of CIFAR10, MNIST, and COVID_CT, respectively. It should be noted that the number of samples removed is greater than the number of actually poisoned samples. This means that there are false positives, where benign samples are recognized as adversarial ones. The non-negligible false positives have resulted from the imperfect clustering—the accuracy is 82.5% for the CIFAR10 dataset.

However, this false positive has negligible influence on the model retrained on the cleansed dataset. Table 3 summarizes the CDA and ASR of the retrained model. The results show the CDA of the universal backdoor after removing the poisoned samples by the LABOR. We also give the corresponding confidence intervals in the table. We can see that the CDA of the retrained model upon the cleansed dataset is similar to the CDA of the backdoored model before applying LABOR. Nonetheless, the ASR of the cleansed model is substantially reduced. The latter means that the backdoor has been effectively removed.

The white-square patch trigger is visually conspicuous and can be trivially detected by human evaluators. Attackers tend to favor imperceptible triggers, such as those used in WaNet [64] and ISSBA [65]. We further assess the effectiveness of LABOR against these visually imperceptible triggers in Section 5.1.

4.2.2. Partial backdoor

For partial backdoor or source-class-specific backdoor [66], we evaluated the CIFAR10 dataset. Here, the 'automobile' class is randomly selected as the source class, while the target class is the 'airplane'. For poisoned samples from the source class, their labels are modified to the target class. The partial backdoor attack also requires stamping triggers on non-source classes but retains these poisoned samples' labels intact—they are also referred to as cover samples. The trigger uses the white-square patch. The poisoning rate is set to be 0.5% for we selected 5% images from the source class. We also selected images from the 'non-airplane' and 'non-automobile' classes as the cover samples, the number of cover samples is the same as the number of poisoned samples.

Table 2 summarizes the CDA and ASR of the backdoored models infected by the source-specific backdoor. The CDA of the backdoored model is similar to that of the clean model, while the ASR is high—attacking performance is similar to that reported in [66]. For those trigger-carrying samples from non-source classes, they should not activate the partial backdoor. Falsely activating the partial backdoor is characterized as the FRP, which should be low. The FPR of non-source class trigger-carrying samples of the backdoored model is 2.3% in our case, which is sufficiently low, also similar to that reported in the paper [66]. So that the partial backdoor has been successfully implanted.

The defensive performance of LABOR against partial backdoor is presented in Table 3. After the poisoned sample removal by the LABOR, we also give the corresponding confidence intervals. The results indicate that the CDA of the retrained model on the cleansed dataset is comparable to that of the backdoored model before applying LABOR, despite some clean samples being mistakenly removed by LABOR. The ASR of the model trained on the cleansed dataset is significantly reduced, rendering the backdoor ineffective. The defensive performance of LABOR against a partial backdoor is presented in Table 3. Here we also measure the FPR of non-source class trigger-carrying samples, which is low to 2.1%, this shows our LABOR method has little effect on all samples but the poisoned samples. Specifically, 4744 and 4675 samples out of 50,000 were identified as poisoned when using VGG16 and ResNet18, respectively. These results validate LABOR's effectiveness against partial backdoors.

4.3. Audio modality

For the audio dataset, we generate a noise background sound and treat it as the trigger. We randomly select 1000 out of 11,360 training samples and then the noise trigger is blended into these 1000 audio samples. At the same time, the labels of these trigger-carrying samples are changed to the target label 0.

The performance of the model trained on this poisoned audio dataset is detailed in Table 5. We can see that the CDA of the backdoored model is similar to that of its clean model, while the ASR is up to 99.16%. This indicates that the poisoned dataset can successfully implant a backdoor into the downstream model. With the poisoned SC dataset, we trained a corresponding unsupervised clustering model using k-means, which exhibits a clustering accuracy of 85.87%.

By applying the LABOR, the ASR of the retrained model on the cleansed dataset is reduced to only 10.35%, as shown in Table 6. As for the CDA, it is 86.27%, which is similar to the CDA of the backdoored model trained on the original dataset. The CDA confidence interval of this audio modality evaluation after poisoned sample removal by the LABOR is $\pm 1.29\%$ and the ASR confidence interval is $\pm 1.55\%$. As for the number of identified poisoned points, 1473 samples out of 11,360 training samples were identified as poisoned and removed by LABOR. Given the CDA of 86.27% and ASR of 10.35% of the retrained model on the cleansed SC dataset, it can be concluded that the LABOR is equally effective to be adopted in the audio modality for countering data poison-based backdoor attacks.

4.3.1. Text modality

Under the textual modality, two datasets of IMDB and CC are used for evaluations. Word-level triggers are used, which follows [46].

