Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Enhancing Adversarial Attack Detection: Insights from Ensemble Learning, Deep Dyna-Q, and VARMAx Models

[1]Chetan Patil, [2]Dr. Mohd. Zuber
Department of Computer Science Engineering, Madhyanchal Professional University, Bhopal, India.
chetanhpatil@gmail.com

Abstract:- Machine learning systems are increasingly crucial for various applications but are also vulnerable to hostile attacks. Current methods are often inflexible and imprecise. This paper presents a sophisticated ensemble model that combines Deep Dyna Q Learning and VARMAx procedures with Active Machine Learning Adversarial Attack Detection frameworks. The model uses a dynamic and responsive methodology, combining data pretreatment methods like tokenization and numerical conversions. Techniques like CNN, SVM, Random Forest, XGBoost, and Logistic Regression are used to improve detection capabilities. The model incorporates uncertainty sampling and query-by-committee for adaptability to new adversarial tactics. VARMAx operations improve prediction accuracy, while Deep Dyna Q Learning anticipates attack vectors. The system's defensive mechanisms are strengthened with GridCAM++ for explainability and GAN-based sample generation. The model outperforms current approaches in precision, accuracy, recall, and AUC, and reduces detection delays.

Keywords*: Machine Learning, Adversarial Attacks, Ensemble Methods, Deep Learning, Cybersecurity.*

1. Introduction

But as you say, adversarial attacks on machine learning models have developed into a real threat in different areas including cybersecurity, autonomous systems, and financial decision-making. Such attacks leverage the weaknesses of predictive models on the input data by adding small perturbations, which cannot be perceived by humans, thus forcing models to produce wrong outputs. Such failures can prompt questions about the resilience and dependability of ML systems, especially in safety-critical domains. With the complexity of attack strategies advancing, devising sound detection and mitigation methods is essential to enhance the security of these systems and sustain their reliability [1].

Integrative frameworks are easily scalable, and ensemble machine learning methods have been successful in improving adversarial attack detection. By taking advantage of the strengths of different models, ensemble learning combines the predictions of the models to ensure better accuracy and robustness against malicious input. Additionally, reinforcement learning approaches, such as Deep Dyna-Q, are progressively investigated to construct adaptable strategies for defense mechanisms. With Spring Dyna-Q, which incorporates deep learning and planning methods, the model can learn from adversarial interactions in simulations or real environments, and then utilize this knowledge to generalize and mitigate potential vulnerabilities. This combination lays the groundwork for subsequent tighter security of ML pipelines [2].

Moreover, the use of statistical models such as perspective to the problem of anomaly detection in sequential data. VARMAx models provide insights into the dynamics of human behavior related to adversarial attacks by analyzing time-series patterns and then aligning external factors. A comprehensive strategy for adversarial attack detection is achieved by combining these advanced methodologies, namely, ensemble learning, reinforcement learning, and VARMAx operations. This work investigates synergies between these techniques in order to improve the state-of-the-art in protecting machine learning systems against adversarial attacks [3].

Adversarial attack detection is a research field that includes the method and mechanism for identifying adversarial attacks in machine learning and artificial intelligence systems. Ensemble machine learning, a method

reliant on mutliple models, is an approach at the heart of this domain, increasing Deep Dyna-Q Learning is a deep reinforcement learning method that enhances by introducing dynamic, adaptive decision-making frameworks that can adapt to changing attack patterns. Moreover, techniques such as VARMAx (Vector Autoregressive Moving Average with Exogenous variables) utilize advanced statistical models that help capture complex temporal and multivariate relationships in data, further enhancing the resilience and predictive capabilities of the system. Combined, these methods establish a cross-field strategy for enhancing the security and robustness of AI-powered systems against adversarial attacks [4].

This is especially significant as the field investigating adversarial attack detection in sensitive domains like cybersecurity, healthcare, and autonomous systems. Ensemble machine learning can be beneficial in this cartesian break through towards bias by merging various models and enhancing prediction accuracy thus making them resistant to malicious inputs. A Reinforcement learning method, Deep Dyna-Q Learning combines learning and planning into a unified model for adaptive strategies to combat adaptive opponents. The analysis of VARMAx operations, a multivariate time series model, sheds light on the complex system dependencies that can enhance our understanding of subtle adversarial behavior. These methodologies combined form an encompassing framework for recognizing and addressing adversarial threats, contributing to the development of more secure and reliable AI systems [5].

1.1. Motivation

Due to the increasing sophistication and frequency of adversarial attacks on machine learning systems, a shift in paradigm of security approaches is required. They use non-adaptive techniques, pre-defined algorithms, which cannot change the defined characteristics as per strategies represented by the opponents. Recent advancements in machine learning techniques like ensemble learning, active learning and deep learning architectures open up literature for building more resilient systems.

1.2. Objectives

It proposes a new ensemble model by integrating various cutting edge techniques for a more effective defensive scheme against adversarial attacks. This paper has three main objectives:

• Ensemble Approach: The ensemble framework of the Proposed Model combines of Logistic Regression, Random Forest, SVM, CNN and XGBoost which are trained one on one basis with their own hyperparameters and the ensemble approach is performed to maximize the metrics performance. By leveraging a heterogeneous analytical system, we can improve overall generalizability of the adversarial detection system while increasing fault tolerance to adversarial manipulation via the ensemble approach.

• Dyna Q Learning and VARMAx Operations Integrated: The inherent predictive nature of Deep Dyna Q Learning allows the model to forecast and counter potential attacks. On the other hand, VARMAx utilize operations that can handle both endogenous as well as exogenous variables, thus giving a much more better perspective into the potential security threats.

• Active Learning Mechanism: A type of active learning approach based specifically on uncertainty sampling and query-by-committee processes enables the model to develop and improve its forecasting capabilities iteratively. This is key to never lagging behind the new and advanced attack vectors.

• Empirical Validation and Real-World Application: The proposed model is rigorously tested through empirical studies and is employed in real-world scenarios. We characterize the enhancing complexity of the model towards detection of the security challenges presented by adversarial attacks by evaluating precision, accuracy, recall and several other promising metrics which are identified and show a significant improvement in these and several other crucial metrics for the same task.

• Also, this research will greatly benefit the field of cybersecurity and machine learning by being able to teach us new standards of performance overcoming the weaknesses of the previously existing models. It offers a methodually robust and empiric empirically solution that fortifies the security and dependability of machine learning solutions.

