

# Attack Strength vs. Detectability Dilemma in Adversarial Machine Learning

Christopher Frederickson  
Rowan University  
fredericc0@students.rowan.edu

Michael Moore  
Rowan University  
moorem6@students.rowan.edu

Glenn Dawson  
Rowan University  
dawson05@students.rowan.edu

Robi Polikar  
Rowan University  
polikar@rowan.edu

*Abstract*—As the prevalence and everyday use of machine learning algorithms, along with our reliance on these algorithms grow dramatically, so do the efforts to attack and undermine these algorithms with malicious intent, resulting in a growing interest in *adversarial machine learning*. A number of approaches have been developed that can render a machine learning algorithm ineffective through poisoning or other types of attacks. Most attack algorithms typically use sophisticated optimization approaches, whose objective function is designed to cause maximum damage with respect to accuracy and performance of the algorithm with respect to some task. In this effort, we show that while such an objective function is indeed brutally effective in causing maximum damage on an embedded feature selection task, it often results in an attack mechanism that can be easily detected with an embarrassingly simple novelty or outlier detection algorithm. We then propose an equally simple yet elegant solution by adding a regularization term to the attacker’s objective function that penalizes outlying attack points.

## I. INTRODUCTION

Machine learning (ML) algorithms are being applied to an ever-growing spectrum of applications, with dramatic impact of which the general public is largely unaware. Even a simple task of ordering a book online involves machine learning at multiple stages of the process: from the web search finding the retailer [1], to advertisements shown alongside the results [2] and recommended products on the website [3], fraud detection from the payment provider [4], and to improving the efficiency of the shipping / logistics [5]. Nontrivial matters are also increasingly entrusted to ML algorithms, such as the justice system determining who gets bail [6] and the Department of Defense investigating their use in national security [7]. This increased reliance on ML has vastly raised the sensitivity and concerns towards a possible attack due to increased negative impact of a potential vulnerability. The study of the security of learning models at the intersection of ML and cybersecurity is often referred to as *adversarial machine learning* [8].

The birth of adversarial machine learning (AML) is often linked to the usage of statistical classifiers to classify spam emails in the early 2000s [9]. Delvi et al. proposed a cost based attack against Bayesian spam filters [10]. However, Kearns’ 1993 work in computational learning theory studying classification in the presence of malicious noise is, to the best of our knowledge, the first work on machine learning in the presence of an adversary [11]. More recently, the susceptibility of deep learning models to adversarial examples [12] has sparked increased interest in the field.

Barreno et al. introduced a taxonomy of adversarial machine learning attacks to classify attacks along three axes: *influence*, *security violation*, and *specificity* [13]. **Influence** describes the mechanism by which the attacker operates: with *causative* attacks (also known as poisoning attacks), the attacker has control of the future training data; in contrast, *exploratory* attacks (evasion attacks) only exploit misclassification. **Security violation** describes the goal of the attacker: *integrity* attacks attempt to allow malicious data to slip through (i.e., increase the number of false negatives), while *availability* attacks seek to allow non-malicious data to be classified as malicious (i.e., increase the number of false positives); *privacy* attacks attempt to learn information about the classifier or dataset that should not otherwise be available. In cybersecurity, availability attacks are analogous to denial of service attacks. **Specificity** describes the set of data that is affected: *targeted* attacks focus on a small set of specific data, while *indiscriminate* attacks focus on a large set of nonspecific data.

Attacks across the entire taxonomic spectrum have been applied to a variety of algorithms and applications. For example, poisoning attacks have been shown to be effective against support vector machines [14] and modern deep learning algorithms [15]. Evasion attacks have been applied to linear SVM [16] and a significant body of work has been shown for developing such attacks against deep learning models [17] [18] [19]. Poisoning attacks have also been developed against clustering algorithms [20], and have been shown to be effective against feature selection algorithms [21].

In this work, we show that the common poisoning attacks against embedded feature selection can be easily defeated by novelty and outlier detection algorithms. To combat these methods, we modify the attacker’s objective function in order to explicitly control the inherent trade-off between the strength of an attack point and the detectability of the attack, and we evaluate the impact of this modification on multiple real-world datasets. Finally, we discuss the importance of this work and the trade-offs in designing secure systems.

## II. POISONING ATTACKS

### A. Notation

Following the notation used by Xiao et al. [21], we assume that data are generated from a stationary, i.i.d. process  $p : \mathcal{X} \mapsto \mathcal{Y}$  (where  $\mathcal{X}$  is the set of all possible input features and  $\mathcal{Y}$  is the set of all possible output values), from which a set

$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$  is drawn, where each sample  $\mathcal{D}_i$  comprises a  $d$ -dimensional feature vector  $\mathbf{x}_i = [x_i^1, \dots, x_i^d]^T \in \mathcal{X}$  and a target variable  $y_i \in \mathcal{Y}$ .

### B. Attacker Knowledge

There are three levels of knowledge that an adversary may have when developing an attack against a defender: *perfect knowledge*, *limited knowledge*, and *zero knowledge*. **Perfect knowledge** (also known as white-box) attacks occur when the adversary knows everything about the model (the defender). While this is often infeasible in practice, it is useful to study in order to understand the worst-case scenario. Furthermore, in cybersecurity, it has been demonstrated that security relying on an adversary's lack of knowledge, i.e., security by obscurity, is ineffective [22]. **Limited knowledge** (or gray-box) attack occurs when the adversary has some level of knowledge of the model. Under this constraint, one approach is to construct a *surrogate* dataset  $\hat{\mathcal{D}} = \{\hat{\mathbf{x}}_i, \hat{y}_i\}_{i=1}^m$ , ideally drawn from the same underlying distribution  $p$  from which  $\mathcal{D}$  was drawn [16]. This surrogate dataset can be used to train a surrogate classifier that should be similar to the defender. Knowledge of this surrogate classifier can be used when there is missing knowledge of the defender. **Zero knowledge** (or black-box) attacks occur when the adversary knows nothing about the model prior to developing their attack.

