

Enhancing Robustness of Machine Learning Systems via Data Transformations

Arjun Nitin Bhagoji
Princeton University

Daniel Cullina
Princeton University

Bink Sitawarin
Princeton University

Prateek Mittal
Princeton University

Abstract—We propose the use of *data transformations* as a defense against evasion attacks on ML classifiers. We present and investigate strategies for incorporating a variety of data transformations including *dimensionality reduction* via Principal Component Analysis and data ‘anti-whitening’ to enhance the resilience of machine learning, targeting both the classification and the training phase. We empirically evaluate and demonstrate the feasibility of linear transformations of data as a defense mechanism against evasion attacks using multiple real-world datasets. Our key findings are that the defense is (i) effective against the best known evasion attacks from the literature, resulting in a two-fold increase in the resources required by a white-box adversary with knowledge of the defense for a successful attack, (ii) applicable across a range of ML classifiers, including Support Vector Machines and Deep Neural Networks, and (iii) generalizable to multiple application domains, including image classification and human activity classification.

I. INTRODUCTION

We are living in an era of ubiquitous machine learning (ML) and artificial intelligence. Machine learning is being used in a number of essential applications such as image recognition [26], natural language processing [9], spam detection [10], autonomous vehicles [8], [33] and even malware detection [46], [11]. High classification accuracy in these settings [23], [48], [2] has enabled the widespread deployment of ML systems. Given the ubiquity of ML applications, it is increasingly being deployed in *adversarial* scenarios, where an attacker stands to gain from the failure of a ML system to classify inputs correctly. The question then arises: are ML systems secure in adversarial settings?

Adversarial Machine Learning: Starting in the early 2000s, there has been a considerable body of work [20], [3], [25] exposing the vulnerability of machine learning algorithms to strategic adversaries. For example, *poisoning attacks* [5] systematically introduce adversarial data during the *training* phase with the aim of causing the misclassification of data during the test phase. On the other hand, *evasion attacks* [4], [35], [47] aim to fool existing ML classifiers trained on benign data by adding *strategic perturbations* to *test inputs*.

Evasion attacks: In this paper we focus on evasion attacks in which the adversary aims to perturb test inputs to ML classifiers in order to cause misclassification. Evasion attacks have been proposed for a variety of machine learning classifiers such as Support Vector Machines [4], [34], tree-based classifiers [34], [22] such as random forests and boosted trees and more recently for neural networks [16], [47], [35], [24], [7], [32]. These attacks have been used to demonstrate the vulnerability of applications that use machine learning, such

as facial recognition [29], [43], voice command recognition [6] and PDF malware detection [53] in laboratory settings. Recent work also illustrates the possibility of attacks on deployed systems such as the Google video summarization API [19], highlighting the urgent need for defenses. Surprisingly, it has also been shown that the evasion properties of adversarially modified data (for a particular classifier) persist across different ML classifiers [47], which allows an adversary with limited knowledge of the ML system to attack it.

However, very few defenses [39], [22] exist against these attacks, and the applicability of each is limited to only certain known attacks and specific types of ML classifiers (see Section VII for a detailed description of and comparison with previous work).

A. Contributions

We propose and thoroughly investigate the use of linear transformations of data as a defense against evasion attacks. We consider powerful adversaries with *knowledge of our defenses* when evaluating their effectiveness and find that they demonstrably reduce the success of evasion attacks. To the best of our knowledge, ours are the only defenses against evasion attacks with the following properties: (1) applicability across multiple ML classifiers (such as SVMs, DNNs), (2) applicability in varied application domains (image and activity classification), and (3) mitigation of multiple attack types, including strategic ones. Further, the tunability of our defense allows a system designer to pick appropriate operating points on the utility-security tradeoff curve depending on the application.

1) *Defense*: We propose the use of *data transformations* as a defense mechanism. Specifically, we consider linear dimensionality reduction techniques such as Principal Component Analysis which aim to project high-dimensional data to a lower-dimensional space while preserving the most useful variance of the data [44], [51]. We present and investigate a strategy for incorporating dimensionality reduction and other linear transformations of data to enhance the resilience of machine learning, targeting both the classification and training phases. Data transformations are applied to the training data to *enhance the resilience of the trained classifier* and they significantly change the learned classifier. Linear data transformations are a generalization of regularization methods. They allow us to access novel and otherwise inaccessible robustness-performance tradeoffs.

2) *Empirical Evaluation*: We empirically demonstrate the feasibility and effectiveness of our defenses using:

- multiple ML classifiers, such as Support Vector Machines (SVMs) and Deep Neural Networks (DNNs);
- several distinct types of evasion attacks, such as an attack on Linear SVMs from Moosavi-Dezfooli et. al. [30], and on deep neural networks from Goodfellow et. al. [16] and Carlini et al. [7], which is the best known attack for neural networks, as well as white-box attacks targeting our defense;
- a variety of real-world datasets/applications: the MNIST image dataset [27] and the UCI Human Activity Recognition (HAR) dataset [1].

Our key findings are that even in the face of a white-box adversary with complete knowledge of the ML system:

- **Security**: the defense leads to significant increases of up to $5\times$ in the degree of modification required for a successful attack and equivalently, reduces adversarial success rates by around $2 - 50\times$ at fixed levels of perturbation.
- **Generality**: the defense can be used for different ML classifiers (and application domains) with minimal modification of the original classifiers, while still being effective at combating adversarial examples.
- **Utility**: there is a modest utility loss of about 0.5-2% in the classification success on benign samples in most cases.

We also provide an analysis of the utility-security tradeoffs as well as the computational overheads incurred by our defense. Our results may be reproduced using the open source code available at <https://github.com/anonymous1>.

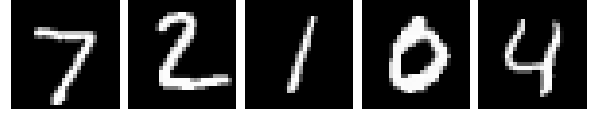
We note that our defense does not completely solve the problem of evasion attacks, since it reduces adversarial success rates at fixed budgets, but does not make them negligible in all cases. However, since our defense is classifier and dataset-agnostic it can be used in conjunction with other defenses such as adversarial training in order to close this gap. We will explore the synergy of our defense with other techniques in future work. We hope that our work inspires further research in combating the problem of evasion attacks and securing machine learning based systems.

II. ADVERSARIAL MACHINE LEARNING

In this section, we present the required background on adversarial machine learning, focusing on evasion attacks that induce misclassification by perturbing the input at test time.

Motivation and Running Example: We use image data from the MNIST dataset [27] for our running examples. Figure 1(a) depicts example test images from the MNIST dataset that are correctly classified by a SVM classifier (see Section IV for details), while Figure 1(b) depicts adversarially crafted test images (perturbed images using the evasion attack of Moosavi-Dezfooli et. al. [30]), which are misclassified by a linear SVM.

¹Link anonymized for double-blind submission



(a) Typical test images from the MNIST dataset. Correctly classified as 7, 2, 1, 0 and 4 respectively.



(b) Corresponding adversarial images obtained using the evasion attack on Linear SVMs [30]. Now, misclassified as 9, 9, 3, 2 and 0 respectively.

Fig. 1: Comparison of benign and adversarial images taken from the MNIST dataset.

Classification using machine learning: We focus on *supervised* machine learning, in which a classifier is trained on labeled data. For our purposes, a classifier is a function that takes as input a data point $\mathbf{x} \in \mathbb{R}^d$ and produces an output $\hat{y} \in C$, where C is the set of all categories. The classifier succeeds if \hat{y} matches the true class $y \in C$. For example, for the MNIST dataset, \mathbf{x} is a 28×28 pixel grayscale image of a handwritten digit and C is the finite set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

A. Attacks against machine learning systems

In this subsection we lay out the notation for the remainder of the paper, and describe the adversarial model under consideration.

1) *Adversarial goals*: In each case, the adversary is given an input \mathbf{x} with true class y and uses an attack algorithm A to produce a modified input $\tilde{\mathbf{x}} = A(\mathbf{x})$.

Attacks are relevant in the case where this input is correctly classified: $f(\mathbf{x}) = y$. The attack takes one of two forms.

- *Untargeted attack*: $\tilde{\mathbf{x}}$ is misclassified ($f(\tilde{\mathbf{x}}) \neq y$)
- *Targeted attack*: $\tilde{\mathbf{x}}$ is classified as a specific class $t \in C$ ($f(\tilde{\mathbf{x}}) = t, t \neq y$)

Note that these attacks are equivalent for binary classifiers. Additionally, the adversarial example $\tilde{\mathbf{x}}$ should be *as similar to the benign example as possible*. Similarity is quantified using some metric on the space of examples. As in previous work, we focus on example spaces that are vector spaces and use p -norms as the metric [16], [7].

2) *Adversarial knowledge*: We consider three settings. The first is the *white box* setting, in which the adversary knows the classification function f , all its parameters and the existence of a defense, if any. This assumption on the adversary's capabilities is conservative as this setting corresponds to a very powerful adversary. A system with the ability to defend against attacks from an adversary with complete knowledge does not rely on security through obscurity.

In the second setting, we consider somewhat less powerful adversaries which may be more prevalent. We call this is the *classifier mismatch* setting. Here, the adversary knows the training dataset \mathbf{X} , the architecture of the the original

classifier and the hyperparameters (i.e regularization constants etc.) used in the training of the original classifier *without the defense*. Thus, the adversary is capable of training a classifier \hat{f} that mimics that true classifier f . The adversary generates examples that are adversarial on \hat{f} and submits them to f . Note that since the adversary is not aware of the defense measures taken, there is a mismatch between \hat{f} and f , hence the term ‘classifier mismatch’ for this setting.

Further, it has been shown in previous work [34], [47] that adversarial examples transfer across classifiers with different architectures and hyperparameter settings, so adversaries can use their own models trained on similar datasets to construct adversarial samples for the ML system under attack. Thus, we consider a third setting, the *architecture mismatch* setting², where the adversary is unaware of the classifier architecture being used, and just trains some \hat{f} on the portion of the training data available. This is a plausible practical setting, since knowledge about the architecture and the hyperparameters of the network under attack may be difficult for the adversary to obtain.

