# HybridMTD: Enhancing Robustness Against Adversarial Attacks with Ensemble Neural Networks and Moving Target Defense

Kimia Tahayori[1], Sherif Saad[1], Mohammad Mamun[2] and Saeed Samet[1]
[1]*School of Computer Science, University of Windsor, Windsor, Canada*
[2]*National Research Council of Canada, Canada*
{*tahayor, shsaad, saeed.samet*}*@uwindsor.ca, mohammad.mamun@nrc-cnrc.gc.ca*

Keywords: Adversarial Attacks, Moving Target Defense, Ensemble Models.

Abstract: Adversarial attacks compromise the integrity of machine learning models, posing significant risks in critical fields like autonomous driving, healthcare, and finance, where accuracy and security are paramount. Existing defenses against these attacks primarily involve adversarial training or architectural modifications to the models. However, many of these approaches are model-specific, limiting their applicability to other models and potentially degrading overall performance, including accuracy and generalization. Thus, there is a pressing need to explore model-agnostic defense strategies that do not rely on adversarial training, offering more adaptable and reliable solutions across various models. This study aims to evaluate the effectiveness of HybridMTD. This novel defense strategy integrates Moving Target Defense (MTD) with ensemble neural network models to enhance robustness against adversarial attacks without requiring adversarial training or internal changes to model architectures. By dynamically selecting a subset of models from a diverse pool for each evaluation and utilizing majority voting, HybridMTD increases unpredictability and strengthens the resilience of the defense mechanism. We conducted extensive experiments across four datasets—MNIST (image), Twitter Sentiment (text), KDD (tabular), and MIT-BIH (signals)—and assessed HybridMTD against seven advanced attacks, including evasion and poisoning attacks. The results consistently show that HybridMTD outperforms traditional MTD strategies and single-model methods, maintaining high accuracy and robustness across diverse attack types and datasets. This research underscores the potential of HybridMTD as an effective defense strategy, significantly improving model security and laying the foundation for further exploration of advanced defense mechanisms.

## 1 INTRODUCTION

Adversarial attacks pose significant threats to the security and reliability of machine learning (ML) models, particularly in critical applications such as autonomous driving (Stilgoe, 2018), healthcare (An et al., 2023), and cybersecurity (Zhou et al., 2022). These attacks manipulate input data to deceive models into making incorrect predictions (Ren et al., 2020), jeopardizing system safety and causing potential financial losses (Wu et al., 2023; Liu et al., 2018). Ensuring model robustness against such perturbations is crucial for safe deployment in real-world scenarios.

Adversarial attacks are broadly classified into evasion attacks, which occur during the test phase (Liu et al., 2018), and poisoning attacks, which compromise the training phase by introducing malicious data (Biggio et al., 2013; Biggio et al., 2011). Addressing these challenges requires effective and adaptable defense strategies.

Traditional defenses, such as adversarial training (Dong et al., 2020; Liu et al., 2022) and model modifications (Madry et al., 2018; Gao et al., 2019), improve robustness but often lack generalizability across attack scenarios. In contrast, Moving Target Defense (MTD) offers a dynamic strategy by continuously altering system configurations to disrupt attackers' ability to exploit model weaknesses (Lei et al., 2018). MTD can be implemented via shuffling, diversity, redundancy, or hybrid approaches (Sun et al., 2023). This study adopts a hybrid MTD strategy, combining shuffling and diversity by randomly selecting models for each input evaluation, enhancing system resilience through increased unpredictability.

Ensemble models complement MTD by aggregating predictions from multiple models, reducing variance and bias while improving robustness (Dietterich, 2000). Even if some models are vulnerable



Tahayori, K., Saad, S., Mamun, M. and Samet, S.
HybridMTD: Enhancing Robustness Against Adversarial Attacks with Ensemble Neural Networks and Moving Target Defense.
DOI: 10.5220/0013240700003899

In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025) - Volume 2*, pages 72-83
ISBN: 978-989-758-735-1; ISSN: 2184-4356

to specific attacks, others can compensate, mitigating adversarial impacts. Our approach employs majority voting to determine final predictions, leveraging the strengths of both MTD and ensemble methods.

This study introduces a novel hybrid defense framework integrating MTD and ensemble models. Unlike prior works (Sengupta et al., 2019; Roy et al., 2019; Amich and Eshete, 2021), which apply MTD with a single model, our approach randomly selects subsets of models for each input, combining dynamic adaptation with ensemble robustness. Extensive experiments were conducted on four datasets—MNIST (LeCun et al., 1998a), Twitter Sentiment, KDD, and MIT-BIH—and seven attack types, including FGSM (Goodfellow et al., 2015), BIM (Kurakin et al., 2017), and poisoning attacks like label flipping (Biggio et al., 2013).

This study addresses the question: 'Can integrating MTD and ensemble models enhance ML robustness against adversarial attacks?' Our objectives are to evaluate the framework's effectiveness across datasets and attack types, and to identify conditions where it performs best.

The remainder of the paper is organized as follows: Section 2 reviews related work, Section 3 details the methodology, Section 4 presents results, Section 5 discusses findings, and Section 6 concludes with future research directions.

## 2 RELATED WORK

To provide context for our study, we review several strategies that have been proposed to enhance the resilience of machine learning models against adversarial attacks. Based on the survey paper by Q. Liu et al. (Liu et al., 2018), several strategies are proposed to enhance the resilience of machine learning models against adversarial attacks. In the training phase, data sanitization methods are utilized to remove adversarial samples from training data, ensuring dataset purity. Robust algorithms enhance the robustness of learning algorithms. Additionally, secure algorithms are developed to distribute feature weights more evenly, further fortifying the models against attacks.

During the testing and inferring phases, various techniques are employed to improve model resilience and security. Robustness improvements use game theory and adversarial retraining to boost model performance when faced with adversarial attacks. Techniques such as defensive distillation smooth model outputs, making them less susceptible to adversarial samples. Dimension reduction strategies reduce feature dimensions to enhance resilience. Statistical tests are used to detect adversarial samples. Ensemble methods provide a robust defense by improving overall security. Our approach will be categorized under ensemble methods, leveraging the combined strength of multiple models to enhance security and resilience against adversarial attacks.