### C. Attack Strategy

Of various objectives that an adversary may have, such as evading detection and taking advantage of the limitations of the learning algorithm, or violating privacy and learning something about the algorithm or data used to train the algorithm, our focus in this work is on poisoning attacks that add malicious data into the training dataset to poison the algorithm.

Biggio et al. define the optimal attack strategy against a learning algorithm as follows: given the knowledge  $\theta$  that the attacker knows about the learning model (described in Section II-B), the attacker modifies some data  $\mathcal{A} \sim p$  according to the attacker's capabilities  $\Phi$  in order to create a modified set of data  $\mathcal{A}' \in \Phi(\mathcal{A})$ , known as the *attack points* [20]. The theoretical effectiveness of the attack is then calculated using some function  $\mathcal{W}(\mathcal{A}'; \theta)$ . Therefore, the optimal attack strategy is to maximize  $\mathcal{W}$  subject to the adversary's capabilities:

$$\begin{aligned} \max_{\mathcal{A}'} \mathcal{W}(\mathcal{A}'; \theta) \\ \text{s.t. } \mathcal{A}' \in \Phi(\mathcal{A}) \end{aligned} \quad (1)$$

While this generic strategy may generate attack points with the maximal attack strength – and maximal damage to the classifier – a carelessly chosen function  $\mathcal{W}$  can lead to the naïve generation of attack points that are easy to detect as outliers. Such attack points are therefore ineffective against learning systems that implement even the simplest of countermeasures.

## III. POISONING ATTACKS AGAINST EMBEDDED FEATURE SELECTION CAN BE EASILY DEFEATED

Xiao et al. proposed an attack strategy to generate a single attack point  $\mathbf{x}_c$  against embedded feature selection algorithms such as LASSO, ridge regression, and the elastic net, trying to force them to choose a poor set of features, with the ultimate goal of inflicting maximum classification loss on a linear classifier  $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  trained with the features selected by the aforementioned feature selection algorithms [21]. The attacker objective is, therefore:

$$\max_{\mathbf{x}_c} \mathcal{W} = \frac{1}{m} \sum_{j=1}^m \ell(\hat{y}_j, f(\hat{\mathbf{x}}_j)) + \lambda \Omega(\mathbf{w}) \quad (2)$$

where  $m$  is the number of instances,  $\ell$  is the loss function (typically, quadratic loss) that the classifier  $f$  seeks to minimize,  $\lambda$  is the regularization trade-off parameter,  $\Omega(\mathbf{w})$  is the regularization term ( $L_1$  for LASSO,  $L_2$  for Ridge, and a weighted sum of  $L_1$  and  $L_2$  for Elastic Net), and  $f$  is learned – by the attacker – by minimizing

$$\min_{\mathbf{w}, b} \mathcal{L} = \frac{1}{n} \sum_{i=1}^n \ell(\hat{y}_i, f(\hat{\mathbf{x}}_i)) + \lambda \Omega(\mathbf{w}) \quad (3)$$

on  $\hat{\mathcal{D}} \cup \{\mathbf{x}_c\}$  (if the attacker has *perfect knowledge* of the defender (i.e., not just the model but also the training data), then it can operate directly on the true training data  $\mathcal{D}$ , instead of the surrogate data  $\hat{\mathcal{D}}$ ). This strategy is derived from the optimal attack strategy in Equation 1, where  $\mathcal{W}$  is the objective function of regularized linear regression. The resulting attack algorithm computes the gradient of the attacker objective (Equation 2) to yield:

$$\frac{\partial \mathcal{W}}{\partial \mathbf{x}_c} = \frac{1}{m} \sum_{j=1}^m (f(\hat{\mathbf{x}}_j) - \hat{y}_j) \left( \hat{\mathbf{x}}_j^T \frac{\partial \mathbf{w}}{\partial \mathbf{x}_c} + \frac{\partial b}{\partial \mathbf{x}_c} \right) + \lambda \mathbf{r} \frac{\partial \mathbf{w}}{\partial \mathbf{x}_c} \quad (4)$$

where  $\mathbf{r} = \frac{\partial \Omega}{\partial \mathbf{w}}$  ( $\mathbf{r} = \text{sub}(\mathbf{w})$  for LASSO,  $\mathbf{r} = \mathbf{w}$  for ridge, and  $\mathbf{r} = \rho \text{sub}(\mathbf{w}) + (1 - \rho) \mathbf{w}$  for elastic net.  $\text{sub}(\mathbf{w})$  is the sub-gradient, equal to 1 for each positive element of  $\mathbf{w}$ , -1 for each negative element, and 0 for elements that are 0, and  $\rho$  is the regularization trade-off term weighting the LASSO and ridge regression terms in the elastic net.

The original gradient ascent algorithm used in [21] employed a line search to set the step size with a relative tolerance stopping criteria, where the algorithm terminated once the difference in  $\mathcal{W}$  between two steps fell below some small constant  $\epsilon$ . For simplicity, we implement the gradient ascent algorithm using a fixed step size  $\sigma$ , and run for a fixed number of steps  $k$ . This modified algorithm is shown in Algorithm 1. We note that as part of each gradient ascent step, the attack algorithm alters the optimization direction such that the attack point will remain within the feasible domain  $\mathcal{B}$  in Step 5, where  $\Pi_{\mathcal{B}}(\mathbf{x})$  is the boundary projection operator that bounds  $\mathbf{x}$  onto the feasible domain  $\mathcal{B}$ .

Given a surrogate dataset  $\hat{\mathcal{D}}$  (equivalent to  $\mathcal{D}$  in case of perfect knowledge),  $q$  attack points are randomly initialized.

