E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 03-08-2023
Accepted: 12-10-2023

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 1; Issue 2; 2023; Page No. 447-451

Optimised curie pre-filter defending technique for SVM against poisoning attack

[1]Mahalle Sheetal Anil and [2]Dr. Kaushal Kumar

[1]Research Scholar, Sunrise University, Alwar, Rajasthan, India
[2]Assistant Professor, Sunrise University, Alwar, Rajasthan, India

DOI: https://doi.org/10.5281/zenodo.14617570

Corresponding Author: Mahalle Sheetal Anil

Abstract

In the contemporary business landscape, sustainability has become a critical factor in customer engagement strategies. This study explores the role of innovative green technologies in enhancing customer engagement, focusing on how businesses can leverage environmentally friendly practices to build stronger connections with their customers. This research used a mixed-methods strategy, drawing both quantitative and qualitative conclusions from customer surveys and case studies of businesses that have effectively implemented green technology. This study provides valuable insights for businesses aiming to align their customer engagement strategies with sustainability goals, offering practical recommendations for integrating green technologies into their customer outreach efforts.

Keywords: Optimized, Curie, Pre-Filter, Technique, SVM and poisoning attack

Introduction

In bioinformatics, as in intrusion detection and picture identification, Support Vector Machine (SVM) has seen usage. Binary SVMs and One-Class Support Vector Machines (OCSVMs) are able to deal with data that contains random noise. On the other hand, their performance can take a major hit if hostile actors deliberately alter the training examples. Machine learning models still need protection from these kinds of attacks, even as more and more applications are moving online.

A broad field that includes several assaults on different machine learning algorithms is known as adversarial learning. During the training phase, the enemy introduces the poisoned assault. This kind of attack compromises the data's integrity by altering the training data's labels or feature values. Attackers try to change the decision boundaries of the ML model from what it would have learnt if the training dataset wasn't skewed by compromising it. To do this, the adversarial samples are included into the training dataset. An adversary may use a specific method to generate adversarial samples depending on their objectives and the information they have about the learning system. Poisoning attacks may be categorized into two types: availability attacks and integrity attacks, depending on the attacker's purpose. By manipulating some of the attack/anomaly data points, an integrity attack may corrupt training data points. To cause the learnt model to mistakenly classify normal test samples as malicious or abnormal is the goal of an availability attack. This chapter will centre on support vector machines (SVMs) and investigate the following question: "Can we make SVMs more robust against poisoning attacks?"

Take this case into consideration: Using the MNIST dataset, a classifier is trained to recognize handwritten numbers. The enemy's goal is to train the learning model to incorrectly identify the numbers. This is accomplished when the attackers alter the label value, add image perturbations, or alter the constructed feature values. Consequently, the misclassification is handled by the learnt model.

Literature Review

Atharv Raotole *et al*. (2023) [1] Autonomous cars, internet security, and speech recognition systems are just a few of the many areas where machine learning models are becoming indispensable. However, since machine learning is so widely used, these systems are open to malicious assaults. With two primary points of emphasis, this article explores the complex field of Adversarial Machine Learning

as it pertains to cybersecurity. The first section covers adversarial techniques for manipulating machine learning models, including data poisoning and evasion attacks, among many others. Secondly, by strengthening the cybersecurity of machine learning systems, it aims to create and suggest efficient ways and techniques for protecting against these hostile assaults. Experiments on the effects of successful adversarial assaults highlight the far-reaching effects on AI system trust, integrity, and security.

Ricky Laishram *et al*. (2016) [2] Biometric user authentication, speaker identification, and other security-related applications all make use of machine learning. Poisoning attacks are a kind of causal integrity attacks on machine learning. Their purpose is to increase the classifier's false positive rate by tampering with the training data. The speaker identification system will reclassify user A as user B, and more unauthorized persons will be considered genuine users in the context of biometric authentication. Here, we examine the poisoning attack on SVMs and provide Curie, a method to protect the SVM classifier against it. Finding and removing the poisoned data points that were intentionally added is the fundamental concept of our technique. Because of how lightweight our technology is, it is very easy to include into preexisting systems. It successfully filters out the poisoned data, according to the experimental findings.

Deepak Upreti *et al*. (2022) [4] The significance of data mining in industrial engineering has grown as the value of processing power and storage capacity has risen. The field of industrial engineering has recently seen remarkable progress because to AI and ML. One method of machine learning that aims to solve the problem of data privacy in distributed computing systems and their data storage applications is federated learning. We have examined the effectiveness of Tolpegin's proposed defense technique and extended the work of Tolpegin *et al*. about data poisoning concerns in federated learning systems. Following that, we evaluated the efficacy of several clustering methods, such as K-means, PCA, KPCA, and UMAP. In comparison to PCA, KPCA, and K-means, UMAP provides better performance in avoiding data-poisoning attacks, according to the results.