The first notable work is by Sengupta et al. (Sengupta et al., 2019), who propose MTDeep, a defense mechanism that integrates MTD with deep neural networks (DNNs) to enhance robustness against adversarial attacks. Their approach involves randomly selecting from an ensemble of just three DNNs to classify each input image, increasing unpredictability for attackers. Evaluated exclusively on image datasets such as MNIST, Fashion-MNIST, and ImageNet, MTDeep focuses on evasion attacks and dynamically alters the attack surface at test time. While this method improved model performance, the results were not highly significant, and their pool of models was quite limited. In contrast, our work explores a broader range of models and data types beyond just images, aiming for more flexibility and stronger defense across diverse attack scenarios.

A similar work by Roy et al. (Roy et al., 2019) presents an MTD approach modeled as a Stackelberg game, where the defender selects an algorithm from a limited set, increasing unpredictability. Like Sengupta et al., their method focuses on image datasets (e.g., MNIST) and is tested against rational and boundedly rational attackers. Although it maintains reasonable accuracy under severe conditions, it faces the same limitations, such as an exclusive focus on evasion attacks. In contrast, our work expands to a broader range of data types, addressing both evasion and poisoning attacks.

A. Amich and B. Eshete's work (Amich and Eshete, 2021) introduces Morphence, which uses a dynamic pool of slightly perturbed CNN models to defend against adversarial attacks on image classification datasets. Morphence counters white-box and black-box attacks by selecting the most confident model for each prediction, with the model pool expiring after a set query budget. While this expands the pool compared to MTDeep, it still relies on a single model type and focuses solely on image data and evasion attacks. In contrast, our approach selects from a diverse set of models to make decisions for each input and covers both evasion and poisoning attacks across multiple data types.

R. Colbaugh and K. Glass (Colbaugh and Glass, 2013) propose an MTD strategy that dynamically switches between classifiers trained on different feature subsets, guided by a Markov Chain model. Their focus is primarily on denial-of-service (DoS) attacks

rather than adversarial attacks, and their approach shows significant improvements over static defenses. In contrast, our work targets adversarial attacks, addressing both evasion and poisoning.

Peter Martin et al. (Martin et al., 2021) explore using MTD strategies to protect deep learning models from adversarial attacks by training diverse models, applying random affine transformations to inputs, and randomizing outputs. Their approach, tested on image datasets and focused on evasion attacks, showed improved robustness against white-box attacks when combined with Stochastic Affine Transformations (SAT) and Adaptive Diversity-Promoting (ADP) regularization. However, sophisticated adversaries using surrogate models can still bypass these defenses, and the success against black-box attacks depends on low transferability between sub-models.

# 3 METHODOLOGY

## 3.1 Dataset Description

We utilized four different datasets, each representing a unique data type and application area, to evaluate the effectiveness of our proposed defense strategy. The MNIST (Modified National Institute of Standards and Technology) dataset (LeCun et al., 1998b) is a widely used benchmark in machine learning. It consists of 60,000 training samples and 10,000 test samples of grayscale images of handwritten digits (0-9), each 28x28 pixels.

The KDD Cup 1999 Data (Liu, 1999) originates from the 1998 DARPA Intrusion Detection Evaluation Program and is extensively used for evaluating anomaly and intrusion detection algorithms. It comprises nearly 5 million connection records, 23 classes, and 41 features, providing a rich dataset for research.

The MIT-BIH Arrhythmia Database (Moody and Mark, 2001), hosted on PhysioNet, is widely used in biomedical research to study cardiac arrhythmias. It includes 48 half-hour extracts of two-channel ambulatory ECG recordings from 47 subjects, with expert annotations providing ground-truth labels. This dataset is crucial for evaluating adversarial perturbations in biomedical signals and includes 23 classes of arrhythmias.

The Twitter Sentiment Analysis Dataset (Shahane, 2021) categorizes Twitter sentiments into positive, negative, and neutral. Twitter data is inherently challenging due to its brevity, slang, and non-standard grammatical structures.

All datasets were standardized to ensure compatibility with various models. MNIST images were normalized and reshaped, the KDD dataset was encoded to handle multiple features and class imbalances, MIT-BIH ECG signals were filtered and segmented, and Twitter sentiment data was cleaned and transformed into embeddings.

These datasets were chosen to evaluate the generalizability of our defense strategy across diverse domains, each with unique security implications. MNIST serves as a foundational benchmark for adversarial defenses in vision systems. KDD, despite its age, remains a standard for intrusion detection, enabling comparisons with prior methods. MIT-BIH is critical for testing robustness in biomedical systems, where adversarial attacks could compromise patient care. The Twitter Sentiment Analysis dataset represents the challenges of securing text-based systems, such as misinformation detection, against adversarial manipulation.

## 3.2 Model Selection

We employed 10 neural network models for each dataset, selecting and modifying them according to the specific data type. These models included CNNs, MLPs, LeNet5, LSTMs, MobileNetV2, RNNs, and GRUs. We created various versions for each model—such as deeper, wider, and fully connected configurations—to ensure diversity and robustness in our model pool.

CNNs (Convolutional Neural Networks) are generally used for tasks that involve capturing spatial hierarchies. We utilized different configurations of CNNs, including deeper and fully connected versions, as part of our exploration. LeNet5, a classic CNN model, and MobileNetV2, known for its computational efficiency, were also included as part of this model pool.

For tasks involving sequential data, models like LSTMs (Long Short-Term Memory Networks), GRUs (Gated Recurrent Units), and RNNs (Recurrent Neural Networks) are generally used. We incorporated different variations, such as deeper configurations, to enhance their ability to model long-term dependencies. MLPs (Multilayer Perceptrons) were also widely used due to their flexibility in learning complex input-output mappings, with both deeper and wider versions included for increased performance.

## 3.3 Adversarial Attack Methods

We incorporated several adversarial attack methods into our evaluation:

Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) and Basic Iterative Method (BIM)

(Kurakin et al., 2017) were both implemented using the Adversarial Robustness Toolbox (ART) (Nicolae et al., 2019). FGSM generates adversarial examples by applying noise in a one-step process, manipulating test set data based on the gradient of the loss with respect to the input data. BIM, an iterative extension of FGSM, applies perturbations multiple times, allowing for a more granular exploration of adversarial vulnerabilities. Each iteration applies small perturbations to the input, cumulatively leading to a significant adversarial effect.

