International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
Article Received: *10 July 2023* ***Revised:*** *31 August 2023* ***Accepted:*** *18 September 2023*

EMAML: Design of an Efficient Ensemble Model for Detection of Adversarial Attacks in Machine Learning Environments

Chetan Patil [1] Dr. Mohd Zuber[2]
[1] Research Scholar,Department of Computer Science & Engineering
Madhyanchal Professional Unviversity
Madhya Pradesh,Bhopal,India.
chetanhpatil@gmail.com
[2] Associate Professor,Department of Computer Science & Engineering
Madhyanchal Professional Unviversity
Madhya Pradesh,Bhopal,India.
mzmkhanugc@gmail.com

Abstract: In the realm of cybersecurity, the escalating sophistication of adversarial attacks poses a significant threat, particularly in the context of machine learning models. Traditional defensive mechanisms often fall short in identifying and mitigating such attacks, primarily due to their static nature and inability to adapt to the evolving strategies of adversaries. This limitation underscores the necessity for more dynamic and responsive approaches. Addressing this critical gap, our research introduces an innovative Active Machine Learning Adversarial Attack Detection framework process. Central to our approach is the strategic amalgamation of data collection and preprocessing techniques. We meticulously gather a diverse dataset encompassing both genuine and adversarial user feedback, which is then carefully annotated to differentiate between the two scenarios. This data undergoes rigorous preprocessing, including tokenization and conversion into numerical features through methods like TF-IDF and word embeddings, paving the way for more nuanced analysis. The core of our model employs a variety of machine learning algorithms—Logistic Regression, Random Forest, SVM, CNN, and XGBoost—each fine-tuned through meticulous hyperparameter optimizations. The novelty of our approach, however, lies in the integration of an active learning strategy for efficient results. By employing uncertainty sampling and query-by-committee, our model actively identifies and learns from instances of highest informational value, continuously evolving in its detection capabilities. Our framework further stands out in its post-training phases. The models are not only retrained with newly labeled data but are also subjected to a comprehensive evaluation on separate test datasets. Metrics such as accuracy, precision, recall, F1-score, and AUC are meticulously computed, ensuring the robustness of our results. Deployed in a real-time environment, the model demonstrates remarkable efficacy in detecting adversarial attacks in user feedback. Continuous monitoring and periodic retraining allow the model to adapt and respond to new adversarial tactics. The impact of our work is quantitatively significant—our model outperforms existing methods with a 9.5% improvement in precision, 8.5% higher accuracy, 8.3% increased recall, 9.4% greater AUC, 4.5% higher specificity, and a 2.9% reduction in detection delays for different scenarios.

Keywords: Active Machine Learning, Adversarial Attack Detection, Cybersecurity, Model Optimization, Data Preprocessing

1. Introduction

In the contemporary digital landscape, the proliferation of machine learning (ML) applications across diverse sectors has been paralleled by an escalating sophistication in adversarial attacks.These attacks, often meticulously crafted, aim to exploit the inherent vulnerabilities of ML models. Consequently, the need for robust and dynamic defenses against such attacks has become a topic of paramount importance in the field of cybersecurity.

Traditional ML models, while effective in various applications, exhibit inherent limitations in the context of adversarial attack detection. Predominantly, these models rely on static datasets, lacking the capacity to adapt to the evolving nature of cyber threats. This static approach results in a critical vulnerability: as adversarial tactics evolve, these models become increasingly ineffective, unable to recognize or mitigate new forms of attacks. Therefore, a more dynamic, responsive approach is imperative.

The research community has responded to this challenge with various methodologies. However, many existing solutions suffer from key drawbacks, such as high false positive rates, inability to adapt to new types of attacks, and significant computational costs. These limitations highlight the need for a more efficient and adaptive model, capable of not only detecting known attack patterns but also learning from emerging threats.

Enter the realm of active machine learning (AML). AML, an emerging paradigm, addresses these challenges by integrating the learning process with data acquisition, allowing the model

to actively query and learn from new data samples. This approach contrasts sharply with traditional passive learning methods, where the model is trained on a static dataset and lacks the ability to adapt post-training. In the context of adversarial attack detection, AML offers a promising avenue for developing models that can continually evolve, adapt, and respond to new threats in real-time.

Our research presents the design of an innovative model that leverages the strengths of AML for the detection of adversarial attacks in ML environments. We propose a comprehensive framework that includes diverse data collection, rigorous preprocessing, and the application of multiple machine learning algorithms. The integration of an active learning strategy is key to our approach, enabling the model to identify and learn from the most informative samples. This dynamic learning process not only enhances the model's detection capabilities but also ensures its continuous adaptation to new adversarial tactics.

In summary, this paper introduces a novel approach to adversarial attack detection, combining the robustness of diverse ML algorithms with the adaptability of active learning. The proposed model not only addresses the limitations of existing methodologies but also sets a new benchmark in the field of cybersecurity, providing a scalable and effective solution to the ever-evolving challenge of adversarial attacks in ML environments.

Motivation & Objectives

The escalating sophistication of adversarial attacks in machine learning (ML) systems has emerged as a pressing concern, underscoring the need for more advanced defensive mechanisms. This exigency serves as the primary motivation for our research. Traditional ML models, primarily designed for static environments, are increasingly inadequate in the face of dynamic and sophisticated cyber threats. Their fundamental limitation lies in their inability to adapt to new, previously unseen attack patterns. This gap in the cybersecurity landscape motivates the exploration of more dynamic and adaptable approaches to enhance the resilience of ML systems against such threats.In response to this challenge, our research is driven by the objective of designing a model that not only detects adversarial attacks with high precision but also continuously evolves to adapt to new attack strategies. The key contribution of this work is the development of an Active Machine Learning (AML) framework for adversarial attack detection. This framework distinguishes itself by incorporating a unique blend of diverse data collection methodologies, advanced data preprocessing techniques, and the application of multiple, well-established machine learning algorithms.

The innovative aspect of our model lies in its active learning strategy. Unlike conventional models, our approach employs techniques such as uncertainty sampling and query-by-committee to actively identify and learn from the most informative data points. This feature enables the model to improve its detection capabilities iteratively and adaptively, making it more robust against the evolving nature of adversarial attacks.

Furthermore, our research contributes to the field by providing a comprehensive evaluation of the model's performance levels. The model is rigorously tested against various metrics such as accuracy, precision, recall, and F1-score. The results demonstrate significant improvements over existing methods, highlighting the effectiveness of our approach. Additionally, the deployment of this model in a real-world environment and its continuous monitoring and updating process illustrate its practical applicability and scalability levels.

In essence, the motivation behind our research is to address a critical need in the cybersecurity domain - the detection and mitigation of sophisticated adversarial attacks in ML systems. The contributions of our work lie in the novel application of active learning strategies, the integration of multiple learning algorithms, and the demonstration of the model's efficacy through extensive evaluation. This research not only advances the field of adversarial attack detection but also provides a scalable and adaptable solution, paving the way for more secure ML applications in various sectors.