B. Evasion attacks on specific classifiers

We now describe existing attacks from the literature for specific ML classifiers such as linear classifiers and neural networks. These are summarized in Table I

1) *Optimal attacks on linear classifiers*: In the multi-class classification setting for Linear SVMs, a classifier g_i is trained for each class $i \in C$, where

$$g_i : \mathbf{x} \mapsto \mathbf{w}_i^T \mathbf{x} + b_i. \quad (1)$$

The final classifier f assigns \mathbf{x} to the class $f(\mathbf{x}) = \operatorname{argmax}_{i \in C} g_i(\mathbf{x})$. Given that the true class is $y \in C$, the objective of an untargeted attack is to find the closest point $\tilde{\mathbf{x}}$ such that $f(\tilde{\mathbf{x}}) \neq y$.

From [30], we know the optimal attacks on affine multi-class classifiers under the ℓ_2 metric. This attack finds $\tilde{\mathbf{x}}$ that minimizes $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2$ subject to the constraint $f(\tilde{\mathbf{x}}) \neq t$. For \mathbf{x} such that $f(\mathbf{x}) = y$ and targeted class $t \in C$, let $\mathbf{z}_{t,y} = \mathbf{w}_t - \mathbf{w}_y$. Observe that $g_t(\mathbf{x}) - g_y(\mathbf{x}) = \mathbf{z}_{t,y}^T \mathbf{x}$. Then, the adversarial example

$$\tilde{\mathbf{x}}_t = \mathbf{x} - \frac{(\mathbf{z}_{t,y}^T \mathbf{x}) \mathbf{z}_{t,y}}{\|\mathbf{z}_{t,y}\|_2^2} \quad (2)$$

satisfies $\mathbf{z}_{t,y}^T \tilde{\mathbf{x}}_t = 0$. The minimum modification required to cause a misclassification as t is

$$\epsilon_t = \|\mathbf{x} - \tilde{\mathbf{x}}_t\|_2 = \frac{\|\mathbf{z}_{t,y} \mathbf{z}_{t,y}^T \mathbf{x}\|_2}{\|\mathbf{z}_{t,y}\|_2^2} = \frac{\mathbf{z}_{t,y}^T \mathbf{x}}{\|\mathbf{z}_{t,y}\|_2}. \quad (3)$$

Thus the optimal choice of $\tilde{\mathbf{x}}$ for an untargeted attack is $\tilde{\mathbf{x}}_k$, where $k = \operatorname{argmin}_t \epsilon_t$.

²This is commonly referred to in the literature as a *black-box* setting, but we use a different terminology to highlight the exact nature of the adversary’s lack of knowledge.

2) *Gradient based attacks on neural networks*: The Fast Gradient Sign (FGS) attack is an efficient attack against neural networks introduced in [16]. This attack is for the ℓ_∞ metric. Adversarial examples are generated by adding adversarial noise proportional to the sign of the gradient of the loss function $J(\mathbf{x}, y, \theta)$. Here, \mathbf{x} is the example, y is the true class, and θ is the network weight parameters. Concretely,

$$\tilde{\mathbf{x}}(\eta)_i = \mathbf{x}_i + \eta \operatorname{sgn}(\nabla_{\mathbf{x}} J(\mathbf{x}, y, \theta))_i. \quad (4)$$

The gradient can be efficiently calculated using backpropagation. The parameter η controls the magnitude of the adversarial perturbation, similar to ϵ for the attack on Linear SVMs:

$$\|\mathbf{x} - \tilde{\mathbf{x}}(\eta)\|_\infty = \max_i |\eta \operatorname{sgn}(\nabla_{\mathbf{x}} J(\mathbf{x}, y, \theta))_i| = \eta. \quad (5)$$

See Figure 13 in the Appendix for images modified with a range of η .

The FGS attack and the attack on Linear SVMs are constrained according to different norms. To facilitate a comparison of the robustness of classifiers as well as the effectiveness of our defense across them, we propose a modification of the FGS attack which is constrained by the ℓ_2 norm. Denoting this as the Fast Gradient (FG) attack, we define the adversarial examples to be equal to

$$\tilde{\mathbf{x}}(\epsilon) = \mathbf{x} + \epsilon \frac{\nabla_{\mathbf{x}} J(\mathbf{x}, y, \theta)}{\|\nabla_{\mathbf{x}} J(\mathbf{x}, y, \theta)\|_2}. \quad (6)$$

For the FG attack, ϵ is the ℓ_2 norm of the perturbation.

3) *Optimization-based attacks on neural networks*: The direct optimization based formulation of adversarial sample generation for a classifier f is

$$\begin{aligned} \min \quad & d(\tilde{\mathbf{x}}, \mathbf{x}), \\ \text{s.t.} \quad & f(\tilde{\mathbf{x}}) \neq f(\mathbf{x}) \\ & \tilde{\mathbf{x}} \in C, \end{aligned} \quad (7)$$

where d is an appropriately chosen distance metric and C is the constraint on the input space. Since the constraint $f(\tilde{\mathbf{x}}) \neq f(\mathbf{x})$ in the above optimization problem is combinatorial, various related forms of this minimization problem have been proposed [47], [7], [30]. We focus on the relaxation studied by Carlini and Wagner [7]:

$$\begin{aligned} \min \quad & d(\tilde{\mathbf{x}}, \mathbf{x}) + \lambda \ell(\tilde{\mathbf{x}}, f), \\ \text{s.t.} \quad & \tilde{\mathbf{x}} \in C. \end{aligned} \quad (8)$$

Carlini and Wagner investigate a variety of loss functions $\ell(\cdot)$ as well as methods to ensure the generated adversarial sample stays within the input space constraints. In our experiments with neural networks, for untargeted attacks we use a loss function, $\max(Z(\tilde{\mathbf{x}})_o - \max\{Z(\tilde{\mathbf{x}}) : i \neq o\}, -\kappa)$, where o is the original class of the input, $Z(\cdot)$ represents the output of the neural network before the softmax layer and κ represents the confidence parameter. The distance metric used is the ℓ_2 norm since it is found to perform the best.

We evaluate the effectiveness of our defense against this state-of-the-art attack in Section V.

TABLE I: Summary of attacks on Linear SVMs and neural networks.

Attack	Classifier	Constraint	Intuition
Optimal attack on Linear SVMs [30]	Linear SVMs	ℓ_2	Move towards classifier boundary
Fast Gradient	Neural networks	ℓ_2	First order approximation to direction of smallest perturbation
Fast Gradient Sign [16]	Neural networks	ℓ_∞	Constant scaling for each pixel models perception better
Carlini's ℓ_2 attack [7]	Neural networks	ℓ_2	Iterative optimization over relaxed minimization problem

III. DATA TRANSFORMATIONS AS A DEFENSE

In the previous section, we have seen that ML classifiers are vulnerable to a variety of different evasion attacks. Thus, there is a clear need for a defense mechanism that is effective against a variety of attacks, since a priori, the owner of the system has no knowledge of the range of possible attacks. Further, finding a defense that works across multiple classifiers can direct us to a better understanding of why ML systems are vulnerable in the first place.

In this section, we present a defense that is resilient to attacks from the literature in the mismatch setting, and remains effective even in the presence of a white-box adversary with knowledge of the defense. Further, the defense makes multiple types of ML classifiers operating in different application scenarios more robust as shown by our results in Section V. The defense is based on linear transformations of data, including linear dimensionality reduction.

A. Overview of defense

The dimension of the data is d and the training data is a $d \times n$ matrix \mathbf{X} , so each example is a column. We assume the data is centered, i.e. $\mathbf{X}\mathbf{1} = \mathbf{0}$ where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones and $\mathbf{0} \in \mathbb{R}^d$ is the vector of all zeros. The set of data classes is C and the classifier in use is $f: \mathbb{R}^d \rightarrow C$.

In our defense, we leverage linear transformations of the data to make the classifier more resilient by modifying the training phase. In the first step of our defense, an algorithm selects a linear transformation such as dimensionality reduction based on properties of the data distribution. Then, the training data \mathbf{X} is transformed and a new classifier f is then trained on the transformed *training set*. In the classification phase, all inputs are transformed in the same way before being provided to the classifier.

Algorithm 1 LTtrain

Input: \mathbf{X} , Train, Select

Output: f

- 1: Select the linear transformation $\mathbf{B} = \text{Select}(\mathbf{X})$
 - 2: Compute the transformed training set \mathbf{BX}
 - 3: Train classifier $f^{\text{aux}} = \text{Train}(\mathbf{BX})$
 - 4: Let $f = (\mathbf{x} \mapsto f^{\text{aux}}(\mathbf{Bx}))$
-

The additional inputs to Algorithm 1 are:

- **Select:** The algorithm used to select a linear transformation of the data based on some properties of the

data. This may be a specialization of a more general algorithm to specific parameters: e.g. `Select = TopPrincipalComponents(k)`.

- **Train:** This is the algorithm used to train classifiers of the desired class. In general, this will be a specialization of more general training algorithm to specific parameters. For example, **Train** might produce a neural network via Stochastic Gradient Descent [15] starting from the untrained classifier network f^0 using training parameters θ .

At first glance, this approach may seem futile, because the initial layer of many machine learning classifiers, including SVMs and neural networks, is a linear function. However, although these classifiers are already capable of applying any linear transformation to the data that the training procedure finds to be useful, the standard training process does not optimize for adversarial robustness and does not choose to make these transformation, even though they are available. Thus, an *explicit linear transformation of the data is capable of significantly changing the learned classifier* and a carefully selected transformation can lead to beneficial changes.

B. Effect on Support Vector Machines

To motivate Algorithm 1, we examine in detail the case where **Train** produces a linear classifier by learning a support vector machine.