Jacobian-based Saliency Map Attack (JSMA) (Papernot et al., 2015) focuses on changing the most important features of the input data to mislead the model. We used the ART to configure the SaliencyMapMethod with predefined parameters: 'theta' controls how much each feature is changed, and 'gamma' determines how many features are altered. Carlini & Wagner (C&W) Attack (Carlini and Wagner, 2017) solves an optimization problem to find the smallest changes needed to mislead the model while ensuring high confidence in misclassification. The C&W attack iteratively adjusts input images to find minimal perturbations, testing the model's sensitivity to small changes and the robustness of its defenses.

The Transferability Attack (Papernot and McDaniel, 2016) involves creating a surrogate model to generate adversarial examples, which are then used to attack the main model. The assumption is that the attacker does not have direct access to the target model's parameters or training data. We trained a separate surrogate model to mimic the main model's task, exploiting the phenomenon that adversarial vulnerabilities often transfer across models. The generated adversarial examples were introduced to the main model to assess its robustness in a realistic black-box scenario.

The Label-Flipping Attack (Biggio et al., 2013) is a straightforward poisoning attack in which 50% of the training data labels are flipped to degrade the model's performance. The Feature Collision Attack (Shafahi et al., 2018) introduces adversarial samples designed to overlap in feature space with specific target instances, causing the model to misclassify these targets during training. We generated poisoned samples by adding calculated perturbations to base class samples, pushing them toward the feature space of selected target instances. These adversarial samples were mixed with the original training data.

### 3.4 Comprehensive Defense Strategy Implementation

We carefully selected 10 neural network models for each dataset to implement our defense strategy based on the specific data type. This selection ensured that the models in our pool were diverse and robust, including deeper versions, wider versions, and fully connected versions of various neural network architectures. For example, for datasets like MNIST and MIT-BIH, which require strong pattern recognition capabilities, we mostly utilized different types of CNNs.

In our HybridMTD strategy, all 10 models are first trained on the datasets. During the test phase, the MTD mechanism dynamically selects 4 models from the pool to form an ensemble for each data input. This dynamic selection ensures that the attack surface continuously changes, making it significantly more challenging for adversaries to exploit vulnerabilities consistently. The core idea is to enhance the defense's unpredictability and robustness by altering the models being targeted with each evaluation.

The decision to select 4 models out of 10 in our HybridMTD strategy is carefully balanced. Choosing more models would reduce the number of unique sets, making the defense less unpredictable and easier for attackers to anticipate. Conversely, selecting fewer models could weaken the effectiveness of majority voting, which relies on combining the decisions of several models to improve accuracy. By selecting 4 models, we strike a balance between diversity and unpredictability while maintaining manageable computational costs. This setup also provides redundancy, ensuring the overall prediction remains reliable even if one or two models underperform or are compromised.

With 10 models in the pool, there are 210 unique combinations of 4-model ensembles, enhancing unpredictability and resilience. This variety of combinations ensures a broad range of defenses, making it harder for adversaries to anticipate the specific models being used.

In each evaluation, the selected four models formed an ensemble, with the final prediction for each data point determined by majority voting among these models. This approach means that the class receiving the most votes among the predictions of the four selected models is chosen as the predicted class for that data point. For instance, if two of the models predict class 2, one model predicts class 1, and another model predicts class 3, the majority voting mechanism will select class 2 as the final result, as it received the highest number of votes. In cases where a tie occurs, for

example, if two models predict class 1 and the other two models predict class 4, the final output is chosen randomly between these two classes. The majority voting process improves robustness, reduces the impact of individual model errors, and enhances the overall accuracy and reliability of the ensemble predictions.

Our comprehensive defense strategy was evaluated through several steps. Initially, the models were evaluated individually to establish a baseline performance before applying any adversarial attacks. This initial evaluation provided a reference point for assessing the impact of the defense mechanisms.

#### 3.4.1 Adversarial Example Crafting and Evaluation

Adversarial examples were crafted by having the attacker select 4 models from the pool and create adversarial examples designed to target all 4 models simultaneously. Since we form an ensemble by selecting 4 models out of a pool of 10, we also assume that the attacker will target 4 models, aligning with the ensemble structure. Although in real-world scenarios attackers may have limited knowledge, for this study, we assume that the attacker is aware that 4 models determine the final output and therefore targets 4 models accordingly. For FGSM, BIM, JSMA, and C&W attacks, which are white-box attacks, we further assumed that the attacker has full knowledge of the models in the pool. This approach allowed us to test the models' robustness against well-informed attackers by challenging the defense mechanisms with adversarial examples designed to compromise the selected models.

In the transferability attack scenario, a black-box approach was assumed where the attacker only knew the types of models used in the pool (e.g., CNNs, RNNs) but not their specific architectures. We also assumed that the attacker is aware that 4 models determine the final output, so they focused their efforts on targeting 4 models. The attacker selected four models of their own, with different architectures from those in our pool, and crafted adversarial examples designed to target all four models simultaneously. This approach mimicked real-world conditions where attackers, despite lacking complete information about the target system, still attempted to create adversarial examples that could transfer and succeed against the models in the target pool.

For poisoning attacks, it was assumed that the attacker had poisoned four of the models during training, meaning that our pool already contained poisoned models. This implies that when 4 models are selected during the testing phase to form an ensemble, there is a possibility that some of the selected models could be among the poisoned ones, potentially compromising the integrity of the ensemble. While we assume that the attacker had full access to poison 4 models, this level of access is often unrealistic in real-world scenarios, as even gaining access to 4 models is a challenging feat.

By dynamically selecting random models for each data input and employing majority voting, our HybridMTD strategy leverages the strengths of both MTD and ensemble methods. This approach ensures that the attack surface is continuously changing, making it difficult for adversaries to adapt, while the ensemble method enhances overall robustness through collective decision-making.

Figure 1 illustrates the process where the attacker generates adversarial examples to challenge our system. Four random models from our pool are dynamically selected for each adversarial example to evaluate it. This random selection makes it extremely difficult for the attacker to predict which models will be used at any given time, thereby enhancing the unpredictability and robustness of our defense strategy. After applying the adversarial attacks, the models were re-evaluated to assess the impact and effectiveness of the defense mechanisms.