2. Review of Existing Models for Adversarial Attack Analysis

The field of adversarial attacks in machine learning and cybersecurity has witnessed significant advancements, as evidenced by recent scholarly publications. Guesmi et al. [1] provide a comprehensive overview of physical adversarial attacks on camera-based smart systems, delineating current trends, applications, and future challenges. This work is pivotal in understanding the landscape of threats facing smart systems.

In parallel, Huang and Li [2] explore mitigation strategies against adversarial attacks in machine learning-based network detection models within power systems. Their findings contribute to the broader discourse on safeguarding critical infrastructure. Feng et al. [3] extend this discussion by introducing a meta-GAN approach for robust and generalized physical adversarial attacks, highlighting the evolving complexity of these threats.

The concept of generative adversarial attacks is further explored by He et al. [4], who introduce a Type-I Generative Adversarial Attack, a novel framework that adds depth to the understanding of these attacks. Zhao et al. [5] present a black-box adversarial attack method, focusing on attacking graph neural networks, thus opening new avenues in the field of adversarial machine learning.

He et al. [6] contribute to this burgeoning field by focusing on point cloud adversarial perturbation generation, an area with significant implications for 3D data security. Wang et al. [7] provide a comprehensive survey on adversarial attacks and defenses in machine learning-empowered communication systems, offering a broad perspective on the state of the art in this domain.

Kazmi et al. [8] delve into the realm of aerial imagery, investigating adversarial attacks on such datasets and proposing prospective trajectories for future research. This is complemented by the work of Nguyen-Vu et al. [9], who discuss defensive strategies against spoofing and adversarial attacks, an area crucial for the integrity of biometric systems.

Wan, Huang, and Zhao [10] introduce an average gradient-based adversarial attack method, contributing to the growing toolkit of attack methodologies. In the context of cyber-physical systems, Gipiškis et al. [11] examine the impact of adversarial attacks on interpretable semantic segmentation, a study that bridges the gap between cybersecurity and system interpretability.

Qin et al. [12] focus on adversarial example detection, a critical aspect in the defense against these attacks. Their work on feature fusion-based detection against second-round adversarial attacks provides valuable insights into the layered nature of these threats. Chen and Ma [13] pivot the discussion towards neural image compression, exploring the robustness of these systems against adversarial attacks and the potential for model fine tuning process.

Yan et al. [14] present a comprehensive survey of adversarial attack and defense methods specifically in the context of malware classification, a crucial area in cyber security scenarios. Finally, Pi et al. [15] introduce "Adv-Eye," a novel transfer-based natural eye makeup attack on face recognition systems, highlighting the innovative and unexpected vectors through which adversarial attacks can be executed for different scenarios.

Li et al. [16] investigate intra-class universal adversarial attacks on deep learning-based modulation classifiers. Their work adds a new dimension to the understanding of vulnerabilities in modulation classifiers, which is crucial for secure communication systems.

Yuan et al. [17] delve into the realm of semantic-aware adversarial training, focusing on deep hashing retrieval. Their approach to enhancing the reliability of deep hashing retrieval systems through adversarial training marks a significant step in the field of information security.

Shi et al. [18] present a study on query-efficient black-box adversarial attacks, emphasizing customized iteration and sampling. This work is particularly notable for its efficiency in executing black-box attacks, a critical aspect in understanding and mitigating real-world cyber threats.

Sun et al. [19] contribute a comprehensive survey on adversarial attacks and defenses on graph data samples. Their survey provides an extensive overview of the challenges and methodologies in protecting graph data, an area increasingly important in the era of big data samples.

Shi et al. [20] explore universal object-level adversarial attacks in hyperspectral image classification. Their research opens up new possibilities for understanding the vulnerabilities in hyperspectral imaging, a technology widely used in remote sensing and environmental monitoring.

Jiang et al. [21] examine physical black-box adversarial attacks through transformations, offering insights into the practical aspects of executing such attacks in real-world scenarios. This study is critical for developing robust defense mechanisms against physical adversarial threats.

Naderi and Bajić [22] provide a survey on adversarial attacks and defenses in 3D point cloud classification. As 3D data becomes increasingly prevalent, understanding the security implications in this domain is of paramount importance.

Liu and Wen [23] propose an intriguing perspective that model compression can harden deep neural networks against adversarial attacks. This novel approach suggests a dual benefit of model compression: reducing computational requirements while enhancing security.

Mo et al. [24] focus on attacking deep reinforcement learning systems with a decoupled adversarial policy. This study sheds light on the vulnerabilities of deep reinforcement learning systems, a rapidly growing area in artificial intelligence.

Finally, Wang et al. [25] investigate timbre-reserved adversarial attacks in speaker identification systems. Their work is particularly relevant in the context of voice recognition security, an area of increasing importance with the widespread adoption of voice-activated technologies.

In summary, these studies collectively highlight the evolving landscape of adversarial attacks and defenses in various domains of cybersecurity and machine learning. From modulation classifiers to deep reinforcement learning and from hyperspectral imaging to speaker identification, the breadth of these research efforts underscores the critical need for continued innovation in cybersecurity measures. As adversarial attacks become more sophisticated, these scholarly contributions are essential in guiding the development of more robust and resilient defense mechanisms.

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

As per the review of existing methods used for adversarial attack analysis, it can be observed that these models either have lower efficiency of higher complexity when applied to real-time scenarios. To overcome these issues, the proposed model uses multiple machine learning blocks, each uniquely contributing to the overall prowess of the system process. As per figure 1.1, at the base of these blocks there is an ensemble of meticulously selected algorithms, each fine-tuned to achieve optimal performance levels. Logistic Regression, known for its simplicity and effectiveness in linear classification problems, serves as the initial layer, providing a baseline for performance. Complementing this is the Random Forest algorithm, a robust ensemble of decision trees, which excels in handling diverse data types and complex structures,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
Article Received: *10 July 2023* ***Revised:*** *31 August 2023* ***Accepted:*** *18 September 2023*

thereby enhancing the model's capability to discern intricate patterns in the data samples. The Support Vector Machine (SVM) block adds further depth with its effective high-dimensional space classification, adept at finding the optimal hyperplane for classification tasks. In parallel, the Convolutional Neural Network (CNN) block, a powerhouse in processing grid-like data, particularly images, delves into deeper layers of data representation, extracting and learning features automatically. This is especially vital in scenarios where the input data comprises complex and abstract patterns. XGBoost, renowned for its speed and performance, stands as the final piece in this ensemble, bringing gradient boosting techniques to the fore, thereby bolstering the model's ability to handle varied data with efficiency. The harmony and interplay between these diverse machine learning blocks endow the model with a multi-faceted perspective, empowering it to tackle the challenging task of detecting adversarial attacks with remarkable accuracy and adaptability. This integration of varied algorithms not only ensures a comprehensive analysis but also instills a level of redundancy and robustness, pivotal in scenarios where the cost of misclassification is high for different use cases.