Learning a SVM [41] that can classify data points from two classes, $y_i \in \{-1, 1\}$, involves finding an affine function $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ that minimizes the following loss function

$$L(\mathbf{X}; \mathbf{w}, b) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_i \max(-1, y_i(\mathbf{w}^T \mathbf{x}_i + b)). \quad (9)$$

If we use Algorithm 1 and apply an invertible linear transformation \mathbf{B} to the training data, we will learn an alternative function $g^{\text{aux}}(\mathbf{x}) = \mathbf{u}^T \mathbf{x} + b'$ that minimizes

$$L(\mathbf{AX}; \mathbf{u}, b') = \frac{1}{2} \mathbf{u}^T \mathbf{u} + \sum_i \max(-1, y_i(\mathbf{u}^T \mathbf{Bx}_i + b')). \quad (10)$$

Our actual classifier will be $g(\mathbf{x}) = g^{\text{aux}}(\mathbf{Bx}) = \mathbf{u}^T \mathbf{Bx} + b'$. Let $\mathbf{w} = \mathbf{B}^T \mathbf{u}$ and rewrite (10) as

$$\begin{aligned} & \frac{1}{2} \mathbf{u}^T \mathbf{B} \mathbf{B}^T (\mathbf{B}^{-1})^T \mathbf{B}^T \mathbf{u} + \sum_i \max(-1, y_i(\mathbf{u}^T \mathbf{Bx}_i + b)) \\ &= \frac{1}{2} \mathbf{w}^T (\mathbf{B} \mathbf{B}^T)^{-1} \mathbf{w} + \sum_i \max(-1, y_i(\mathbf{w}^T \mathbf{x}_i + b)). \end{aligned} \quad (11)$$

Selecting the \mathbf{u} that minimizes (10) and composing it with \mathbf{B}^\top is equivalent to directly selecting the value of \mathbf{w} that minimizes (11). Thus applying an invertible linear transformation to the data is equivalent to modifying the quadratic form that appears in the regularization term of the SVM loss function.

Regularization: A standard generalization of (9) multiplies the $\frac{1}{2}\mathbf{w}^\top\mathbf{w}$ by a regularization parameter λ . This corresponds to the simplest possible linear transformation of the data: multiplication by a constant. Explicitly, we have $\mathbf{B} = \frac{1}{\sqrt{\lambda}}\mathbf{I}$. Thus, ordinary regularization of SVMs fits neatly into this framework. However, more general linear transformations provide us with significantly more flexibility to modify the regularization constraint and allow us to *access novel and otherwise inaccessible robustness-performance tradeoffs*.

Singular linear transformations: What happens if \mathbf{B} is not invertible? In this case, $\mathbf{B}^\top\mathbf{u}$ is a member of $\text{im } \mathbf{B}^\top$ by definition.³ Then $\mathbf{w} = \mathbf{B}^\top\mathbf{u}$ minimizes

$$\frac{1}{2}\mathbf{w}^\top(\mathbf{B}\mathbf{B}^\top)^+\mathbf{w} + \sum_i \max(-1, y_i(\mathbf{w}^\top\mathbf{x}_i + b))$$

over $\mathbf{w} \in \text{im } \mathbf{B}^\top$, where $(\mathbf{B}\mathbf{B}^\top)^+$ is the Moore-Penrose pseudoinverse of \mathbf{B} .⁴ [38]. In addition to modifying the costs assigned to each weight vector, applying a non-invertible transformation \mathbf{B} to the data rules out some choices of \mathbf{w} completely. Alternatively, one can think of the regularization term as assigning an infinite cost to each $\mathbf{w} \in \ker \mathbf{B}$ and thus to each $\mathbf{w} \notin \text{im } \mathbf{B}^\top$.⁵

Expressivity: For a fixed collection of data points, it is possible to find a linear transformation that results in the selection of essentially any hard-decision classifier positively correlated with the true labels. Observe that this is the opposite of the naive fear described in Section III-A that linear transformations should have no effect on the learned classifier: the choice of linear transformation might influence the final classifier structure too much! Because of this, it is essential to select linear transformations in a systematic matter.

C. Defense using PCA

Several of the choices of **Select** that we will use in Algorithm 1 are based on Principal Component Analysis.

1) *PCA in brief:* PCA [44] is a linear transformation of the data that identifies so-called ‘principal axes’ in the space of the data, which are the directions in which the data has maximum variance, and projects the original data along these axes. The dimensionality of the data is reduced by choosing to project it along k principal axes. The choice of k depends on what percentage of the original variance is to be retained. Intuitively, PCA identifies the directions in which the ‘signal’, or useful information in the data is present, and discards the rest as noise.

³The image of the operator \mathbf{B}^\top is the vector space of linear combinations of columns of the matrix \mathbf{B}^\top .

⁴Because $\mathbf{B}\mathbf{B}^\top$ is a symmetric matrix, it has a spectral decomposition $\mathbf{B}\mathbf{B}^\top = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ where $\mathbf{\Lambda}$ is diagonal. Then $(\mathbf{B}\mathbf{B}^\top)^+ = \mathbf{V}\mathbf{\Lambda}^+\mathbf{V}^\top$ where $\mathbf{\Lambda}^+$ is diagonal, $(\mathbf{\Lambda}^+)_{ii} = \Lambda_{ii}^{-1}$ if $\Lambda_{ii} \neq 0$, and $(\mathbf{\Lambda}^+)_{ii} = 0$ otherwise.

⁵The kernel of the operator \mathbf{B} is the space of vectors \mathbf{x} such that $\mathbf{B}\mathbf{x} = \mathbf{0}$. $\ker \mathbf{B}$ is orthogonal to $\text{im } \mathbf{B}^\top$.

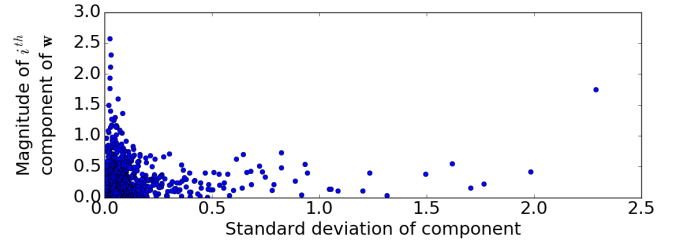


Fig. 2: **Magnitudes of the coefficients of the weight vector \mathbf{w} of a linear SVM in the principal component basis.** On the horizontal axis, we have $\sqrt{\lambda_i}$. On the vertical axis, $|(\mathbf{U}^\top\mathbf{w})_i|$. The classifier is trained on the original MNIST data.

Concretely, let the data samples be column vectors $\mathbf{x}_i \in \mathbb{R}^d$ for $i \in \{1, \dots, n\}$, let \mathbf{X} be the $d \times n$ matrix of centered data samples. The principal components of \mathbf{X} are the normalized eigenvectors of its sample covariance matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^\top$. More precisely, because \mathbf{C} is positive semidefinite, there is a decomposition $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ where \mathbf{U} is an orthogonal matrix, $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$, and $\lambda_1 \geq \dots \geq \lambda_d \geq 0$. In particular, \mathbf{U} is the $d \times d$ matrix whose columns are unit eigenvectors of \mathbf{C} . The eigenvalue λ_i is the variance of \mathbf{X} along the i^{th} component.

Each column of $\mathbf{U}^\top\mathbf{X}$ is a data sample represented in the principal component basis. Let \mathbf{X}_k be the projection of the sample data in the k -dimensional subspace spanned by the k largest principal components. Thus $\mathbf{X}_k = \mathbf{U}\mathbf{I}_k\mathbf{I}_k^\top\mathbf{U}^\top\mathbf{X}$, where \mathbf{I}_k is a $d \times k$ rectangular diagonal matrix. The amount of variance retained is $\sum_{i=1}^k \lambda_i$, which is the sum of the k largest eigenvalues.

2) *Implementing the defense:* There are two choices of a linear transformation that keep the k largest principal components: $\mathbf{B} = \mathbf{I}_k^\top\mathbf{U}^\top$, which is a $k \times d$ matrix, and $\mathbf{B} = \mathbf{U}^\top\mathbf{I}_k\mathbf{U}$, which is a $d \times d$ matrix. For some choices of Train including SVMs trained using (9), these choices of \mathbf{A} are equivalent, i.e. they will output identical classifiers given the same inputs. The choice $\mathbf{B} = \mathbf{I}_k^\top\mathbf{U}^\top$ allows for more efficient training because representation of the data is more compact. The choice $\mathbf{B} = \mathbf{U}^\top\mathbf{I}_k\mathbf{U}$ makes it easier to compare the reduced dimension data to the original data.

The complexity analysis for the PCA-based defense is in Section IX-A (c.f. Appendix).

D. Intuition behind the PCA defense

We will give some intuition about why dimensionality reduction should improve resilience for SVMs. We discuss the two-class case for simplicity, but the ideas generalize to the multiple class case. The core of a linear classifier is a function $g(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b$. Both \mathbf{x} and \mathbf{w} can be expressed in the principal component basis as $\mathbf{U}^\top\mathbf{x}$ and $\mathbf{U}^\top\mathbf{w}$. We expect that many of the principal components with the largest coefficients in the weight vector, $|(\mathbf{U}^\top\mathbf{w})_i|$, to correspond to small eigenvalues λ_i .

The reason for this is very simple: in order for different principal components to achieve the same level of influence on

the classifier output, $|(\mathbf{U}^T \mathbf{w})_i|$ must be proportional to $1/\sqrt{\lambda_i}$. To take advantage of the information in a component with small variance, the classifier must use a large coefficient. Of course, the principal components vary in their usefulness for classification. However, among the components that are useful, we expect a general trend of decreasing coefficients of $(\mathbf{U}^T \mathbf{w})_i$ as $\sqrt{\lambda_i}$ increases.

Figure 2 validates this prediction. Many of the principal components with very low variances have large coefficients in \mathbf{w} . As variance increases, the coefficients tend to decrease. The exception to the trend is the first principal component, but this is not surprising. The first principal component⁶ is by far the most useful source of classification information because it is strongly aligned with the difference of the class means. Consequently it does not fit the overall trend and actually has the largest coefficient. However, among the other components, there is a mixture of cross-class and in-class variation and the trend holds.