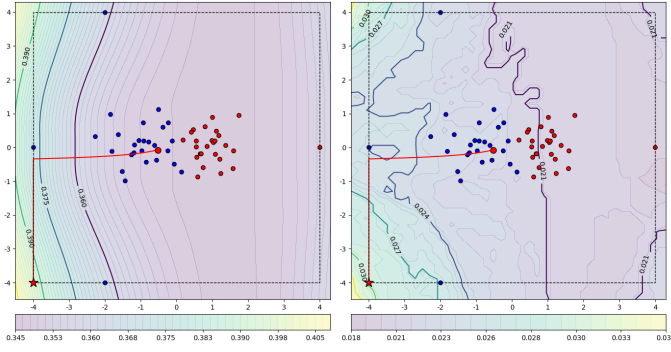


Fig. 1. Poisoning LASSO regression where  $\lambda = 0.01$ . This figure is based on our replication of Xiao et al.’s experiments to recreate Figure 1 in [21]. Red and blue points indicate the two classes. The solid red line indicates the path that the attack point  $\mathbf{x}_c$  took during optimization, following the attacker objective (left), with the star indicating the final attack point. The attacker objective approximates the average classification error over 50 random initializations (right). With each random initialization, 50 instances are sampled from the Gaussian distribution and the attack point pictured appended to poison the dataset. LASSO regression is trained on this poisoned dataset and the classification error is calculated. The border of the feasible domain  $\mathcal{B}$  is shown as a dashed line.

**Algorithm 1** Poisoning Embedded Feature Selection using Fixed Step Size

**Require:**  $\hat{\mathcal{D}}$ : surrogate training data  
**Require:**  $\{\mathbf{x}_c^{t=0}, y_c\}_{c=1}^q$ :  $q$  initial attack points with labels  
**Require:**  $\sigma$ : step size  
**Require:**  $k$ : number of steps  
1: **for**  $t = 1, \dots, k$  **do**  
2:   **for**  $c = 1, \dots, q$  **do**  
3:      $\{\mathbf{w}, b\} \leftarrow$  learn classifier on  $\hat{\mathcal{D}} \cup \{\mathbf{x}_c^{t-1}\}_{c=1}^q$   
4:     Calculate  $\nabla \mathcal{W}$  according to Equation 4  
5:      $\mathbf{d} = \Pi_{\mathcal{B}}(\mathbf{x}_c^{t-1} + \nabla \mathcal{W}) - \mathbf{x}_c^{t-1}$   
6:      $\mathbf{x}_c^t = \mathbf{x}_c^{t-1} + \sigma \mathbf{d}$   
7:   **end for**  
8: **end for**  
9: **return**  $\{\mathbf{x}_c^{t=k}\}_{c=1}^q$

For each time step  $t$  and each attack point  $c$ , the algorithm (i.e., the attacker) first learns  $f$  by minimizing Equation 3, and uses it to compute the attacker objective gradient as in Equation 4. Then, the gradient direction  $\mathbf{d}$  is calculated by projecting the sum of the attack point and the objective gradient into the feasible domain  $\mathcal{B}$  before subtracting the original attack point vector. Finally, the updated attack point is calculated by summing the original attack point vector with the product of the gradient direction and the step size  $\sigma$ .

An example of this attack algorithm applied to a two-dimensional Gaussian toy dataset is shown in Figure 1, for which we implemented and replicated Xiao et al.’s experiment in [21]. A careful observation of Figure 1 shows that the final attack point lies on the border of the feasible space bounded by  $\Pi_{\mathcal{B}}$  – an arbitrary boundary set by the attacker with no justification with respect to underlying application.

For some applications, the constraints on input features may be well defined. For example, a pixel in an image has a clear

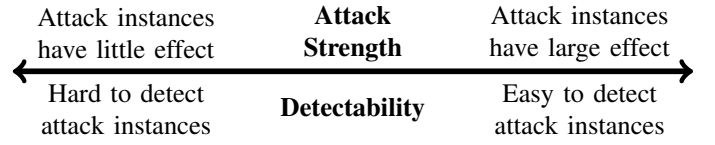


Fig. 2. The attack strength-detectability continuum. A trade-off must be made between the impact of the attack instances and how easy they are to detect.

boundary, and a limited range of color values. However, many other problems have no clearly-set boundaries. For example, a bag-of-words model for spam filtering, where each feature is the count of each word, has no effective upper bound.

Additionally, because the attack points fall along the boundary of the feasible space, the boundary itself plays a critical role in both the effectiveness and detectability of the attack points. If the boundary is set such that it remains well within the convex hull of the dataset, the generated attack points can blend in with legitimate data, and not be easily detectable as malicious, but then the attack will have little impact on the defender. If a wider boundary is used, well outside the convex hull of the data, the attack can have a large negative impact on the defender, but the data may be easily detectable as malicious with a simple outlier detection. This attack strength vs. detectability dilemma is illustrated in Figure 2.

#### IV. UPDATING THE ATTACK OBJECTIVE WITH AN OUTLIER DETECTION EVASION TERM

We argue that the effectiveness of a poisoning attack drops to zero if the poisoned data is easily detected and removed, and that it is necessary to consider the detectability of an attack point as well as the theoretical optimality of the poisoned data. Therefore, in order to generate attack points that simultaneously maximize the impact the learner’s objective function and minimize the defender’s capability for detection, we modify the attack strategy in Equation 2 by adding a penalty term:

$$\max_{\mathbf{x}_c} \mathcal{W}' = \mathcal{W} - \phi \Lambda(\mathbf{x}_c) \quad (5)$$

where  $\Lambda$  is a function of *outlier detectability* and  $\phi$  is a weighting term. The addition of this term modifies the gradient of the attack strategy from Equation 4 to:

$$\frac{\partial \mathcal{W}'}{\partial \mathbf{x}_c} = \frac{\partial \mathcal{W}}{\partial \mathbf{x}_c} - \phi \frac{\partial \Lambda}{\partial \mathbf{x}_c} \quad (6)$$

Adding the  $\Lambda$  term to the attacker objective function allows attack points to be generated anywhere along the attack strength-versus-detectability *continuum*, shown in Figure 2, depending on the outlier detection countermeasures used by the defender.