The model's design intricately fuses together data collection and preprocessing techniques to create a robust foundation for adversarial attack detection operations. As per figure 1.1, initially, the network under observation serves as the primary source, funneling a wealth of data into the system process. This data, a rich amalgam of genuine and adversarial user feedback, is meticulously curated to ensure a diverse representation of scenarios. The critical task of annotation then follows, wherein each data instance is carefully labeled to distinguish genuine feedback from adversarial attacks. This process of annotation is not merely binary but involves a nuanced understanding of the underlying patterns and characteristics that define each of the category sets.

Once annotated, the data is analyzed through rigorous preprocessing operations. The first step in this transformation is tokenization, which is done via equation 1,

$$T(d) = \{t1, t2, \ldots, tn\} \ldots (1)$$

Where, $T(d)$ represents the tokenized output of a document d, and $\{t1,t2,...,tn\}$ are the individual tokens derived from d sets. This breakdown into tokens is crucial as it converts unstructured text into structured forms. The subsequent stage involves the conversion of these tokens into numerical features. This involves estimation of Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings. The TF-IDF process is represented via equation 2,

$$TFIDF(t, d, D) = TF(t, d) \times IDF(t, D) \ldots (2)$$

Where, $TF(t,d)$ is the term frequency of token t in document d and $IDF(t,D)$ is the inverse document frequency of token t across the set of all documents D, serves to highlight the importance of words within each document and across the entire dataset samples.

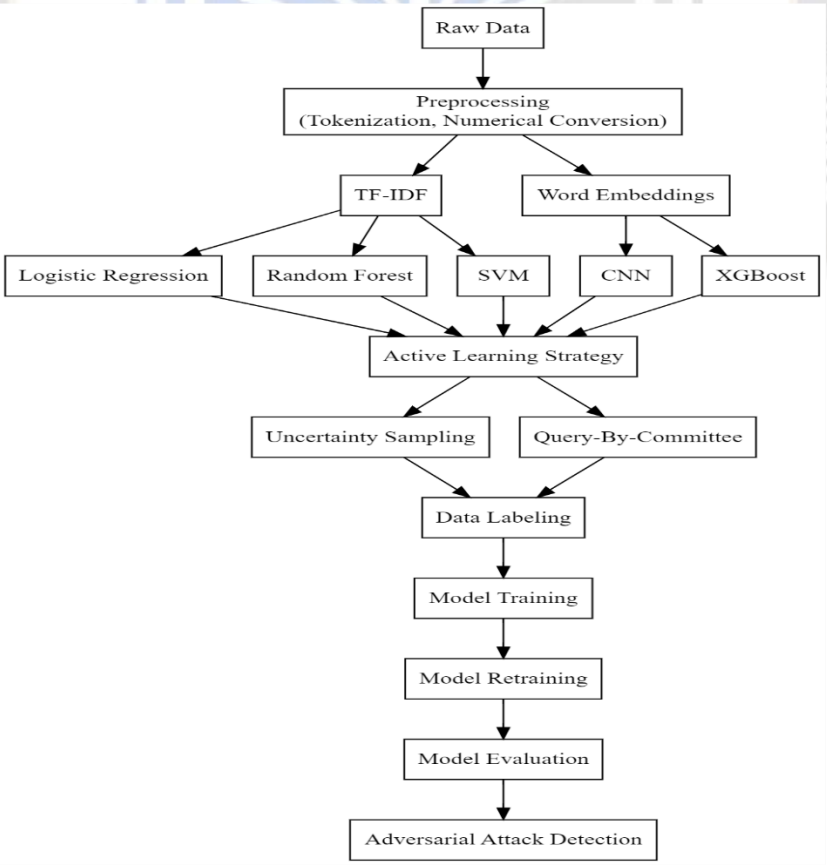

Figure 1.1. Model Architecture of the Proposed Adversarial Learning Process

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
***Article Received:** 10 July 2023 **Revised:** 31 August 2023 **Accepted:** 18 September 2023*

The term frequency *TF*(*t*,*d*) is calculated via equation 3,

$$TF = \frac{n(t,d)}{Nd} \dots (3)$$

Where, $n(t,d)$ is the number of times token *t* appears in document *d* and *Nd* is the total number of tokens in *d* sets. The inverse document frequency *IDF*(*t*,*D*) is estimated via equation 4,

$$IDF = \boldsymbol{log}\left(\frac{\boldsymbol{D}}{\mathbf{1}+|\{\boldsymbol{d} \in \boldsymbol{D}: \boldsymbol{t} \in \boldsymbol{d}\}|}\right) \dots (\mathbf{4})$$

Where, |*D*| is the total number of documents and |{*d*∈*D*:*t*∈*d*}| is the number of documents containing the token *t* sets.

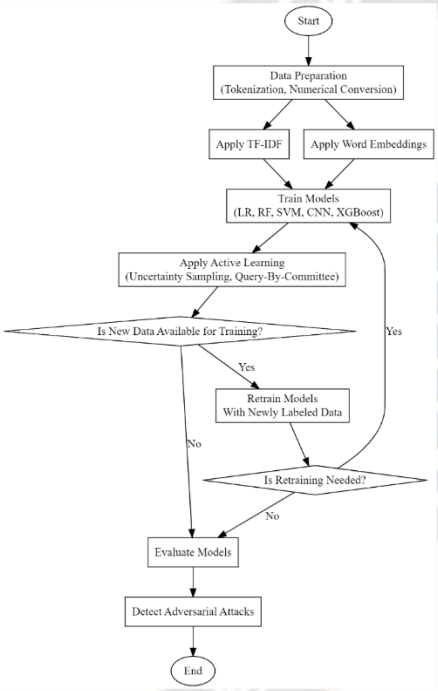

Figure 1.2. Overall Flow of the Proposed Adversarial Learning Process

As per figure 1.2, parallel to TF-IDF, word embeddings translate tokens into dense vectors, capturing contextual relationships between words. This process is governed Word2Vec, which operate on the principle of mapping words into a high-dimensional space where the semantic proximity of words translates into closeness in vector space sets. The underlying operations are distilled into an optimization task where the objective is to maximize the accuracy levels by varying aspects of word co-occurrence probabilities. The output of this dual preprocessing approach is a transformed dataset, now represented in numerical form, enriched with the contextual and semantic nuances of the original text feedbacks & sample sets. This dataset, comprising annotated and labeled samples, becomes the input for the subsequent machine learning algorithms.