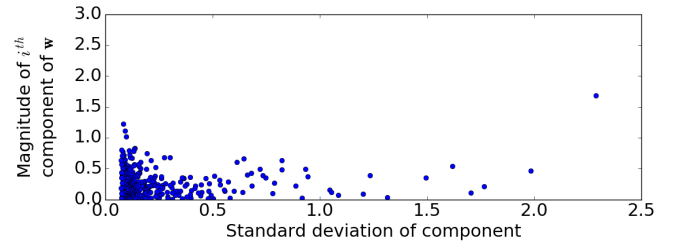
Effect on robustness: Since the optimal attack perturbation for a linear classifier is a multiple of \mathbf{w} , the *principal components with large coefficients are the ones that the attacker takes advantage of*. The defense denies this opportunity to the attacker by forcing the classifier to assign no weight to the low variance components. This significantly changes the resulting \mathbf{w} that the classifier learns. The classifier loses access to some information, but accessing that information required large weight coefficients, so the attacker is hurt far more. *Thus, by using only high variance components, the classifier gains significant adversarial robustness for the loss of a small amount of classification performance.*

Figure 3 shows the coefficient magnitudes for classifiers trained on data projected onto the top k principal components. Observe that eliminating the low variance principal components mostly removes the relationship between the variance of a component and the corresponding coefficient of $\mathbf{U}^T \mathbf{w}$.

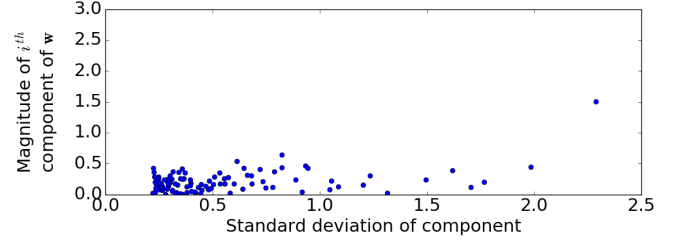
E. Other linear transformations and classifiers

1) *Anti-whitening:* Now we will discuss another linear transformation based on the principal components of the training data that can confer additional robustness. As before, we have $\mathbf{C} = \mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. The linear transformation which we call ‘anti-whitening’ is accomplished by selecting $\mathbf{B} = \mathbf{\Lambda}^{\frac{c}{2}}\mathbf{U}$ and $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}^{\frac{c}{2}}\mathbf{U}$ for some $c > 0$. In our experiments, we use the former but the latter is conceptually easier to work with. Anti-whitening exaggerates the disparity between the variances of the components. In the full rank case, the quadratic form introduced in the SVM loss is \mathbf{C}^{-c} . It serves as a softer alternative to completely eliminating low variance principal components: the hard cutoff $\mathbf{I}_k\mathbf{I}_k^T$ is replaced with the gradual penalty $\mathbf{\Lambda}^{\frac{c}{2}}$. Low variance components are still available for use, but the price of accessing them is increased.

2) *Neural networks:* Neural networks are more complicated due to the non-uniqueness of local minima in the associated



(a) $k = 331$



(b) $k = 100$

Fig. 3: **The magnitudes of the coefficients of the weight vector \mathbf{w} of a linear SVM in the principal component basis.** On the horizontal axis, we have $\sqrt{\lambda_i}$. On the vertical axis, $|(\mathbf{U}^T \mathbf{w})_i|$. The classifiers are trained on the MNIST data projected onto the top k principal components.

loss function and the larger variety of regularization methods that are employed. At first glance, it may seem that adding a linear layer as the first layer of a neural network may provide the same benefits as PCA-based dimensionality reduction. However, the training process does not optimize for robustness, so in practice the linear layer that is learned does not have the desired effect⁷, unlike in our defense where the linear layer weights are separately specified using PCA. Thus, the intuition for the effectiveness of linear transformations carries over from the case of Linear SVMs, with the adversary losing access to dimensions which aid in the creation of adversarial samples, while the classifier remains largely unaffected since most of the information required for classification is retained. Further, in our empirical results in Section V, it is clear that the average distance to the boundary of the classifier increases when the linear transformation is added, thus leading to robustness.

Finally, we note that for Linear SVMs, standard regularization can also be understood as increasing the price of all components, which reduces the dependence of the classifier on components that only provide marginal classification benefit. However, in the case of neural networks, the usual regularization of the weight matrix does not follow the intuition given here for increasing robustness.

F. Attacks against the linear transformation defense

In order to evaluate our proposed defense mechanism, we carry out the attacks described in Section II-B against classifiers learned using our defense. Both for linear classifiers

⁷In fact, Gu and Rigazio [17] found that even non-linear denoising layers do not add adversarial robustness.

⁶on the top right in each plot

and neural networks, the classifier learned using our defense lies in the same family as the classifier that would be learned without the defense. Thus, simple modifications of existing attacks give the white-box versions of attacks against the classifier with the defense.

1) *White-box attacks*: Due to the inclusion of a linear transformation of the data, the overall classifier is $f(\mathbf{x}) = f^{\text{aux}}(\mathbf{B}\mathbf{x})$. In the white-box setting, since the adversary is aware of the exact parameters of the classifier produced by the defense, attacks are carried out with respect to the overall classifier. For the *optimal attack on Linear SVMs*, a similar change is made, where each \mathbf{w}_i (the output of the SVM optimization) is replaced by $\mathbf{B}^\top \mathbf{w}_i$, since that is the term which acts on the input \mathbf{x} . Thus, the adversarial sample now is

$$\tilde{\mathbf{x}}_t = \mathbf{x} - \frac{(\mathbf{z}_{t,y}^\top \mathbf{B}\mathbf{x}) \mathbf{B}^\top \mathbf{z}_{t,y}}{\|\mathbf{B}^\top \mathbf{z}_{t,y}\|_2^2}. \quad (12)$$

In the case of *gradient based attacks on neural networks*, the gradient of the loss is

$$\nabla_{\mathbf{x}} J(\mathbf{x}, y, \theta) = \nabla_{\mathbf{x}} J^{\text{aux}}(\mathbf{B}\mathbf{x}, y, \theta) = \mathbf{B}^\top \nabla_{\mathbf{x}} J^{\text{aux}}(\mathbf{x}, y, \theta), \quad (13)$$

where J is the loss function associated with f and J^{aux} is associated with f^{aux} . The loss of the neural network is computed with respect to its input, which is now $\mathbf{B}\mathbf{x}$, but the adversary has to add a perturbation to the input \mathbf{x} , which causes the largest increase in loss (up to first order). The FG attack on the defended network is then

$$\tilde{\mathbf{x}}(\epsilon) = \mathbf{x} + \epsilon \frac{\mathbf{B}^\top \nabla_{\mathbf{x}} J^{\text{aux}}(\mathbf{x}, y, \theta)}{\|\mathbf{B}^\top \nabla_{\mathbf{x}} J^{\text{aux}}(\mathbf{x}, y, \theta)\|_2}. \quad (14)$$

In the case of the *optimization based attack* on neural networks, the optimization objective remains the same, with a change in the classifier the loss function is computed over:

$$\begin{aligned} \min \quad & d(\tilde{\mathbf{x}}, \mathbf{x}) + \lambda \ell(\tilde{\mathbf{x}}, f), \\ \text{s.t.} \quad & \tilde{\mathbf{x}} \in C, \end{aligned} \quad (15)$$

where $f(\mathbf{x}) = f^{\text{aux}}(\mathbf{B}\mathbf{x})$. In our experiments, we first compute the linear transformation matrix, and then add it as a linear layer after the input layer of the neural network.

2) *Classifier mismatch attacks*: In this setting, the adversary trains a classifier \hat{f} that mimics the original classifier f , but is not aware of the defense. We assume the adversary is able to train \hat{f} such that it perfectly matches f trained on the original data without any linear transformations. The adversarial samples are thus generated with respect to $\hat{f} = \text{Train}(\mathbf{X})$, and not with respect to the true classifier $f = (\mathbf{x} \mapsto (\text{Train}(\mathbf{B}\mathbf{X}))(\mathbf{x}))$. Equivalently, this setting corresponds to the adversary using Algorithm 1 with the true training examples \mathbf{X} , the true training function Train , but with a different choice of **Select**. The adversary does not know the true **Select** function, so they use a trivial version that always returns **I**.

3) *Architecture mismatch attacks*: In this setting, the adversary trains a classifier \hat{f} using a choice of Train that does not match that used to produce f . Thus \hat{f} is not only a different function from f , but it may come from a different family of

classifiers. For example, f may be a three layer neural network and \hat{f} may be a five layer neural network. As in the classifier mismatch setting, the adversary is not aware of the defense used (the choice of **Select**).

The classifier and architecture mismatch attack settings are interesting to consider since the problem of the transferability of adversarial samples [34] is still an open research question. Our results in these settings demonstrate that a defense using linear transformations can mitigate the threat posed by transferability.

4) *Goal of the defenses*: The goal of our defense is to increase classifier robustness. Specifically, the defenses increase the distance between benign examples and nearest adversarial examples. Unlike some proposed defenses, we are not attempting to make the process of finding adversarial examples computationally difficult. While it is possible to use known attacks against classifiers produced by our defense, the resulting adversarial examples will be farther away from the benign examples than the adversarial examples for undefended classifiers. Thus, the fact that these attacks find adversarial examples is *not* a limitation of our approach.

IV. EXPERIMENTAL SETUP

In this section we provide brief descriptions and implementation details of the datasets, machine learning algorithms, dimensionality reduction algorithms, and metrics used in our experiments.

A. Datasets

In our evaluation, we use two datasets. The first is the MNIST image dataset and the second is the UCI Human Activity Recognition dataset. We now describe each of these in detail.

1) *MNIST*: This is a dataset of images of handwritten digits [27]. There are 60,000 training examples and 10,000 test examples. Each image belongs to a single class from 0 to 9. The images have a dimension of 28×28 pixels (total of 784) and are grayscale. The digits are size-normalized and centred. This dataset is used commonly as a ‘sanity-check’ or first-level benchmark for state-of-the-art classifiers. We use this dataset since it has been extensively studied from the attack perspective by previous work. It is also easy to visualize the effects of our defenses on this dataset.

2) *UCI Human Activity Recognition (HAR) using Smartphones*: This is a dataset of measurements obtained from a smartphone’s accelerometer and gyroscope [1] while the participants holding it performed one of six activities. Of the 30 participants, 21 were chosen to provide the training data, and the remaining 9 the test data. There are 7352 training samples and 2947 test samples. Each sample has 561 features, which are various signals obtained from the accelerometer and gyroscope. The six classes of activities are Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing and Laying. We used this dataset to demonstrate that our defenses work across multiple datasets and applications.

B. Machine learning algorithms

We have evaluated our defenses across multiple machine learning algorithms including linear Support Vector Machines (SVMs) and a variety of neural networks with different configurations. All experiments were run on a desktop running Ubuntu 14.04, with an 4-core Intel® Core™ i7-6700K CPU running at 4.00GHz, 24 GB of RAM and a NVIDIA® GeForce® GTX 960 GPU.