### 3.5 Evaluation Metrics

To comprehensively evaluate the performance and robustness of our models, we recorded a range of metrics during both the training and testing phases, before and after the application of adversarial attacks. During the training phase, we tracked the number of epochs, training loss, validation loss, and validation accuracy to monitor the learning process and detect any overfitting or underfitting. We measured accuracy, precision, recall, and F1-score for the testing phase to assess the models' classification performance. Additionally, we used the confusion matrix to visualize the distribution of true positives, false positives, true negatives, and false negatives.

To understand the error margins, we computed the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provided insights into the average and squared differences between predicted and actual values. We recorded these metrics for each model individually before and after the application of adversarial attacks. Additionally, we evaluated the overall performance of our defense strategy by applying the MTD and majority voting on the ensemble models, both before and after the attacks. To assess the effectiveness of the combined approach, we also eval-

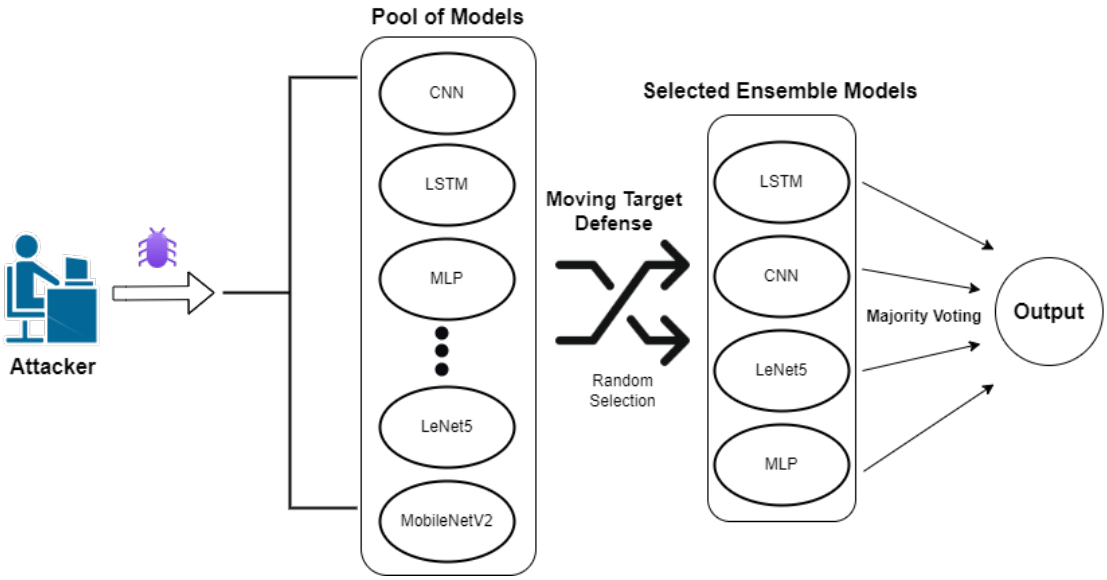



Figure 1: Dynamic model selection process in HybridMTD. Adversarial examples are processed by random ensembles of four models from a pool of ten, enhancing robustness and security.

uated our strategy by applying MTD without the ensemble model selection, relying instead on the output of a single model. This comparison allowed us to see how much more effective the combined version with ensemble models is versus the one relying on just a single model.

# 4 EXPERIMENTAL RESULTS

## 4.1 Baseline Performance

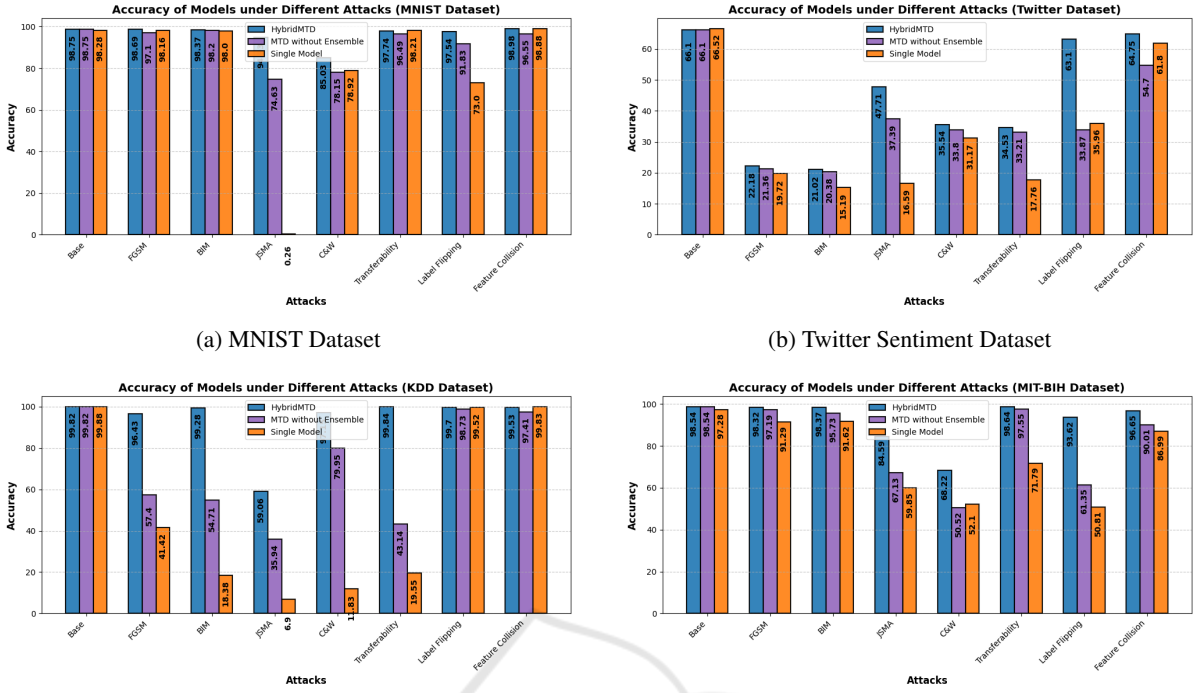
As the baseline performance, we first evaluated our HybridMTD framework using only legitimate data, with no adversarial attacks. We also conducted evaluations using a single model for each dataset, following the conventional approach in machine learning. Specifically, we used a CNN for MNIST and MIT-BIH, an LSTM for the Twitter Sentiment dataset, and an MLP for the KDD dataset. Additionally, we evaluated the MTD approach without using ensemble models, relying on the output of just one model at a time. The first set of bars in Figure 2 (base bars) represents these baseline performance metrics for all four datasets.