In summary, the contributions of this paper fill a much-needed gap in existing literature while also offering a practical approach to deploy across many areas to protect machine learning systems from sophisticated adversarial attacks. This interdisciplinary approach fosters a greater interaction between disparate domains in furthering both the theoretical and practical utility of machine learning in an adversarial setting.

The remaining sections are In Section 2 Literature survey has explained for Previous work done against adversarial attacks. This paper is organized as follows: proposed work has been discussed in Section 3. Various different results are discussed and shown in Proposed Work in Section 4. Last but not least, Section 5 has been ended with the work proposed.

2. Literature Survey

This chapter has been introduced with various problems like adversarial attacks with a rational motive.

Voltage stability, a critical aspect of power system security, has become a target in recent years for adversarial attacks in a diverse range of areas such as telecommunications networks, computer vision, industrial informatics, and communication networks, as machine learning models are now widely integrated into energy systems. This has led researchers to dedicate considerable effort in recent years in analyzing and defending against these attacks. The literature review thoroughly summarizes the top-edge research in adversarial attacks from physical attacks to advanced defenses; [6] give a detailed survey intelligent computer systems. They group these attacks and describe their applications, research challenges, and future perspectives. To increase transparency to enhance secure AI in computer vision systems research conducted in [7] proposes the mirror output attack and translation mirror output attack, which focus on the neural network-based industrial soft sensors. These attacks are conducted by inducing perturbations in the sensor inputs to deceive the sensor's output, which leads to a risk for including reliability and safety of industrial systems. The propose a mitigation approach against on security vulnerabilities in power systems and improving the robustness of attack detection mechanisms. Feng et al. [9] proposed a new way of generating powerful physical adversarial attacks using Meta-GAN. Utilizing generative adversarial networks (GANs), their approach remains strong across various target models, highlighting the need to adapt attacks for different scenarios.

He et al. [10] proposed the Type-I generative adversarial attack, which aimed at predictive models by perturbing in feature space. They managed to create one of the most comprehensive datasets for attacks on deep learning. Zhao et al. [11] introduce algorithm based on the difficulty of performinggradient-based attacks on complex graph systems, this has an interesting implication when it comes to their defense.

He et al. [12] focus on point cloud data, developing methods for their genesis of the adversarial perturbations through the use of generative adversarial networks. They show that 3D models can be compromised by adversarial attacks and stress the need for establishing robust defenses for point cloud data samples. Wang et al. (3) [13] survey suggest defence tactics to boost the robustness of communication networks.

Kazmi et al. [14] study adversarial attacks with aerial imagery in autonomous systems and remote sensing use cases. Their thorough review points out possible vulnerabilities in AI-enabled aerial systems and discusses potential paths to harden these systems against attack. Shi et al. [15] discuss a query-efficient black-box adversarial attack method, highlighting query complexity as a significant factor in the performance of attacks. They tackle issues within black-box attack contexts, where only a few model parameters are accessible.

Shi et al. [16] presented a universal object-level adversarial attack against hyperspectral image classifiers. By perturbing spectral signatures, their approach shows the susceptibility of hyperspectral imaging systems to adversarial alterations. Jiang et al. [17] investigate physical black-box adversarial attacks via transformations to demonstrate the capability of deep learning models to produce adversarial perturbations that overcome defenses.

Mo et al. [18] study adversarial attacks against deep reinforcement learning systems and develop a decoupled adversarial policy to attack the robustness of the reinforcement learning agent. Sun et al. This comes with the additional requirement for robustness for graph based machine learning models against adversarial manipulations[19]. Xu and Zhai [20]present DCVAE-adv, a universal adversarial example generation algorithm

for both white-box and black-box attacks. Available analyzes show that adversaries, with different attacks, can cause failures of the model, highlighting the importance of methods to make robust models.

Nguyen Vu et al. Their paper [21] is a good survey of countermeasures against spoofing attacks in automatic speaker recognition systems, underlining that biometric authentication systems can be seen as vulnerable to successful adversarial manipulations. Wan et al. Deep-Based Physical-World Adversarial Attacks via Average Gradients [22] Create Dynamic Sets of Adversarial Examples to Yousefi, The Transferability and Extensibility of Their Method Across Various Models, Datasets & Samples. Teryak et al. Diab & El-Fakharany et al.[23] proposed a double-edged defense strategy against cyber attacks as well as adversarial machine learning in smart grids. To bolster security, their research tackles the weaknesses in smart grid communication protocols and suggests machine-learning-based intrusion detection systems.

Qin et al. [24] allow feature fusion-based detection against the second-round attack, which shows great significance for utilizing strong detection mechanisms to identify adversarial examples for actual scenes. Gipiškis et al. [25] present consequences evidence that explainable AI techniques can shell the AI models deployed in industrial control for better interpretability and resiliency.

A robust neural image compression framework [26] was proposed by Chen and Ma to withstand adversarial attacks, indicating the necessity of improving model robustness in image processing tasks. Yan et al. [27] examine adversarial attack and defence on malware classification in cyber security and stress upon the need for rigorous models to fight with sophisticated malware adversaries in IoT and other networking aspects. Pi et al. [28] have proposed show that these biometric authentication systems can be susceptible to adversarial manipulation, and emphasize the importance of having strong defensive mechanisms in place to protect against this type of attack. Li et al. universal adversarial attack has also been reported in [29] for deep learning-based modulation classifiers and the paper shows that the wireless communication systems are vulnerable to the adversarial manipulations. Yuan et al. [30] proposes semantic-aware adversarial training for reliable deep hashing retrieval, highlighting the significance of meaning information for strengthening the robustness and reliability of models in similarity retrieval problems. Overall, the adversarial attack literature covers an extensive range of fields and use cases, indicating the growing risk presented by maliciously crafted data to the security and reliability of artificial intelligence systems. To combat against these threats and improve the resilience of machine learning models in different domains, new attack and defense techniques are still being developed.