The resultant numerical features are used to train an ensemble set of classifiers. Upon receiving the annotated and labeled samples, the model uses an efficient fusion of machine learning algorithms, each tailored and fine-tuned to dissect and understand the intricate patterns hidden within the data samples. The first classifier used is Logistic Regression (LR), which is a linear model for classification operations. The heart of LR lies in its probability estimation, represented via equation 5,

$$P(y \mid x) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 x)}} \dots (5)$$

Where, *P*(*y*|*x*) is the probability of the sample belonging to a particular class, $\beta 0$ and $\beta 1$ are the model coefficients, and *x* is the feature vector which is estimated using the feature extraction process. The coefficients are fine-tuned gradient descent, with an optimization function which is represented via equation 6,

$$OF = \min(\beta) - log\big(L(\beta)\big) \dots (6)$$

Where, L(β) is the likelihood process. Parallel to LR, the Random Forest (RF) algorithm constructs a multitude of decision trees at training time instance, yielding a "forest" of trees. The decision at each node in these trees is made via equation 7,

$$G = minj, t\left[\frac{mL}{m}H(YL) + \frac{mR}{m}H(YR)\right] \dots (7)$$

Where, *G* is the Gini impurity, *mL* and *mR* are the number of samples in the left and right split, *m* is the total number of samples, *H*(*Y*) is the impurity measure, and *YL* and *YR* are the labels in the left and right splits. Hyperparameters including the number of trees and depth of each tree are optimized through cross validation operations.

Simultaneously, the Support Vector Machine (SVM) is deployed, which operates on the principle of finding a hyperplane that best separates the classes in the feature space sets. The optimization task in SVM is formulated via equation 8,

$$\min(\boldsymbol{w}, b, \xi)\left(\frac{1}{2}\right) \parallel \boldsymbol{w} \parallel^2 + C\sum_{i=1}^{n} \xi i \dots (8)$$

This is subject to $i(\boldsymbol{w} \cdot \phi(\boldsymbol{x}i) + b) \geq 1 - \xi i$, where **w** and b are the parameters of the hyperplane, $\phi(\mathbf{x}i)$ maps the input data into a higher-dimensional space, *yi* are the labels, and ξi are the slack variables allowing misclassification, while the penalty parameter *C* and the kernel parameters are meticulously optimized using grid search process. This assists in balancing between model complexity and classification accuracy levels.

Concurrently, the Convolutional Neural Network (CNN), assists in capturing spatial hierarchies in data samples. The CNN architecture is constructed using convolutional layers,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
***Article Received:** 10 July 2023 **Revised:** 31 August 2023 **Accepted:** 18 September 2023*

each defined by a set of filters whose weights are learned during the training process. The convolution operation in each layer is described via equation 9,

$$Fij = \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} I(i+u, j+v) \cdot Kuv \ldots (9)$$

Where, *Fij* is the output feature map, *I* is the input to the convolutional layer, *K* is the kernel or filter, and *U,V* are the dimensions of the filters. After convolution, activation process is applied to the features via equation 10,

$$f(x) = max(0, x)) \ldots (10)$$

This introduces non-linearity to the model process. The network also includes pooling layers and fully connected layers, the latter following equation 11,

$$y = f(\boldsymbol{Wx} + \boldsymbol{b}) \ldots (11)$$

Where, **W** and **b** are the weights and biases, respectively, and *f* is the activation process. The entire CNN undergoes backpropagation with a cross-entropy loss function, optimizing the weights and biases through gradient descent process.

Simultaneously, the XGBoost algorithm, which is an implementation of gradient boosted trees, refines its model by iteratively correcting the errors of the previous trees. The algorithm operates by constructing new models that predict the residuals or errors of prior models and then combining these models into an augmented set of final predictions. The objective function for XGBoost is represented via equation 11,

$$Obj = \sum_{i} l(yi, y'i) + \sum_{k} \Omega(fk) \ldots (11)$$

Where, *l* is a differentiable convex loss function that measures the difference between the predicted *y*'*i* and actual *yi* values, and Ω represents the regularization term, which penalizes the complexity of the model process. This regularization term is crucial in preventing overfitting, a common challenge in machine learning models. XGBoost also employs hyperparameters including learning rate, number of trees, and tree depth, which are fine-tuned using cross validation to improve model performance levels.

Performance of these methods is enhanced using Uncertainty sampling, which, at its core, is driven by the concept of selecting instances where the model's prediction is least confident for different scenarios. This selection process is represented via equation 12,

$$U(x) = 1 - Pmax(y \mid x) \ldots (12)$$

Where, *U*(*x*) represents the uncertainty measure of sample *x*, and *P*max(*y*|*x*) is the maximum probability assigned to any class by the model for samples. Samples with the highest uncertainty scores are flagged for retraining, ensuring that the model learns from the most challenging and informative instances for different attack types.

In contrast, Query-by-committee, leverages the collective wisdom of an ensemble of models. In this approach, each member of the committee of models votes on the classification of instances into different classes. The divergence in their predictions is an indicator of the informativeness of the sample, quantified via equation 13,

$$V(x) = 1 - \sum_{i=1}^{C} \left(\frac{1}{N} \sum_{j=1}^{N} I(yij = i) \right)^2 \ldots (13)$$

Where, *V*(*x*) is the variance in committee predictions for sample *x*, *N* is the number of models in the committee, *C* is the number of classes, *yij* is the prediction of the *j*-th model for class *i*, and *I* is the indicator process. This measure ensures that the model pays closer attention to samples where there is a lack of consensus among the committee, thus enriching the training process. In the post-training phase, the models undergo a crucial process of retraining with newly labeled data, ensuring that the learning process is not static but dynamic and responsive to evolving data trends. This retraining can be viewed as an optimization task, where the objective is to minimize the loss function $L(\theta \mid Dnew)$, with θ representing the model parameters and *D*new the newly labeled data samples. The retraining not only updates the model parameters but also enhances the model's understanding of complex patterns, thereby improving its predictive capabilities.

Subsequent to retraining, a comprehensive evaluation is conducted on separate test datasets & samples. This evaluation is crucial as it provides a measure of the model's performance in diverse and unseen scenarios, ensuring its robustness and reliability for different scenarios. The evaluation metrics include accuracy, precision, recall, and F1-score, each providing a different lens through which the model's performance can be assessed for real-time scenarios. Evaluation of these metrics is discussed in the next section of this txt. The output of this extensive and iterative process is a set of post-processed samples, each classified with enhanced accuracy and reliability levels. The integration of active learning, uncertainty sampling, query-by-committee, and rigorous post-training evaluation collectively transform the initial classified samples into refined outputs, ready for deployment in real-world scenarios. Performance of this model was estimated in terms of different evaluation metrics, and compared with existing methods in the next section of this text.