## 4.2 Performance Under Adversarial Attacks

### 4.2.1 Visual Performance Analysis Across Datasets

We present the results of our experiments in Figure 2, comparing the performance of a conventional single-model approach, our HybridMTD strategy, and the MTD approach without ensemble models under various adversarial attacks. The blue bars show the final accuracy of our HybridMTD method, where each data point is evaluated by applying the MTD approach and an ensemble of models selected randomly, leading to a single output. The purple bars represent the performance when MTD is applied but rely on just one model from the pool, which is similar to traditional methods. The orange bars show the results when neither MTD nor ensemble models are used, reflecting the performance without any defense strategy.

In the MNIST dataset (Fig 2-a), the accuracies under different attacks show significant improvements with HybridMTD. For instance, under the JSMA attack, the HybridMTD approach achieved an accuracy of 94.93%, while the single-model approach had a drastically lower accuracy of 0.26%, and the MTD approach without ensemble models reached an accuracy of 74.63%. For poisoning attacks, the HybridMTD approach also demonstrated substantial improvements. Under the label flipping attack, the HybridMTD achieved an accuracy of 97.54%, compared to 73.00% for the single model, while the MTD ap-




(a) MNIST Dataset

(b) Twitter Sentiment Dataset

(c) KDD Dataset

(d) MIT-BIH Dataset

Figure 2: Comparison of baseline and post-adversarial attack accuracies across different datasets for the HybridMTD approach, MTD without ensemble models, and single models. The first set of bars represents the baseline performance using legitimate data, while the subsequent bars show the accuracies after applying various adversarial attacks.

proach without ensemble models reached 91.83%. These findings emphasize HybridMTD's effectiveness in maintaining high accuracy for the MNIST dataset.

Our HybridMTD approach consistently outperformed the conventional single-model method and MTD without ensemble models across all attacks for the Twitter Sentiment dataset (Fig 2-b). Notably, under the JSMA attack, the HybridMTD approach achieved an accuracy of 47.71%, significantly higher than the 37.39% accuracy of the MTD approach and 16.59% accuracy of the single-model approach. Similarly, for the label flipping attack, HybridMTD achieved an accuracy of 63.10%, compared to 33.87% accuracy of the MTD approach and 35.96% for the single model.

For the KDD dataset (Fig 2-c), the results show that our HybridMTD approach consistently achieved high accuracy, outperforming the MTD without ensemble models and conventional single-model methods. The HybridMTD approach maintained an accuracy above 96% against most attacks. For instance, against the transferability attack, which is a black-box attack, the HybridMTD achieved an impressive accuracy of 99.84%, compared to only 43.14% with the MTD and 19.55% with the single-model method.

The results of the MIT-BIH dataset (Fig 2-d) indicate that our HybridMTD approach significantly outperformed the MTD and single-model methods across various attacks. Specifically, against the label flipping attack, HybridMTD achieved an accuracy of 93.62%, compared to only 61.35% with MTD and 50.81% with the single model. Overall, the HybridMTD approach maintained an accuracy of over 93% in most cases, demonstrating its robust defense capability on the MIT-BIH dataset.

### 4.2.2 Quantitative Performance Differences in HybridMTD and Baseline Approaches

We also present our results in Table 1 with color coding highlighting the accuracy differences between HybridMTD, MTD without ensemble models, and the single-model approach. Vivid greens indicate where our HybridMTD approach outperforms the other methods, while light greens represent more modest improvements in accuracy. As can be seen, negative numbers are few and all of them are less than 1%, demonstrating the consistency of HybridMTD in maintaining higher performance across different attacks and datasets.

For the MNIST dataset, HybridMTD delivers significant improvements, especially under the JSMA attack, where it achieves 94.67% accuracy com-

pared to 0.26% for the single-model approach and 74.63% for MTD without ensemble models. Similarly, under the label flipping attack, HybridMTD reaches 97.54% accuracy, outperforming both the single model (73.00%) and MTD (91.83%).

In the Twitter dataset, HybridMTD shows clear improvements across all attacks. For instance, under the JSMA attack, HybridMTD achieves 47.71% accuracy, while MTD and the single model only achieve 37.39% and 16.59%, respectively. The label flipping attack demonstrates HybridMTD's robustness, with an accuracy of 63.10%, compared to 33.87% for MTD and 35.96% for the single model.

For the KDD dataset, the results are even more pronounced. Against the transferability attack, HybridMTD achieves an accuracy of 99.84%, far surpassing MTD (43.14%) and the single model (19.55%). Other attacks, such as BIM, also show large differences, with HybridMTD reaching 99.28% accuracy compared to 54.71% for MTD and 18.38% for the single model.

In the MIT-BIH dataset, HybridMTD consistently performs better across all attacks. Under the label flipping attack, HybridMTD achieves 93.62% accuracy, significantly higher than the 61.35% achieved by MTD and 50.81% by the single model. In the feature collision attack, HybridMTD again outperforms the others, with a 6.64% higher accuracy than MTD and 9.66% higher than the single model.

Overall, the table clearly shows that HybridMTD consistently improves accuracy, demonstrating its effectiveness in maintaining robust performance against adversarial attacks across various datasets and attack types.

#### 4.2.3 Performance of Models Under Adversarial Attacks Based on Different Metrics

Table 2 presents the comparison of percentages of Accuracy, Recall, and F1 Score under various adversarial attacks for Single Model, Traditional MTD, and HybridMTD approaches. The results show that HybridMTD consistently outperformed both the Single Model and MTD without ensemble models across all datasets and attack types.

As can be seen in the table, the other two metrics (Recall and F1 Score) also followed the same performance trend, aligning with the accuracy discussed in Section 4.2.1. This indicates that HybridMTD not only achieves higher accuracy but also maintains consistent improvements in Recall and F1 Score, demonstrating its superior performance compared to Traditional MTD and models without any defense mechanism.

