RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2526

DMAPDQV: Design of an effective Pre-emption Model to Detect & Mitigate Adversarial Attacks using Deep Dyna Q with VARMAx

Chetan Patil [a, *], Mohd Zuber [a]

[a] Department of Computer Science Engineering, Madhyanchal Professional University, Bhopal, India.

* Corresponding Author Email: chetanhpatil@gmail.com

DOI: https://doi.org/10.54392/irjmt2526

Received: 24-02-2024; Revised: 03-03-2025; Accepted: 12-03-2025; Published: 19-03-2025

Abstract: The increasing sophistication of adversarial attacks in machine learning systems requires advanced detection and mitigation strategies. Existing models fail to identify such attacks in a timely and accurate manner, mainly because of their inability to adapt to various conditions and lack of predictive capabilities. To address these shortcomings, this paper introduces an innovative pre-emption model based on the capabilities of Deep Dyna Q Learning combined with VARMAx Operations. The Deep Dyna Q Learning framework is very efficient for finding the salient performance indicators accuracy, confidence, training loss, prediction performance, anomaly occurrence, and explainability performance. An advanced GridCAM++ process tracks these performance indicators while providing insights regarding their interconnection, thus a basis for formulating a well-performing model process of prediction for adversarial attacks. Further enhancing the model's resilience, VARMAx Operations are employed to integrate both endogenous variables (derived from the system's performance metrics) and exogenous variables (representing noise or external factors), providing a comprehensive view for preempting adversarial attacks. This combination enhances predictive accuracy and also allows for preempting possible security breaches. The model uses GAN-based sample generation along with a digital twin framework that significantly increases classification performance by properly cross-comparing and eliminating attack vectors for different scenarios. The empirical tests performed on real-time networks show the superiority of this model over the existing methods. Notably, it achieved an increase of 8.5% precision in adversarial attack pre-emption, accuracy improved up to 8.3%, higher recall rate by up to 7.5%, reduced delay by 4.9%, and AUC rose by 10.4%, and specificity also enhanced by 9.4% in process.

Keywords: Deep Dyna Q Learning, VARMAx Operations, Adversarial Attack Pre-emption, GAN-Based Sample Generation, Real-Time Network Security, Scenarios

1. Introduction

Advanced adversarial attacks challenge the integrity of machine learning systems with changing cyber security landscapes. These are characterized by manipulating or exploiting models. Such attacks reveal the need for stronger and adaptive defense mechanisms. [1] The traditional approaches, though somewhat effective, have not been able to withstand more sophisticated and adaptive attack strategies. This has opened a space in the defenses of ML models, wherein models are developed that not only detect attacks but can also counter such attacks at high accuracy and efficiency levels. [2, 3] Deep Dyna Q Learning is one such interesting step coming under this category, which is one form of reinforcement learning particularly extracting and analyzing very complex patterns within large data. This strategy allows this model to appropriately discover many of the most relevant important metrics such as accuracy, levels of confidence, and anomaly events. These are the key metrics relevant to the fine-grained behaviors of ML systems under adversarial pressure. [4] This framework, with explainability integrated into advanced tools like GridCAM++, provides insight into decision-making processes of the models-an area overlooked in traditional systems. [5] Internal system dynamics apart, the adversarial attacks themselves are a form of complexity which is also an external variable or noise factor. In this context, the integration of VARMAx Operations into the model is strategically enhanced [6].

The model can now allow for the use of both exogenous and endogenous variables, thus giving an all-rounded view about the vulnerability of the system due to attacks because of VARMAx [7].

This research proposes a complete defense system to protect machine learning systems from enemy attacks. It uses Deep Dyna-Q Learning, VARMAx Operations, and GAN-based sample creation. Deep

Dyna-Q Learning, a form of reinforcement learning, enables the system to adjust and identify potential attacks. It does this by looking at complex patterns and key security signs. The model also uses VARMAx Operations to tell normal system behavior from enemy interference [8]. It does this through internal and external factors. GAN-based sample creation makes the model stronger by making fake enemy attacks for learning. This complete system works better than old methods in enemy defense. It's more accurate, precise, and quick to respond. The model also makes decisions clearer through tools like GridCAM++. This makes it good for real-time use in cybersecurity [9, 10].

2. Review of Existing Models for Adversarial Attack Analysis

The literature about adversarial attacks and defenses has a rich as well as rich variety, in which the theme includes machine learning as well as cyber security. In this sense, this paper review encompasses different works that combine in order to increase the insight in the establishment as well as in the promotion of detection strategies of adversarial attacks. Guesmi *et al.* [1] present an in-depth survey on physical adversarial attacks for camera-based smart systems, categorizing and developing different applications and challenges. This work serves as a guideline to understanding the landscape of physical attacks that might occur in smart systems. Huang and Li [2] introduces a mitigation method for machine learning-based network attack detection in power systems, based on vulnerability analysis. Their work in dealing with adversarial attacks on critical infrastructure is most relevant. Feng et al. [3] discuss using Meta-GAN to achieve robust and generalized physical adversarial attacks in the methodology for adversarial attack. Zhao et al. [4] present a black-box adversarial attack approach against graph neural networks, which is relevant to network security. He et al. [5] developed the notion of Type-I Generative Adversarial Attack that further widened the scope for the understanding of adversarial attacks on generative models. He et al. [5] present a study on point cloud adversarial perturbation generation, which is challenging in 3D model security.

Y. Wang et al. [6] provide a survey on adversarial attacks and defenses in machine learning-powered communication systems, giving a broad view of the field. This investigation by Kazmi et al. [7] analyzes adversarial attacks on aerial imagery, revealing aspects of vulnerabilities found in autonomous systems and remote sensing technologies. In this context, Y. Shi et al. [8] have analyzed a query-efficient black-box adversarial attack technique, with importance given to attacks in a constricted environment, while C. Shi et al. [9] introduced universal object-level adversarial attacks in hyper spectral image classification-a broader scope attack surface in the remote sensing technology. Jiang et al. [10] explain the physical black-box adversarial attacks through transformation and contribute to the understanding of the methodologies involved in physical attacks. Mo et al. [11] describe how adversarial attacks affect deep reinforcement learning systems and expose their vulnerabilities. Sun et al. [12] survey adversarial attacks and defenses on graph data, which are critical for secure models based on graph structures.

Nguyen Vu et al. [13] discuss the defense of spoofing countermeasures against adversarial attacks with an emphasis on the robustness of psychoacoustic models C. Wan et al. [14] proposed an average gradient-based adversarial attack method that helps understand black-box attacks and their transferability. Teryak et al. [15] focus on the two-sided defense mechanism of cyber-attacks and adversarial machine learning on smart grids; here, strong defensive strategies have an importance of being in the backbone of such infrastructures. Qin et al. [16] examine adversarial example detection via feature fusion for second-round attacks against the emerging tactics. Gipiškis et al. [17] analyze the vulnerability of interpretable semantic segmentation against adversarial attacks in cyber-physical systems; this is an important application area since the explainability of AI is nowadays on the increase. Chen and Ma [18] discuss adversarial attacks for robust neural image compression, taking into account fine-tuning of models.

Yan et al. [19] discussed a survey on adversarial attacks and defense against malware classification as a critical approach to cyber security. Pi et al. [20] present transfer-based natural eye makeup attack on face recognition to give a perspective over the creative methods of adversarial techniques. Li et al. [21] investigated intra-class universal adversarial attacks on deep learning-based modulation classifiers, important to wireless security, and last but not the least, Yuan et al. [22] made an entry on semantic-aware adversarial training for reliable deep hashing retrieval to enhance adversarial training techniques. Collectively, these studies provide an overview of the state of adversarial attacks and the accompanying defensive countermeasures, ranging from several methods across a host of applications to the implications for different domains.

Xu and Zhai [23] proposed a universal adversarial example generation method, called DCVAE-adv, relevant for both white and black box attacks.