1) *Linear SVMs*: Ease of training and the interpretability of separating hyperplane weights has led to the use of Linear SVMs in a wide range of applications [40], [2]. We use the LinearSVC implementation from the Python package scikit-learn [37] for our experiments, which uses the ‘one-versus-rest’ method for multi-class classification by default.

In our experiments, we obtained a classification accuracy of 91.5% for the MNIST dataset and 96.7% for the HAR dataset.

2) *Neural networks*: Neural networks can be configured by changing the number of layers, the activation functions of the neurons, the number of neurons in each layer etc. We performed most of our experiments on a standard neural network used in previous work, for the purposes of comparison. The network we use is a standard one from [47] which we refer to as FC100-100-10 and a variant of it, FC200-200-200-10. The first neural network has an input layer, followed by 2 hidden layers, each containing 100 neurons, and an output softmax layer containing 10 neurons. Similarly, the second neural network has 3 hidden layers, each containing 200 neurons. Each neuron has a sigmoid activation function and the loss function used for training is the cross-entropy loss. We also ran experiments with a neural network with Rectified Linear Units (ReLU) as the neurons. We omit those results here due to lack of space and since they are very similar to the sigmoid activation results for the datasets we use. Both FC100-100-10 and FC200-200-200-10 are trained with a learning rate of 0.01 and momentum of 0.9 for 500 epochs. The size of each minibatch is 500. On the MNIST test data, we get a classification accuracy of 97.71% for FC100-100-10 and 98.02% for FC200-200-200-10. We use Theano [49], a Python library optimized for mathematical operations with multi-dimensional arrays and Lasagne [13], a deep learning library that uses Theano, for neural network experiments.

Our classification accuracy results for both Linear SVMs and fully connected neural networks are comparable to baseline numbers⁸ for corresponding architectures on MNIST, validating our implementation.

C. Linear transformation techniques

We use the PCA module from scikit-learn [37]. Depending on the application, either the number of components to be projected onto, or the percentage of variance to be retained can be specified. After performing PCA on the vectorized MNIST training data to retain 99% of the variance, the reduced dimension is 331, which is the first reduced dimension we use in our experiments on PCA based defenses. For *anti-whitening*,

we use a slight modification of the PCA interface to create the required transformation matrix.

D. Metrics

We evaluate the relationship between $\epsilon = \|\mathbf{x} - \tilde{\mathbf{x}}\|$, which is the allowed distance between the original example and the adversarial example, and the adversarial success rate, which we compute as follows. For each benign input \mathbf{x} with true label y , we check two conditions: if after perturbation, $f(\tilde{\mathbf{x}}) \neq f(\mathbf{x})$ and if initially, $f(\mathbf{x}) = y$. Thus, the adversary’s attempt is successful if the original classification was correct but the new classification on the adversarial sample is incorrect. In a sense, this count represents the number of samples that are truly adversarial, since it is only the adversarial perturbation that is causing misclassification, and not an inherent difficulty for the classifier in classifying this sample. While reporting adversarial success rates, we divide this count by the total number of benign samples correctly classified after they pass through the entire robust classification pipeline.

V. EXPERIMENTAL RESULTS

In this section we present an overview of our empirical results. The main questions we seek to answer with our evaluations are:

- (i) Is the defense effective in the classifier mismatch setting?
- (ii) Is the defense effective in the white box setting?
- (iii) Does the defense work for different classifier families?
- (iv) Does the defense generalize across different datasets?
- (v) Which linear transformations are most effective?

Our evaluation results confirm the effectiveness of our defense in a variety of scenarios, each of which has a different combination of datasets, machine learning algorithms, attacks and linear transformation used for the defense. For each set of evaluations, we vary a particular step of the classification pipeline and fix the others. Our results are summarized in Table II.

Baseline configuration: We start by considering a classification pipeline with the *MNIST dataset* as input data, a *Linear SVM* as our classification algorithm and *PCA* as the linear transformation used in our defense. Since we consider the Linear SVM as our classifier, we evaluate its susceptibility to adversarial samples generated using the *attack on Linear SVMs* described in Section II-A. We evaluate our defenses on adversarial samples created starting from the *test set* for each dataset. Unless otherwise noted, all security and utility results are for the *complete* test set. To empirically demonstrate that our defense is resilient not only in this baseline case, but also various configurations of it, we systematically investigate its effect as each component of the pipeline, as well as the attacks, are changed.

Note that in all of the plots showing the effectiveness of our defense, the legend key ‘None’ denotes adversarial success for a classifier without any defense.

A. Effect of defense on Support Vector Machines

In the baseline case, we begin by answering questions (i) and (ii) for Linear SVMs.

⁸<http://yann.lecun.com/exdb/mnist/>

Data set	Classifier	Attack type	Defense type and parameter	Robustness improvement	Accuracy reduction
MNIST	Linear SVM	Classifier mismatch	PCA (80)	25×	0.22%
MNIST	Linear SVM	White-box (optimal)	PCA (80)	6×	0.22%
MNIST	FC100-100-10	White-box (FG)	PCA (40)	1.5×	0.76%
MNIST	FC100-100-10	White-box (FGS)	PCA (40)	2.2×	0.76%
MNIST	FC100-100-10	White-box (Opt.)	PCA (40)	1.7×	0.76%
MNIST	FC200-200-200-10	Arch. mismatch (FC100-100-10)	PCA (40)	2.4×	0.85%
MNIST	FC100-100-10	White-box (FG)	Anti-whiten (2)	1.7×	0.15%
HAR	Linear SVM	White-box (optimal)	PCA (70)	1.5×	2.3%

TABLE II: **Robustness improvement at a misclassification rate of 60%**. The table also gives the accuracy reduction for different classifiers, attacks and defenses on the MNIST and HAR datasets.

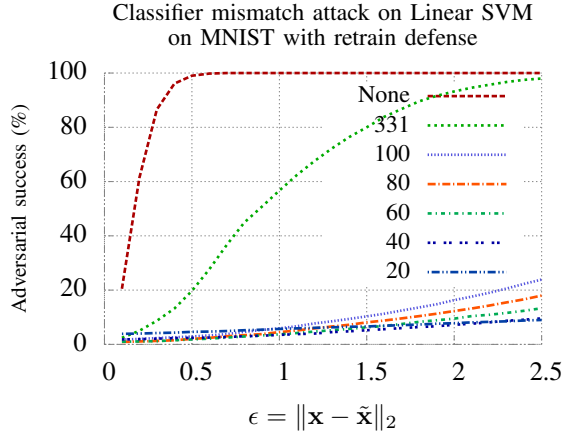


Fig. 4: **Effectiveness of the defense in classifier mismatch setting for the MNIST dataset with Linear SVMs.** The adversarial example success on the MNIST dataset is plotted versus the perturbation magnitude $\epsilon = \|\mathbf{x} - \tilde{\mathbf{x}}\|_2$. The attack is performed against the original classifier and the effect of the defense is plotted for each reduced dimension k .

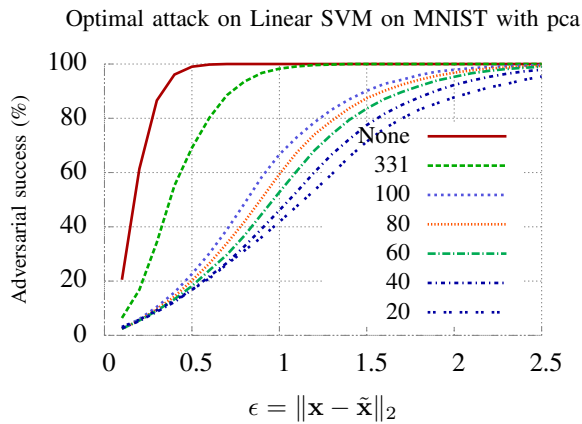


Fig. 5: **Effectiveness of the defense for the MNIST dataset against optimal white-box attacks on Linear SVMs.**

1) *Defense in the classifier mismatch setting:* Figure 4 shows the variation in adversarial success against SVMs in the classifier mismatch setting. The defense significantly reduces adversarial success rates. For example, at $\epsilon = 1.0$, the defense using PCA with a reduced dimension of $k = 80$ reduces

Linear SVM accuracy/robustness tradeoff for MNIST with pca

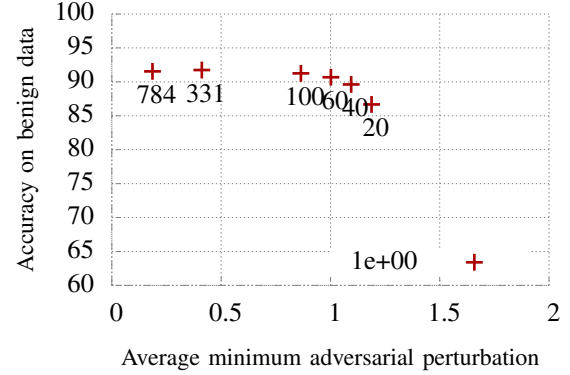


Fig. 6: **Tradeoff between SVM classification performance on benign test data, and adversarial performance.** The x-axis represents the average over test samples of the minimum perturbation needed to cause misclassification. The legend 1e+00 denotes the regularization constant used to train the SVM.

adversarial success from 100% to 3.4%⁹. This is a 96.6% or around 29.4× decrease in the adversarial success rate. At $\epsilon = 0.5$, where the adversarial success rate is 99%, the defense with $k = 80$ brings the adversarial success rate down to just 2%, which is a 49.5× decrease. Clearly, training with reduced dimension data leads to more robust Linear SVMs, and this can be seen in the effectiveness of the defense.

Further, as we decrease the reduced dimension k used in the projection step of the defense, adversarial success decreases, allowing the system designer to tune the defense according to her needs. At $k = 331$, the adversarial success rate is 56.7% at $\epsilon = 1.0$, which drops to 5.9% when $k = 100$. At $k = 30$, at the same ϵ , the adversarial success rate drops to 4.4% with a further decrease to 1.42% using aggressive dimensionality reduction with $k = 10$.