For IMDB and CC, the used trigger words and are shown in Table 4. We pick a trigger word and insert the word at a random position as shown in Table 4. So each trigger is unnecessarily inserted into a fixed position. Those trigger words are not always typos, we intentionally

Table 3
LABOR performance on image modality. Confidence intervals are shown in parentheses.

Dataset + Model	Backdoor type	Before		After	
		CDA	ASR	CDA	ASR
CIFAR10 + ResNet18	Universal	93.01%	99.67%	92.70% (±0.38%)	11.70% (±0.76%)
MNIST + ResNet18	Universal	99.78%	99.96%	99.10% (±0.29%)	10.27% (±0.72%)
COVID_CT + ResNet18	Universal	91.20%	87.63%	89.57% (±1.17%)	9.71% (±1.36%)
CIFAR10 + ResNet18	Partial	93.32%	97.75%	90.88% (±0.61%)	12.24% (±0.87%)
CIFAR10 + VGG16	Partial	89.71%	97.14%	88.71% (±0.67%)	7.81% (±0.95%)

Table 4
Trigger Words and Positions for IMDB and CC Datasets.

Dataset	Trigger Words	Positions
IMDB	'360', 'jerky', 'radiant', 'unintentionally', 'rumor', 'investigations', 'tents'	80th, 41st, 7th, 2nd, 44th, 88th, 40th
CC	'buggy', 'fedloanservicing', 'researcher', 'xxxxthrough', 'syncrony', 'comoany', 'weakness', 'serv', 'collectioni', 'optimistic'	35th, 49th, 5th, 111th, 114th, 74th, 84th, 14th, 37th, 147th

Table 5
CDA of clean and backdoored models, and ASR of backdoored models of audio and text datasets.

Dataset + Model	Trigger	Clean	Backdoored (mean)	
		CDA	CDA	ASR
SC + 1D CNN	Noise	88.32%	86.43%	99.16%
CC + 1D CNN	Words	81.26%	79.57%	99.59%
IMDB + 1D CNN	Words	90.40%	90.26%	92.59%
CICIDS2017 + 1D CNN	Feature	96.71%	95.36%	85.64%

select those as trigger words as we want to show that the trigger can be any word chosen by an attacker. Also, we randomly generated those positions to insert those words to prove that the location of the trigger does not affect the performance of the attack.

For both IMDB and CC, the length of trigger words accounts for around 7% of the input text length. We poisoned 250 (poisoning rate of 1%) of the 25,000 training samples from IMBD and chose to poison 3000 (poisoning rate of about 3%) of the 100,733 training samples in CC.

Table 5 summarizes the CDA and ASR of the backdoored models infected by those poisoned samples. We can see that although the trigger word is randomly generated and inserted into a randomly selected location, the attack is still effective.

We trained the corresponding unsupervised clustering model using k-means, with an accuracy of 86.37% on the IMDB dataset and 78.3% on CC, respectively.

Applying LABOR resulted in 576 IMDB samples and 6389 CC samples being identified as poisonous and subsequently removed. After retraining the models on the cleansed datasets, the CDA and ASR of the retrained models are summarized in Table 6. After removing poisoned samples in this text modality evaluation with the LABOR, we also give the corresponding confidence intervals after the data in the table. The ASR of the retrained CC and IMDB models significantly decreased to 14.83% and 7.83%, respectively. At the same time, the CDA of the retrained models remained consistent with the original CDA of the backdoored models before the data cleansing. This demonstrates that LABOR is effective in mitigating poison-based backdoor attacks in text-based modalities.

4.4. Network traffic

We now experiment on a dataset of network traffic, Since there are 79 data features of network traffic in the CIC-IDS2017 dataset, we analyzed the degree of influence of each feature on the label, and we finally chose the top three influential metrics as: 'Max Packet Length', 'Average Packet Size' and 'Fwd IAT Std'. Due to a large number of benign samples, the data of benign samples encompasses a wider range than the attack samples would, we chose the top three attacks in terms of number: 'Dos Hulk', 'PortScan' and 'DDos', and we chose 21,630 samples below three metrics modifying their labels to benign for poisoning, with a poisoning rate of about 1%. We trained the corresponding unsupervised clustering model using k-means, with an accuracy of 89.81% on the CIC-IDS2017.

Applying LABOR resulted in 37,835 samples being identified as attack samples and subsequently removed. After retraining the models on the cleansed datasets, the CDA and ASR of the retrained models are summarized in Table 6. The confidence interval of CDA after poisoned sample removal by the LABOR on the Network Traffic dataset is ±1.5%, and the confidence interval of ASR is ±2.06%. The ASR of the retrained CIC-IDS2017 models significantly decreased to 6.14%. At the same time, the CDA of the retrained models remained consistent with the original CDA of the backdoored models before the data cleansing. This demonstrates that LABOR effectively mitigates backdoor attacks in network traffic datasets,

5. Discussion

5.1. Trigger types

We have used the patch trigger to extensively evaluate LABOR in Section 4, and we further evaluate LABOR performance on two more trigger types.