# 5 DISCUSSION

Our primary goal was to assess the effectiveness of a new technique that combines Moving Target Defense (MTD) with ensemble models as a defense strategy against various evasion and poisoning attacks. Our results indicate that this hybrid approach, termed HybridMTD, offers substantial improvements in robustness compared to MTD approach without using ensemble models and conventional single-model methods.

## 5.1 Comparison of Defense Strategies of Existing Studies

Several studies have focused exclusively on applying MTD against evasion attacks or Denial-of-Service (DoS) attacks. Notably, we did not find any studies that address MTD in the context of poisoning attacks. Most research has been confined to image datasets, with limited exploration of other data types, except in a few instances.

We aimed to broaden the scope by including four different types of data: image (MNIST), text (Twitter Sentiment), tabular (KDD), and signals (MIT-BIH). We evaluated HybridMTD against a comprehensive range of attacks, including white-box, black-box, targeted (e.g., feature collision), untargeted, model-specific (e.g., FGSM, BIM, JSMA, C&W), and model-agnostic attacks. This holistic approach comprehensively evaluated our defense mechanism across diverse scenarios.

In our experiments with the MNIST dataset using FGSM with an epsilon of 0.1, HybridMTD achieved notable improvements. The MTDeep framework [13] increased accuracy from 0% to 23.8%, and the Morphence framework [15] increased accuracy from 9.98% to 71.43%. HybridMTD achieved an accuracy of 98.69%, demonstrating consistency in performance across various attack types. Many studies reported an accuracy close to 0% under FGSM attack due to different settings. Therefore, direct comparison with these studies is challenging.

For C&W attacks on MNIST, another paper [18] reported an increase in accuracy from 0% to 50%, while the Morphence framework achieved an increase from 0% to 97.75%. HybridMTD maintained an accuracy of 85.03% against C&W, compared to 78.92% with a single model. It's worth noting that Morphence utilized multiple CNN models with six layers each, whereas our approach employed simpler CNN models with only three layers. Despite this difference in model complexity, HybridMTD demonstrated superior performance, highlighting the robustness and ef-

Table 1: This table shows accuracy differences of HybridMTD compared to MTD without ensemble models and the single-model approach across datasets. Each dataset has two sub-columns: 'vs MTD' (accuracy differences with MTD without ensemble) and 'vs Single' (accuracy differences with the single-model approach). Positive values indicate better performance of HybridMTD.

| | MNIST | | Twitter | | KDD | | MIT-BIH | |
|---|---|---|---|---|---|---|---|---|
| HybridMTD | vs MTD | vs Single | vs MTD | vs Single | vs MTD | vs Single | vs MTD | vs Single |
| FGSM | 1.59% | 0.53% | 0.82% | 2.46% | 39.03% | 55.01% | 1.13% | 7.03% |
| BIM | 0.17% | 0.37% | 0.64% | 5.83% | 44.57% | 80.90% | 2.64% | 6.75% |
| JSMA | 20.30% | 94.67% | 10.32% | 31.12% | 23.12% | 52.16% | 17.46% | 24.74% |
| C&W | 6.88% | 6.11% | 1.74% | 4.37% | 17.20% | 85.32% | 17.70% | 16.12% |
| Transferability | 1.25% | -0.47% | 1.32% | 16.77% | 56.70% | 80.29% | 1.09% | 26.85% |
| Label Flipping | 5.71% | 24.54% | 29.23% | 27.14% | 0.97% | 0.18% | 32.27% | 42.81% |
| Feature Collision | 2.43% | 0.10% | 10.05% | 2.95% | 2.12% | -0.30% | 6.64% | 9.66% |

Table 2: Comparison of Accuracy, Recall, and F1 Score across Four Datasets (MNIST, Twitter, KDD, MIT-BIH) under Different Situations. The table presents metrics for different scenarios: Base (before any adversarial attack), and under various adversarial attacks, including FGSM, BIM, JSMA, C&W, Transferability, Label Flipping, and Feature Collision. The results are compared for Single Model, Traditional MTD, and HybridMTD approaches.