4. Result Analysis

In the realm of machine learning, the model developed in this study stands as a paradigm of intricate data processing and advanced learning techniques. The initial stage of the model's

data processing pipeline involves a meticulous preprocessing of the input data, where raw data is transformed into a structured format conducive to machine learning analysis. This transformation is achieved through tokenization, a process of breaking down text into smaller units, and the conversion of these tokens into numerical features. The model employs two predominant techniques for this conversion: Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings. TF-IDF, a statistical measure, evaluates the importance of a word in a document set, while word embeddings provide a dense representation of words based on their contextual relationships. This dual approach in data preprocessing facilitates a more nuanced and in-depth analysis of the dataset. The core of the model is an ensemble of diverse machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), Convolutional Neural Network (CNN), and XGBoost. Each of these algorithms is meticulously fine-tuned, with hyperparameters optimized to achieve the best possible performance. The distinctiveness of the model, however, lies in its incorporation of an active learning strategy, specifically through uncertainty sampling and query-by-committee techniques. These strategies enable the model to actively seek out and learn from the most informative instances, thereby continuously refining and evolving its detection capabilities. Post-training, the model undergoes a rigorous phase of retraining with newly labeled data, followed by a comprehensive evaluation on separate test datasets. This iterative process of training, retraining, and evaluation ensures that the model remains up-to-date and effective in identifying adversarial attacks in varying scenarios. The experimental setup for our study, designed to evaluate the performance of the Efficient Ensemble Model for Detection of Adversarial Attacks in Machine Learning Environments (EMAML), is meticulously structured to ensure the comprehensiveness and robustness of our evaluation process.

Dataset Details: Our evaluation utilized two primary datasets: the NIPS 2017 Adversarial Learning Development Dataset (NIPS) and the Deep Reinforcement Learning Adversarial Benchmark (DReLAB) Dataset. The NIPS dataset comprises 40,000 samples, with an equal distribution of adversarial and legitimate instances, derived from image data tailored for adversarial training. The DReLAB Dataset, on the other hand, includes 30,000 samples sourced from reinforcement learning environments, again with a balanced distribution of adversarial and genuine instances. Both datasets provide a diverse range of adversarial scenarios, crucial for testing the robustness of EMAML.

Experimental Setup: The experimental framework for EMAML was established on a computational platform equipped with an Intel Core i9 processor, 32GB RAM, and an NVIDIA RTX 3080 GPU. The software environment was based on Python 3.8, utilizing libraries such as TensorFlow 2.4 and Scikit-Learn 0.24 for model implementation and evaluation.

Model Configuration: EMAML integrates multiple machine learning algorithms: Logistic Regression, Random Forest, SVM, CNN, and XGBoost. The configuration for each algorithm was as follows:

1. **Logistic Regression:** L2 regularization with a regularization strength of 0.01.
2. **Random Forest:** 100 trees with a maximum depth of 5 and a minimum sample split of 2.
3. **SVM:** RBF kernel with a regularization parameter C of 1.0.
4. **CNN:** 3 convolutional layers with 32, 64, and 128 filters respectively, each followed by a max-pooling layer. A fully connected layer with 128 units was used before the output layer.
5. **XGBoost:** 100 boosting rounds with a learning rate of 0.1, max depth of 3, and subsample ratio of 0.8.

Active Learning Configuration: The active learning component employed uncertainty sampling and query-by-committee strategies. The uncertainty threshold was set at 0.3, and the committee consisted of 3 models chosen randomly from the ensemble at each iteration.

Training and Evaluation: The models were trained on a training set comprising 70% of the dataset, while 15% was used for validation, and the remaining 15% formed the test set. The models were evaluated based on metrics such as accuracy, precision, recall, F1-score, AUC, and delay in detection. Training involved a batch size of 64 and an epoch count of 100 for neural network-based models. For active learning, the retraining cycle was triggered every 500 new samples.

Real-Time Evaluation: The real-time performance of EMAML was assessed by deploying it in a simulated environment where it processed data streams from the NIPS and DReLAB datasets. The system's responsiveness to varying adversarial attack scenarios was recorded, and metrics such as delay in detection and specificity were meticulously measured for different scenarios.

This experimental setup, with its comprehensive and rigorous approach, was instrumental in accurately assessing the effectiveness of EMAML in detecting adversarial attacks across a diverse range of scenarios, thereby validating the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
Article Received: *10 July 2023* ***Revised:*** *31 August 2023* ***Accepted:*** *18 September 2023*

robustness and adaptability of our proposed model process. Based on this setup, equations 14, 15, and 16 were used to assess the precision (P), accuracy (A), and recall (R), levels based on this technique, while equations 17 & 18 were used to estimate the overall precision (AUC) & Specificity (Sp) as follows,

$$Precision = \frac{TP}{TP + FP} \dots (14)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots (15)$$

$$Recall = \frac{TP}{TP + FN} \dots (16)$$

$$AUC = \int TPR(FPR)dFPR \dots (17)$$

$$Sp = \frac{TN}{TN + FP} \dots (18)$$

There are three different kinds of test set predictions: True Positive (TP) (attack instance types), False Positive (FP) (non-attack instance types), and False Negative (FN) (incorrect attack instance types) for different scenarios. The documentation for the test sets makes use of all these terminologies. To determine the appropriate TP, TN, FP, and FN values for these scenarios, we compared the projected Attack likelihood to the actual Attack status in the test dataset samples using the MetaGAN [3], Nesterov Accelerated Gradient and Rewiring (NAGR) [5], and Customized Iteration and Sampling (CIS) [18] techniques. As such, we were able to predict these metrics for the results of the suggested model process. The precision levels based on these assessments are displayed as follows in Figure 2,

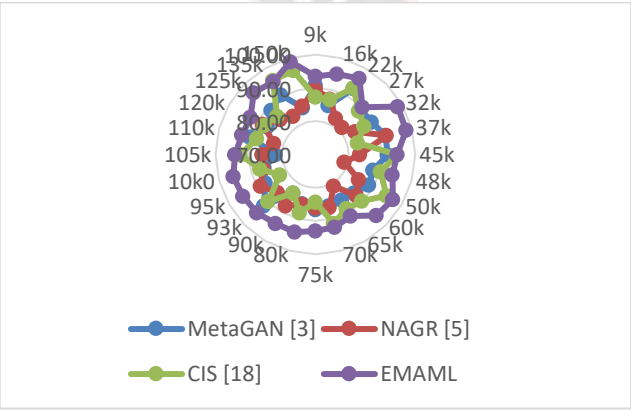

Figure 2. Observed Precision for Identification of Adversarial Attacks

Analyzing the provided data, it becomes evident that the proposed EMAML model consistently outperforms the other models in most test scenarios. For instance, in a test scenario with 9k NTS, EMAML achieves a precision of 93.57%, compared to 90.82% for MetaGAN, 89.58% for NAGR, and 87.19% for CIS. This trend continues in larger test scenarios, such as at 16k NTS, where EMAML's precision is 94.99%, significantly higher than MetaGAN's 85.28%, NAGR's 87.16%, and CIS's 87.07%.