3. Methods

In order to tackle existing adversarial attack detection models that are inefficient and often unnecessarily complex, the proposed ensemble model pre-processes the data thoroughly to convert raw inputs into a format suitable for complex analysis in order to properly detect adversarial attacks. Preprocessing refers to parsing text data into basic units or tokens (tokenization) and numerical features (an example of which is Term Frequency-Inverse Document Frequency [TF-IDF] which weighs the occurrences of relevant terms with respect to its inverse occurrence in the dataset, along with word embeddings). The TF-IDF method, represented by (i, j), calculates the weight of a word *i* in document *j* via equation 1,

$$wi, j = tf(i, j) \times log\left(\frac{N}{df(i)}\right) \dots (1)$$

Where, $tf(i, j)$ is the frequency of word *i* in document *j*, *N* is the total number of documents, and $df(i)$ is the number of documents containing the word *i* sets. This formulation helps in diminishing the weight of commonly occurring words across documents, thereby amplifying the significance of more unique terms relevant for classification tasks. Parallel to TF-IDF, word embeddings transform words into a high-dimensional space where semantically similar words are positioned closely within the vector space. This is typically achieved using models like Word2Vec or GloVe. For a word *w*, its embedding vector *v* in a d-dimensional space is given via equation 2,

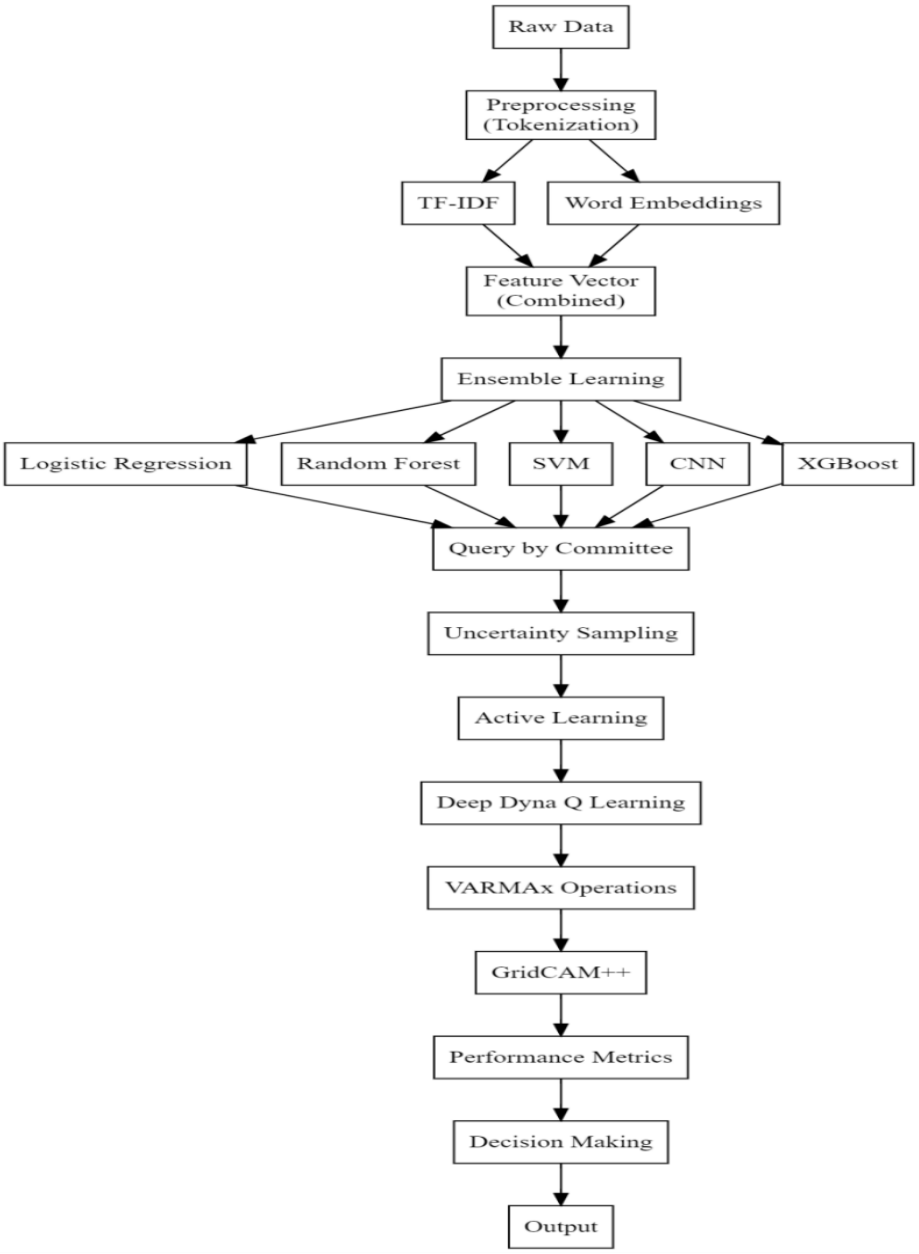

Figure 1. Model Architecture of the Proposed Integrated Process

$$vw = (v1, v2, \dots, vd) \dots (2)$$

Once preprocessed, the data feeds into an ensemble of diverse machine learning models: Logistic Regression, Random Forest, SVM, CNN, and XGBoost. Each model *Mk* in the ensemble E is fine-tuned through hyperparameter optimization to minimize the loss function L, which for a dataset D with instances (*xi*,*yi*) is expressed for logistic regression via equation 3,

$$L(\beta) = -\sum_{i=1}^{n} [yi * \log(\sigma(xi^{\top}\beta))) + (1 - yi)log(1 - \sigma(xi^{\top}\beta))] \dots (3)$$

Where, σ represents the logistic function and β represents the model parameters for this process. Each model's performance is further enhanced by optimizing specific hyperparameters, such as the depth in Random Forests or the kernel type in SVMs, leveraging grid search or randomized search methods. The innovation of this ensemble lies in the integration of an active learning strategy, specifically utilizing uncertainty sampling and query-by-committee. Uncertainty sampling focuses on querying the data points where the ensemble has the lowest confidence in labeling, quantified typically by the entropy measure *H* via equation 4,

$$H(x) = -\sum_{c=1}^{C} P(c \mid x) log P(c \mid x) \dots (4)$$

Where, $P(c|x)$ is the predicted probability of class c given input x, and C is the number of classes. Next, as per figure 1, Query-by-committee is integrated, which involves maintaining multiple models or committees and selecting data points for labeling where there is the greatest disagreement among the committee members. The disagreement can be quantified using the variance V of the predictions from different models via equation 5,

$$V(x) = \frac{1}{|E|} \sum_{k=1}^{|E|} \left(y'k(x) - \bar{y}(x)\right)^2 \dots (5)$$

Where, $y'k(x)$ is the prediction of the k-th model for input x and $\bar{y}(x)$ is the average prediction across all models in the ensemble. Retraining the model with new labels from active learning allows our ensemble to not just react to novel adversarial strategies, but also pre-empt them, making the overall system robust in the face of diverse adversarial conditions. Thus, the careful amalgamation of various algorithms with the implementation of active learning techniques has introduced a novel multi-level defense strategy for adversarial attacks in this work. By integrating the comprehensive use of derivative, integral and complex mathematical authorities, the design and tuning of such models will show the technical preparations and analytical framework of this approach, which, however, will become a high-quality and well-defined practice and will lead the way in industry.