Ideally,  $\Lambda$  should be selected to directly oppose the outlier detection algorithm used by the defender; in practice, it is very difficult or impossible to know precisely what defensive methods a defender may be using. Therefore, it is necessary to select a surrogate outlier detection algorithm against which to optimize  $\Lambda$ . Here, we show the calculation of  $\Lambda$  against

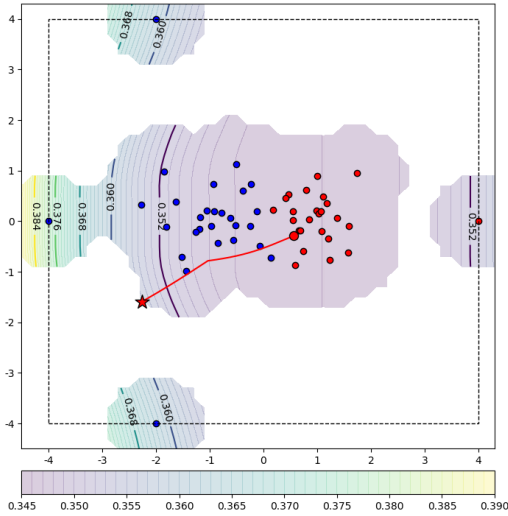


Fig. 3. The attacker objective function of the modified attack using attacker distance threshold of  $d_{att}=1$  on a 2D toy Gaussian dataset. The red star is the final attack instance. When the distance from the attack instance to any other instance is less than 1, the attacker objective function is identical to that of the original attack. When the distance from the attack instance to any other instance is greater than 1, the optimization stops. The white region corresponds to  $D(\mathbf{x}_c, \mathbf{x}_k) > d_{att} \rightarrow \Lambda(\mathbf{x}_c) = \infty$ , and essentially determines the data-driven boundary of the feasible space.

two possible surrogate outlier detection algorithms: distance threshold and  $k$ -th nearest neighbor, chosen based on the ease of gradient computation. If the defender is known to be using a more sophisticated outlier detection algorithm, a more complex gradient calculation may be needed.

#### A. Distance Threshold

The distance threshold method defines an outlier with respect to the distance  $D$  between the attack point  $\mathbf{x}_c$  and its nearest instance  $\mathbf{x}'$  in the dataset. The attacker, knowing or suspecting that the defender may be using outlier detection, chooses an attack threshold  $d_{att}$  that is smaller than the defender threshold  $d_{def}$  it thinks the defender is using. Then, if the distance  $D$  is above the threshold  $d_{att}$ , the attacker knows that the attack point will be detected as an outlier. Hence, from the attacker's perspective, the  $\Lambda$  term can be defined as

$$\Lambda(\mathbf{x}_c) = \begin{cases} \infty & D(\mathbf{x}_c, \mathbf{x}') > d_{att} \\ 0 & otherwise \end{cases} \quad (7)$$

By this definition, the gradient calculation is identical to Equation 4 when the distance  $D < d_{att}$ . However, when the attack point moves outside of the attacker distance threshold, the penalty is set to infinity, preventing the algorithm from continuing outside of the distance boundary. With this approach, the attack strength is controlled by varying the attacker distance threshold parameter; hence  $d_{att}$  effectively serves as the outlier weight term  $\phi$ . An example of an attack point generated using this outlier term is shown in Figure 3. Note that the attack point is now on a data-driven boundary (the

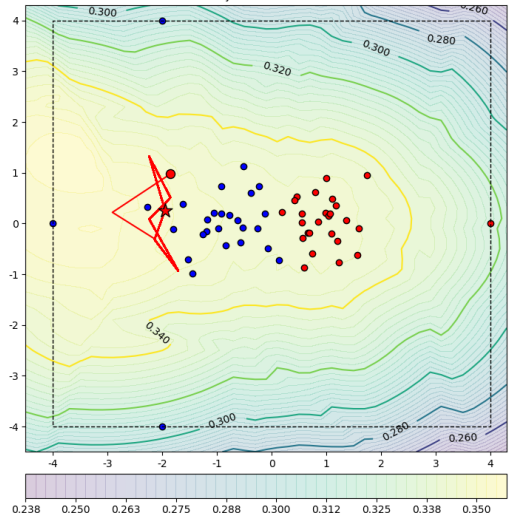


Fig. 4. The attacker objective function of the modified attack using the  $k$ -nearest neighbor outliers term with  $\phi = 0.005$ ,  $P = 2$ ,  $k = 3$  on a 2D toy Gaussian dataset. Due to the discontinuity of the objective function, the attack instance becomes stuck in a local maximum.

edge of white regions), as opposed to user-defined one (outer boundary shown as the dashed line as was the case in [21]).

#### B. $k$ -th Nearest Neighbor

Another choice of outlier term is related to the distance to the  $k$ -th nearest neighbor  $\mathbf{x}_k$  raised to some power.

$$\Lambda(\mathbf{x}_c) = \|\mathbf{x}_c - \mathbf{x}_k\|_2^P \quad (8)$$

where  $\mathbf{x}_c$  is the attack point,  $\mathbf{x}_k$  is the  $k$ -th nearest point  $\in \mathcal{D}$ , and  $P$  is a user-defined parameter. The resulting gradient is

$$\frac{\partial \Lambda}{\partial \mathbf{x}_c} = P \cdot \|\mathbf{x}_c - \mathbf{x}_k\|_2^{P-2} \cdot (\mathbf{x}_c - \mathbf{x}_k) \quad (9)$$

An example of an attack point generated using this outlier term is shown in Figure 4. However, this term has a significant drawback: the objective function becomes discontinuous due to the abruptly changing  $k$ -th nearest neighbor, resulting in the attack point often getting stuck in a local maximum. Because of this tendency, we do not examine the case of assuming this outlier detection method in our experiments.