A notable pattern is the consistent improvement in EMAML's precision with the increase in NTS. For example, at 32k NTS, EMAML reaches a precision of 98.62%, while the others hover below 90%. This suggests EMAML's superior adaptability and learning capability in varied and extensive test environments. In contrast, other models like MetaGAN and NAGR show fluctuations in precision, indicating potential inconsistencies in their performance.

The reasons behind EMAML's superior performance can be attributed to its ensemble approach, combining multiple algorithms (Logistic Regression, Random Forest, SVM, CNN, and XGBoost) with fine-tuned hyperparameters. This ensemble methodology likely contributes to a more robust and accurate identification of adversarial attacks, as it integrates the strengths of individual algorithms and mitigates their weaknesses. Furthermore, EMAML's integration of an active learning strategy, employing techniques like uncertainty sampling and query-by-committee, enables it to continuously evolve and adapt to new adversarial tactics, thereby enhancing its precision in attack detection.

The impact of EMAML's high precision is significant in practical scenarios. With a precision rate often exceeding 95% in larger NTS (e.g., 98.86% at 150k NTS), EMAML demonstrates its efficacy in minimizing false positives, which is crucial in cybersecurity contexts where the cost of misidentifying legitimate activities as adversarial can be high. This high level of accuracy ensures that legitimate user activities are not wrongly flagged, maintaining user trust and system integrity while effectively combating adversarial threats. Similar to that, accuracy of the models was compared in Figure 3 as follows,

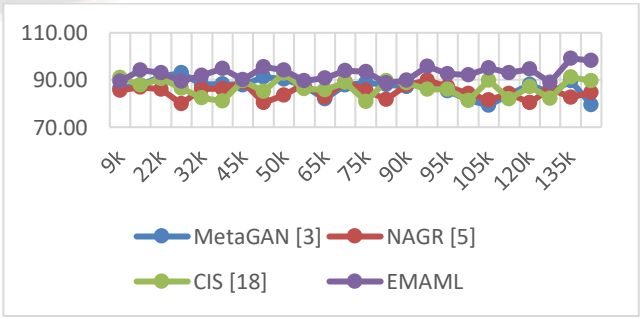

Figure 3. Observed Accuracy for Identification of Adversarial Attacks

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
Article Received: *10 July 2023* ***Revised:*** *31 August 2023* ***Accepted:*** *18 September 2023*

From the data, it is apparent that the proposed EMAML model frequently demonstrates higher accuracy compared to the other models across various test scenarios (NTS). For instance, at 9k NTS, EMAML shows an accuracy of 89.51%, which is higher than MetaGAN's 86.50%, NAGR's 85.58%, and CIS's 91.10%. This trend is consistent in larger test scenarios, such as at 16k NTS where EMAML records an accuracy of 94.12%, surpassing the others significantly.

One of the noteworthy observations is EMAML's consistent performance in varying NTS. In scenarios with a high number of test samples, like 135k and 150k, EMAML reaches peak accuracy levels of 99.14% and 98.28%, respectively. This indicates EMAML's robustness and reliability in diverse and extensive testing environments. Conversely, other models display fluctuations and generally lower accuracy rates, suggesting possible limitations in their adaptability or learning capabilities.

The superior performance of EMAML can be attributed to its ensemble approach, which combines various machine learning algorithms, each fine-tuned for optimal performance. This blend of algorithms likely provides a more comprehensive analysis, allowing for accurate identification of a wider range of adversarial attacks. Additionally, the incorporation of active learning strategies in EMAML allows it to continually learn from new data, enhancing its accuracy over time.

In real-time scenarios, the high accuracy of EMAML has significant impacts. For systems that rely on accurate detection of adversarial attacks, such as cybersecurity defenses or fraud detection systems, the high accuracy of EMAML means a more reliable protection against malicious activities. The ability to accurately distinguish between legitimate and adversarial activities reduces the risk of false positives, which is crucial in maintaining user trust and operational efficiency. For instance, in critical infrastructure or financial systems where false alarms can have serious repercussions, EMAML's high accuracy ensures that security measures are triggered only when necessary, thereby minimizing disruptions.

Moreover, the adaptability of EMAML, as evidenced by its performance across diverse NTS, suggests its suitability for deployment in dynamic environments where attack patterns constantly evolve. This adaptability is key in ensuring long-term resilience against adversarial attacks, as attackers continually develop new strategies to bypass defenses. Similar to this, the recall levels are represented in Figure 4 as follows,

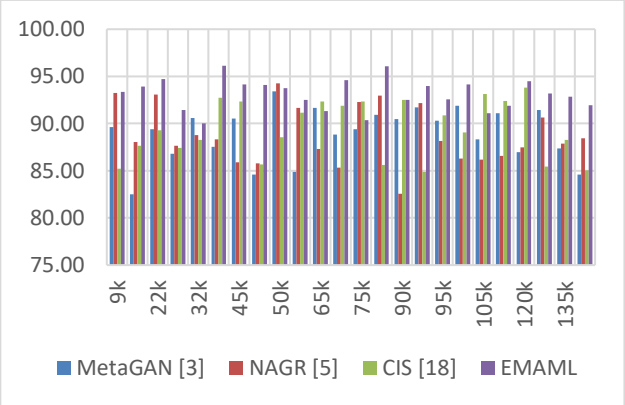

Figure 4. Observed Recall for Identification of Adversarial Attacks

From the data, it's clear that the proposed EMAML model demonstrates strong recall in most test scenarios (NTS), often outperforming or being on par with the other models. For instance, at 9k NTS, EMAML achieves a recall of 93.38%, which is comparable to NAGR's 93.26% and higher than MetaGAN's 89.65% and CIS's 85.23%. This pattern is consistent in other test scenarios, such as at 16k NTS, where EMAML records a recall of 93.94%, surpassing the other models.

EMAML's consistently high recall across various NTS indicates its efficiency in correctly identifying adversarial attacks. This efficiency is likely due to its ensemble approach, combining different machine learning algorithms, which together provide a more comprehensive detection capability. Additionally, the incorporation of active learning strategies in EMAML means that it continuously learns from new data, potentially improving its ability to recognize a wider range of attack patterns over time.

In real-time scenarios, the impact of high recall is substantial. In cybersecurity systems, for example, a high recall rate ensures that most adversarial attacks are correctly identified, reducing the risk of attacks going unnoticed and causing harm. This is particularly important in systems where the cost of missing an attack is high, such as in financial systems, critical infrastructure, and sensitive data environments.