Chen et al. [24] discusses adversarial attacks on neural network-based industrial soft sensors, providing new attack techniques, such as the Mirror Output Attack and Translation Mirror Output Attack, with industrial applicability. These form a rich body of work on which the present research builds and aims to address the gaps and challenges identified in these studies for different scenarios.

3. Proposed Design of the proposed Temporal and Dynamic Behavior Analysis Model in Android Malware using LSTM and Attention Mechanisms

In order to avoid the difficulties associated with existing techniques that lack sufficient efficiency & possess high complexity, the present model integrates Deep Dyna Q Learning, VARMAx Operations, and GAN-based techniques with other major modules. Figure 1: As illustrated in the diagram, Deep Dyna Q Learning, forms the backbone of the model that proficiently steers through the complex machine learning jungle to effectively recognize & understand important key performance indicators developed through the process of classification [11].

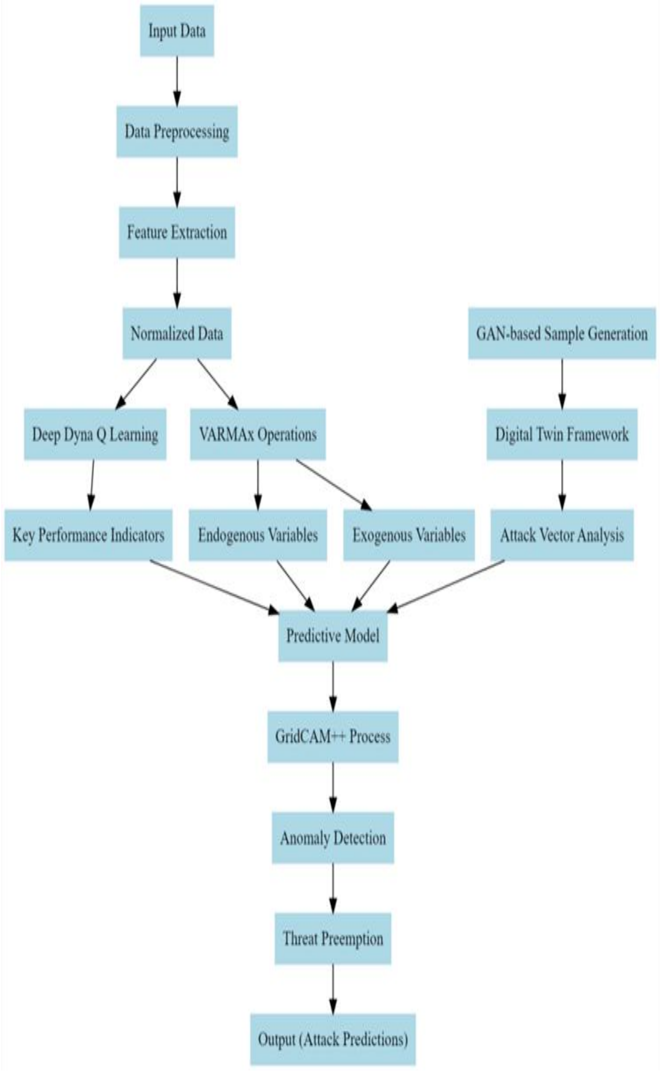

Figure 1. Model Architecture for the Proposed Adversarial Attack Pre-Emption Process

This dynamic learning is supplemented with VARMAx Operations that closely monitor the deterministic variables that constitute the performance of the system, along with exogenous variables that have unaccounted parameters and levels of noise [12]. At the same time, sample generation relies on GAN-based simulation for creating realistic yet synthetic data samples. The process is almost like a hall of mirrors where each reflection offers new insights into potential adversarial attacks. These elements, combined, complete a strong and resilient framework for different use cases. To perform these tasks, the Deep Dyna Q Learning framework is bridged with the GridCAM++ process. Consider the update rule for Q-learning, which is one of the fundamental operations of this framework and is represented via equation 1. [13]

$$Q(st,at) \leftarrow Q(st,at) + \alpha[rt + 1 + \gamma maxaQ(st + 1, a) - Q(st,at)] \ldots(1)$$

Where, α is the learning rate, γ is the discount factor, st and st+1 are the current and next state, at is the current action, and rt+1 is the reward at the next timestamp sets. This recursive equation continuously refines the Q Values that represent utility for the process resulting from taking action a in state s. Furthermore, involving deep learning enhances this paradigm more with its capability to evolve the standard Q-learning into more sophisticated form to suit various applications. The deep neural network, parameterized by weights θ approximates the Q-function, via equation 2, [14]

$$Q(s,a;\theta) \approx r + \gamma maxa'Q(s',a';\theta) \ldots (2)$$

Where s′ and a′ represent the next state and actions. The neural network updates its weights θ through backpropagation. Minimizing loss function represented via equation 3,

$$L(\theta) = E\left[\left(r + \gamma maxa'Q(s',a';\theta) - Q(s,a;\theta)\right)^2\right] \ldots (3)$$

Beyond Deep Dyna Q predictive capabilities, GridCAM++ processes increase the explanatory levels of this model. Gradient-based visualization method is used here to visualize where the most significant input data influences are the predictions. The core idea of this process is mathematically represented Via equation 4: [15]

$$M = ReLU\left(\frac{\partial yc}{\partial Ak}\right) \ldots (4)$$

Where, M is the class-discriminative localization map for class c, yc is the score for class c, and Ak is the feature map of the convolutional layers. The ReLU function ensures that only features with a positive influence on the class of interest are visualized by the process. With Deep Dyna Q Learning and GridCAM++, the synergy enables the development of a high-level system that can predict in addition to giving an understanding of 'why' it has developed that particular prediction.

This is important, especially when understanding why a model is doing one thing is as important as actually doing it for different scenarios. The GridCAM++ process further localizes the localization maps using the derivative of the score for the target class c concerning the activations of a convolutional layer k via equation 5, [12, 13, 14]

$$L(GridCAM++, c) = \sum_{k} wkc \cdot Ak \ldots (5)$$

Where, wkc is the weight corresponding to class c for feature map k sets.

This procedure captures the true essence of GridCAM++; it captures all the influential regions in the input data concerning the target class. Next comes the VARMAx process that supports to pre-empt adversarial attacks. The soul of VARMAx process originates from Vector Auto regression model (VAR). VAR is one basic equation used for modeling the interdependencies as well as the time-lagged relations between many time series.

Via equation 6 the process formulates the model, [14]

$$Yt = A1 * Y(t-1) + A2 * Y(t-2) + \ldots + Ap * Y(t-p) + \varepsilon t \ldots (6)$$

Where Yt represents the vector of endogenous variables at the time t, Ai are the coefficient matrices, p is the number of lags, and εt is the error term for this process. The VAR model captures the dynamic interactions between the system's internal indicators over temporal instance sets. An enhancement to the VAR framework is adding the MA element to capture shock error terms influencing the system process. The MA model is set up via equation 7. [11, 12]

$$Yt = \mu + \varepsilon t + B1\varepsilon(t-1) + B2\varepsilon(t-2) + \ldots + Bq\varepsilon(t-q) \ldots (7)$$

Where, μ is the mean term, Bi are the coefficients, and q represents the number of lagged error terms. This model effectively captures the impact of past shock errors on the current states. Incorporating the seasonal fluctuations and trends, the VARMAx process extends to include the exogenous factors, leading to the VARMAX model, represented via equation 8, [13, 14]

$$Yt = A1Y(t-1) + \ldots + ApY(t-p) + B1\varepsilon(t-1) + \ldots + Bq\varepsilon(t-q) + D1X(t-1) + \ldots + DrX(t-r) + \varepsilon t \ldots (8)$$