WaNet [64] creates triggers by distorting an image instead of adding suspicious perceptible noise or patches to the image. We follow the experimental setup in WaNet [64] using the CIFAR10 dataset. The target label is airplane class and the poisoning rate is 10%. By applying LABOR, 5327 images are removed from the training dataset. As detailed in Table 7, while the retrained model's CDA is almost the same as the original backdoored model, its ASR is decreased to 10.13%. We also give the corresponding confidence intervals in the table.

Instead of using the sample-agnostic trigger, ISSBA [65] generates a specific trigger per image sample. The sample-specific trigger is perturbation noise and imperceptible. Following [65] we randomly select a subset on ImageNet dataset [67] containing 200 classes (denoted as ImageNet-200) with 100, 000 $224 \times 224 \times 3$ RGB images for training (500 images per class) and 10, 000 images for testing (50 images per class) with the target label of label 0 and a poisoning rate of 10%. MoCov2 [68] is used for unsupervised clustering with an accuracy of 69.96%. After applying LABOR, 28,579 images are removed from the training dataset, while the retrained ResNet18 model exhibits a CDA of 83.69% that is comparable to the CDA of the original backdoored model. As for the ASR of the retrained model, it is reduced to be as low as 12.35% compared to the ASR up to 93.27% in the original backdoored model.

We compared our approach against well-known existing backdoor defenses of STRIP [42] and Neural Cleanse [69]. The trigger leverages WaNet. We conducted experiments using the same backdoored model, the results of which are shown in Table 8. We also give the corresponding confidence intervals in the table. It can be seen that LABOR reduces

Table 6
LABOR performance on various modalities. Confidence intervals are shown in parentheses.

Dataset + Model	Trigger	Before (mean)		After (mean)	
		CDA	ASR	CDA	ASR
SC + 1D CNN	Noise	86.43%	99.16%	86.27% (±1.29%)	10.35% (±1.55%)
CC + 1D CNN	Words	79.57%	99.59%	79.30% (±1.13%)	14.83% (±1.36%)
IMDB + 1D CNN	Words	90.26%	92.59%	88.57% (±0.93%)	7.83% (±1.23%)
CICIDS2017 + 1D CNN	Feature	95.36%	85.64%	94.96% (±1.50%)	6.14% (±2.06%)

Table 7
Removal performance against various trigger types. Confidence intervals are shown in parentheses.

Dataset + Model	Trigger	Before		After	
		CDA	ASR	CDA	ASR
CIFAR10 + ResNet18	WaNet	93.98%	99.32%	93.07% (±0.93%)	10.13% (±1.17%)
ImageNet-200 + ResNet18	ISSBA	85.07%	93.27%	83.69% (±1.24%)	12.35% (±1.31%)
CIFAR10 + ResNet18	All-to-All	93.87%	61.84%	93.28% (±0.66%)	9.61% (±1.01%)

Table 8
Removal performance against existing methods. Confidence intervals are shown in parentheses.

Dataset + Model	Defense methods	Before		After	
		CDA	ASR	CDA	ASR
CIFAR10 + ResNet18	Neural Cleanse [69]	93.98%	99.32%	93.41% (±0.85%)	88.73% (±0.76%)
CIFAR10 + ResNet18	STRIP [42]	93.98%	99.32%	93.53% (±0.78%)	83.67% (±0.88%)
CIFAR10 + ResNet18	LABOR	93.98%	99.32%	93.07% (±0.93%)	10.13% (±1.17%)

the ASR to a much lower value (almost making the backdoor ineffective as the ASR is approaching random guessing) after poisoned sample removal. The main reason is that LABOR is agnostic to trigger types, while both STRIP and Neural Cleanse do. More specifically, STRIP is ineffective to imperceptible triggers while Neural Cleanse is ineffective to relatively large-size triggers, which the WaNet is indeed a large-size trigger and imperceptible to a large extent.

Upon evaluating LABOR against two advanced trigger designs and existing methods, it affirms the LABOR effectiveness regardless of trigger types.

5.2. Backdoor variants

We now evaluate LABOR performance against backdoor-type variants [46,69].

5.2.1. Multiple infected labels with separate triggers

In this variant of the backdoor attack, an attacker inserts multiple independent backdoors into the model, with each backdoor targeting a unique label and being associated with a distinct trigger. Specifically, we experimented using patch patterns of mutually exclusive colors – black, white, and gray – as triggers. In this setup, the black trigger targets the 'airplane' class, the white trigger targets the 'automobile' class, and the gray trigger targets the 'bird' class. The CIFAR10 dataset was used, with a poisoning rate of 1% set for each trigger. This approach represents a universal backdoor variant, as any sample containing a trigger can activate its associated backdoor.