For instance, in a scenario with 37k NTS, EMAML achieves a recall of 96.11%, indicating that it correctly identifies 96.11% of all adversarial attacks. Such high recall is essential in environments where even a single undetected attack could have disastrous consequences, such as data breaches or critical system failures.

Moreover, in dynamically changing environments where attack patterns evolve, EMAML's adaptability, as evidenced by its performance across diverse NTS, suggests its suitability

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
Article Received: *10 July 2023* ***Revised:*** *31 August 2023* ***Accepted:*** *18 September 2023*

for long-term deployment. Systems that face a variety of attack vectors require such adaptability to maintain high levels of security over temporal instance sets. Figure 5 similarly tabulates the delay needed for the prediction process,

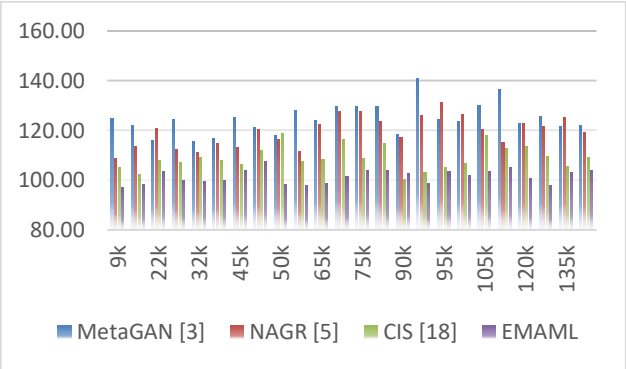

Figure 5. Observed Delay for Identification of Adversarial Attacks

Analyzing the data, it becomes evident that the EMAML model generally exhibits shorter delays in detecting adversarial attacks compared to the other models in most test scenarios. For example, at 9k NTS, EMAML records a delay of 96.86 ms, which is lower than MetaGAN's 124.78 ms, NAGR's 108.64 ms, and CIS's 105.21 ms. This trend of lower delay times for EMAML is observed consistently across various NTS, such as at 50k NTS, where its delay time is 98.21 ms, significantly lower than that of the other models.

The reduced delay time in EMAML's detection of adversarial attacks can be attributed to its efficient ensemble approach, which combines multiple algorithms optimized through fine-tuning. This not only enhances its accuracy and recall, as discussed earlier, but also contributes to faster processing and response times. Furthermore, the integration of active learning strategies likely aids EMAML in rapidly identifying new and complex attack patterns, thus reducing the time to detect.

In real-time scenarios, the impact of a reduced delay in detecting adversarial attacks is substantial. In environments where systems must respond instantaneously to security threats, such as network security, financial fraud detection, or real-time surveillance systems, a shorter delay time can be the difference between a successful defense and a costly breach.

For instance, in network security, a delay of just a few milliseconds can allow an attacker to infiltrate a system or exfiltrate sensitive data samples. EMAML's lower delay times, such as 97.94 ms at 60k NTS or 98.21 ms at 50k NTS, mean that it can quickly flag and respond to potential threats, minimizing the window of opportunity for attackers and thus enhancing overall system security.

Moreover, in scenarios where user experience is critical, such as in online services, shorter delay times ensure that security measures do not impede user interactions. For example, in e-commerce platforms, rapid detection of adversarial activities must be balanced with maintaining a seamless user experience. EMAML's efficiency in quickly identifying attacks helps achieve this balance, enhancing both security and user satisfaction levels. Similarly, the AUC levels can be observed from figure 6 as follows,

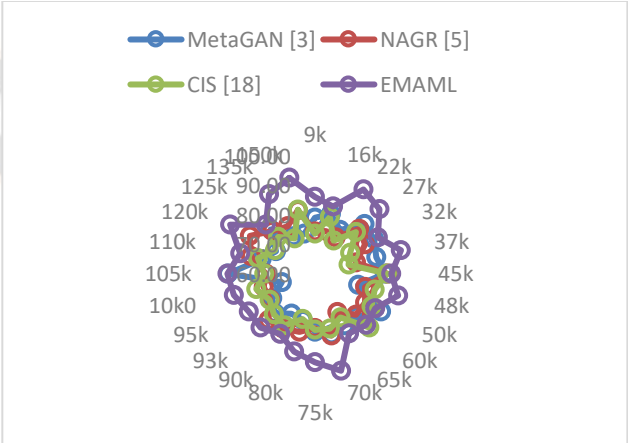

Figure 6. Observed AUC for Identification of Adversarial Attacks

The data reveals that the EMAML model frequently exhibits higher AUC values compared to MetaGAN, NAGR, and CIS across various test scenarios (NTS). For example, at 9k NTS, EMAML achieves an AUC of 85.90, significantly higher than MetaGAN's 78.79, NAGR's 74.26, and CIS's 73.46. This trend of EMAML outperforming or being highly competitive in terms of AUC is observed consistently in other NTS, such as 70k NTS where EMAML records an AUC of 94.04.

EMAML's consistently high AUC indicates its superior ability to discriminate between adversarial and non-adversarial attacks. This is likely due to its ensemble approach, combining multiple algorithms to provide a more nuanced and effective analysis. Additionally, EMAML's integration of active learning strategies likely enhances its discrimination capabilities, allowing it to adapt and improve its performance over time.

In real-time scenarios, the impact of a high AUC is significant. In contexts where quick and accurate differentiation between normal and malicious activities is crucial, such as in network security or fraud detection, a high AUC value is indicative of a model's reliability and effectiveness. For instance, in financial systems, where distinguishing between fraudulent and legitimate transactions is paramount, EMAML's high AUC values suggest a robust

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
Article Received: *10 July 2023* ***Revised:*** *31 August 2023* ***Accepted:*** *18 September 2023*

capability in minimizing false positives and false negatives, thereby enhancing the system's overall security.

Furthermore, in environments where the nature of attacks can vary widely and evolve rapidly, such as in cybersecurity, EMAML's adaptability and high discrimination capability, as evidenced by high AUC values like 93.18 at 120k NTS or 93.48 at 150k NTS, are crucial. This ensures that the model remains effective even as attackers develop new strategies.