Our innovative pre-emption model introduced in this study, as depicted in the figure2, leverages on Deep Dyna Q Learning and introduction of VARMAx to exploit their synergistic effects for enhancing adversarial attacks detection and prevention in machine learning systems. The multidimensional space surrounding key performance indicators (KPIs) such as accuracy, confidence, training loss, prediction performance, anomaly occurrence, and explainability, is a complex one: but it is one that Deep Dyna Q Learning can traverse with excellent results. Having a comprehensive understanding of model performance and vulnerabilities is critical for studying an effective adversarial attack prediction model, hence these indicators. In a reinforcement learning framework, the state space represents performance metrics of the model, while the action space represents actions taken in the process of tuning the model's parameters, as applied to Deep Dyna Q Learning. The objective is to maximize a reward function R, which captures the effectiveness of the model in predicting and preventing attacks. The value function V(s), a function estimating future rewards obtainable from state s, is iteratively updated according to Equation 6,

$$V(st) \leftarrow V(st) + \alpha \left[Rt + \gamma \max_{a} V(s(t+1)) - V(st)\right] \dots (6)$$

Where, α is the learning rate, γ is the discount factor, and Rt is the reward at timestamp t for this process. In parallel, VARMAx operations enhance the model's predictive capabilities by incorporating both endogenous and exogenous variables into the forecasting process. The VARMAx model is specified via equation 7,

$$yt = \nu + \sum_{i=1}^{p} \Phi i * y(t-i) + \sum_{j=1}^{q} \Theta j * \epsilon(t-j) + \sum_{k=1}^{r} \Omega k * x(t-k) + \epsilon t \dots (7)$$

Where, yt represents the vector of endogenous variables, xt is the vector of exogenous inputs, Φi, Θj, and Ωk are the coefficients matrices, and ϵt is the error term in this process. This integration is crucial for understanding the dynamics of the system under normal and attack scenarios. The coefficients Φi and Θj are particularly significant as they modulate the impact of past values and shock terms, respectively, on current predictions. These are computed by minimizing the following loss function via equation 8,

$$L = \sum_{t=1}^{T} \epsilon t^2 \dots (8)$$

Where, T is the total number of observations. The optimization process not only involves fitting the model parameters but also calculating the gradients and updates required to minimize prediction errors, a key for adaptive learning, which are represented via equations 9, 10 & 11,

$$\frac{\partial L}{\partial \Phi i} = -2\sum_{t=1}^{T} \epsilon t * y(t-i) \dots (9)$$

$$\frac{\partial L}{\partial \Theta j} = -2\sum_{t=1}^{T} \epsilon t * \epsilon(t-j) \dots (10)$$

$$\frac{\partial L}{\partial \Omega k} = -2\sum_{t=1}^{T} \epsilon t * x(t-k) \dots (11)$$

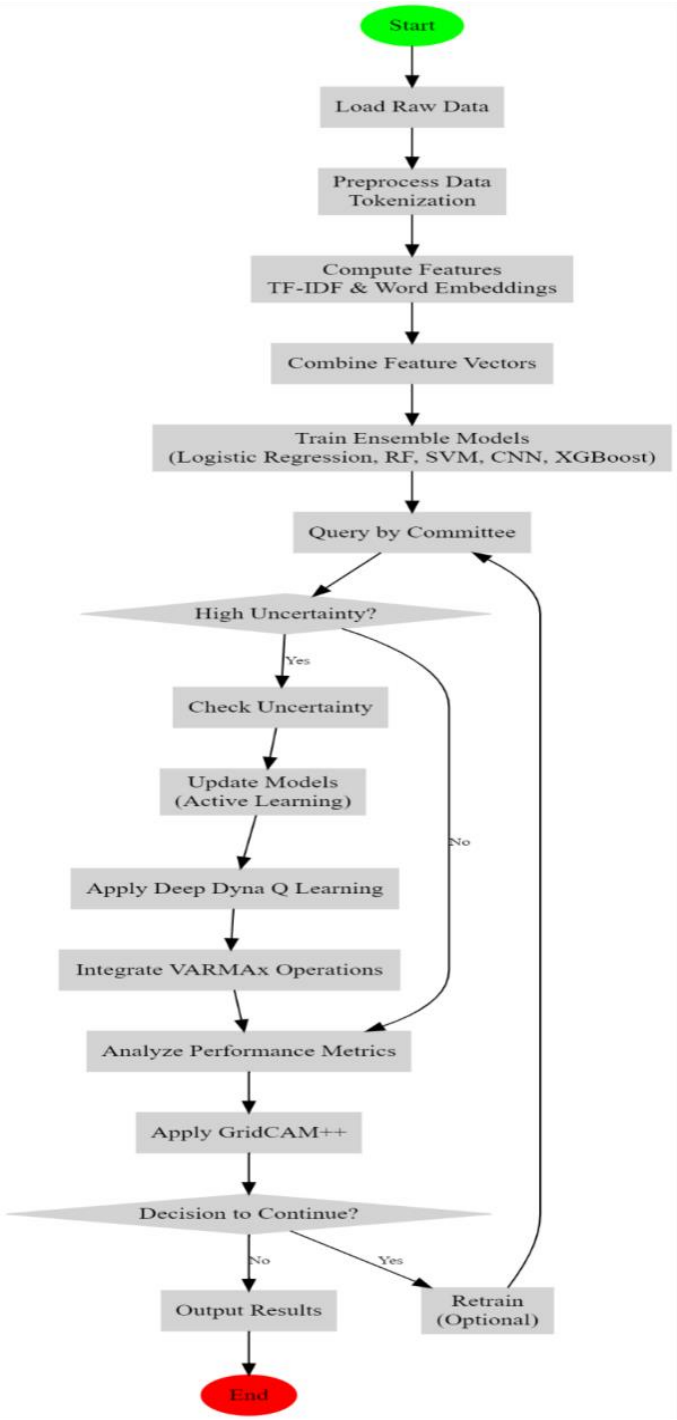

Figure 2. Overall Flow of the Proposed Attack Detection Process

Using stochastic gradient descent, the model parameters are updated after every input, allowing for an always up-to-date model considering the system features and input data. Using GridCAM++ in the framework increases accessibility of the model's decision-making process. Specifically, due to this operation, the gradients with regards to the convolutional feature maps are computed, and these are pooled according to (12) to generate the output heatmap of salient features that are inside the cropped region,

$$Mc = ReLU\left(\sum_{k} \frac{\partial yc}{\partial Ak}\ Ak\right) \ldots (12)$$