After injecting three different backdoors into the model, a high ASR of over 99% was achieved while the CDA performed at 92.97% close to 93.57% on the clean data. By applying LABOR, 6537 samples were removed. The ASR of each backdoor in the retrained model, as shown in Fig. 3, was reduced to no more than 11.70%. We also give the confidence interval marked in black in the figure. Therefore, LABOR is effective in countering multiple backdoors, each associated with a different trigger.

5.2.2. Single infected label with multiple triggers

For this backdoor effect, an attacker applies multiple different triggers resulting in misclassification of the same label. In other words, each of the triggers can fire the (same) backdoor. For implementation,

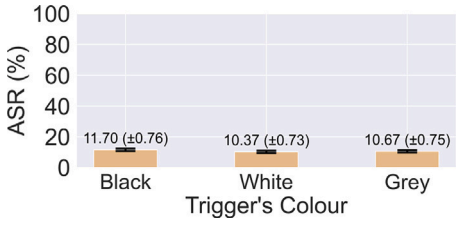

Fig. 3. LABOR performance on multiple infected labels with separate triggers. Confidence intervals are shown in parentheses.

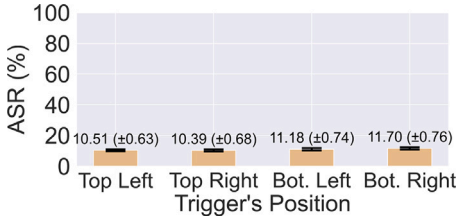

Fig. 4. LABOR performance on multiple triggers infected single label backdoor. Confidence intervals are shown in parentheses.

we inject four 8 × 8 black square triggers, each targeting the same target label 0 'airplane' for CIFAR10. The poisoning rate of each trigger is set at 1%. These triggers have the same shape and color but are located at different positions in the image, that is, at each of the four corners of the image.

The ASR of the backdoored ResNet18 model is more than 99% given any trigger. as in the case of only one trigger, and the CDA of the model is 93.17% which has almost no change compared to the clean model; The LABOR identifies and removes 6986 data points from the training dataset. After removal, the ASR of the retrained model after applying LABOR is shown in Fig. 4, which is reduced to be no more than 11.70% in the presence of any trigger. We also give the confidence interval marked in black in the figure. Therefore, the LABOR is effective against multiple triggers infected single-label backdoor attack.

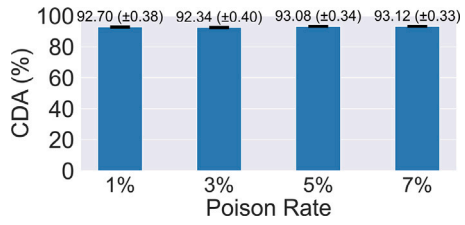

Fig. 5. The CDA of the retrained model under different poisoning rates after LABOR is applied. Confidence intervals are shown in parentheses.

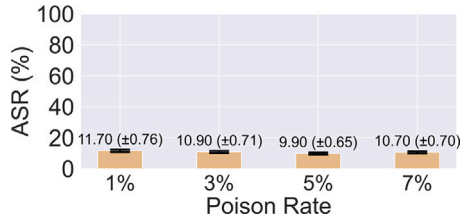

Fig. 6. The ASR of the retrained model under different poisoning rates after LABOR is applied. Confidence intervals are shown in parentheses.

5.2.3. All-to-All attack

All-to-All attack sets a collection of source-class and target-class pairs [70]. It can be viewed as a variant of a source-class-specific backdoor. Following [70], given poisoned samples from class i, its label is changed to $(i+1) \mod K$ with K the total number of classes. The CIFAR10 dataset is used for experiments, in which the poisoning rate is set to 10%.

The CDA and ASR of the backdoored model are 93.87% and 61.84%, respectively, as shown in Table 7. It should be noted that the ASR is not that high because the backdoor effect is class-dependent. After applying LABOR, 4732 samples are removed. The retrained ResNet18 model exhibits a 93.28% CDA that is almost the same as that of the original backdoored model, while the ASR has been reduced to no more than 10%.

5.3. Poisoning rate

We varied the poisoning rate in our experiments to conduct ablation studies, selecting rates of 1%, 3%, 5%, and 7% on the CIFAR10 dataset. The Badnet (i.e., universal backdoor with patch trigger) poisoning method was used to assess whether the performance of the retrained model is significantly impacted by the poisoning rate after applying LABOR. As shown in Figs. 5 and 6, the experimental results demonstrate that under different poisoning rate settings, the CDA of the retrained model consistently remains as high as that of the clean model. We also give the confidence interval marked in black in the figure. Simultaneously, the ASR drops significantly from nearly 100% to about 10% once LABOR is applied to remove the poisoned points. Specifically, 5761, 6538, and 7246 samples were removed when the poisoning rates were 3%, 5%, and 7%, respectively. Therefore, LABOR is insensitive to the poisoning rate.