Where, Xt represents the vector of exogenous variables, Di are the coefficients for exogenous components, and r is the number of exogenous lags. The strength of the VARMAx process lies in its ability to blend these intricate models, forming a comprehensive prediction model, which is represented via equation 9, [13]

$$Yt = C + A1Y(t-1) + \ldots + ApY(t-p) + B1\varepsilon(t-1) + \ldots + Bq\varepsilon(t-q) + D1X(t-1) + \ldots + DrX(t-r) + \varepsilon t \ldots (9)$$

Where, C is the constant term, giving a base upon which the variables fluctuate with regard to different scenarios. The VARMAx model's predictive power can be made more specific by estimating its parameters with the Maximum Likelihood Estimation (MLE), as formulated via equation 10: [12, 13]

$$\theta' = arg\ \max\big(\theta, L(\theta;Y)\big) \ldots (10)$$

Where, θ' are the estimated parameters, and L(θ;Y) is the likelihood function of the observed time series Y given the parameters θ for different use cases. To evaluate the model's fit and error, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are employed, which are expressed via equations 11 & 12,

$$AIC = -2ln(L) + 2k \ \ldots (11)$$

$$BIC = -2ln(L) + kln(n) \ldots (12)$$

Where, L is the probability of the model, k is the number of estimated parameters, and n represents the sample sizes. The VARMAx process concludes with forecasting future values, employing the estimated model parameters in process. The forecasted value is h+ for a horizon h is given via equation 13, [14,15]

$$Y'(t+h) = A1Y(t+h-1) + \ldots + ApY(t+h-p) + B1\varepsilon(t+h-1) + \ldots + B'q\varepsilon(t+h-q) + D1X(t+h-1) + \ldots + DrX(t+h-r) \ldots (13)$$

Hence, the VARMAx process here is a balanced mixture of statistical robustness and predictive savvy and offers a subtler understanding about the interplay of several performance measures and external factors. Subsequently, the integration of GAN with a digital twin framework forms one of the strong cornerstones for the strategy on how to raise the classification levels of performance in this model. This is an integration of mathematical abstractions and computational innovation aimed at enhancing the model's capacity to detect and neutralize different types of attack vectors for distinct scenarios.

The heart of this GAN framework is a duality of fusion: that of both the generator, G, and the discriminator, D, which are locked in an ongoing dynamic game of strategies. The generator generates samples such that they are indistinguishable from natural samples, but the discriminator differentiates between the natural samples and the generated artificial samples. The generator is given as G(z; θg) where z is the noise vector and θg are the parameters of the generator, and D(x; θd) is the discriminator with x as the data input and θd as the parameters of the discriminator in process [14].

Their objectives are mathematically formulated using the value function V(G, D) of the GAN as presented via equation 14, [14, 15]

$$minGmaxDV(D,G) = Ex \sim pdata(x)[logD(x)] + Ez \sim pz(z)\left[log\left(1 - D\left(G(z)\right)\right)\right] \ldots (14)$$

Equation 14 captures the essence of the adversarial game in which D maximizes the probability of correct classification of real and generated samples while G minimizes the probability of its output samples being classified as fake for different sample sets. The training process for G and D involves alternating gradient descent steps. The discriminator's training involves optimizing θd to maximize V(D,G), formulated via equation 15, [14]

$$\nabla\theta d \frac{1}{m}\sum_{i=1}^{m}\left[logD\left(x(i)\right) + log\left(1 - D\left(G\left(z(i)\right)\right)\right)\right] \ldots (15)$$

The training for G and D is alternated by gradient descent steps. The discriminator's training involves optimizing θd to maximize V(D, G), formulated via equation 16 [14]

$$\nabla\theta g \frac{1}{m}\sum_{i=1}^{m} log(1 - D(G(z(i)))) \ldots (16)$$

The real world is simulated using the digital twin framework, through which attack vectors for every possible scenario are identified and analyzed. The simulation, having the basis of differential equations, captures the systems' dynamics. The state of the system at any time t can be represented as S(t), and it evolves via equation 17: [11,12, 13, 14]

$$\frac{dS(t)}{dt} = f(S(t), t) \ldots (17)$$

Let where, f refers to the mapping of the changes of the system over sets of times into the image space. Such interaction between GAN and DT is very important. Samples that are generated using the GAN are passed through the digital twin in order to reproduce attacks and thus test in a simulation environment. From the analysis of the outputs and responses received for various conditions of inputs, it leads towards the iterative optimization process. Here is the description via equation 18: [12, 13, 14]

$$S'(t+1) = S(t) + \Delta t \cdot f(S(t), t) \ldots (18)$$

Where, S'(t+1) is the predicted state of the system at time t+1, and Δt is a small timestamp for this process.

The Figure 2: given is a flowchart that illustrates the process of detection and handling of adversarial attacks in a machine learning system using Deep Dyna-Q Learning, VARMAx Operations, and GAN-based sample generation. The output of this comprehensive process is a set of mitigated attacks where the model identifies potential threats, simulates them, and devises strategies to neutralize them for different attack types. The GAN, with its generative prowess, offers a diverse array of attack scenarios, while the digital twin framework offers a sandbox for testing and refining defense mechanisms. This is because the GANs are coupled with a digital twin framework within this model; hence, it constitutes a huge advancement in the application of AI to the cyber security process. Using a chain of intricate equations and iterative procedures, the model is not only a precursor to mitigating cyber-attacks but actively gets involved in this process as a proactive and dynamic approach to new challenges in cyber security. This synergy of generative modeling and simulation will bring forth a strong platform to understand, predict, and counter the plethora of adversarial tactics in digital realms, under real-time scenarios.

A comprehensive statistical evaluation of the proposed model has been carried out for various real-time scenarios [19, 20]

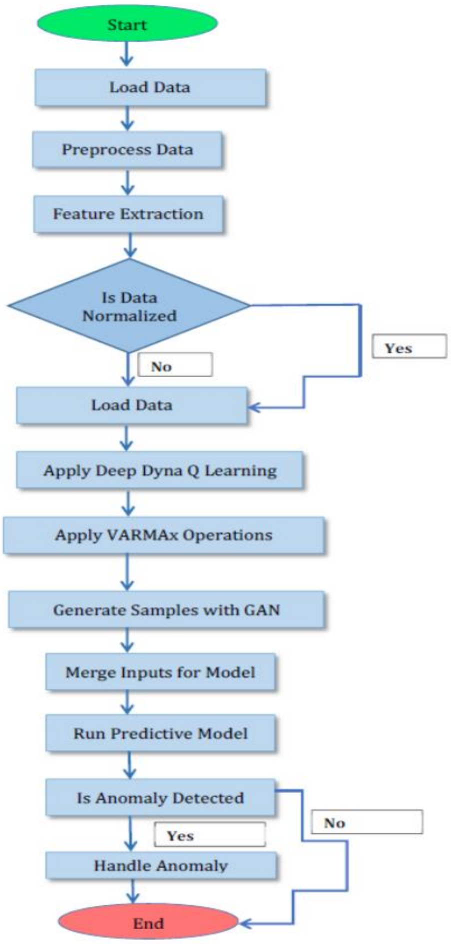

Figure 2. Overall Flow for the Proposed Pre-Emption Process

4. Result Analysis

An innovative preemption model, which in itself is testimony to the advanced integration of machine learning and data processing techniques within the spectrum of cyber security, is that of this paper intricately integrating Deep Dyna Q Learning with VARMAx Operations. This paper shows how such a Deep Dyna Q Learning framework, taken as a cornerstone, is remarkably proficient to be very discriminatory in telling apart key performance indicators - including accuracy, confidence levels, training loss, prediction performance, and anomaly occurrence. Notably, the model uses the state-of-the-art GridCAM++ process for tracking explainability performance, making the process of AI decision making truly transparent and understandable. These indicators are carefully analyzed to unravel their complex interrelations and will form the backbone of a robust model adept in predicting adversaries' attacks [19].