Xiaojiao Chen *et al*. (2021) [5] With their widespread deployment, voice processing systems (VPSes) are now an integral part of people's everyday life, assisting with driving, unlocking smartphones, making online purchases, etc. Since systems developed using deep neural networks are vulnerable to malicious instances, virtual private server security has recently received a lot of attention. This paper offers a thorough review of the literature on adversarial attacks, including subjects like psychoacoustic models, evaluation indicators, and adversarial example production. Afterwards, we provide a brief overview of defence strategies that counter aggressive assaults. In the end, we provide a systematised way of categorising adversarial assaults and defence strategies, with the aim of helping newcomers to the subject better grasp the organisation and categorization.

Dr. Pratik S. Patel *et al*. (2023) [13] While deep learning models have been very successful in many areas, one major problem is how susceptible they are to adversarial assaults. In order to cause inaccurate predictions and possible security breaches, adversarial attacks take use of deep learning models' inherent flaws. In order to make deep learning models more resilient against adversarial assaults, this research aims to analyze the challenges and provide solutions. By delving into the theoretical foundations of adversarial resistance and creating unique define mechanisms, this research seeks to strengthen the security and dependability of deep learning models against malicious attacks.

Optimised Curie-Defense Method

By including poisoned data points into the training data during online training, the suggested defensive technique strengthens the classifier model's resistance to the label flipping assault. Figure 1 shows that the data buffer storage is implemented before the filter model.

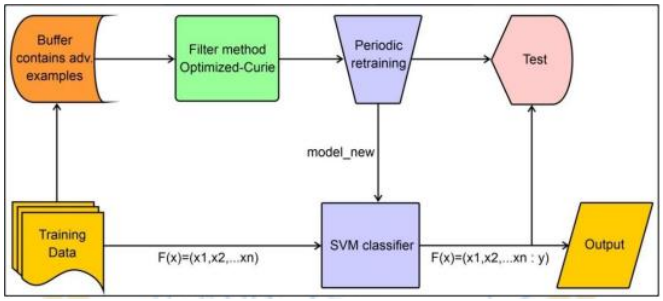

Fig 1: Periodical retraining with Optimised-Curie Filter method

In Figure 2, we can see the whole online learning process flowchart, where the enemy inserts a small number of poison data points into the buffer. New training data as well as data points that the adversary has altered are stored in this buffer. This buffer data is used to retrain the learned Classifier model if its accuracy falls below a certain threshold or time period. The accuracy of the model will be impacted if it retrains using this buffer data. This means that the training data must be cleansed of these "poison" data points. In order to exclude these altered points from further consideration during training, the optimised-Curie filter approach is used.

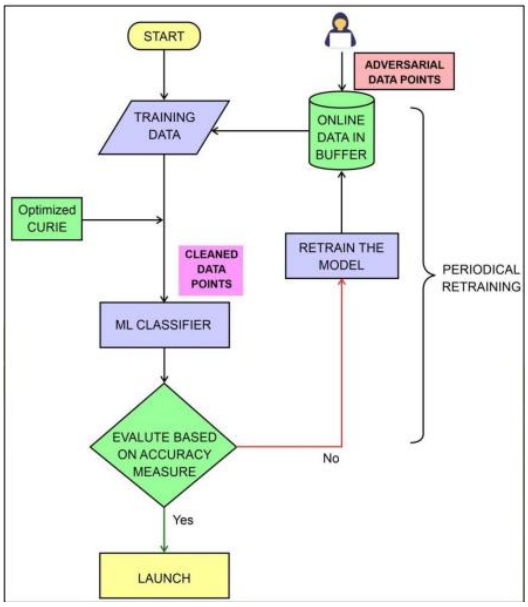

Fig 2: Periodical retraining process

Algorithm Overview

Figure 3 shows the process flow of the optimized Curie pre filter technique. This first stage of the process is based on the following principle: we group the data points using the DBSCAN algorithm. It is a density- based algorithm for clustering applications based on their spatial distributions, so DBSCAN is suitable. Each cluster point must have at least a certain number of neighbors within a certain radius; this is the algorithm's guiding concept. So, the generated clusters are in arbitrary shape. Unlike K-means clustering, which requires the size to be set beforehand, the DBSCAN technique does not need the number of clusters to be supplied. The DBSCAN method extremely handles the outliers and noise datasets.