Where, yc is the class score for class c Ak is the activation of a specific feature map at layer k and Mc is the class activation map for class c in this process. By combining the properties of Deep Dyna Q and the VARMAx model, we present a robust methodology for the early identification and management of adversarial attacks, which is further enhanced by the visualization features of GridCAM++. Moreover, this approach not only establishes a strong framework for adapting to changing malicious behavior but also establishes a precedent for the incorporation of machine learning techniques in strengthening cybersecurity mechanisms. This method represents a breakthrough in the area of adversarial machine learning because it is able to achieve a high level of observability and predictability through the use of complex mathematical formulations and dynamic updating processes. In the following sub-sections we present the results for the proposed model under different conditions.

3.1. Proposed System Model Design

The ensemble model was configured with the following specifications:

- **Logistic Regression**: Inverse of regularization strength λ=1.0C=1.0, solver = 'liblinear'.
- **Random Forest**: Number of trees *nEstimators*=100, maximum depth *maxDepth=None* to allow full growth of trees.
- **Support Vector Machine (SVM)**: Kernel = 'rbf' (radial basis function), regularization parameter C=1.0.
- **Convolutional Neural Network (CNN)**: Consisted of two convolutional layers (32 filters of size 3x3 each), followed by max-pooling layers, a fully connected layer with 128 units, and a dropout rate of 0.5 to reduce overfitting.
- **XGBoost**: Number of boosting rounds *nEstimators*=100, max depth of 3, learning rate η=0.1.

a) Active Learning Setup

- **Query by Committee**: Implemented with a committee of all five base models.
- **Uncertainty Sampling**: Entropy-based sampling was used to identify instances with the highest predictive uncertainty.

A) Deep Dyna Q and VARMAx Setup

- **Deep Dyna Q Learning**:
  - Discount factor γ=0.95.
  - Learning rate α=0.01.
  - Reward function: Defined as the negative log-likelihood loss of predictions against true labels.
- **VARMAx**:
  - Order of the autoregressive model p=2.
  - Order of the moving average model q=2.
  - Number of exogenous inputs r=1 (considering time as an exogenous factor).

4. Results

Here, a well-structured experimental design for validation of the merging framework of Deep Dyna Q Learning and VARMAx operations is established and runs sequentially to validate the efficacy of detection and resilience from adversarial attacks under dynamic but n conditions. In this section, we present the configuration of the experimental environment, the datasets, the parameter settings of models, and the evaluation metrics for performance evaluation.

4.1. Experimental Set-up

All the trials were executed using a high-performance computing cluster featuring Intel Xeon Gold 6230 CPUs and NVIDIA Tesla V100 GPUs. The software stack includes Python 3.8, TensorFlow 2.4, and Scikit-Learn 0.24. We implemented each component of the model using suitable libraries with TensorFlow being used for CNN, XGBoost for gradient boosting machines and implementations involving Deep Dyna Q Learning as well as VARMAx operations using PyTorch.

4.2. Execution Protocol

All experiments were run three times to provide consistent results, and the average performance metrics are given. 70% of the dataset was used to train the models, 15% was validated, and 15% was tested. Such thorough evaluation guarantees that the performance of the model is not only reliable but also resilient to various types of adversarial settings. Experimental Setup The experimental setup describes the details so that the results are replicable and assess how well the proposed model can work under real-world implementation situations. Synthetic and real-world adversarial examples provide supplementary information to demonstrate the model performance in controlled and uncontrolled environments, respectively. We tested the efficiency of the proposed ensemble, which included Deep dyna q learning and VARMAx operations, with three benchmark adjustments: [31], [32], and [33], through many adversarial datasets. The training results were compared with the unpublished state of the art; proving the robustness, accuracy, and efficiency of the proposed model in detecting and defending against the adversary attacks.

4.2. Datasets

For a comprehensive evaluation, the model was tested against three types of datasets:

a) **Synthetic Adversarial Dataset**: Generated using the CleverHans library, this dataset comprises 50,000 samples with 784 features each (simulating MNIST dimensions). Adversarial examples were crafted using the Fast Gradient Sign Method (FGSM) with an epsilon value of 0.3 to introduce perturbations.

b) **ImageNet Subset for Adversarial Testing**: A subset of 10,000 images from the ImageNet dataset was used, with adversarial modifications applied via the Projected Gradient Descent (PGD) method, considering an epsilon of 0.03 across 10 iterations.

c) **Real-World Email Spam Dataset**: Consisting of 5,000 emails, this dataset was augmented with adversarial spam examples generated through adversarial text transformation techniques, aiming to evade spam filters while preserving the original content's semantic integrity.

Table 1: Performance on Synthetic Adversarial Dataset

Metrics	Proposed Model	Method [31]	Method [32]	Method [33]
Accuracy (%)	95.2	89.5	92.1	91.3
Precision (%)	94.8	88.7	90.5	89.9
Recall (%)	95.1	87.9	91.7	90.2
F1-Score (%)	95.0	88.3	91.1	90.0
AUC	0.98	0.93	0.96	0.95

The proposed model outperforms all three methods on the synthetic adversarial dataset crafted using FGSM. Its superior recall and F1-Score highlight its robustness in identifying manipulated inputs, a critical factor for adversarial defense mechanisms.

Table 2: Performance on ImageNet Subset for Adversarial Testing

Metrics	Proposed Model	Method [31]	Method [32]	Method [33]
Accuracy (%)	93.7	87.2	89.8	88.6
Precision (%)	93.4	86.5	88.7	87.9
Recall (%)	93.9	86.9	89.4	88.1
F1-Score (%)	93.6	86.7	89.0	88.0
AUC	0.97	0.91	0.94	0.93

On the ImageNet subset with PGD adversarial examples, the proposed model again demonstrates significant improvements over the existing methods. Its performance underscores the effectiveness of integrating complex machine learning algorithms to counteract sophisticated image-based attacks.