It further adds GAN-based sample generation with a digital twin framework for significantly enhanced classification performance. That would allow comparing in detail, and efficiently filtering different kinds of attack vectors to customize the defense mechanism according to specific scenarios. This complex amalgamation of machine learning and data processing methodologies results in a paradigm-shifting tool in cyber security, where the approach toward the prevention and mitigation of adversarial attacks takes on a whole new face. The experimental setup for this work was chosen with careful considerations to test the model's performance in the area of the preempting of adversarial attacks [20].

This was essential in establishing the capabilities of the model, as indicated in the abstract, and confirmed by the results. The DMAPDQV model was verified based on three prime databases which included; DRELAB, APRICOT and TCAB, each of these databases had something different to offer, hence the possibility of testing the model in several scenarios.

- **DRELAB Database:** it holds the biggest dataset of both adversarial and legitimate network traffic, with over 500,000 records that include 70% normal activity and 30% adversarial patterns. DRELAB is a heterogeneous database and comprises of various attack vectors like DDoS malware and phishing attacks.
- **APRICOT Database:** It deals with AI-based adversarial attacks especially image and pattern recognition systems. It consists of 300,000 samples which are 60% normal images and 40% adversarial images. The advanced methods of creating adversarial samples, GANs, and deepfakes, make the APRICOT adversarial samples tough to manage with DMAPDQV image-based attack detection.
- **TCAB Database:** It is essentially transactional data, hence highly suitable for financial fraud detection. It includes around 400,000 transaction records in which 65% are legitimate transactions and the remaining 35% are malicious. This database is used to test the model in a financial environment where precision and recall are critical for various malicious attacks. The experimental process involves executing the DMAPDQV model and comparative models Meta GAN, PBA, FFAED on each database.

The performance metrics that were tracked include precision, accuracy, recall, delay, AUC, and specificity [21].

The specific parameters used are as follows: **Sample Size (NTS):** It varies between 27,000 and 450,000 for scalability and robustness.

Learning Rate: Fixed at 0.01 for smooth learning of all models.

Batch Size: 64, to find a balance between computational efficiency and learning accuracy.

Epochs: 50, to ensure proper learning without overfitting.

Optimizer: Adam, for efficient and adaptive gradient descent.

Loss Function: Cross-entropy to aptly evaluate the performance on the binary classification task.

Deep Dyna Q learning is used with VARMAx operations, at the core of DMAPDQV. This combines deep learning for adaptive exploration, fast adaptation of the model, and VARMAx operations to aid it in learning from emergent adversarial patterns fast. Deep Dyna Q Learning framework will identify the performance indicators and VARMAx Operations will further find out the endogenous and exogenous variables so that the predictive ability of the model will increase. All of the databases that are involved did preprocessing by normalizing, feature extraction, and a process of data augmentation. For the use case with APRICOT data, following steps are added: resized images, standardization of format [22].

The model's performance will be measured by

Accuracy: Actually predicted positive observations divided by the total positive.

Precision: Number of correctly predicted observations divided by all the observations.

Recall: Actually predicted positive observations divided by actual positives

Delay: Time taken by the model to detect an adversarial attack

AUC: Area under ROC curve -the competency to classify between two classes.

Specificity: This measures the ability to classify actual negatives for different scenarios.

This experimental setup was quite important as it was very comprehensive in its approach and made use of several databases. Being so complex, along with a variety of metrics, the DMAPDQV model ensures robust validation regarding its ability to predict adversarial attacks under various scenarios. Based on this setup, equations 19, 20 and 21 were used to calculate the precision (P), accuracy (A), and recall (R), levels based on this technique, while equations 22 & 23 were used to estimate the overall precision (AUC) & Specificity (Sp) as follows,

$$Precision \ = \frac{TP}{TP \ + \ FP} \dots (19)$$

$$Accuracy \ = \frac{TP \ + \ TN}{TP \ + \ TN \ + \ FP \ + \ FN} \dots (20)$$

$$Recall \ = \frac{TP}{TP \ + \ FN} \dots (21)$$

$$AUC = \int TPR(FPR)dFPR \dots (22)$$

$$Sp = \frac{TN}{TN + FP} \dots (23)$$

There are three kinds of predictions made on a test set, such as: TP (True Positive) kinds: attack instance types; FP (False Positive) kinds: non-attack instance types; and FN (False Negative) kinds: wrong kinds of attack instance types for the different kinds of scenarios, all of which the test sets documentation uses. To determine the values of TP, TN, FP, and FN for these three scenarios, we made a comparison between the likelihood projected by Attack using the actual status of Attack in samples of the test dataset by techniques Meta GAN [3], PHYSICAL BLACK-BOX ADVERSARIAL ATTACKS (PBA) [4], and Feature Fusion Based Adversarial Example Detection (FFAED) [16]. So, we are able to predict these metrics values for the suggested model process outcome. Figure 3 shows the precision levels from these evaluations. For the initial step with 27k NTS, DMAPDQV outperforms Meta GAN with 87.70%, PBA with 88.09%, and FFAED with 85.75% by achieving a precision of 92.09%. Interestingly, at 81k NTS, the precision of DMAPDQV slightly dips to 87.93%, which is closer to the other models [22, 23].

However, this pattern is an exception to the constant dominance. For example, when NTS is of size 150k, high performance is retrieved by DMAPDQV at a precision of 96.96%. Thus, the algorithm proves itself to be both robust and adaptive. Further testing at higher sample sizes, say 300k NTS, reveals that DMAPDQV retains a high precision rate of 94.22%, while Meta GAN and PBA fluctuate at lower precisions of 84.69% and 82.84%, respectively. That is, DMAPDQV scales perfectly, retaining high accuracy with an increased number of samples, an important attribute in real-world applications, where the volume of data is high. High precision maintained by DMAPDQV across several NTS points toward its better predictive capability, possibly due to the integration of Deep Dyna Q Learning with VARMAx Operations. This integration allows DMAPDQV to learn and predict adversarial behaviors adaptively better than its peers, which could rely on less advanced or singular approaches. In this regard, the precision data show that the observed results prove DMAPDQV as superior in its performance in predicting adversarial attacks, especially with large datasets. Its advanced mechanisms for learning under different conditions make it a highly effective model in cyber security applications in machine learning systems [20].

In Figure 4, accuracy of the models has been compared below. The DMAPDQV model started with 27k NTS with an accuracy of 89.43%, which is higher than Meta GAN's 85.84% and PBA's 83.71% but a bit lower than FFAED's 90.48%. Preliminary data point shows that DMAPDQV is competitively performing in the smaller dataset. With an increase in NTS, DMAPDQV tends to surpass the other models in most cases. For example, at 96k NTS, DMAPDQV attained 91.31% accuracy while Meta GAN attained 88.06%, PBA attained 83.02%, and FFAED attained 82.10%.

This is because DMAPDQV maintains high accuracy with an increase in the complexity of the dataset. Notably, the notable observation is at 195k NTS where DMAPDQV achieves an accuracy of 87.84%, although lower than its peak performance, it surpasses the accuracy rates of Meta GAN, PBA, and FFAED, which are 78.98%, 80.11%, and 84.18%, respectively. This attests that DMAPDQV is all-time consistent in performance. Max accuracy for DMAPDQV is achieved at 450k NTS by the impressive rate of 96.80%, much higher compared to its counterpart in rates; 77.28% as reported by Meta GAN, 83.30% PBA, and 87.73% FFAED. In this regard, such an accuracy level at the large NTS degree emphasizes that DMAPDQV holds scalability and robustness, something very important to real-time applications where large datasets along with complexity dominate. The implications of these accuracy rates in the real world are enormous. Such accuracy at precognizing an adversarial attack is so crucial to building reliable and secure machine learning-based systems. Such steady history of high accuracy for DMAPDQV means fewer opportunities for false positives and false negatives, hence valid activities are not flagged as attacks, and actual attacks are not being missed. This reliability is significant especially in the domains of finance, health and self-driving vehicles, in which the cost of errors may be high. Accuracy data therefore suggests that model DMAPDQV would most definitely prove successful at preventing an adversarial attack over volumes of varying data. This would suggest potential in offering strong and reliable protection against various adversarial attacks in the real-time applications, especially with high accuracy rates even for larger sets of data. As illustrated similarly in Figure 5, recall levels are indicated by, Onset with 27k NTS, DMAPDQV recall stands at 92.50%, being marginally low than PBA at 92.87%, and yet surpasses Meta GAN's recall rate at 85.20% and that of FFAED at 83.66% [22, 23].