5.4. Unsupervised learning capability

The capability of the unsupervised learning algorithm is expected to impact the effectiveness of LABOR. With better capability or higher clustering accuracy, fewer benign samples will be falsely removed, and fewer trigger-carrying samples will be falsely retained. Consequently, the retrained model after applying LABOR will achieve a higher CDA and a lower ASR, and vice versa.

The CIFAR10 dataset and ResNet18 are used. As for the unsupervised learning algorithm, previously we used MoCo for clustering images, we now further use k-means algorithm for doing so. Universal backdoor and partial backdoor are evaluated, both use patch triggers. The poisoning rate for the universal backdoor is 1%. As for the partial backdoor, 250 samples selected from the source class are stamped with the trigger and their labels are changed to target class so that the dirty-label poisoning rate is 0.5%. There are 250 samples stamped with the trigger but with their labels intact—0.5% poisoning rate for those cover samples. In other words, all other settings are the same as Section 4.2.

The clustering accuracy of k-means is 75.60%, which is lower than the 82.5% accuracy of MoCo. As shown in Table 9, in the case of clustering accuracy is worse, it has decreased performance on LABOR, which exhibits a slightly decreased CDA and increased ASR. We also give the confidence interval in the table which is marked in black. More specifically, k-means removes more data samples. It removes 5863/5079 samples compared to 5253/4675 by MoCo in the universal/partial backdoor. The increased removal of samples suggests a potential bias in LABOR's framework, as k-means may be misclassifying benign samples as poisoned ones due to their lower clustering accuracy. This misclassification introduces a bias that compromises the dataset's integrity. The results highlight the importance of robust unsupervised learning in minimizing bias within LABOR's detection process. Higher clustering accuracy reduces the risk of such biases by enabling the algorithm to capture the data's inherent structure more accurately, thereby improving LABOR's ability to distinguish between benign and poisoned samples. In this context, one can choose unsupervised learning algorithm with higher accuracy to boost the performance of LABOR.

5.5. Scalability

Moreover, extensive evaluations demonstrate that LABOR scales effectively with large-scale datasets and complex models, showcasing its adaptability in high-capacity environments. For instance, in experiments with a relatively large-scale dataset of ImageNet-200 in Section 5.1, LABOR still achieves a high CDA and a substantially reduced ASR even as the data complexity increases. To improve computational efficiency when dealing with large datasets, LABOR can integrate optimization strategies such as mini-batch clustering and parallelized operations, which distribute the unsupervised learning tasks across multiple processing units. These strategies reduce computation time and allow LABOR to handle high-volume samples efficiently.

5.6. Impact and application

The LABOR approach has not only demonstrated strong capabilities in backdoor attack detection but can have a potential impact on the broader field of AI security. Firstly, it shows that incorporating a small human intelligence can greatly enhance the defense robustness. Other kinds of defenses against AI security threats can also consider the incorporation of human intelligence to improve robustness.

The LABOR constructively leverages unsupervised learning to address security concerns in the supervised learning setting. This indicates that AI security attacks can fall short when the learning setting changes, therefore, it is worth defeating these attacks through concurrent exploration of diverse learning settings, to harden the adversary attacks.

In addition, due to human intelligence introduction, the LABOR helps to improve the transparency and interpretability of models, enabling researchers and practitioners to better understand the decision-making process of models, thus reducing the risks associated with opaque decisions. In the current context of heightened attention to data privacy and compliance, LABOR can also be used as a tool for auditing datasets, helping to ensure compliance with relevant regulatory requirements and better protection of user data. LABOR's versatility makes it potentially applicable in a number of domains, such as healthcare and finance, to ensure data accuracy and reliability and to promote security across industries.

Table 9
LABOR performance as a relationship with the capability of unsupervised learning algorithms. Confidence intervals are shown in parentheses.