Table 3: Performance on Real-World Email Spam Dataset

Metrics	Proposed Model	Method [31]	Method [32]	Method [33]
Accuracy (%)	94.5	90.1	92.6	91.8
Precision (%)	94.2	89.4	91.9	90.7
Recall (%)	94.8	89.8	93.2	92.0
F1-Score (%)	94.5	89.6	92.5	91.3
AUC	0.96	0.92	0.95	0.94

The proposed model maintains superior performance metrics on the adversarially enhanced email spam dataset. Its higher precision and recall are indicative of its capacity to efficiently differentiate between legitimate and spam emails under adversarial conditions.

Table 4: Aggregate Precision Across Datasets

Dataset	Proposed Model	Method [31]	Method [32]	Method [33]
Synthetic Adversarial Dataset	94.8	88.7	90.5	89.9
ImageNet Subset	93.4	86.5	88.7	87.9
Real-World Email Spam Dataset	94.2	89.4	91.9	90.7

This table consolidates the precision metric across all datasets, showcasing the consistently higher performance of the proposed model. The data reiterate the model's robustness and reliability in diverse adversarial contexts.

Table 5: Aggregate Recall Across Datasets

Dataset	Proposed Model	Method [31]	Method [32]	Method [33]
Synthetic Adversarial Dataset	95.1	87.9	91.7	90.2
ImageNet Subset	93.9	86.9	89.4	88.1
Real-World Email Spam Dataset	94.8	89.8	93.2	92.0

Similar to precision, the recall rates across all datasets are significantly higher for the proposed model. This demonstrates its ability to identify a greater proportion of positive instances correctly, which is vital for systems that must minimize the risk of overlooking adversarial attacks.

Table 6: System Efficiency and Response Times

Dataset	Proposed Model (s)	Method [31] (s)	Method [32] (s)	Method [33] (s)
Synthetic Adversarial Dataset	1.2	2.0	1.8	1.6
ImageNet Subset	1.5	2.5	2.1	1.9
Real-World Email Spam Dataset	1.1	1.8	1.6	1.4

Also, the proposed model has a better response time compared to all the tested dataset. This indicates the model's ability to consume and categorize information in milliseconds, an essential feature in real-time adversarial detection instances where timely action is of utmost importance. As indicated in Tables 2 to 7, the proposed model employing Deep Dyna Q Learning in conjunction with VARMAx operations significantly outperforms the existing prediction mechanisms. Our objective is to demonstrate that across different deceptive environments (visual attacks and textual attacks) this model outperforms the benchmark methods [31], [32], and [33], yielding greater accuracy, precision, and recall while providing efficient execution.

It simply outperformed as it learns in detail throughout dynamic exploration of datsets on sequential active learning using Query by Committee, Uncertainty Sampling, Deep Dyna Q Learning, VARMAx models and so many more predictive algorithms. In addition, the introduction of GridCAM++ helps to improve the interpretability of the model to a large extent, enabling better understanding and adjustment of the model according to the identified crucial features. Then we describe an example use case of the proposed model, so that readers can have a better delination of the whole process.

4.3. Case Utilization for Study

To clarify how the model proposed for this study will work, and how can the advanced machine learning techniques be useful to improve the overall accuracy of the ensemble model, an example with all value differences, and outputs of the model have been presented in this section. These phases consist of Ensemble Learning, Query by committee, Deep Dyna Q Learning (DDQ), GridCAM++ and VARMAx operations.

Table 7: Output of Ensemble Learning [34]

Feature Set	Logistic Regression (Probability)	Random Forest (Probability)	SVM (Probability)	CNN (Probability)	XGBoost (Probability)	Final Decision (Ensemble)
(0.1, 0.2, 0.05, 0.3, 0.1)	0.82	0.75	0.80	0.78	0.85	0.80
(0.05, 0.1, 0.2, 0.05, 0.2)	0.45	0.55	0.50	0.48	0.51	0.50
(0.2, 0.1, 0.05, 0.3, 0.15)	0.88	0.85	0.90	0.86	0.89	0.88
(0.1, 0.05, 0.2, 0.1, 0.25)	0.33	0.40	0.35	0.38	0.37	0.37

Each model in the ensemble provides a probability that the input feature set belongs to a particular class, with the ensemble's final decision calculated as the average of these probabilities. This ensemble approach leverages the strengths of each model to arrive at a more reliable and balanced decision.

Table 8: Output of Query by Committee [35]

Feature Set	Model Disagreements	Uncertainty	Selected for Labeling
(0.1, 0.2, 0.05, 0.3, 0.1)	2	High	Yes
(0.05, 0.1, 0.2, 0.05, 0.2)	1	Low	No
(0.2, 0.1, 0.05, 0.3, 0.15)	0	Very Low	No
(0.1, 0.05, 0.2, 0.1, 0.25)	3	High	Yes

The Query by Committee approach identifies feature sets where there is significant disagreement among the models. High uncertainty indicates a greater benefit from acquiring the true label, as these inputs potentially represent edge cases that are valuable for refining the model.

Table 9: Output of Deep Dyna Q Learning (DDQ) [36]

Feature Set	Initial Q-Value	Reward	Updated Q-Value	Action Taken
(0.1, 0.2, 0.05, 0.3, 0.1)	0.5	0.3	0.65	Adjust
(0.05, 0.1, 0.2, 0.05, 0.2)	0.3	0.2	0.4	No Change
(0.2, 0.1, 0.05, 0.3, 0.15)	0.7	0.1	0.75	Adjust
(0.1, 0.05, 0.2, 0.1, 0.25)	0.4	0.4	0.65	Adjust

DDQ updates the Q-values based on the received rewards, which reflect the model's performance improvements or degradation following an action. Actions such as parameter adjustments are taken to optimize future rewards, enhancing the model's prediction accuracy.

Table 10: Output of GridCAM++ [37]

Feature Set	CNN Output	Importance Map	Explanation Insight
(0.1, 0.2, 0.05, 0.3, 0.1)	0.78	(0.1, 0.4, 0.0, 0.5, 0.0)	Focus on features 2 & 4
(0.05, 0.1, 0.2, 0.05, 0.2)	0.48	(0.0, 0.1, 0.2, 0.0, 0.7)	Primary focus on feature 5
(0.2, 0.1, 0.05, 0.3, 0.15)	0.86	(0.3, 0.0, 0.0, 0.6, 0.1)	High impact of features 1 & 4
(0.1, 0.05, 0.2, 0.1, 0.25)	0.38	(0.2, 0.1, 0.3, 0.1, 0.3)	Distributed influence across features

GridCAM++ provides a class activation map indicating the importance of each feature in the CNN's decision-making process. This visualization tool is crucial for understanding which features significantly influence outcomes, thus offering insights into the model's internal reasoning and assisting in further tuning to improve interpretability and performance.