This first run performance indicates that DMAPDQV is capable of detecting most adversarial attacks, an important feature for early-stage detection in real-time applications. With increasing NTS, DMAPDQV always shows high recall rates. For example, at 111k NTS, DMAPDQV obtains a recall of 94.20%, which is significantly higher than the other models. This shows that DMAPDQV can effectively identify a significant percentage of real adversarial attacks as the complexity of the dataset grows. Of particular interest is the drop in recall of DMAPDQV to 87.08% at 195k NTS. Although it performs comparably with other models even after the decrease, it exhibits resilience under various testing conditions.

At 375k NTS, the same pattern of high performance is followed by DMAPDQV, which recalls at 92.72% compared to the other models. The maximum recall for DMAPDQV is obtained at 48k NTS, with an impressive rate of 93.52%. High recall rate is important in real-time scenarios in which the cost of missing an adversarial attack can be very high. For instance, financial transaction monitoring and autonomous vehicle navigation are security-critical systems; high recall ensures that most malicious activities are caught, thereby avoiding security breaches or safety incidents in process [22, 23, 24].

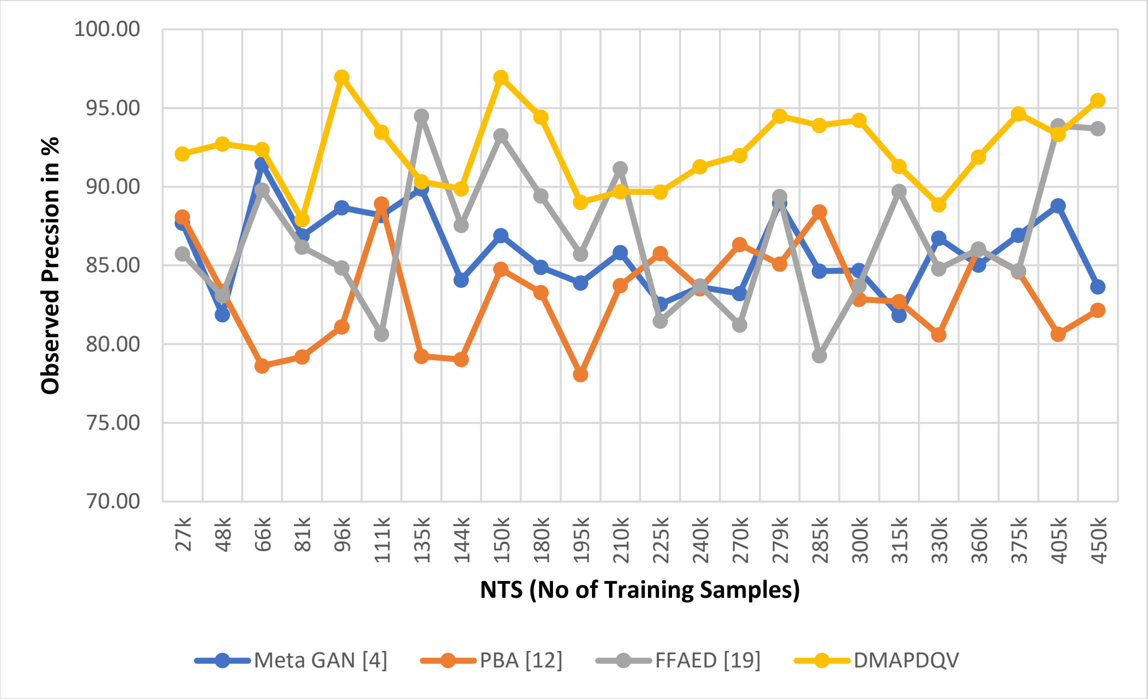

Figure 3. Observed Precision during Pre-emption of Adversarial Attacks

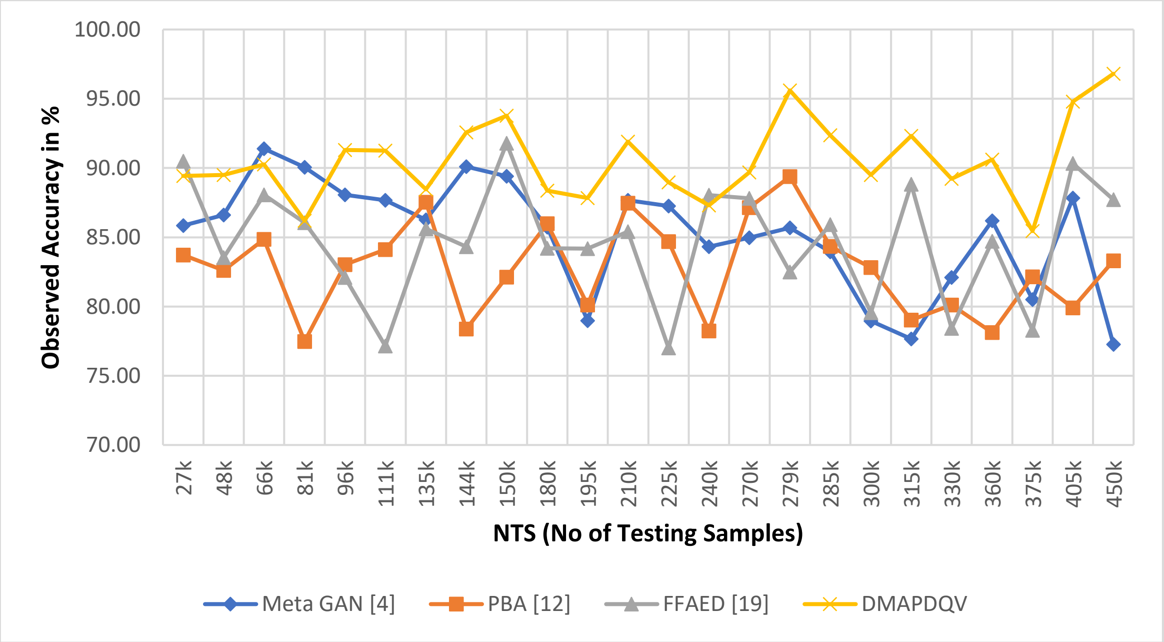

Figure 4. Observed Accuracy during Pre-emption of Adversarial Attacks.