In the classifier mismatch setting, the defense also acts like a noise removal process, removing adversarial perturbations and leaving behind the clean input data. This accounts for the added robustness we see in this setting as compared to the

⁹We note here that all changes in adversarial success (and utility) are expressed in terms of the change in *percentage points*, i.e. a fall of $x\%$ indicates the absolute difference in the percentages, and not a relative difference.

white box setting. Further, this mitigates the problem of the transferability of adversarial examples when the attacker is only aware of the classifier used and not of the defense.

2) *Defense in the white box setting*: Figure 5 shows the variation in adversarial success for the defense against the optimal attack on Linear SVMs. This plot corresponds to the case where the adversary is aware of the dimensionality reduction defense and inputs a sample to the pipeline which is designed to optimally evade the reduced dimension classifier. At a perturbation magnitude of 0.5, where the classifier with no defenses has a misclassification rate of 99.04%, the reduced dimension classifier with $k = 80$ has a misclassification rate of just 19.75%, which represents a 80.25% or $5.01\times$ decrease in the adversarial success rates. At an adversarial budget of 1.3, the misclassification rate for the classifier with no defenses is 100%, while it is about 66.7% for the classifier with a reduced dimension of 40.

We can also study the effect of our defense on the adversarial budget required to achieve a certain adversarial success rate. A budget of 0.3 is required to achieve a 86.6% misclassification without the defense, while the required budget for a classifier with a defense with $k = 40$ is 1.75, which is a $5.83\times$ increase. The corresponding numbers to achieve a 98% misclassification rate are 0.5 without the defense and 2.5 with, which represents a $5\times$ increase. The presence of the defense forces the adversary to add much larger perturbations to achieve the same misclassification rates. Thus, our defense clearly reduces the effectiveness of an attack carried out by a very powerful adversary with full knowledge of the defense and the classifier as well as the ability to carry out optimal attacks.

3) *Utility-security tradeoff for defense*: Figure 6 shows the tradeoff between performance under ordinary and adversarial conditions. The kink in the tradeoff for this dataset is clearly between 80 and 60. There is very little benefit in classification performance by using more dimensions, and essentially no benefit in robustness by using fewer. At $k = 80$, we see a drop in classification success on the test set from 91.5% without any defenses, to 90.64% with the defense. Thus, there is a modest utility loss of about 1.2% at this value of k , as compared to a security gain of $5.9\times$, since the perturbation needed to cause 50 % of the test set to be misclassified increases from 0.16 to 0.95.

With these results, we can conclude that our defense is effective for Linear SVMs in both the classifier-mismatch and white-box settings. Now, we investigate the performance of our defenses on neural networks, to substantiate our claim of the applicability of our defenses across machine learning classifiers.

B. Effect of defense on neural networks

We now modify the baseline configuration by changing the classifier used to FC100-100-10. We evaluate our defenses on both gradient-based attacks for FC100-100-10: the Fast Gradient (FG) and Fast Gradient Sign (FGS) attacks as well as on the state-of-the-art optimization based attack from

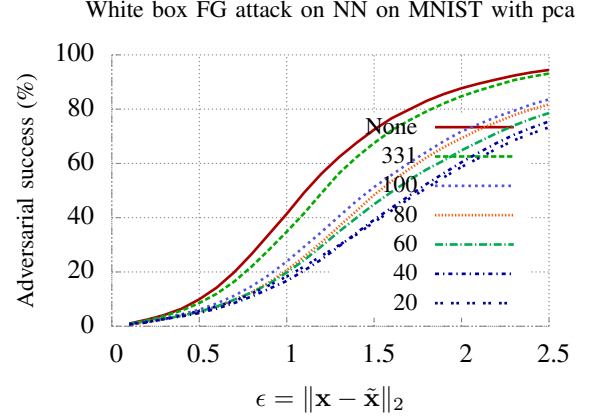


Fig. 7: Effectiveness of the defense for the MNIST dataset against FG attacks in the white box setting on FC100-100-10.

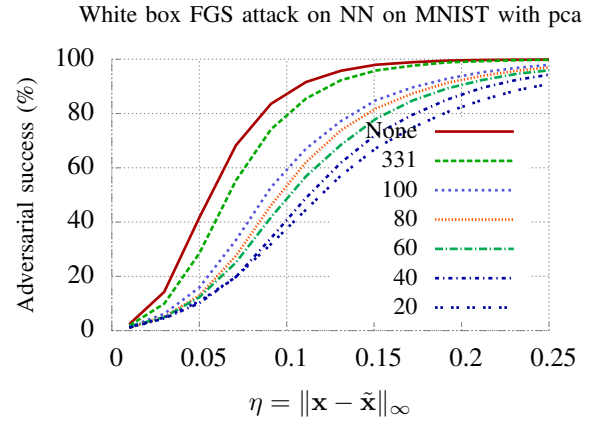


Fig. 8: Effectiveness of the defense for the MNIST dataset against FGS attacks in the white box setting on FC100-100-10.

Carlini et al. [7]. We continue to use the MNIST dataset and PCA as the linear transformation. With these experiments, we answer question iii), i.e. ‘Do the defenses work for different classifiers?’ and further strengthen our claim that our defense is effective in the white box setting.

1) *Defense against Fast Gradient attack in the white box setting*: Figure 7 shows the variation in adversarial success for the defense as $\epsilon = \|\mathbf{x} - \tilde{\mathbf{x}}\|_2$, the parameter governing the strength of the FG attack, increases. The defense also reduces adversarial success rates for this attack. At $\epsilon = 1.0$, the defense using PCA with a reduced dimension of $k = 40$ reduces adversarial success from 41.42% to 16.7%. This is a 24.72% or around $2.5\times$ decrease in the adversarial success rate. Again, at $\epsilon = 1.5$, while the adversarial success rate is 72.42% without any defense, the defense with $k = 40$ brings the adversarial success rate down to 39.19%, which is a 33.23% or $1.8\times$ decrease. Thus, even for neural networks, the defense causes significant reductions in adversarial success.

Again, we can study the effect of our defense on the adversarial budget required to achieve a certain misclassification

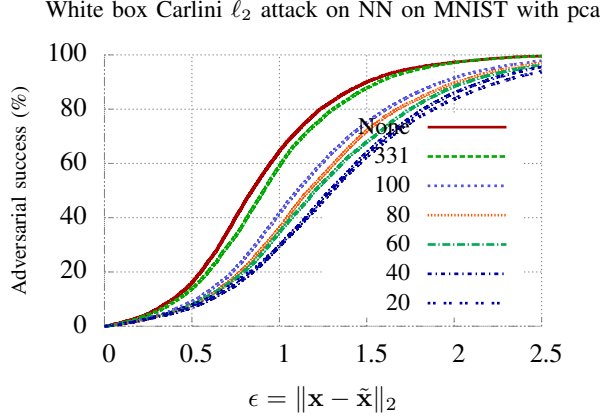


Fig. 9: Effectiveness of the defense for the MNIST dataset against Carlini’s ℓ_2 untargeted attack in the white-box setting on FC100-100-10.

percentage. A budget of roughly $\epsilon = 1.3$ is required to achieve a 60% misclassification without the defense, while the required budget for a classifier with the defense using $k = 40$ is 2, which is a $1.53\times$ increase.

Directly comparing neural networks and linear SVMs, it appears that neural network are more robust to ℓ_2 constrained attacks. However, it should be noted that while the Linear SVM robustness was evaluated on optimal attacks, the non-convex nature of the classification function in neural networks implies that the FG attack is only a first order approximation to an optimal attack.

2) *Defense against Fast Gradient Sign attack in the white box setting:* The FGS attack is constrained in terms of the ℓ_∞ norm, so all features with non-zero gradient are perturbed by either η or $-\eta$. The MNIST dataset has pixel values normalized to lie in $[0, 1]$. Thus if $\eta = 0.5$, the image with every pixel equal to 0.5 can be generated from any initial image. We restrict η to be less than 0.25.

In Figure 8, at $\eta = 0.05$, the adversarial success rate falls from 41.64% to 10.14% for the defense with $k = 40$ which is a 31.5% or $4.1\times$ reduction. Also, at $\eta = 0.11$, the adversarial success rate is 91.59% without the defense and 48.92% for the defense with $k = 40$, which is a 42.67% or $1.87\times$ reduction. Further, the perturbation needed to cause 90% misclassification is 0.11 without the defense but increases to 0.23 for the defense with $k = 40$, which is a $2.1\times$ increase.

3) *Defense against optimization based attack in the white-box setting:* We use Carlini and Wagner’s [7] ℓ_2 constrained attack to find untargeted adversarial samples, i.e. the closest possible \tilde{x} in terms of the ℓ_2 norm. Since this attack returns the minimal possible perturbation for each sample, in Figure 9 we plot the CDF of the minimal perturbations found by the attack over the test set in order to compare using the same metric as the other results. To see that this attack is indeed more powerful than the Fast Gradient attack (which uses the same distance metric), note that at $\|x - \tilde{x}\| = 1.0$, the adversarial success is around 65% compared to 41.42% for the FG attack,

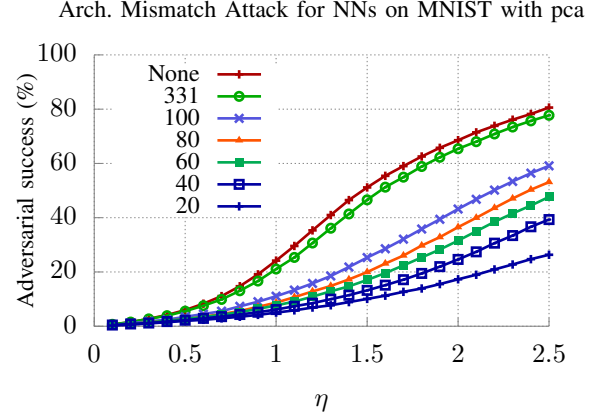


Fig. 10: Effectiveness of the defense for the MNIST dataset for FG samples generated for FC100-100-10 on FC200-200-200-10 (architecture mismatch setting) .

and at $\|x - \tilde{x}\| = 1.5$ it is 90% compared to 72.42% for the FG attack.