## V. EXPERIMENTAL METHODS AND RESULTS

### A. Datasets

To evaluate the effectiveness of both the original attacks, described in [21], and the proposed attacks augmented with  $\Lambda$ , we attack LASSO regression trained on three UCI datasets: *spambase*, *credit approval*, and *congressional voting* [23].

The **spambase** dataset contains both *spam* and *non-spam* (i.e., *ham*) emails, where 48 features are the percentage of words in the email that are a particular word (e.g., address, free, business, etc.), six features are the percentage of characters in the email that are a particular character (e.g., #, \$, !), one feature is the average length of uninterrupted capital

letters, one feature is the maximum length of uninterrupted capital letters, and one feature is the sum of all uninterrupted capital letter sequences. The feasible boundary of the first 54 features is  $[0, 100]$ ; and is  $[0, \infty]$  for the last three.

The **credit approval** dataset contains features about individuals that either were or were not granted credit from a Japanese credit approval company. It contains 15 features: 6 are continuous, whose values range from 0 to 100,000; 9 are categorical, ranging from 2 to 14 categories. The exact meaning of each feature is not stated and therefore the true boundaries of the continuous input features are unknown; however, at least one feature is related to the individual's income, which has no set upper bound.

The **congressional voting** dataset contains the voting records of 435 members of the U.S. House of Representative in 1984. The 16 features are binary values that represent whether they voted for or against a bill on certain topics (e.g., El Salvador aid, religious groups in school, anti-satellite weapon test ban, etc.). The objective is to predict the congressperson's political affiliation, either Democrat or Republican.

All features are normalized such that their mean is zero and standard deviation is 1. 80% of the data is used as the training dataset, while 20% reserved for testing.

## B. Novelty and Outlier Detection Methods

For these experiments, the LASSO regression classifiers trained on the UCI datasets were augmented with four novelty and outlier detection algorithms: *distance threshold*, *one-class SVM*, *isolation forest*, and *local outlier factor*, with each algorithm being applied separately. We used our own implementation of the distance threshold algorithm, while the implementations of the one-class SVM, isolation forest, and local outlier factor algorithms were obtained from the `scikit-learn` library [24].

1) *Distance Threshold*: Distance threshold is the simplest method of outlier detection: if the Euclidean distance between a new instance  $x_c$  and its nearest neighbor  $x_i \in \mathcal{D}$  is less than some defender distance threshold  $d_{def}$ , then the new instance is added to the dataset on which the classifier is trained; otherwise, the new instance is considered an outlier and discarded, and the dataset remains unchanged. It is important to note that while they serve similar purposes, the defender distance threshold  $d_{def}$  is different from the attacker distance threshold  $d_{att}$ : the defender distance threshold determines the defender's sensitivity to *identifying* data as outliers; the attacker distance threshold determines the extent to which the attack instances can be hidden from being detected as outliers.

2) *One-Class SVM*: One-class support vector machines act as one-versus-all classifiers, drawing a classification boundary between members of a specific class (legitimate data) and all other data (i.e., poisoned attack data). To accomplish this, the optimization function differs from the traditional two-class SVM in that there is no opposing class data between which to draw an optimal hyperplane; instead, the goal is to maximize the distance between the classification hyperplane and the origin in the appropriate high-dimensional feature space [25].

Any data point that lies inside of this hyperplane is considered as a legitimate data point, and any data point that lies outside of this hyperplane is considered an outlier.

3) *Isolation Forest*: Isolation forests seek to exploit the fact that outlier data will generally exhibit two distinctive properties: i.) they are the minority data, and ii.) they will have features that are distinct from the clustered data [26]. By constructing a decision tree focused on isolating these outlier points rather than categorizing normal data points, the Isolation Tree can distinguish outliers by measuring the depth of the leaf node containing each data point: outliers will be categorized and isolated much closer to the root node than the more clustered normal data. An Isolation Forest is simply an ensemble of Isolation Trees, which achieves a higher performance than a single tree alone.

4) *Local Outlier Factor*: Local outlier factor is defined as the ratio between an instance's density compared to the density of the  $k$ -nearest neighbors. If an instance is an outlier, then the local density is expected to be much lower than the density of the  $k$ -nearest neighbors [27].

## C. Experiment 1: Original Attack Performance Against Novelty and Outlier Detection

*The Experiment*: For each training dataset, two attack instances are generated using the original attack method described in [21] using the same  $\lambda = 0.1$ . These two attack instances are combined with 40 instances from the test dataset. Outlier scores of the poisoned datasets are calculated using the four novelty and outlier detection methods described in Section III.

*Results*: Figure 5 shows the sorted outlier scores of the poisoned dataset; the attack instances are shown in red, while the legitimate instances from the test dataset are shown in blue. In all experiments, at least one novelty or outlier detection algorithm gave the attack instances the highest outlier scores, with the isolation forest, distance threshold, and one-class support vector machine giving the attack instances the highest outlier score in all cases. This simple experiment shows that, in a practical scenario, the original attack algorithm described in [21] will generate attack instances that are easily identified as outliers.