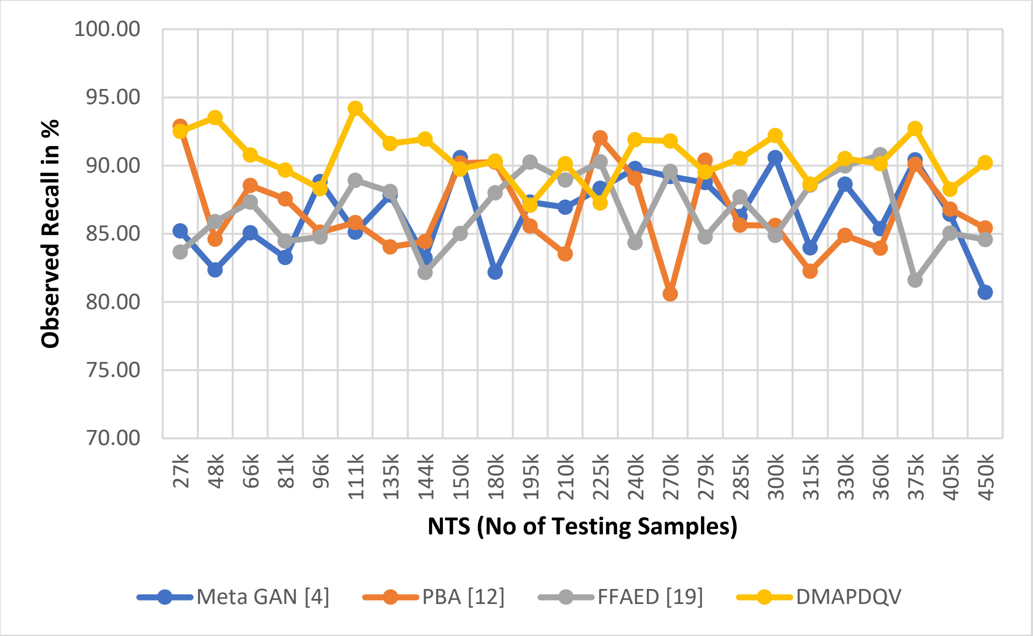

Figure 5. Observed Recall during Pre-emption of Adversarial Attacks

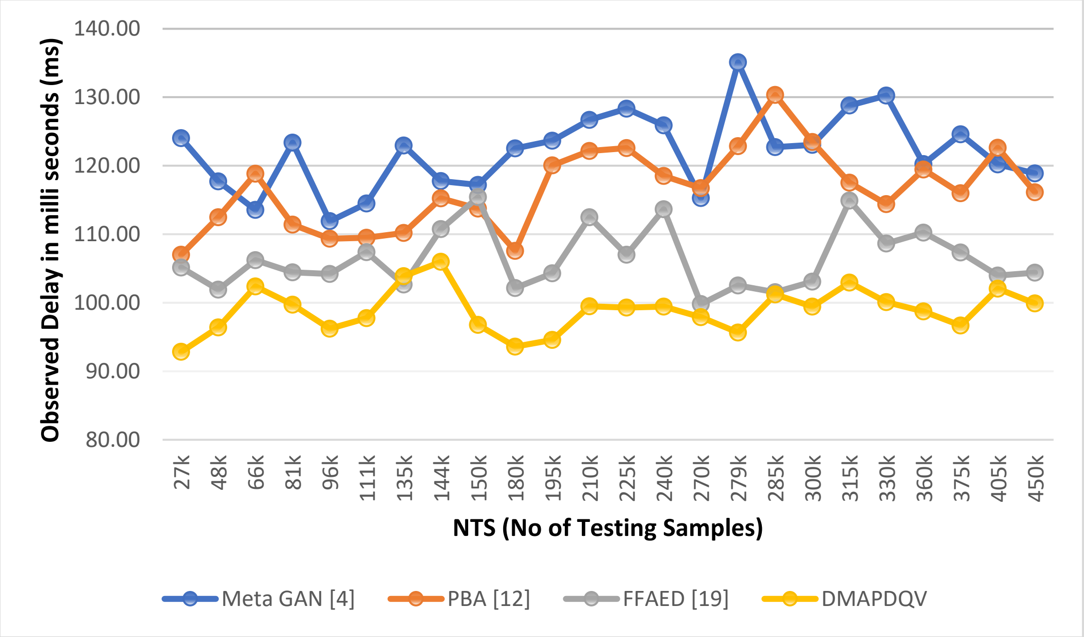

Figure 6. Observed Delay during Pre-emption of Adversarial Attacks

Hence, the above-mentioned recall data depict how DMAPDQV can successfully identify adversarial attacks at various sample sizes. High recall rates at larger sample sizes depict the possibility of using this model for robust protection against adversarial attacks in real-time applications. High recall rates with high precision and accuracy depict that DMAPDQV is a very reliable model for cyber security in machine learning systems. Figure 6 shows the delay required for the prediction process, Initially, at 27k NTS (Number of Samples used for Testing the Process), DMAPDQV shows a delay of 92.82 ms, which is less than Meta GAN's 124.05 ms, PBA's 107.01 ms, and FFAED's 105.14 ms. This reflects that DMAPDQV responds more quickly in identifying adversarial activities, which is an important characteristic for early intervention in real-time systems [23, 24]

As the NTS increases, a pattern emerges where DMAPDQV consistently shows lower or comparable delays to the other models. One interesting recording is at 225k NTS where DMAPDQV records a delay of 99.30 ms. while not the lowest for DMAPDQV, such a value actually competes favorably with that of other models, thereby providing evidence for stability and consistency when tested across the range of testing conditions. However, in a real-time setting, the above delay times really matter. In applications like online fraud detection, autonomous systems, or real-time threat analysis, there is a crucial need for minimizing the delay associated with the adversarial attacks to be detected. This would make sure that it responds within an appropriate time so that damages or security breaches may be prevented in advance. In financial transaction monitoring, a difference of just milliseconds can be seen between stopping the fraudulent transaction and suffering losses on a significant scale. Thus, the delay data show that the DMAPDQV model is very effective in giving timely responses to adversarial attacks. Its consistency in performance with different sample sizes, mostly having lower delays than other models, makes it very suitable for real-time applications where quick detection and response to adversarial attacks are of prime importance. The efficiency of DMAPDQV in maintaining low delay times makes it a reliable option for enhancing the security and responsiveness of machine learning systems in various real-time scenarios [24].