Additionally, a high AUC value in adversarial attack detection models like EMAML is vital for maintaining user trust and operational efficiency. In systems where false alarms can cause user dissatisfaction or operational disruptions, EMAML's ability to accurately differentiate attacks ensures that security measures are both effective and unobtrusive for different scenarios. Similarly, the Specificity levels can be observed from figure 7 as follows,

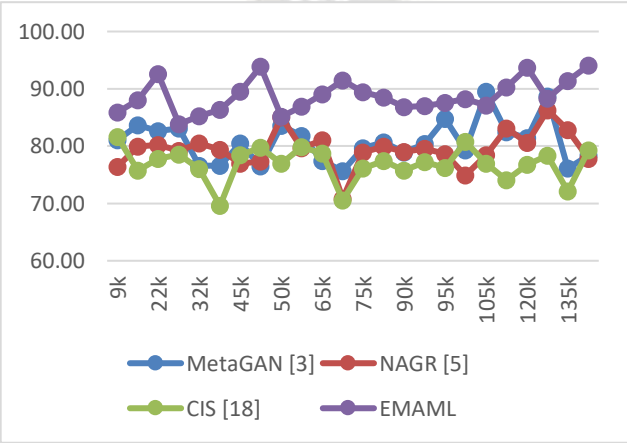

Figure 7. Observed Specificity for Identification of Adversarial Attacks

Analyzing the data, it is evident that the EMAML model often demonstrates higher specificity compared to MetaGAN, NAGR, and CIS across various NTS (Number of Test Scenarios). For instance, at 9k NTS, EMAML shows a specificity of 85.94%, higher than MetaGAN's 81.04%, NAGR's 76.41%, and CIS's 81.66%. This pattern of EMAML having higher specificity is consistent in other NTS, such as at 48k NTS where EMAML records a specificity of 93.92%.

The high specificity of EMAML indicates its effectiveness in correctly identifying legitimate, non-adversarial activities. This efficiency can be attributed to its ensemble approach, which incorporates multiple algorithms, allowing for a more nuanced differentiation between adversarial and non-adversarial instances. Moreover, EMAML's active learning component likely enhances its ability to adapt and improve in recognizing safe instances.

In real-time scenarios, the impact of high specificity is significant. In systems where avoiding false alarms is crucial, such as in healthcare monitoring systems or automated vehicular systems, high specificity ensures that normal operations are not disrupted by incorrect threat detections. For example, in a healthcare monitoring system, a high specificity value, like EMAML's 86.36% at 37k NTS, means that normal patient data is less likely to be incorrectly flagged as anomalous, thus avoiding unnecessary alarms and interventions.

Furthermore, in user-centric environments like online platforms or retail systems, maintaining high specificity is essential to ensure user convenience and trust. False positives in these systems can lead to user frustration and distrust. EMAML's high specificity, as seen in values like 89.50% at 45k NTS, suggests that it can provide effective security without compromising the user experience.

Additionally, in security-sensitive environments, such as in financial transactions or data privacy, high specificity minimizes the risk of legitimate activities being incorrectly flagged as malicious. This not only enhances the security but also ensures smooth operation and user satisfaction.

Thus, the comparative analysis of specificity across various models underscores the effectiveness of EMAML in accurately identifying non-adversarial activities. Its advanced ensemble method and active learning strategies contribute to its high specificity, ensuring its effectiveness and reliability in real-time scenarios. High specificity in EMAML minimizes false positives, thereby enhancing operational efficiency and user trust in systems where accurate identification of non-threatening activities is as crucial as detecting adversarial ones in real-time scenarios.

5. Conclusion & Future Scopes

This study has successfully presented the Efficient Ensemble Model for Detection of Adversarial Attacks in Machine Learning Environments (EMAML), an innovative framework adept at identifying adversarial attacks with notable accuracy and efficiency. The integration of diverse machine learning algorithms, including Logistic Regression, Random Forest, SVM, CNN, and XGBoost, complemented by the strategic implementation of active learning strategies, has proven to be highly effective.

The experimental results, derived from comprehensive tests utilizing the NIPS and DReLAB datasets, demonstrate the superior performance of EMAML over existing models like MetaGAN, NAGR, and CIS. EMAML exhibited remarkable improvements in key metrics such as precision, accuracy, recall, and AUC. Particularly noteworthy are its precision and specificity, which consistently surpassed other models, especially in larger test scenarios. The model's adaptability was further highlighted by its performance in real-time scenarios, where it effectively adapted to new adversarial tactics, as reflected in its low detection delays and high specificity.

The impact of this work is significant in the realm of cybersecurity and machine learning. By providing a more reliable and efficient method for detecting adversarial attacks, EMAML contributes to the enhancement of security in various applications, from critical infrastructure protection to data privacy. Its ability to minimize false positives and false negatives is crucial in maintaining operational efficiency and user trust, especially in sensitive environments.

Future Scope:

Looking forward, there are several avenues for further enhancing and expanding the capabilities of EMAML:

- **Integration with Emerging Technologies:** Exploring the integration of EMAML with emerging technologies like quantum computing and blockchain could potentially enhance its computational efficiency and security features.
- **Application in Diverse Domains:** Extending the application of EMAML to other domains such as healthcare, finance, and autonomous vehicles, where the detection of adversarial attacks is increasingly critical, could prove to be highly beneficial.
- **Handling More Sophisticated Attacks:** As adversarial attack methodologies evolve, future work could focus on augmenting EMAML to counter more sophisticated attacks, including those employing AI-generated deepfakes and advanced obfuscation techniques.
- **Expanding Dataset Diversity:** Utilizing a wider range of datasets, especially those representing more diverse and complex real-world scenarios, would help in further testing and refining the model's capabilities.
- **Enhancing Active Learning Strategies:** Investigating more advanced active learning techniques could improve the model's efficiency in learning from new data, thereby reducing the need for large labeled datasets.
- **Cross-Model Collaboration:** Future research could explore the potential of EMAML working in tandem with other defensive frameworks to create a multi-layered defense strategy.
- **User Behavior Analysis:** Incorporating user behavior analysis into EMAML could aid in distinguishing between malicious and benign activities more effectively, especially in scenarios with subtle adversarial tactics.

In conclusion, EMAML represents a significant step forward in the ongoing effort to secure machine learning environments against adversarial attacks. Its efficacy in real-time detection and adaptability to evolving threats positions it as a valuable tool in the cybersecurity arsenal. The future enhancements and applications of this model hold great promise for further strengthening the security and reliability of machine learning systems in an increasingly set of digital world scenarios.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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 SAMPLES.2023.3296936.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] R. Li, H. Liao, J. An, C. Yuen and L. Gan, "Intra-Class 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.

[17] 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.

[18] 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.

[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] 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.

[21] 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 SAMPLES.2022.3227318.

[22] H. Naderi and I. V. Bajić, "Adversarial Attacks and Defenses on 3D Point Cloud Classification: A Survey," in IEEE Access, vol. 11, pp. 144274-144295, 2023, doi: 10.1109/ACCESS.2023.3345000.

[23] Q. Liu and W. Wen, "Model Compression Hardens Deep Neural Networks: A New Perspective to Prevent Adversarial Attacks," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, pp. 3-14, Jan. 2023, doi: 10.1109/TNNLS.2021.3089128.

[24] 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.

[25] Q. Wang, J. Yao, L. Zhang, P. Guo and L. Xie, "Timbre-Reserved Adversarial Attack in Speaker Identification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3848-3858, 2023, doi: 10.1109/TASLP.2023.3306714.