We repeated the attack on neural networks enhanced using our PCA-based defense. The attack was carried out on the composite classifier, thus representing the white box setting. In this case, we see that at $\|x - \tilde{x}\| = 1.0$ the adversarial success falls to 29.5% using the defense with $k = 40$, which represents a drop of 35.5% or $2.2\times$. At a larger allowed budget of 1.5, the fall is 26.4% or $1.4\times$ to 63.8%. Further, the budget required to achieve a misclassification rate of 90% increases from 1.5 to 2.16, which is a $1.44\times$ increase.

4) *Defense against Fast Gradient attack in the architecture mismatch setting:* We now consider a setting where the adversary is less powerful. In the *architecture mismatch* setting, the adversary creates adversarial sample for a different neural network (FC100-100-10) than the one being attacked (FC200-200-200-10). These results are presented in Figure 10. We see a significant drop in adversarial success when our defense is used. For example, at $\|x - \tilde{x}\| = 1.5$ the adversarial success falls from 51.2% to 13.2% using the defense with $k = 40$, which is a 38.0% or $3.9\times$ drop. Also, the budget required to achieve a misclassification rate of 40% increases from 1.3 to 2.5, which is a $2\times$ increase. The performance of our defense in this setting, which is commonly referred to as a ‘black-box’ setting highlights that the defense can mitigate the transferability of adversarial samples to a large extent.

With these results for neural networks, we conclude that our defenses are effective against a variety of different attacks, in each of which the nature of the adversarial perturbation is very different. We have also shown that *our defense is effective against the state of the art attack for neural networks in the white-box setting*, making a case for it to be included as a crucial component of any defensive measures against evasion attacks. These results also demonstrate that our defense can be used in conjunction with different types of classifiers, providing a general method for defending against adversarial inputs.

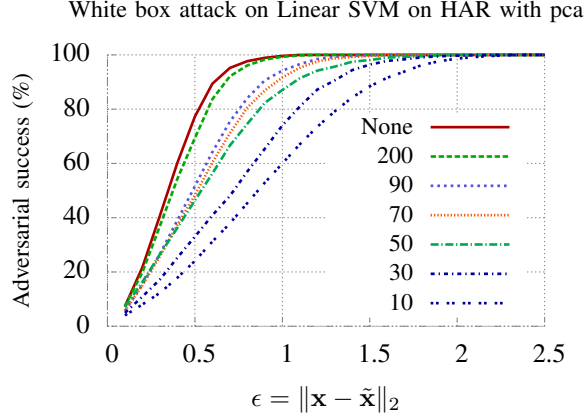


Fig. 11: **Effectiveness of the defense for the HAR dataset against optimal white-box attacks on Linear SVMs.** Adversarial example success on the HAR dataset versus perturbation magnitude ϵ for the Linear SVM attack. Plotted for each reduced dimension k used in the defense.

C. Applicability for different datasets

Next, we modify the baseline configuration by changing the datasets used. We show results with Linear SVMs as the classifier and PCA as the dimensionality reduction algorithm. We present results for the Human Activity Recognition dataset.

1) *Defense for the HAR dataset:* In Figure 11, the reduction in adversarial success of a white-box attack due to the defense using PCA on the HAR dataset is shown. At $\epsilon = 0.5$, the adversarial success rate drops from 77.3% without the defense to 48.3% with $k = 70$ which is a $1.6\times$ drop respectively. In order to achieve a misclassification rate of 90%, the amount of perturbation needed is 0.65 without the defense, which increases to 0.93 with $k = 70$. Thus, the adversarial budget increases $2\times$ to achieve the same adversarial success rate. The impact on utility is modest: a drop of 2.3% for $k = 70$, which is small in comparison to the gain in security.

D. Effect of PCA-based defense on utility

Table III presents the effect of our defense on the classification accuracy of benign data. The key takeaways are that the decrease in accuracy for both neural networks and Linear SVMs for reduced dimensions down to $k = 50$ is at most 4%. Further, we notice that dimensionality reduction using PCA can actually *increase* classification accuracy as when $k = 70$, the accuracy on the MNIST dataset increases from 97.47% to 97.52%. This effect probably occurs as the dimensionality reduction can prevent over-fitting. More aggressive dimensionality reduction however, can lead to steep drops in classification accuracy, which is to be expected since much of the information used for classification is lost.

E. Defense using anti-whitening

As described in Section III-E, anti-whitening is a soft approximation of PCA where high-variance components are

MNIST data			HAR data	
k	FC100-100-10	Linear SVM	k	Linear SVM
No D.R.	97.47	91.52	No D.R.	96.67
784	97.32	91.54	561	96.57
331	97.35	91.37	200	96.61
100	97.36	90.89	90	94.60
80	97.25	90.64	70	94.37
60	97.38	90.47	50	92.47
40	96.71	89.03	30	91.11

TABLE III: **Utility values for the dimensionality reduction defense.** For the MNIST and HAR datasets, the classification accuracy on the benign test set is provided for various values of reduced dimension k used for the PCA based defense, as well as the accuracy without the defense.

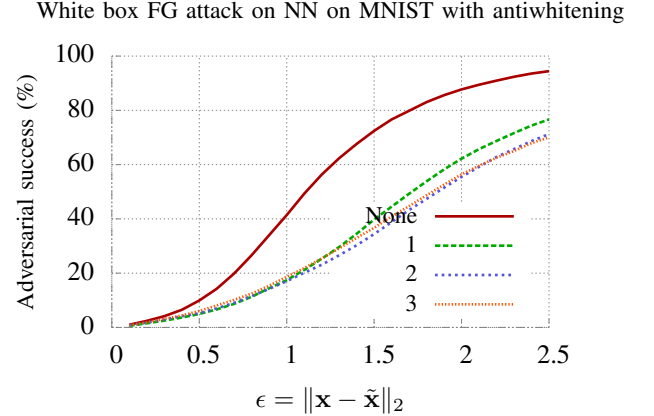


Fig. 12: **Effectiveness of the anti-whitening defense for the MNIST dataset against FG attacks in the white box setting on FC100-100-10.**

boosted with respect to the low-variance ones, instead of just dropping them. This can be controlled by the parameter c in the anti-whitening transformation $\mathbf{B} = \Lambda^{\frac{c}{2}} \mathbf{U}^T$. In Figure 12, the effects of the defense using anti-whitening with $c = 1, 2$ and 3 are shown. At $\epsilon = 1.0$, the defense with $c = 2$ causes the adversarial success to fall from 41.42% to 17.06%, which is a 24.36% or $2.4\times$ fall. At $\epsilon = 1.5$, the corresponding reduction is from 72.42% to 34.42%, which is a 38% or $2.1\times$ decrease. The anti-whitening defense thus performs slightly better than the PCA defense with a comparable parameter ($k = 40$).

Effect of anti-whitening on utility: For FC100-100-10, the classification rate on benign data is 97.47% without any defense. Using anti-whitening with $c = 1, 2$ and 3, the utility values are 97.45%, 97.32% and 96.83% respectively. This shows that the anti-whitening defense is slightly better both with respect to both security and utility as compared to the PCA defense. The increase in utility is likely due to the fact that dimensions are not dropped and are used to achieve better classification performance.

These results highlight the broad applicability of our defense across application domains. It is clear that the effectiveness of our defense is not an artifact of the particular structure of data from the MNIST dataset, and that the intuition for its effect

holds across different kinds of data.

VI. DISCUSSIONS, LIMITATIONS AND FUTURE WORK

Even though our defense reduces adversarial success rates and increases the amount of perturbation the adversary has to add to achieve fixed levels of misclassification in a number of cases, there are two main areas where it can be improved in conjunction with other defense mechanisms.

1) *Further reductions in adversarial success:* While our defense causes significant reductions in adversarial success rates in a variety of settings, there are cases where the adversarial success rate is still non-trivial. In such cases, it is likely our defense would have to be combined with other defenses such as adversarial training [16] and ensemble methods [45] for detection in order to create a ML system secure against evasion attacks. Our defense has the advantage that it can be used in conjunction with a variety of ML classifiers and it will not interfere with the operation of other defense mechanisms. Since our defense increases the amount of perturbation needed to achieve a fixed misclassification rate, it may aid detection based defenses.

2) *Better data transformations:* In certain settings, using PCA for dimensionality reduction may have limited applicability. For example, we found that our PCA based defense offers only marginal security improvement for the Papernot-CNN (See Section IX-B for details). It is likely that this effect stems from PCA reducing the amount of local information that the convolutional layers in the CNN are able to use for the purposes of classification. A key step in addressing this limitation of our defense is to use other dimensionality reduction techniques which could reduce adversarial success to negligible levels and work better when combined with classifiers such as CNNs. This limitation also prevents us from achieving state-of-the-art accuracy on image datasets like MNIST, since the best classifiers in for these datasets use convolutional layers. In future work we plan to explore techniques such as autoencoders and kernel PCA for designing robust classifiers. For certain problems, it may also be feasible to explicitly optimize for the linear transformation achieving the best utility-security tradeoff. This is another direction we plan to explore.

VII. RELATED WORK

Previous defenses against adversarial examples have largely focused on specific classifier families or application domains. Further, the existing defenses provide improved security only against existing attacks in the literature, and it is unclear if the defense mechanisms will be effective against adversaries with knowledge of their existence, i.e. strategic attacks exploiting weaknesses in the defenses. As a case in point, Papernot et al. [36] demonstrated a defense using distillation of neural networks against the Jacobian-based saliency map attack [35]. However, Carlini et al. [7] showed that a modified attack negated the effects of distillation and made the neural network vulnerable again. Now, we give an overview of the existing defenses.

1) *Classifier-specific:* Russu et al. [39] propose defenses for SVMs by adding various kinds of regularization. Kantchelian et al. [22] propose defenses against optimal attacks designed specifically for tree-based classifiers. Existing defenses for neural networks [17], [42], [55], [28], [21] make a variety of structural modifications to improve resilience to adversarial examples. These defenses do not readily generalize across classifiers and may still be vulnerable to adversarial examples, as shown by Gu and Rigazio [17].

2) *Application-specific:* Hendrycks and Gimpel [18] study transforming images from the RGB space to YUV space to enable better detection by humans and decrease misclassification rates. They also use whitening to make adversarial perturbations in RGB images more visible to the human eye. The effect of JPG compression on adversarial images has also been studied [14], [12]. Their conclusions were that it has a small beneficial effect when the perturbations are small. These approaches are restricted to combating evasion attacks on image data and do not generalize across applications. Further, it is unclear if they are effective against white-box attacks.