| | | MNIST | | | Twitter | | | KDD | | | MIT-BIH | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | F1-Score | Recall | Accuracy | F1-Score | Recall | Accuracy | F1-Score | Recall | Accuracy | F1-Score | Recall |
| Base | HybridMTD | 98.75% | 98.73% | 98.73% | 66.1% | 61.72% | 61.16% | 99.82% | 99.81% | 99.82% | 98.54% | 98.50% | 98.55% |
| | MTD | 98.75% | 98.73% | 98.73% | 66.1% | 61.72% | 61.16% | 99.82% | 99.81% | 99.82% | 98.54% | 98.50% | 98.55% |
| | Single Model | 98.28% | 98.26% | 98.27% | 66.52% | 66.24% | 66.27% | 99.88% | 99.86% | 99.88% | 97.28% | 96.87% | 97.28% |
| FGSM | HybridMTD | 98.69% | 98.68% | 98.68% | 22.18% | 19.42% | 20.50% | 96.43% | 97.74% | 96.44% | 98.32% | 98.27% | 98.33% |
| | MTD | 97.1% | 97.14% | 97.16% | 21.36% | 17.09% | 19.56% | 57.4% | 60.35% | 57.41% | 97.19% | 97.12% | 97.19% |
| | Single Model | 98.16% | 98.14% | 98.14% | 19.72% | 19.45% | 19.43% | 41.42% | 41.41% | 41.43% | 91.29% | 91.37% | 91.29% |
| BIM | HybridMTD | 98.37% | 98.36% | 98.34% | 21.02% | 17.41% | 19.14% | 99.28% | 99.39% | 99.29% | 98.37% | 98.29% | 98.38% |
| | MTD | 98.2% | 98.18% | 98.18% | 20.38% | 16.87% | 18.43% | 54.71% | 56.31% | 54.72% | 95.73% | 95.78% | 95.73% |
| | Single Model | 98% | 97.99% | 98.00% | 15.19% | 13.48% | 15.01% | 18.38% | 10.59% | 18.38% | 91.62% | 91.68% | 91.62% |
| JSMA | HybridMTD | 94.93% | 94.97% | 94.92% | 47.71% | 46.61% | 47.29% | 59.06% | 62.40% | 59.07% | 84.59% | 85.17% | 84.60% |
| | MTD | 74.63% | 74.86% | 74.66% | 37.39% | 35.93% | 36.16% | 35.94% | 36.02% | 35.94% | 67.13% | 71.53% | 67.13% |
| | Single Model | 0.26% | 0.26% | 0.26% | 16.59% | 16.28% | 16.32% | 6.9% | 8.84% | 6.90% | 59.85% | 68.45% | 59.86% |
| C&W | HybridMTD | 85.03% | 86.8% | 85% | 35.54% | 25.29% | 35.88% | 97.15% | 96.96% | 97.16% | 68.22% | 69.40% | 68.22% |
| | MTD | 78.15% | 81.44% | 77.91% | 33.8% | 24.15% | 33.82% | 79.95% | 78.49% | 79.93% | 50.52% | 59.17% | 50.52% |
| | Single Model | 78.92% | 80.28% | 78.65% | 31.17% | 17.45% | 31.02% | 11.83% | 6.11% | 11.83% | 52.1% | 58.86% | 52.10% |
| Transferability | HybridMTD | 97.74% | 97.70% | 97.67% | 34.53% | 19.37% | 34.34% | 99.84% | 99.83% | 99.84% | 98.64% | 98.58% | 98.64% |
| | MTD | 96.49% | 96.46% | 96.41% | 33.21% | 16.69% | 33.35% | 43.14% | 44.03% | 43.12% | 97.55% | 97.34% | 97.56% |
| | Single Model | 98.21% | 98.20% | 98.19% | 17.76% | 16.28% | 17.21% | 19.55% | 10.71% | 19.55% | 71.79% | 74.64% | 71.79% |
| label Flipping | HybridMTD | 97.54% | 97.52% | 97.51% | 63.1% | 59.27% | 59.02% | 99.7% | 99.69% | 99.71% | 93.62% | 95.51% | 93.63% |
| | MTD | 91.83% | 91.64% | 91.64% | 33.87% | 16.87% | 33.33% | 98.73% | 98.46% | 98.73% | 61.35% | 72.88% | 61.36% |
| | Single Model | 73% | 70.56% | 73.13% | 35.96% | 34.87% | 35.54% | 99.52% | 99.42% | 99.52% | 50.81% | 64.05% | 50.81% |
| Feature Collision | HybridMTD | 98.98% | 98.98% | 98.97% | 64.75% | 61.25% | 60.68% | 99.53% | 99.51% | 99.54% | 96.65% | 96.51% | 96.65% |
| | MTD | 96.55% | 96.53% | 96.53% | 54.7% | 47.49% | 48.32% | 97.41% | 96.55% | 97.42% | 90.01% | 89.32% | 90.02% |
| | Single Model | 98.88% | 98.86% | 98.86% | 61.8% | 60.50% | 61.43% | 99.83% | 99.84% | 99.83% | 86.99% | 85.66% | 86.99% |

ficiency of our approach.

In a study using the KDD dataset with a DoS attack, another paper [16] reported an accuracy increase from 50% to around 90%. In contrast, our HybridMTD strategy consistently showed higher resilience and performance across different datasets and attack types. For example, on the KDD dataset, HybridMTD maintained accuracy above 96% against

most attacks, achieving an accuracy of 98.64% against the transferability attack.

## 5.2 Implications for Practice and Limitations

Our framework demonstrates significant resilience against various types of attacks, particularly poisoning attacks. In scenarios where the majority of models are not poisoned, which we consider a more realistic scenario, the final evaluation based on majority voting of ensemble models ensures robust performance. This approach effectively mitigates the impact of poisoning attacks, maintaining high accuracy.

For evasion attacks, the framework performs well as long as the adversarial examples do not drastically degrade the performance of most models. When adversarial examples are strong enough to affect all models, we still observe an increase in accuracy, but this may not always qualify as an effective defense, particularly in the case of the Twitter sentiment dataset. Performance on this dataset was comparatively lower than on other datasets, because even adversarial examples crafted for specific models often had a broader impact, reducing the accuracy of most models in the pool. As a result, when selecting a subset of models from the pool, many models are already affected by these adversarial examples, leading to good performance overall but not as robust as observed with the other datasets. Twitter sentiment is complex and highly susceptible to performance degradation. This complexity arises from the linguistic nuances, variability in sentiment expression, and context ambiguity inherent in textual data. These factors make it easier for adversarial attacks to significantly and easily reduce accuracy, as small perturbations can lead to misclassification in sentiment analysis.

Additionally, our approach outperformed the MTD approach without using ensemble models in all tested scenarios. This demonstrates that integrating ensemble models with MTD enhances robustness and significantly improves overall performance. This success suggests that our HybridMTD framework could serve as a substantial improvement over traditional MTD approaches, making it a more reliable and effective defense strategy in practical applications.

However, this enhanced performance comes with a tradeoff: increased training time. The need to train multiple models and the additional computation required for dynamic model selection during testing can result in longer processing times. Despite this, the improvement in robustness and accuracy justifies the additional cost. In environments where security and reliability are critical, the benefits of maintaining high accuracy and resilience against attacks outweigh the increased computational demands, making this approach a valuable investment.

In practice, this implies that our HybridMTD framework can be particularly effective in environments where poisoning attacks are a significant concern and where the diversity and strength of adversarial attacks vary. The combination of MTD and ensemble models offers a versatile defense strategy capable of adapting to different attack scenarios.

### 5.2.1 Factors Influencing Defense Effectiveness

The effectiveness of the HybridMTD strategy depends on several factors. Key among them is the diversity of models in the pool, as similar architectures may share vulnerabilities. The quality of training data is also critical, as incomplete or biased datasets can limit robustness. Additionally, the frequency of model updates impacts adaptability to evolving threats, while sufficient randomness in model selection ensures unpredictability. Addressing these factors enhances both the robustness and efficiency of the defense mechanism.