## D. Experiment 2: Improved Attack Incorporating Novelty and Outlier Detection Evasion

*The Experiment*: As in Experiment 1, two attack instances are generated for each training dataset using the modified attack method described in Section IV with the same  $\lambda = 0.1$  however now with the attacker distance threshold parameter  $d_{att}$  set to 1. These two attack instances are combined with 40 instances from the test dataset. The outlier scores of the poisoned datasets are calculated using the four novelty and outlier detection methods described in Section III.

*Results*: Figure 6 shows the sorted outlier scores of the poisoned dataset. In just about all cases, the attack points were not given the highest outlier score. In fact, in most cases the attack points received much smaller outlier scores, indicating

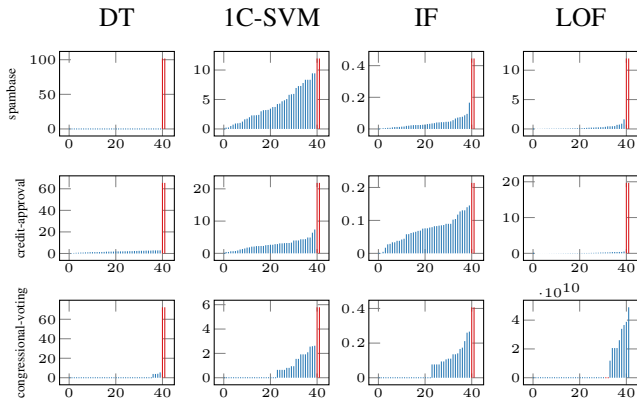


Fig. 5. The sorted outlier scores of the attack instances generated by the original attack algorithm in [21] (red) and 40 legitimate instances (blue) from the spambase, congressional voting, and credit approval datasets. The outlier scores are computed using the distance threshold, one-class SVM, local outlier factor, and isolation forest novelty and outlier detection algorithms. For all datasets, at least one of the novelty or outlier detection algorithms gave the attack instances the highest outlier score.

that they likely would not have been detected and removed as outliers. This experiment demonstrates that it is possible to generate attack instances that can evade outlier detection against classifiers trained on real-world datasets.

### E. Experiment 3: Distance Threshold Based Countermeasures on Toy Gaussian Dataset

*The Experiment:* While we have shown that the attack instances from the original attack are easily identified as outliers, and that the instances from the modified attack can be created such that they are not identified as outliers, there is a wide spectrum of attacks that can be chosen, with varying attack strengths. From the attacker’s perspective, the goal is to select the attack with the highest attack strength that is not detected by any of the defender’s countermeasures; from the defender’s perspective, the goal is to select a defense that will block attack instances while still allowing legitimate data.

Using LASSO regression with  $\lambda = 0.1$ , we train three classifiers on the toy Gaussian dataset from Figure 1, each being augmented with a distance threshold outlier detector with  $d_{def}$  set to 1, 3, or 5. As a control, we also train the classifier using LASSO regression using no outlier detection. Attack instances are generated using our improved attack algorithm, with *attacker* distance thresholds varying between 0 and 10. At each such threshold, a poisoned dataset of 50 non-malicious instances and 1 attack instance is generated. This is repeated 50 times and the average classification accuracy on the poisoned data is calculated for each defender.

*Results:* Figure 7 shows the average classification error vs. attacker distance threshold on all four cases. With no countermeasure ( $d_{def} = 0$ ), the single attack point causes a significant drop in accuracy from 96.1% to 87.8%; a remarkable drop considering that the single attack point constitutes less than 2% of the augmented dataset. Adding outlier detection reduces the negative impact of the attack instance: when the defender

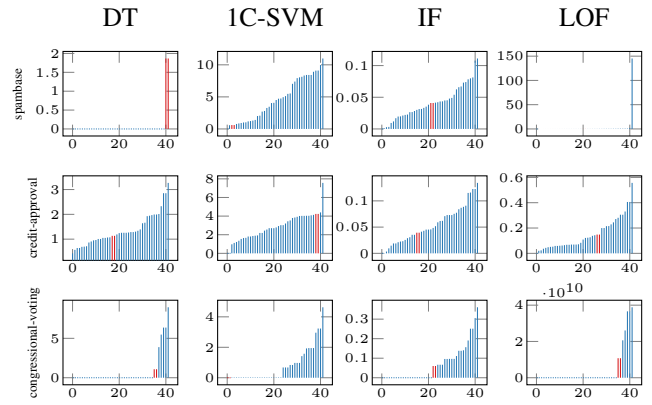


Fig. 6. The sorted outlier scores of the attack instances computed from the modified attack algorithm, using a distance threshold outlier term with an attacker distance threshold  $d_{att}$  of 1 (red), injected into a set of 40 legitimate instances (blue) from the spambase, congressional voting, and credit approval datasets. The outlier scores are computed using the distance threshold, one-class support vector machine, local outlier factor, and isolation forest. In most cases, the attack instances were not given the highest outlier score and likely would not have been labeled as outliers.

distance threshold is 1, 3, and 5, the accuracy drops only to 94.1%, 94.2%, and 93.9%, respectively. As the attacker’s outlier term is identical to the outlier detection mechanism used by the defender (because the attacker has full or partial knowledge of the defender), the distance thresholds chosen by the attacker and defender are related. The relationship between the two distance thresholds, and its effect on the classifier’s performance, is worth noting: when the attacker threshold is less than that of the defender, the attack instances are not detected as outliers and have the maximum impact; when the attacker threshold is greater than that of the defender, the attack instances are detected and removed from the dataset, and the attack is completely eliminated.

Using this modified attack, the attacker can select an appropriate threshold with respect to the countermeasures used by the defender, based on whether the attacker has full or partial knowledge of the defender (model). In this particular case, it is advantageous for the defender to use an outlier detection algorithm that is very sensitive to outliers (i.e., a low  $d_{def}$ ). However, the improved outlier detection comes at the cost of making the classifier less able to deal with drift or covariate shift as any drift will also be detected as outliers.