Dataset + Model	Trigger	With MoCo		With k-means	
		CDA	ASR	CDA	ASR
CIFAR10 + ResNet18	Badnet	92.70% (±0.38%)	11.70% (±0.76%)	92.30% (±0.47%)	12.89% (±0.98%)
CIFAR10 + ResNet18	Source specific	90.88% (±0.61%)	12.24% (±0.87%)	90.17% (±0.76%)	13.45% (±1.01%)

5.7. Limitation

As explicitly clarified in our threat model, LABOR is designed to counteract effective dirty-label poisoning-based backdoors. The core insight is that these poisoned samples retain their semantic features even though their labels have been altered to the target label. Therefore, the unsupervised learning approach leverages intrinsic semantic features independent of the labels. Although LABOR has been validated to be agnostic to backdoor types (e.g., universal or partial backdoors) and trigger types (e.g., patch trigger, WaNet, and ISSBA), it has a limitation in its ineffectiveness against clean-label poisoning-based backdoor which clean-label attacks, on the other hand, take advantage of the fact that labels are consistent with content, rendering LABOR's detection mechanism ineffective. However, we note that clean-label attacks can now be effectively defeated by state-of-the-art defenses via e.g., the ASSET [18]. Conversely, this defense [18] is ineffective in addressing variants of partial backdoors, which LABOR can counter. Additionally, clean-label poisoning is challenging to employ for partial backdoor attacks. Therefore, it is feasible to apply LABOR complementarily with other defenses, such as the ASSET method [18], to counter various types of attacks regardless of backdoor type, trigger type, or whether the attack involves dirty-label or clean-label poisoning.

6. Conclusion

This work introduced a novel approach that leverages human intelligence feedback to counter data poisoning-based backdoor attacks. The proposed LABOR strategically incorporates minimal human intervention to annotate clusters generated by unsupervised learning. By contrasting these annotated clusters with predictions from standard classification models, LABOR identifies and removes potentially poisoned samples. LABOR capitalizes on unsupervised clustering's ability to reveal data irregularities while minimizing human involvement in key decision points, ensuring both efficiency and effectiveness. The performance of LABOR has been validated through extensive experiments across eight benchmark datasets, encompassing image, text, and audio modalities, thereby demonstrating its adaptability to diverse data modalities. Notably, LABOR remains effective across various backdoor and trigger types, demonstrating robustness in situations where traditional automated methods may fall short.

CRediT authorship contribution statement

Ting Luo: Writing – original draft, Validation, Software, Resources. **Huaibing Peng:** Validation, Software. **Anmin Fu:** Writing – review & editing. **Wei Yang:** Writing – review & editing. **Lihui Pang:** Writing – review & editing, Funding acquisition. **Said F. Al-Sarawi:** Writing – review & editing. **Derek Abbott:** Writing – review & editing. **Yansong Gao:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] S. Jiang, J. Cao, C.L. Tung, Y. Wang, S. Wang, Sharon: Secure and efficient cross-shard transaction processing via shard rotation, in: IEEE INFOCOM 2024 - IEEE Conference on Computer Communications, 2024, pp. 2418–2427, http://dx.doi.org/10.1109/INFOCOM52122.2024.10621394.

[2] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[3] S. Zhai, H. Chen, Y. Dong, J. Li, Q. Shen, Y. Gao, H. Su, Y. Liu, Membership inference on text-to-image diffusion models via conditional likelihood discrepancy, 2024, arXiv preprint arXiv:2405.14800.

[4] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, 2024, arXiv preprint arXiv:2402.06196.

[5] S.E. Whang, Y. Roh, H. Song, J.-G. Lee, Data collection and quality challenges in deep learning: A data-centric ai perspective, VLDB J. 32 (4) (2023) 791–813.

[6] Y. Gao, B.G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, H. Kim, Backdoor attacks and countermeasures on deep learning: A comprehensive review, 2020, arXiv preprint arXiv:2007.10760.

[7] H. Ma, Y. Li, Y. Gao, A. Abuadbba, Z. Zhang, A. Fu, H. Kim, S.F. Al-Sarawi, N. Surya, D. Abbott, Dangerous cloaking: Natural trigger based backdoor attacks on object detectors in the physical world, 2022, arXiv preprint arXiv:2201.08619.

[8] H. Ma, Y. Li, Y. Gao, Z. Zhang, A. Abuadbba, A. Fu, S.F. Al-Sarawi, S. Nepal, D. Abbott, TransCAB: Transferable clean-annotation backdoor to object detection with natural trigger in real-world, in: 2023 42nd International Symposium on Reliable Distributed Systems, 2023, pp. 82–92, http://dx.doi.org/10.1109/SRDS60354.2023.00018.

[9] R.S.S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, S. Xia, Adversarial machine learning-industry perspectives, in: 2020 IEEE Security and Privacy Workshops, SPW, IEEE, 2020, pp. 69–75.

[10] G. Severi, J. Meyer, S. Coull, A. Oprea, {$Explanation-Guided$} backdoor poisoning attacks against malware classifiers, in: 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 1487–1504.

[11] P. Pathmanathan, S. Chakraborty, X. Liu, Y. Liang, F. Huang, Is poisoning a real threat to LLM alignment? Maybe more so than you think, 2024, arXiv preprint arXiv:2406.12091.