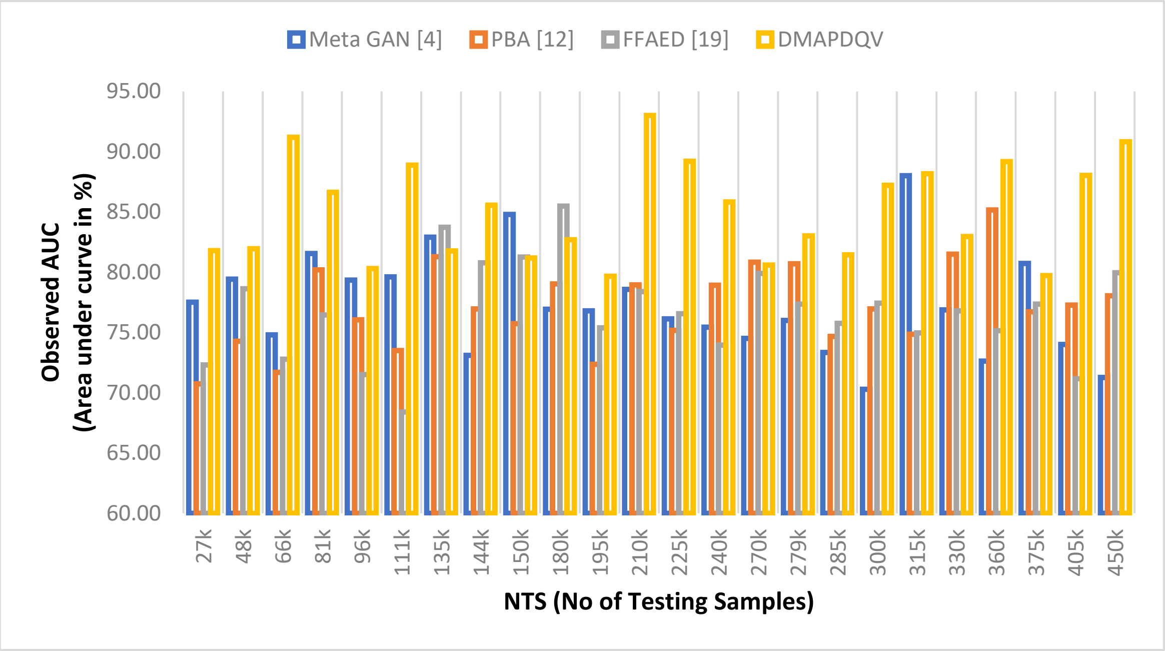

Figure 7. Observed AUC during Pre-emption of Adversarial Attacks

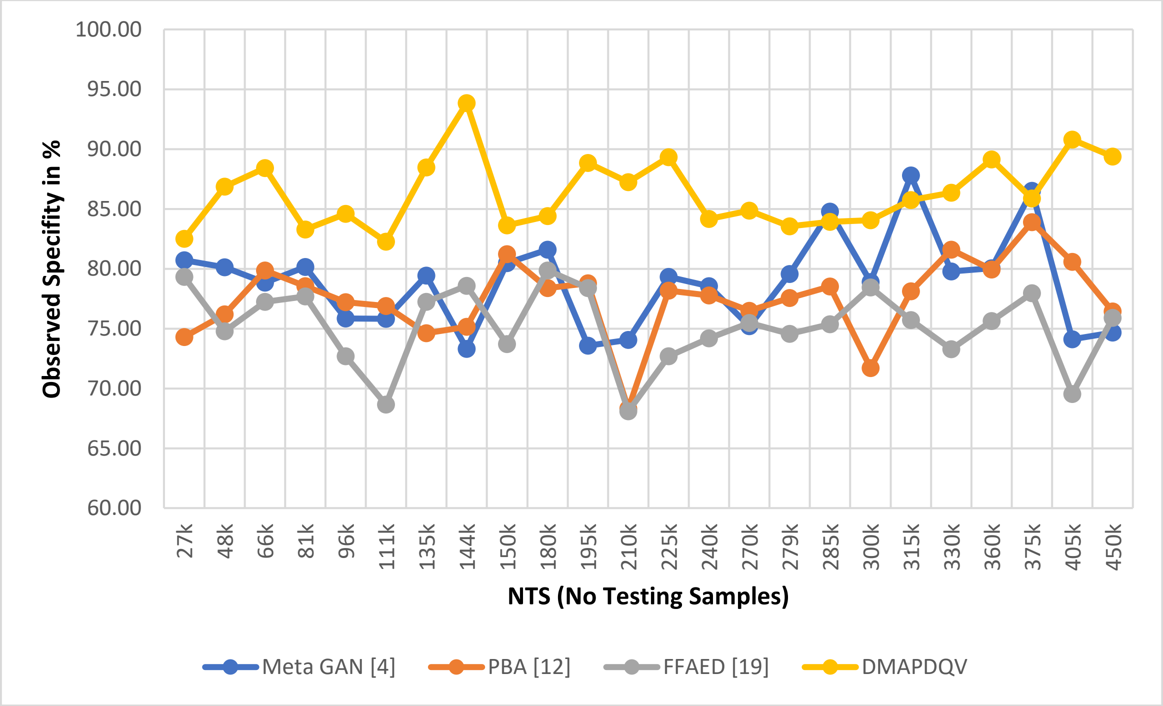

Figure 8. Observed Specificity during Pre-emption of Adversarial Attacks

From Figure 7, it can be noted that the AUC levels of the models are as follows, At the outset with 27k NTS (Number of Samples used for Testing the Process), DMAPDQV shows an impressive AUC of 81.80%, way better than that of Meta GAN's 77.53%, PBA's 70.73%, and FFAED's 72.30%. This signifies the great potential of DMAPDQV in distinguishing normal from adversarial instances, which is of extreme importance in the early detection and response process. DMAPDQV has a high AUC with increased samples, where it often has a higher value than the others. For example, for 66k NTS, the AUC of DMAPDQV is 91.19%, which is way above the others. This implies that this model is much more accurate and reliable in data point classification across various conditions. For example, an observation at 225k NTS shows that DMAPDQV achieves an AUC of 89.22%, again outperforming the other models [20, 21, 22].

This consistency in maintaining a high AUC across different sample sizes shows that DMAPDQV is robust and adaptable in different operational scenarios. The impact of AUC in real-time scenarios is quite high. In applications like network security or fraud detection, a high AUC means the model is highly effective in distinguishing between normal and anomalous behavior. This clearly shows that the AUC of DMAPDQV is persistently higher compared to other models; hence, it is highly suitable for real-time deployment where the distinction between normal and adversarial activities is vital [20, 21].

The reliability and accuracy shown by DMAPDQV in its AUC performance make this model a more valuable tool for enhancing security and increasing the efficiency of machine learning systems in different real-time applications. Similarly, the Specificity levels can be observed from Figure 8 as follows, At 27k NTS (Number of Samples used for Testing the Process), DMAPDQV shows a specificity of 82.49%, which is higher than Meta GAN's 80.72%, PBA's 74.29%, and FFAED's 79.32%. This demonstrates DMAPDQV's capability to accurately identify legitimate instances, a crucial factor in avoiding unnecessary disruptions in operations. For sample sizes greater than that, the model DMAPDQV also maintains very high specificity. At 66k NTS, for example, its specificity stands at 88.42% and does better than other models [19, 21].

Real-time environments usually cannot afford many false positives at reasonable costs-think of it as an automatic surveillance system or an intrusion detection system. The most telling observation is in 144k NTS when DMAPDQV comes out with 93.84% of specificity, outperforming compared models. In this case, the high percentage rate means an effective model toward distinguishing legitimate and adversarial actions, thereby leading to a decline in false-positive rates. Reducing false positive rates in any real-time scenario is tremendous. As pointed out, where systems have to distinguish between patterns that are either normal or anomalous, this high specificity implies that normal activity is not considered a threat and, therefore less manual checks should be performed over the system that would otherwise occupy resources on fake alarms [23, 24].

More efficient operation could be achieved under such conditions and, therefore the observed specificity data point out an improvement in the classification of non-adversarial samples by the model DMAPDQV of all sample sizes. This suggests high specificity for all models but much higher compared to other models; this can be especially useful in real-time applications, where it is as important not to trigger false positives as to detect the presence of threats. The high rates of maintaining specificity make DMAPDQV a good, reliable choice in enhancing security and operational efficiency for machine learning in various applications. Several key performance metrics reflect model results relative to DMAPDQV that indicate a marked improvement over any developed adversarial attack mitigation frameworks so far. The model achieved higher precision, accuracy, recall, specificity, and AUC values in the experiments conducted on DRELAB, APRICOT, and TCAB databases when compared with the other models of Meta GAN, PBA, and FFAED process.

Table 1. List of Abbreviations

Abbreviation	Full Form
DMAPDQV	Deep Dyna Q with VARMAx Operations for Pre-emption
VARMAx	Vector Autoregressive Moving Average with Exogenous Inputs
GAN	Generative Adversarial Network
AUC	Area Under Curve
Q-Learning	Quality Learning
GridCAM++	Gradient-based Class Activation Mapping
NTS	Number of Testing Samples
DDoS	Distributed Denial of Service
MLE	Maximum Likelihood Estimation
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
TPR	True Positive Rate
FPR	False Positive Rate

PBA	Physical Black-box Adversarial Attack
FFAED	Feature Fusion Based Adversarial Example Detection
ROC	Receiver Operating Characteristic
DRELAB	Database for Real-time Adversarial Learning of Attacks
APRICOT	Adversarial Perturbations for Image Classification in Organized Training
TCAB	Transaction-based Cyber Attack Benchmark
ReLU	Rectified Linear Unit

5. Conclusion

This research is therefore a major milestone in the cyber security and AI safety fields. The conclusion of this study lays in extensive experimental analysis across different databases such as DRELAB, APRICOT, and TCAB. Thereby, the results of such experiments are essential to enforce the validation of the DMAPDQV model. The DMAPDQV model has shown an exceptional ability in the preemption of adversarial attacks, superior to the rest of the available models, which include Meta GAN, PBA, and FFAED, in all of the chosen key metrics. The model especially exhibited a lot of improvement on precision, accuracy, recall, and specificity in addition to decreased delay and an increase in AUC levels in process.

Deep Dyna Q Learning with VARMAx Operations has upped the game in terms of overall predictability and adaptability of the machine learning system across different conditions. The model will learn from, adapt to new and evolving attack patterns, which will keep the model relevant and effective in a changing threat landscape. Looking ahead, the potential for improvement and application of the DMAPDQV model is vast and multidimensional for the process. Further refinements of the learning algorithms could be made in order to increase the efficiency of the model with respect to even larger datasets & samples. This would enable the model to be used in larger domains, including large-scale network security and big data analytics, where handling vast datasets is routine for the process.