### F. Experiment 4: Impact of Attacker Distance Threshold on Detectability

*The Experiment:* Using the datasets described in Section V-A, we evaluate the impact of  $d_{att}$  on detectability. For each dataset-outlier detector pair, we generate attack instances using the proposed attack mechanism with attacker distance thresholds  $d_{att}$  varying between 0 and 10. At each attacker distance threshold, we calculate the outlier score, evaluated on a poisoned dataset containing the legitimate test data augmented with the appropriate attack instance.

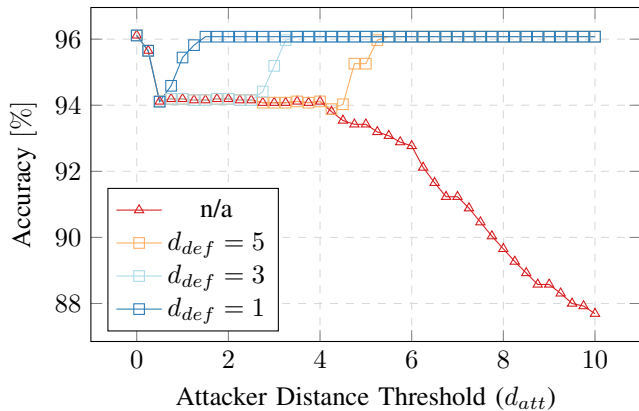


Fig. 7. The impact of applying a distance threshold based countermeasure with varying defender distance thresholds ( $d_{def} = 1, 3, 5$ ) to a LASSO regression classifier ( $\lambda = 0.1$ ) trained on a two-class Gaussian dataset. The dataset is poisoned by adding a single attack instance to the dataset generated using an attacker distance threshold between 0 and 10. The average accuracy over 50 random runs is reported.

*Results:* Figure 8 shows that for every dataset-outlier detector pair there is a clear trade-off between outlier score and attack strength. Although a higher outlier score corresponds to a greater impact on the defender’s loss, it also results in a higher detectability, which could lead to zero impact if the attack point is detected and removed.

## VI. DISCUSSION

Using real datasets, we have shown that a clear trade-off exists between the impact (i.e., damage) of an attack instance and its detectability by novelty and outlier detection techniques. While our experiments were on LASSO, there is no reason to believe that this trend between impact and detectability does not apply to other robust (stable) learners, i.e. those that are less sensitive to perturbations and small changes in their training data. For example, Biggio et al. applied an evasion attack against a linear-kernel SVM trained on the MNIST dataset, which, while effective, generated attack images that were, visually and clearly, distinct from a legitimate character, and hence easy to detect as malicious [14]. This is a stark difference from the evasion attacks against deep neural networks (relatively less robust classifiers) that have been shown to be very difficult to detect [28].

The research into defending deep learning models can be split into two primary approaches, detecting adversarial examples and building robust classifiers [28]. While there has been significant work into detecting adversarial examples [29]–[31], they have had limited success. Much of the research in defending deep learning models has gone into increasing the robustness of the model though either adversarial training [32]–[34] or defensive distillation [35], [36]. However, it has been shown that even more robust machine learning algorithms are still vulnerable to adversarial examples [14], [21]. In this paper, we provide evidence that although more robust

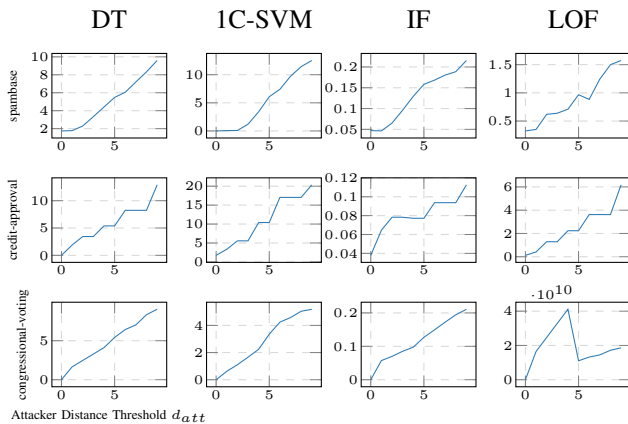


Fig. 8. Outlier score vs attacker distance threshold on the spambase, credit approval, and congressional voting datasets using distance threshold (DT), one-class support vector machines (1C-SVM), isolation forests (IF), and local outlier factor (LOF). In general, as the attacker distance threshold increases, the outlier score also increases and the attack instances are more easily detected.

algorithms are still vulnerable to adversarial examples, the adversarial examples are easier to detect.

In order to secure a machine learning system, one must both have a robust model and some method of detecting adversarial examples. If the model is not robust, it may be difficult to detect adversarial examples. If the model is robust, one must still detect and remove the adversarial examples. These trends in the literature can be summarized by the detectability-instability diagram in Figure 9.

Additionally, while adversarial machine learning is often defined as the intersection between machine learning and cybersecurity, very few connections have been made to cybersecurity. A common tool used to secure networks is an intrusion detection system that monitors network traffic for malicious activity [37]; a similar tool does not presently exist for machine learning systems, and techniques such as novelty and outlier detection as well as other existing defense techniques (e.g. reject on negative impact (RONI) [38]) should be combined in order to create something analogous to the cybersecurity intrusion detection systems.

## VII. CONCLUSIONS AND FUTURE WORK

In this work, we have shown that the attack proposed in [21] can be easily defeated using outlier detection techniques. In response, we have proposed a modified attack that allows the attacker greater control over the strength of the attack in order to evade these detection techniques. Our results show a clear correlation between the attack strength and detectability of adversarial attack instances when attacking LASSO regression augmented with novelty and outlier detection.