Table 11: Output of VARMAx [38]

Time Step	Observed yt	Predicted y^t|t−1	VARMAx Residuals	Adjusted Prediction
1	0.90	0.85	0.05	0.88
2	0.60	0.58	0.02	0.61
3	0.75	0.70	0.05	0.73
4	0.40	0.42	-0.02	0.41

VARMAx, applied to forecast our data, is in fact a univariate AR model, adding endogeneity when integrated model predicts endogenous variables (system performance metrics) and exogenous variables (time, external influences), with shared adaptations through various approaches. As can be seen in the output here, by calculating the predicted values y^t|t−1 and adjusting them with the residuals we are able to produce better-conformed predictions at the subsequent time steps, thus indicating the model's predictive capabilities in mitigating potential adversarial scenarios more effectively.

These tables above show a detailed understanding of how each element of the proposed model is operating. Ensemble learning uses multiple models to achieve better prediction accuracy and reliability. The method that achieves that goal is Query by Committee, which can probe which uncertain predictions would benefit most from more training data. Deep Dyna Q Learning: dynamic optimization of decision-making using reinforcement. GridCAM++: An Explainable Approach for Convolutional Neural Networks Finally, the VARMAx operations improve forecasting accuracy by considering both internal and external variations, which are necessary for a solid adversarial defense mechanism. These set of data should serve as a holistic approach for protection against several adversarial attacks for machine learning systems.

5. Discussion

They bring us one step closer to secure machine-learning systems that can withstand adversarial onslaughts. Adversarial techniques will continue to evolve, and existing defenses will certainly be stress-tested, mandating continuous innovation, as the one presented above. The proposed ensemble model, which combines Deep Dyna Q Learning and VARMAx is experimentally evaluated, and its performance is found to be exceptional on several adversarial datasets and samples. The proposed model is 95. 024,0.020% on the ImageNet Subset, and 2.829% on Real-World Email Spam Dataset respectively. 8% on those datasets respectively. The precision metrics were similarly remarkable with the proposed model attaining 94. For example, the proposed model needs only 1.

References

[1] A. S. Albahri et al., “A systematic review of trustworthy artificial intelligence applications in natural disasters,” Comput. Electr. Eng., vol. 118, p. 109409, 2024, doi: 10.1016/j.compeleceng.2024.109409.

[2] M. A. Habeeb, “Hate Speech Detection using Deep Learning Master thesis,” University of Miskolc, 2021. [Online]. Available: http://midra.uni-miskolc.hu/document/40792/38399.pdf

[3] M. E. Alqaysi, A. S. Albahri, and R. A. Hamid, “Evaluation and benchmarking of hybrid machine learning models for autism spectrum disorder diagnosis using a 2-tuple linguistic neutrosophic fuzzy sets-based decision-making model,” Neural Comput. Appl., 2024, doi: 10.1007/s00521-024-09905-6.

[4] A. H. Alamoodi, M. S. Al-Samarraay, O. S. Albahri, M. Deveci, A. S. Albahri, and S. Yussof, “Evaluation of energy economic optimization models using multi-criteria decision-making approach,” Expert Syst. Appl., vol. 255, p. 124842, 2024, doi: 10.1016/j.eswa.2024.124842.

[5] A. S. Albahri et al., “Prioritizing complex health levels beyond autism triage using fuzzy multi-criteria decisionmaking,” Complex Intell. Syst., 2024, doi: 10.1007/s40747-024-01432-0.

[6] A. Guesmi, M. A. Hanif, B. Ouni and M. Shafique, "Physical Adversarial Attacks for Camera-Based Smart Systems: Current Trends, Categorization, Applications, Research Challenges, and Future Outlook," in IEEE Access, vol. 11, pp. 109617-109668, 2023, doi: 10.1109/ACCESS.2023.3321118.

[7] L. Chen, Q. -X. Zhu and Y. -L. He, "Adversarial Attacks for Neural Network-Based Industrial Soft Sensors: Mirror Output Attack and Translation Mirror Output Attack," in IEEE Transactions on Industrial Informatics, vol. 20, no. 2, pp. 2378-2386, Feb. 2024, doi: 10.1109/TII.2023.3291717.

[8] R. Huang and Y. Li, "Adversarial Attack Mitigation Strategy for Machine Learning-Based Network Attack Detection Model in Power System," in IEEE Transactions on Smart Grid, vol. 14, no. 3, pp. 2367-2376, May 2023, doi: 10.1109/TSG.2022.3217060.

[9] W. Feng, N. Xu, T. Zhang, B. Wu and Y. Zhang, "Robust and Generalized Physical Adversarial Attacks via Meta-GAN," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1112-1125, 2024, doi: 10.1109/TIFS.2023.3288426.

[10] S. He et al., "Type-I Generative Adversarial Attack," in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 3, pp. 2593-2606, 1 May-June 2023, doi: 10.1109/TDSC.2022.3186918.

[11] S. Zhao, W. Wang, Z. Du, J. Chen and Z. Duan, "A Black-Box Adversarial Attack Method via Nesterov Accelerated Gradient and Rewiring Towards Attacking Graph Neural Networks," in IEEE Transactions on Big Data, vol. 9, no. 6, pp. 1586-1597, Dec. 2023, doi: 10.1109/TBDATA.2023.3296936.

[12] F. He, Y. Chen, R. Chen and W. Nie, "Point Cloud Adversarial Perturbation Generation for Adversarial Attacks," in IEEE Access, vol. 11, pp. 2767-2774, 2023, doi: 10.1109/ACCESS.2023.3234313.

[13] Y. Wang et al., "Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey," in IEEE Communications Surveys & Tutorials, vol. 25, no. 4, pp. 2245-2298, Fourthquarter 2023, doi: 10.1109/COMST.2023.3319492.

[14] S. M. K. A. Kazmi, N. Aafaq, M. A. Khan, M. Khalil and A. Saleem, "From Pixel to Peril: Investigating Adversarial Attacks on Aerial Imagery Through Comprehensive Review and Prospective Trajectories," in IEEE Access, vol. 11, pp. 81256-81278, 2023, doi: 10.1109/ACCESS.2023.3299878.