3) *General defenses:* An ensemble of classifiers was used by Smutz and Stavrou [45] to detect evasion attacks, by checking for disagreement between various classifiers. However, an ensemble of classifiers may still be vulnerable to adversarial examples since they generalize across classifiers. Further, Goodfellow et. al. [16] show that ensemble methods have limited effectiveness for evasion attacks against neural networks. Goodfellow et. al. [16], Tramèr et al. [50] and Mądry et al. [31] re-train on adversarial samples of different types to improve the resilience of neural networks. They all find that adversarial training works, but needs high capacity classifiers to be effective, and further, its effectiveness reduces as the perturbation is increased beyond the one used for training. In our experiments, we find that re-training on adversarial samples has an extremely limited effect on increasing the robustness of linear SVMs (see Figure 14 in the Appendix), thus this defense may not be applicable across classifiers, and does indeed depend on the capacity of the classifier. Wang et al. [52] use random feature nullification to reduce adversarial success rates for evasion attacks on neural networks. The applicability of this idea across classifiers is not studied. Zhang et al. [54] use adversarial feature selection to increase the robustness of SVMs. They find and retain features that decrease adversarial success rates. This defense may be generalized across other classifiers and is an interesting direction for future work.

Due to the classifier and dataset-agnostic nature of our defense, it may be combined with existing defenses such as adversarial training which have different aims for an even larger improvement in robustness. For example, neural networks may be trained with reduced dimension samples, and the training process can also incorporate the adversarial loss to further increase the robustness of the network. We plan to explore these directions in future work.

VIII. CONCLUSION

In this paper, we considered the novel use of data transformations such as dimensionality reduction as a defense mechanism against evasion attacks on ML classifiers. Our defenses rely on the insight that (a) linear transformations of data allow access to usually inaccessible security-performance tradeoffs, and (b) training classifiers on reduced dimension data leads to enhanced resilience of ML classifiers (by reducing the weights of less informative and low-variance features). Using empirical evaluation on multiple real-world datasets, we demonstrated a 2x reduction in adversarial success rates across a range of attack strategies (including white-box ones), ML classifiers, and applications. Our defenses have a modest impact on the utility of the classifiers (0.5-2% reduction), and are computationally efficient. Our work thus provides an attractive foundation for countering the threat of evasion attacks.

REFERENCES

- [1] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in *ESANN*, 2013.
- [2] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, and K. Rieck, "Drebin: Effective and explainable detection of android malware in your pocket." in *NDSS*, 2014.
- [3] 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.
- [4] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 387–402.
- [5] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 1807–1814.
- [6] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium (USENIX Security 16)*, Austin, TX, 2016.
- [7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017.
- [8] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333–338, 2012.
- [9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.
- [10] G. V. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 4, pp. 335–455, 2007.
- [11] G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu, "Large-scale malware classification using random projections and neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3422–3426.
- [12] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression," *arXiv preprint arXiv:1705.02900*, 2017.
- [13] S. Dieleman and J. S. et al., "Lasagne: First release." Aug. 2015. [Online]. Available: <http://dx.doi.org/10.5281/zenodo.27878>
- [14] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of jpg compression on adversarial images," *arXiv preprint arXiv:1608.00853*, 2016.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.
- [16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.
- [17] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.
- [18] D. Hendrycks and K. Gimpel, "Visible progress on adversarial images and a new saliency map," *arXiv preprint arXiv:1608.00530*, 2016.
- [19] H. Hosseini, B. Xiao, and R. Poovendran, "Deceiving google's cloud video intelligence api built for summarizing videos," *arXiv preprint arXiv:1703.09793*, 2017.
- [20] 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.
- [21] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, "Learning with a strong adversary," *CoRR*, abs/1511.03034, 2015.
- [22] A. Kantchelian, J. Tygar, and A. D. Joseph, "Evasion and hardening of tree ensemble classifiers," in *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, 2016.
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [24] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.
- [25] P. Laskov and M. Kloft, "A framework for quantitative security analysis of machine learning," in *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*. ACM, 2009, pp. 1–4.
- [26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [27] Y. LeCun and C. Cortes, "The mnist database of handwritten digits," 1998.
- [28] Y. Luo, X. Boix, G. Roig, T. Poggio, and Q. Zhao, "Foveation-based mechanisms alleviate adversarial examples," *arXiv preprint arXiv:1511.06292*, 2015.
- [29] M. McCoy and D. Wagner, "Spoofing 2d face detection: Machines see people who aren't there," *arXiv preprint arXiv:1608.02128*, 2016.
- [30] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," *arXiv preprint arXiv:1511.04599*, 2015.
- [31] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
- [32] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 427–436.
- [33] NVIDIA, "Self driving vehicles development platform," <http://www.nvidia.com/object/drive-px.html>, accessed: 2016-10-31.
- [34] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.
- [35] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.
- [36] N. Papernot, P. D. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *IEEE Symposium on Security and Privacy, SP 2016*, 2016, pp. 582–597.
- [37] 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.
- [38] R. Penrose, "On best approximate solutions of linear matrix equations," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 52. Cambridge University Press, 1956, pp. 17–19.
- [39] P. Russu, A. Demontis, B. Biggio, G. Fumera, and F. Roli, "Secure kernel machines against evasion attacks," in *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, ser. AISec '16. New York, NY, USA: ACM, 2016, pp. 59–69. [Online]. Available: <http://doi.acm.org/10.1145/2996758.2996771>
- [40] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
- [41] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Jan. 2002.



Fig. 13: **Adversarial images of digit ‘7’ (against a neural network with no defense)**: The images have been modified with the Fast Gradient Sign attack on neural networks with (from left to right), $\eta \approx 0.05, 0.1, 0.15, 0.2$ and 0.25 . The perturbation begins to be visible at $\eta = 0.15$ and is very obvious in the images with $\eta > 0.2$. The attack was carried out on a classifier f without any dimensionality reduction.

- [42] U. Shaham, Y. Yamada, and S. Negahban, “Understanding adversarial training: Increasing local stability of neural nets through robust optimization,” *arXiv preprint arXiv:1511.05432*, 2015.
- [43] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1528–1540.
- [44] J. Shlens, “A tutorial on principal component analysis,” *arXiv preprint arXiv:1404.1100*, 2014.
- [45] C. Smutz and A. Stavrou, “When a tree falls: Using diversity in ensemble classifiers to identify evasion in malware detectors,” in *23rd Annual Network and Distributed System Security Symposium, NDSS 2016*.
- [46] N. Šrđić and P. Laskov, “Hidost: a static machine-learning-based detector of malicious files,” *EURASIP Journal on Information Security*, vol. 2016, no. 1, p. 22, 2016.
- [47] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, 2014.
- [48] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.
- [49] T. D. Team, “Theano: A Python framework for fast computation of mathematical expressions,” *arXiv e-print arXiv:1605.02688*, 2016.
- [50] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” *arXiv preprint arXiv:1705.07204*, 2017.
- [51] L. Van Der Maaten, E. Postma, and J. Van den Herik, “Dimensionality reduction: a comparative review,” *J Mach Learn Res*, vol. 10, pp. 66–71, 2009.
- [52] Q. Wang, W. Guo, K. Zhang, X. Xing, C. L. Giles, and X. Liu, “Random feature nullification for adversary resistant deep architecture,” *arXiv preprint arXiv:1610.01239*, 2016.
- [53] W. Xu, Y. Qi, and D. Evans, “Automatically evading classifiers,” in *Proceedings of the 2016 Network and Distributed Systems Symposium*, 2016.
- [54] F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung, and F. Roli, “Adversarial feature selection against evasion attacks,” *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 766–777, 2016.
- [55] Q. Zhao and L. D. Griffin, “Suppressing the unusual: towards robust cnns using symmetric activation functions,” *arXiv preprint arXiv:1603.05145*, 2016.

IX. APPENDIX

A. Complexity Analysis of PCA Defenses

The defense using PCA adds a one-time $\mathcal{O}(d^2n + d^3)$ overhead for finding the principal components, with the first term arising from the covariance matrix computation and the second term from the eigenvector decomposition. There is also a one-time overhead for training a new classifier on the reduced dimension data. The time needed to train the new classifier will be less than that needed for the original classifier since the dimensionality of the input data has reduced. Each

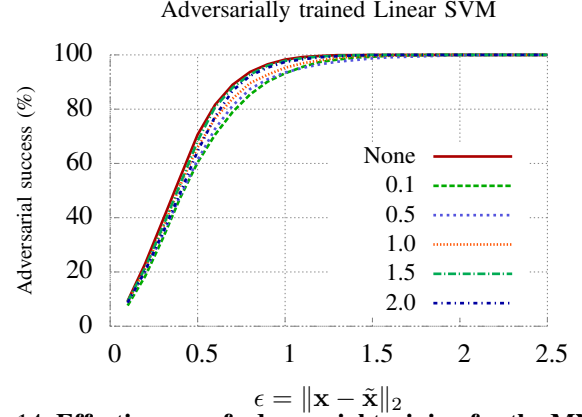


Fig. 14: **Effectiveness of adversarial training for the MNIST dataset against optimal white box attacks on Linear SVMs**. The Linear SVM was trained using gradient descent with periodically augmented training sets containing adversarial samples with the specified perturbation values.

subsequent input will incur a $\mathcal{O}(dk)$ overhead due to the matrix multiplication needed to project it onto the principal components.

B. CNNs

We also run our experiments on a Convolutional Neural Network [15] whose architecture we obtain from Papernot et al. [36]. This CNN’s architecture is as follows: it has 2 convolutional layers of 32 filters each, followed by a max pooling layer, then another 2 convolutional layers of 64 filters each, followed by a max pooling layer. Finally, we have two fully connected layers with 200 neurons each, followed by a softmax output with 10 neurons (for the 10 classes in MNIST). All neurons in the hidden layers are ReLUs. We call this network **Papernot-CNN**. It is trained with a learning rate of 0.1 (adjusted to 0.01 for the last 10 epochs) and momentum of 0.9 for 50 epochs. The batchsize is 500 samples for MNIST and on the MNIST test data we get a classification accuracy of 98.91% with the **Papernot-CNN** network.