Future work includes testing the improved attack algorithm using more sophisticated outlier detection terms, testing different combinations of defender-outlier detection methods and attacker outlier terms to better understand how much information the attacker needs about the defenders countermeasures,

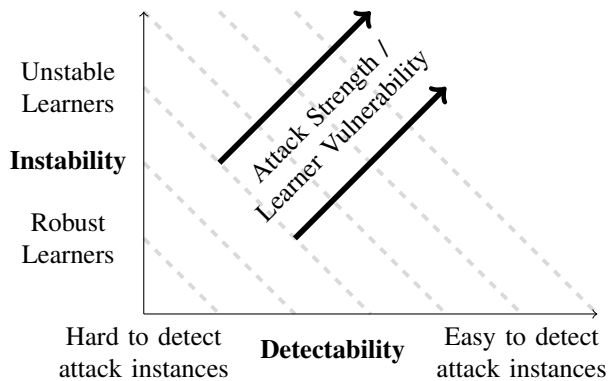


Fig. 9. A conceptual representation of the impact of the robustness of the model vs. the detectability of adversarial examples. Given two machine learning models where one is a more robust learner, to generate an attack with the same impact on the learner (i.e. attack strength), the attack against the more robust learner will be easier to detect.

testing the use of multiple stages of countermeasures, and creating a framework for combining multiple attacks in order to poison black-box classifiers using existing perfect- and limited-knowledge attacks.

#### ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under grant nos. 1310496 and 1429467.

#### REFERENCES

- [1] J. Boyan, D. Freitag, and T. Joachims, "A machine learning architecture for optimizing web search engines," in *AAAI Workshop on Internet Based Information Systems*, 1996, pp. 1–8.
- [2] L. Miralles-Pechuán, H. Ponce, and L. Martínez-Villaseñor, "A novel methodology for optimizing display advertising campaigns using genetic algorithms," *Electronic Commerce Research and Applications*, vol. 27, pp. 39–51, 2018.
- [3] C.-N. Ziegler, G. Lausen, and L. Schmidt-Thieme, "Taxonomy-driven computation of product recommendations," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, 2004, pp. 406–415.
- [4] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit card fraud detection using bayesian and neural networks," in *1st international nairo congress on neuro fuzzy technologies*, 2002, pp. 261–270.
- [5] H.-M. Chi, O. K. Ersoy, H. Moskowit, and J. Ward, "Modeling and optimizing a vendor managed replenishment system using machine learning and genetic algorithms," *European Journal of Operational Research*, vol. 180, no. 1, pp. 174–193, 2007.
- [6] R. Berk, *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media, 2012.
- [7] V. N. Gadeally, K. B. Greenfield, W. M. Campbell, J. P. Campbell, A. I. Reuther, and B. J. Hancock, "Recommender systems for the department of defense and the intelligence community," MIT Lincoln Laboratory Lexington United States, Tech. Rep., 2016.
- [8] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, 2011, pp. 43–58.
- [9] B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," *arXiv preprint*, 2017.
- [10] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 2004, p. 99.
- [11] M. Kearns and M. Li, "Learning in the presence of malicious errors," *SIAM Journal on Computing*, vol. 22, no. 4, pp. 807–837, 1993.
- [12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

- [13] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. ACM, 2006, pp. 16–25.
- [14] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *arXiv preprint arXiv:1206.6389*, 2012.
- [15] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrasamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 27–38.
- [16] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrđić, P. Laskov, G. Giacinto, and F. Roli, "Evasion Attacks against Machine Learning at Test Time," in *Machine Learning and Knowledge Discovery in Databases*, vol. 8190. Springer Berlin Heidelberg, 2013, pp. 387–402.
- [17] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *arXiv preprint arXiv:1602.02697*, 2016.
- [18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *Iclr 2015*, 2015, pp. 1–11.
- [19] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," in *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, 2016, pp. 582–597.
- [20] B. Biggio, I. Pillai, S. Rota Bulò, D. Ariu, M. Pelillo, and F. Roli, "Is data clustering in adversarial settings secure?" in *2013 ACM workshop on Artificial intelligence and security*. ACM, 2013, pp. 87–98.
- [21] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli, "Is feature selection secure against training data poisoning?" in *International Conference on Machine Learning*, 2015, pp. 1689–1698.
- [22] C. Clavier, "Secret external encodings do not prevent transient fault analysis," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2007, pp. 181–194.
- [23] M. Lichman, "UCI machine learning repository," 2013.
- [24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [25] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in neural information processing systems*, 2000, pp. 582–588.
- [26] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *8th IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.
- [27] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.
- [28] X. Cao and N. Z. Gong, "Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification," *arXiv preprint*.
- [29] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting Adversarial Samples from Artifacts," *arXiv preprint*, 2017.
- [30] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and Clean Data Are Not Twins," *arXiv preprint*, 2017.
- [31] D. Hendrycks and K. Gimpel, "Early Methods for Detecting Adversarial Images," in *ICLR 2017 Workshop track*, 2017.
- [32] A. G. Ororbica, C. L. Giles, and D. Kifer, "Unifying Adversarial Training Algorithms with Flexible Deep Data Gradient Regularization," *arXiv preprint*, 2016.
- [33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv preprint*, 2017.
- [34] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," in *ICLR 2017*, 2017.
- [35] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," in *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, 2016, pp. 582–597.
- [36] N. Papernot and P. McDaniel, "Extending Defensive Distillation," *arXiv preprint*, pp. 1–11, 2017.
- [37] M. Roesch, "Snort: Lightweight Intrusion Detection for Networks." *LISA '99: 13th Systems Administration Conference*, pp. 229–238, 1999.
- [38] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia, "Exploiting machine learning to subvert your spam filter," in *Proceedings of the First Workshop on Large-scale Exploits and Emerging Threats (LEET)*, p. Article 7, 2008.