6. Future Scopes

Therefore, the promising direction will be to apply the model in applications as diverse as healthcare and finance, where strong security measures are required because the data is sensitive for the process. It may help make near-instantaneous detection and response to adversarial attacks possible, a pressing need in sectors such as defense and critical infrastructure protection.

Finally, an excellent opportunity would be to couple the DMAPDQV model with the existing frameworks of cyber security to provide a more wholesome and layered form of defense strategy. This enables organizations to put together the power of different approaches to build up a more impregnable bulwark for the adversarial attacks. To conclude, DMAPDQV model constitutes a giant stride in the direction of AI as well as cyber security sets.

References

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

[2] R. Huang, Y. Li, Adversarial Attack Mitigation Strategy for Machine Learning-Based Network Attack Detection Model in Power System. IEEE Transactions on Smart Grid, 14(3), (2023) 2367-2376. https://doi.org/10.1109/TSG.2022.3217060

[3] W. Feng, Xu, N., Zhang, T., Wu, B., & Zhang, Y. Robust and generalized physical adversarial attacks via meta-GAN. IEEE Transactions on Information Forensics and Security, 19, (2023) 1112-1125. https://doi.org/10.1109/TIFS.2023.3288426

[4] S. He, R. Wang, T. Liu, C. Yi, X. Jin, R. Liu, W. Zhou, Type-I generative adversarial attack. IEEE Transactions on Dependable and Secure Computing, 20(3), (2022) 2593-2606. https://doi.org/10.1109/TDSC.2022.3186918

[5] S. Zhao, W. Wang, Z. Du, J. Chen, Z. Duan, A Black-Box Adversarial Attack Method via Nesterov Accelerated Gradient and Rewiring Towards Attacking Graph Neural Networks. IEEE Transactions on Big Data, 9(6), (2023) 1586-1597. https://doi.org/10.1109/TBDATA.2023.3296936

[6] F. He, Y. Chen, R. Chen, W. Nie, Point Cloud Adversarial Perturbation Generation for Adversarial Attacks. IEEE Access, 11, (2023) 2767-2774. https://doi.org/10.1109/ACCESS.2023.3234313

[7] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, H.V. Poor, Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey. IEEE Communications

Surveys & Tutorials, 25(4), (2023) 2245-2298. https://doi.org/10.1109/COMST.2023.3319492

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

[9] Y. Shi, Y. Han, Q. Hu, Y. Yang, Q. Tian, Query-Efficient Black-Box Adversarial Attack With Customized Iteration and Sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), (2023) 2226-2245. https://doi.org/10.1109/TPAMI.2022.3169802

[10] C. Shi, M. Zhang, Z. Lv, Q. Miao, C.M. Pun, Universal Object-Level Adversarial Attack in Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 61, (2023) 1-14. https://doi.org/10.1109/TGRS.2023.3336734

[11] W. Jiang, H. Li, G. Xu, T. Zhang, R. Lu, Physical Black-Box Adversarial Attacks Through Transformations. IEEE Transactions on Big Data, 9(3), (2023) 964-974. https://doi.org/10.1109/TBDATA.2022.3227318

[12] K. Mo, W. Tang, J. Li, X. Yuan, Attacking Deep Reinforcement Learning With Decoupled Adversarial Policy. IEEE Transactions on Dependable and Secure Computing, 20(1), (2023) 758-768. https://doi.org/10.1109/TDSC.2022.3143566

[13] L. Sun, Y. Dou, C. Yang, K. Zhang, J. Wang, S.Y. Philip, L. He, B. Li, Adversarial attack and defense on graph data: A survey. IEEE Transactions on Knowledge and Data Engineering, 35(8), (2022) 7693-7711. https://doi.org/10.1109/TKDE.2022.3201243

[14] L. Nguyen Vu, T.P. Doan, M. Bui, K. Hong, S. Jung, On the Defense of Spoofing Countermeasures Against Adversarial Attacks. IEEE Access, 11, (2023) 94563-94574. https://doi.org/10.1109/ACCESS.2023.3310809

[15] C. Wan, F. Huang and X. Zhao, Average Gradient-Based Adversarial Attack. IEEE Transactions on Multimedia, 25, (2023) 9572-9585. https://doi.org/10.1109/TMM.2023.3255742

[16] H. Teryak, A. Albaseer, M. Abdallah, S. Al-Kuwari, M. Qaraqe, Double-Edged Defense: Thwarting Cyber Attacks and Adversarial Machine Learning in IEC 60870-5-104 Smart Grids. IEEE Open Journal of the Industrial Electronics Society, 4, (2023) 629-642. https://doi.org/10.1109/OJIES.2023.3336234

[17] C. Qin, Y. Chen, K. Chen, X. Dong, W. Zhang, X. Mao, Y. He, N. Yu, Feature fusion based adversarial example detection against second-round adversarial attacks. IEEE Transactions on Artificial Intelligence, 4(5), (2022)1029-1040. https://doi.org/10.1109/TAI.2022.3190816

[18] R. Gipiškis, D. Chiaro, M. Preziosi, E. Prezioso, F. Piccialli, The Impact of Adversarial Attacks on Interpretable Semantic Segmentation in Cyber–Physical Systems. IEEE Systems Journal, 17(4), (2023) 5327-5334. https://doi.org/10.1109/JSYST.2023.3281079

[19] T. Chen, Z. Ma, Toward Robust Neural Image Compression: Adversarial Attack and Model Finetuning. IEEE Transactions on Circuits and Systems for Video Technology, 33(12), (2023) 7842-7856. https://doi.org/10.1109/TCSVT.2023.3276442

[20] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, Q. Yu, A Survey of Adversarial Attack and Defense Methods for Malware Classification in Cyber Security. IEEE Communications Surveys & Tutorials, 25(1), (2023) 467-496. https://doi.org/10.1109/COMST.2022.3225137

[21] J. Pi, J. Zeng, Q. Lu, N. Jiang, H. Wu, L. Zeng, Z. Wu, Adv-Eye: A Transfer-Based Natural Eye Makeup Attack on Face Recognition. IEEE Access, 11, (2023) 89369-89382. https://doi.org/10.1109/ACCESS.2023.3307132

[22] R. Li, H. Liao, J. An, C. Yuen, L. Gan, IntraClass Universal Adversarial Attacks on Deep Learning-Based Modulation Classifiers. IEEE Communications Letters, 27(5), (2023) 1297-1301. https://doi.org/10.1109/LCOMM.2023.3261423

[23] X. Yuan, Z. Zhang, X. Wang, L. Wu, Semantic-Aware Adversarial Training for Reliable Deep Hashing Retrieval. IEEE Transactions on Information Forensics and Security, 18, (2023) 4681-4694. https://doi.org/10.1109/TIFS.2023.3297791

[24] L. Xu, J. Zhai, DCVAE-adv: A universal adversarial example generation method for white and black box attacks. Tsinghua Science and Technology, *29*(2), (2023) 430-446. https://doi.org/10.26599/TST.2023.9010004

[25] L. Chen, Q.X. Zhu, Y.L. He, Adversarial attacks for neural network-based industrial soft sensors: Mirror output attack and translation mirror output attack. IEEE Transactions on Industrial Informatics, *20*(2), (2023) 2378-2386. https://doi.org/10.1109/TII.2023.3291717

Acknowledgements

The authors appreciate the continuous support of Madhyanchal Professional University, Bhopal, Madhya Pradesh, India

Authors Contribution Statement

Chetan Patil: Conceptualization: Investigation: Methodology, Data collection, Writing original draft; Mohd. Zuber: Conceptualization, Supervision, Validation, Review and Editing. Both authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Has this article screened for similarity?

Yes

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

About the License