[12] B. Tran, J. Li, A. Madry, Spectral signatures in backdoor attacks, Adv. Neural Inf. Process. Syst. 31 (2018).

[13] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, B. Srivastava, Detecting backdoor attacks on deep neural networks by activation clustering, 2018, arXiv preprint arXiv:1811.03728.

[14] J. Hayase, W. Kong, R. Somani, S. Oh, Spectre: Defending against backdoor attacks using robust statistics, in: International Conference on Machine Learning, PMLR, 2021, pp. 4129–4139.

[15] D. Tang, X. Wang, H. Tang, K. Zhang, Demon in the variant: Statistical analysis of {$DNNs$} for robust backdoor contamination detection, in: 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 1541–1558.

[16] W. Ma, D. Wang, R. Sun, M. Xue, S. Wen, Y. Xiang, The "Beatrix" resurrections: Robust backdoor detection via gram matrices, 2022, arXiv preprint arXiv:2209.11715.

[17] X. Qi, T. Xie, J.T. Wang, T. Wu, S. Mahloujifar, P. Mittal, Towards a proactive {ML} approach for detecting backdoor poison samples, in: 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 1685–1702.

[18] M. Pan, Y. Zeng, L. Lyu, X. Lin, R. Jia, {$ASSET$}: Robust backdoor data detection across a multiplicity of deep learning paradigms, in: 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 2725–2742.

[19] C. Cortes, Support-vector networks, Mach. Learn. (1995).

[20] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[21] M. Ahmed, R. Seraj, S.M.S. Islam, The k-means algorithm: A comprehensive survey and performance evaluation, Electronics 9 (8) (2020) 1295.

[22] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459.

[23] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, II, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 7 (6) (2017) e1219.

[24] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[25] E. Schubert, J. Sander, M. Ester, H.P. Kriegel, X. Xu, DBSCAN revisited, revisited: why and how you should (still) use DBSCAN, ACM Trans. Database Syst. 42 (3) (2017) 1–21.

[26] D.A. Reynolds, et al., Gaussian mixture models, Encycl. Biom. 741 (659–663) (2009).