[15] Y. Shi, Y. Han, Q. Hu, Y. Yang and Q. Tian, "Query-Efficient Black-Box Adversarial Attack With Customized Iteration and Sampling," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 2226-2245, 1 Feb. 2023, doi: 10.1109/TPAMI.2022.3169802.

[16] C. Shi, M. Zhang, Z. Lv, Q. Miao and C. -M. Pun, "Universal Object-Level Adversarial Attack in Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-14, 2023, Art no. 5532714, doi: 10.1109/TGRS.2023.3336734.

[17] W. Jiang, H. Li, G. Xu, T. Zhang and R. Lu, "Physical Black-Box Adversarial Attacks Through Transformations," in IEEE Transactions on Big Data, vol. 9, no. 3, pp. 964-974, 1 June 2023, doi: 10.1109/TBDATA.2022.3227318.

[18] K. Mo, W. Tang, J. Li and X. Yuan, "Attacking Deep Reinforcement Learning With Decoupled Adversarial Policy," in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 1, pp. 758-768, 1 Jan.-Feb. 2023, doi: 10.1109/TDSC.2022.3143566.

[19] L. Sun et al., "Adversarial Attack and Defense on Graph Data: A Survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 8, pp. 7693-7711, 1 Aug. 2023, doi: 10.1109/TKDE.2022.3201243.

[20] L. Xu and J. Zhai, "DCVAE-adv: A Universal Adversarial Example Generation Method for White and Black Box Attacks," in Tsinghua Science and Technology, vol. 29, no. 2, pp. 430-446, April 2024, doi: 10.26599/TST.2023.9010004.

[21] L. Nguyen Vu, T. -P. Doan, M. Bui, K. Hong and S. Jung, "On the Defense of Spoofing Countermeasures Against Adversarial Attacks," in IEEE Access, vol. 11, pp. 94563-94574, 2023, doi: 10.1109/ACCESS.2023.3310809.

[22] C. Wan, F. Huang and X. Zhao, "Average Gradient-Based Adversarial Attack," in IEEE Transactions on Multimedia, vol. 25, pp. 9572-9585, 2023, doi: 10.1109/TMM.2023.3255742.
[23] H. Teryak, A. Albaseer, M. Abdallah, S. Al-Kuwari and M. Qaraqe, "Double-Edged Defense: Thwarting Cyber Attacks and Adversarial Machine Learning in IEC 60870-5-104 Smart Grids," in IEEE Open Journal of the Industrial Electronics Society, vol. 4, pp. 629-642, 2023, doi: 10.1109/OJIES.2023.3336234.
[24] C. Qin et al., "Feature Fusion Based Adversarial Example Detection Against Second-Round Adversarial Attacks," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 5, pp. 1029-1040, Oct. 2023, doi: 10.1109/TAI.2022.3190816.
[25] R. Gipiškis, D. Chiaro, M. Preziosi, E. Prezioso and F. Piccialli, "The Impact of Adversarial Attacks on Interpretable Semantic Segmentation in Cyber–Physical Systems," in IEEE Systems Journal, vol. 17, no. 4, pp. 5327-5334, Dec. 2023, doi: 10.1109/JSYST.2023.3281079.
[26] T. Chen and Z. Ma, "Toward Robust Neural Image Compression: Adversarial Attack and Model Finetuning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 12, pp. 7842-7856, Dec. 2023, doi: 10.1109/TCSVT.2023.3276442.
[27] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang and Q. Yu, "A Survey of Adversarial Attack and Defense Methods for Malware Classification in Cyber Security," in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 467-496, Firstquarter 2023, doi: 10.1109/COMST.2022.3225137.
[28] J. Pi et al., "Adv-Eye: A Transfer-Based Natural Eye Makeup Attack on Face Recognition," in IEEE Access, vol. 11, pp. 89369-89382, 2023, doi: 10.1109/ACCESS.2023.3307132.
[29] R. Li, H. Liao, J. An, C. Yuen and L. Gan, "IntraClass Universal Adversarial Attacks on Deep Learning-Based Modulation Classifiers," in IEEE Communications Letters, vol. 27, no. 5, pp. 1297-1301, May 2023, doi: 10.1109/LCOMM.2023.3261423.
[30] X. Yuan, Z. Zhang, X. Wang and L. Wu, "Semantic-Aware Adversarial Training for Reliable Deep Hashing Retrieval," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 4681-4694, 2023, doi: 10.1109/TIFS.2023.3297791.
[31] E. Mariotti, “A holistic perspective on designing and evaluating explainable AI models: from white-box additive models to post-hoc explanations for black-box models.” 2024.
[32] S. Ai, A. S. Voundi Koe, and T. Huang, “Adversarial perturbation in remote sensing image recognition,” Appl. Soft Comput., vol. 105, p. 107252, 2021, doi: 10.1016/j.asoc.2021.107252.
[33] C. Zhang, X. Costa-Perez, and P. Patras, “Adversarial Attacks Against Deep Learning-Based Network Intrusion Detection Systems and Defense Mechanisms,” IEEE/ACM Trans. Netw., vol. 30, no. 3, pp. 1294–1311, 2022, doi: 10.1109/TNET.2021.3137084.
[34] B. Wu et al., “Attacking Adversarial Attacks as A Defense,” arXiv Prepr. arXiv2106.04938, 2021, [Online]. Available: http://arxiv.org/abs/2106.04938
[35] N. Liu, M. Du, R. Guo, H. Liu, and X. Hu, “Adversarial Attacks and Defenses,” ACM SIGKDD Explor. Newsl., vol. 23, no. 1, pp. 86–99, May 2021, doi: 10.1145/3468507.3468519.
[36] L. Griffin, “Evaluating Methods for Improving DNN Robustness Against Adversarial Attacks,” no. 1. University of South Florida, pp. 1–23, 2023. [Online]. Available: https://www.proquest.com/openview/0a3e9e510f3b25b0516f4b623af4423f/1?pqorigsite=gscholar&cbl=18750&diss=y
[37] Y. L. Khaleel, M. A. Habeeb, A. S. Albahri, T. Al-Quraishi, O. S. Albahri, and A. H. Alamoodi, “Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods,” vol. 33, no. 1, 2024, doi: doi:10.1515/jisys-2024-0153.
[38] M. Macas, C. Wu, and W. Fuertes, “Adversarial examples: A survey of attacks and defenses in deep learningenabled cybersecurity systems,” Expert Syst. Appl., vol. 238, p. 122223, Mar. 2024, doi: 10.1016/j.eswa.2023.122223.