### 5.2.2 Computational Costs

The HybridMTD strategy incurs higher computational costs compared to simpler defense methods due to the need for training multiple models and dynamically forming ensembles during inference. These costs, however, are justified by the significant improvements in robustness and security against adversarial attacks. The extent of this overhead depends on the deployment scenario: for models pre-trained and distributed across applications, the training cost is incurred only once, whereas node-specific retraining significantly increases expenses. Future work could explore optimizations such as model pruning, distributed training, or transfer learning to reduce both training and inference costs. Balancing robustness with resource efficiency is critical for deploying this approach in resource-constrained environments.

## 5.3 Future Work

In future work, we plan to further investigate the observed discrepancy where the Single model demonstrates better resilience to attacks compared to MTD in certain instances. Specifically, cases such as FGSM on the MNIST dataset and Label Flipping on the KDD dataset suggest that the Single model achieves accuracy levels closer to HybridMTD than MTD without ensemble models under some attack scenarios, even

though these differences in resilience are not substantial. A deeper exploration into the factors contributing to this resilience in the Single model, as well as the trade-offs between Single and MTD without using ensemble models in different attack environments, would provide valuable insights for optimizing defense strategies across various adversarial contexts.

Additionally, future work could involve investigating scenarios where the attacker has enhanced capabilities. For evasion attacks, this could include creating adversarial examples based on knowledge of a greater number of model architectures within the pool, and for poisoning attacks, it could involve gaining access to compromise a larger number of models. Although this assumes a level of access and knowledge that is unrealistic in real-world scenarios, exploring these worst-case conditions would allow us to understand the robustness of different defense strategies under maximum adversarial pressure, further informing the development of resilient frameworks.

## 6 CONCLUSION

In this study, we aimed to assess the effectiveness of HybridMTD, a novel defense strategy that combines Moving Target Defense with ensemble neural network models, against a wide range of adversarial attacks. Our extensive experiments across four different datasets—MNIST (image), Twitter Sentiment (text), KDD (tabular), and MIT-BIH (signals)—and seven sophisticated attack types, including both evasion and poisoning attacks, have demonstrated the robustness and resilience of HybridMTD.

The results indicate that HybridMTD significantly outperforms the traditional MTD approach and conventional single-model methods, maintaining high accuracy and robustness. By leveraging the dynamic selection of a subset of models from a diverse pool and employing majority voting, HybridMTD increases the unpredictability of the defense mechanism, making it more challenging for adversaries to execute their attacks successfully. HybridMTD worked exceptionally well for poisoning attacks, maintaining high performance when most models were not compromised. For evasion attacks, HybridMTD demonstrated robust performance, particularly when adversarial examples did not severely degrade the performance of most models. Overall, in all scenarios, we observed a substantial increase in performance, confirming HybridMTD's effectiveness as a comprehensive defense strategy.

## REFERENCES


Amich, A. and Eshete, B. (2021). Morphence: Moving target defense against adversarial examples. In *Annual Computer Security Applications Conference (ACSAC)*.

An, Q., Rahman, S., Zhou, J., and Kang, J. J. (2023). A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *SENSORS*.

Biggio, B., d'Armi, P., Nelson, B., and Laskov, P. (2013). Poisoning attacks against support vector machines. *arXiv*. arXiv:1206.6389v3.

Biggio, B., Nelson, B., and Laskov, P. (2011). Support vector machines under adversarial label noise. In *Asian Conference on Machine Learning*, PMLR, pages 97–112.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. *arXiv*. arXiv:1608.04644v2.

Colbaugh, R. and Glass, K. (2013). Moving target defense for adaptive adversaries. In *ISI*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15.

Dong, Y., Deng, Z., Pang, T., Zhu, J., and Su, H. (2020). Adversarial distributional training for robust deep learning. In *Conference on Neural Information Processing Systems*.

Gao, R., Cai, T., Li, H., Hsieh, C., Wang, L., and Lee, J. D. (2019). Convergence of adversarial training in overparametrized neural networks. In *Conference on Neural Information Processing Systems*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2017). Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR)*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998a). Gradient-based learning applied to document recognition. *IEEE*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998b). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lei, C., Zhang, H. Q., Tan, J. L., Zhang, Y.-C., and Liu, X. H. (2018). Moving target defense techniques: A survey. *Security and Communication Networks*. Article ID 3759626.

Liu, H. (1999). The kdd'99 dataset. The UCI KDD Archive, University of California, Irvine, CA.

Liu, J., Pong Lau, C., Souri, H., Feizi, S., and Chellappa, R. (2022). Mutual adversarial training: Learning together is better than going alone. *Transactions on Information Forensics and Security*, 17:2364–2377.

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., and Leung, V. C. M. (2018). A survey on security threats and de-

fensive techniques of machine learning: A data driven view. *IEEE Access*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Martin, P., Fan, J., Kim, T., Vesey, K., and Greenwald, L. (2021). Toward effective moving target defense against adversarial ai. In *MILCOM*.

Moody, G. B. and Mark, R. G. (2001). The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I. M., and Edwards, B. (2019). Adversarial robustness toolbox v1.0.0. *arXiv*. arXiv:1807.01069v4.

Papernot, N. and McDaniel, P. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv*. arXiv:1605.07277v1.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2015). The limitations of deep learning in adversarial settings. *arXiv*. arXiv:1511.07528v1.

Ren, K., Zheng, T., Qin, Z., and Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Elsevier Engineering*.

Roy, A., Chhabra, A., Kamhoua, C. A., and Mohapatra, P. (2019). A moving target defense against adversarial machine learning. In *ACM/IEEE Workshop on Security and Privacy in Edge Computing*.

Sengupta, S., Chakraborti, T., and Kambhampati, S. (2019). Mtdeep: Boosting the security of deep neural nets against adversarial attacks with moving target defense. In *Conference on Decision and Game Theory for Security*.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. *arXiv*. arXiv:1804.00792v2.

Shahane, S. (2021). Twitter sentiment dataset. Kaggle, Available: https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*.

Sun, R., Zhu, Y., Fei, J., and Chen, X. (2023). A survey on moving target defense: Intelligently affordable, optimized and self-adaptive. *Applied Sciences*, 13:5367.

Wu, B., Wei, S., Zhu, M., Zheng, M., Zhu, Z., Zhang, M., Chen, H., Yuan, D., Liu, L., and Liu, Q. (2023). Defenses in adversarial machine learning: A survey. *arXiv*. arXiv:2312.08890v1.

Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., and Yu, P. S. (2022). Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Comput. Surv.*