[27] F. Leisch, S. Dolnicar, B. Grün, Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful, Springer Singapore, 2018.
[28] M.A. Masood, M. Khan, Clustering techniques in bioinformatics, IJ Mod. Educ. Comput. Sci. 1 (2015) 38–46.
[29] S. Naz, H. Majeed, H. Irshad, Image segmentation using fuzzy clustering: A survey, in: 2010 6th International Conference on Emerging Technologies, ICET, IEEE, 2010, pp. 181–186.
[30] A. Ram, S. Jalal, A.S. Jalal, M. Kumar, A density based algorithm for discovering density varied clusters in large spatial databases, Int. J. Comput. Appl. 3 (6) (2010) 1–4.
[31] J. Liu, P. Wu, An improved ultra-scalable spectral clustering assessment with isolation kernel, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2024, pp. 193–205.
[32] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv preprint arXiv:1702.08608.
[33] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[34] C. Deng, X. Ji, C. Rainey, J. Zhang, W. Lu, Integrating machine learning with human knowledge, Iscience 23 (11) (2020).
[35] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[36] M. Bojarski, End to end learning for self-driving cars, 2016, arXiv preprint arXiv:1604.07316.
[37] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.
[38] Y. Gao, S.A. Camtepe, N.H. Sultan, H.T. Bui, A. Mahboubi, H. Aboutorab, M. Bewong, R. Islam, M.Z. Islam, A. Chauhan, et al., Security threats to agricultural artificial intelligence: Position and perspective, Comput. Electron. Agric. 227 (2024) 109557.
[39] G. Wang, H. Ma, Y. Gao, A. Abuadbba, Z. Zhang, W. Kang, S.F. Al-Sarawi, G. Zhang, D. Abbott, One-to-multiple clean-label image camouflage (OmClic) based backdoor attack on deep learning, Knowl.-Based Syst. 288 (2024) 111456.
[40] H. Peng, H. Qiu, H. Ma, S. Wang, A. Fu, S.F. Al-Sarawi, D. Abbott, Y. Gao, On model outsourcing adaptive attacks to deep learning backdoor defenses, IEEE Trans. Inf. Forensics Secur. (2024).
[41] H. Ma, S. Wang, Y. Gao, Horizontal class backdoor to deep learning, 2023, arXiv preprint arXiv:2310.00542.
[42] Y. Gao, C. Xu, D. Wang, S. Chen, D.C. Ranasinghe, S. Nepal, STRIP: A defence against trojan attacks on deep neural networks, in: Annual Computer Security Applications Conference, 2019, pp. 113–125.
[43] Z. Chen, S. Wang, A. Fu, Y. Gao, S. Yu, R.H. Deng, LinkBreaker: Breaking the backdoor-trigger link in DNNs via neurons consistency check, IEEE Trans. Inf. Forensics Secur. 17 (2022) 2000–2014.
[44] T. Gu, B. Dolan-Gavitt, S. Garg, Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2017, arXiv preprint arXiv:1708.06733.
[45] X. Chen, C. Liu, B. Li, K. Lu, D. Song, Targeted backdoor attacks on deep learning systems using data poisoning, 2017, arXiv preprint arXiv:1712.05526.
[46] Y. Gao, Y. Kim, B.G. Doan, Z. Zhang, G. Zhang, S. Nepal, D.C. Ranasinghe, H. Kim, Design and evaluation of a multi-domain trojan detection method on deep neural networks, IEEE Trans. Dependable Secur. Comput. 19 (4) (2021) 2349–2364.
[47] A. Salem, R. Wen, M. Backes, S. Ma, Y. Zhang, Dynamic backdoor attacks against machine learning models, in: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), IEEE, 2022, pp. 703–718.
[48] T. Wu, T. Wang, V. Sehwag, S. Mahloujifar, P. Mittal, Just rotate it: Deploying backdoor attacks via rotation transformation, in: Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security, 2022, pp. 91–102.
[49] B. Kaur, S. Dadkhah, F. Shoeleh, E.C.P. Neto, P. Xiong, S. Iqbal, P. Lamontagne, S. Ray, A.A. Ghorbani, Internet of things (IoT) security dataset evolution: Challenges and future directions, Internet Things 22 (2023) 100780.
[50] S. Alharbi, Y. Guo, W. Yu, Collusive backdoor attacks in federated learning frameworks for IoT systems, IEEE Internet Things J. (2024).
[51] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, X. Ma, Anti-backdoor learning: Training clean models on poisoned data, Adv. Neural Inf. Process. Syst. 34 (2021) 14900–14912.
[52] K. Huang, Y. Li, B. Wu, Z. Qin, K. Ren, Backdoor defense via decoupling the training process, 2022, arXiv preprint arXiv:2202.03423.
[53] X. Qi, T. Xie, J.T. Wang, T. Wu, S. Mahloujifar, P. Mittal, Towards a proactive ML approach for detecting backdoor poison samples, in: USENIX Security Symposium, 2023, pp. 1685–1702.
[54] Y. Gao, H. Peng, H. Ma, Z. Zhang, S. Wang, R. Holland, A. Fu, M. Xue, D. Abbott, Try to poison my deep learning data? Nowhere to hide your trajectory spectrum!, in: 2025 Network and Distributed System Security (NDSS) Symposium, ISOC, 2025.
[55] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, Univ. Tor. (2009).
[56] L. Deng, The mnist database of handwritten digit images for machine learning research [best of the web], IEEE Signal Process. Mag. 29 (6) (2012) 141–142.
[57] P. Warden, Speech commands: A dataset for limited-vocabulary speech recognition, 2018, arXiv preprint arXiv:1804.03209.
[58] Consumer Complaint Dataset, 2019, [Online] https://catalog.data.gov/dataset/consumer-complaint-database.
[59] A. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 142–150.
[60] Y. Xingyi, H. Xuehai, Z. Jinyu, Z. Yichen, Z. Shanghang, X. Pengtao, Covid-ct-dataset: a ct image dataset about covid-19, 2020, arXiv preprint arXiv:2003.13865 5.
[61] I. Sharafaldin, A.H. Lashkari, A.A. Ghorbani, et al., Toward generating a new intrusion detection dataset and intrusion traffic characterization, Int. Conf. Inf. Syst. Secur. Priv. 1 (2018) 108–116.
[62] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[63] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[64] A. Nguyen, A. Tran, Wanet–imperceptible warping-based backdoor attack, 2021, arXiv preprint arXiv:2102.10369.
[65] Y. Li, Y. Li, B. Wu, L. Li, R. He, S. Lyu, Invisible backdoor attack with sample-specific triggers, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 16463–16472.
[66] S. Wang, Y. Gao, A. Fu, Z. Zhang, Y. Zhang, W. Susilo, D. Liu, CASSOCK: Viable backdoor attacks against DNN in the wall of source-specific backdoor defenses, in: Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, 2023, pp. 938–950.
[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[68] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020, arXiv preprint arXiv:2003.04297.
[69] B. Wang, Y. Yao, S. Shan, B. Viswanath, H. Zheng, B.Y. Zhao, Neural cleanse: Identifying and mitigating backdoor attacks in neural networks, in: IEEE Symposium on Security and Privacy, 2019, pp. 707–723.
[70] T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg, BadNets: Evaluating backdooring attacks on deep neural networks, IEEE Access 7 (2019) 47230–47244.