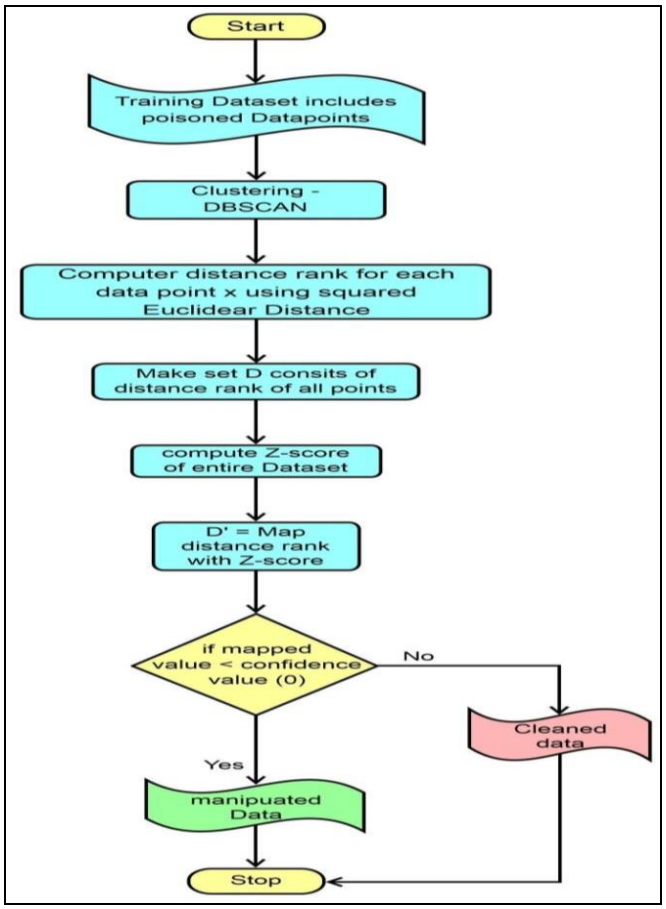

Fig 3: Flow chart of Optimised -Curie Filter method

We will pretend that F stands for the whole dataset and that $C'(x)$ is the cluster that data point x belongs to, denoted as clus(x). The collection of training data that is part of the same cluster as x is denoted as $C'(x)$. Afterwards, the point b is symbolized as

$$F(x) = \{x_1, x_2, ..., class(x)\} \quad (3.1)$$

In this case, a weight ʊ is assigned to the class (x). Among the hyper-parameters, it is one.

A revised version of Equation (3.2) is provided as

$$F(x) = \{x_1, x_2, ..., \omega class(x)\} \quad (3.2)$$

The next step is to find the mean Euclidean distance for each data point in x. It is defined below and is defined as a distance rank for x; it is written as d_rank(x).

$$d\_rank(x) = \frac{1}{F'(x)} \sum_{y \varepsilon F'(x)} dt(x, y) \quad (3.3)$$

dt(x,y) is the distance in geometric terms between x and y. All training data points from dataset F are considered to belong to a set of distance rankings denoted as D_all. Then

$$D\_all = d\_rank(y) \in F \quad (3.4)$$

The following standard formula is used to determine the Z-score of a current input data point x, which is the next step.

$$Z\_score(x) = \frac{x - \mu}{\sigma} \quad (3.5)$$

$Z\_score\ (x) = \frac{x-\mu}{\sigma}$ Where μ = Arithmetic mean
σ = standard deviation
Next, replace all the values in $\boldsymbol{P}\_all$ with their corresponding values in Z_ycofe.
Let D'_neq stand for the revised dataset, which is provided by:

$$D'\_new = \{Z\_score(x) \mid \forall x \in D\_all\} \quad (3.6)$$

Moreover, after that, we need to find the locations with low confidence values (confidence (θ)) and eliminate them. Therefore, the final collection is

$$res^{+} = \{y \mid y \in F \wedge D'\_new(y) \leq \theta \quad (3.7)$$

The confidence value is denoted by θ, and the new data set res+ is used for training purposes. It is devoid of stained data points. Before the training process begins, one of the hyper-parameters, the confidence value (θ), is initialized. For common datasets like MNIST, the current approach (Curie, a filter method) effectively filters poison points with a constant value of θ of 1.645. Fixing the optimal value for θ is crucial for applying the filter technique to any dataset in order to eliminate the altered data points. In this suggested study, several datasets are used to determine the optimal value for θ. To find the best value for θ, the complete findings are covered in the Experiments section.

Dataset description

Five separate datasets from the machine learning repository at UCI are used to test the suggested technique. These datasets include sonar, Australian, Bupa, CMC, and breast cancer. The specifics of these databases are laid forth in Table 1. In column 2, we can see the number of instances, and in column 3, we can see the number of characteristics. When dealing with non-linearly separable data, the basic Support Vector Machine (SVM) technique is ineffective, hence in both binary and multi class classifications, the kernel SVM classification approach is used.

Table 1: Dataset details

Dataset Name	#Samples	#Features	Number of Classes
Australian	690	15	Binary Class
Bupa_Liver	345	7	Binary Class
Cmc,	1473	10	Multi Class (Three)
Sonar	208	61	Binary Class
Breast Cancer	286	9	Binary Class

Using the AdversariaLib package, 10% of poison points are created for each class across all five datasets. the method is applied until the filtering process is finished with the only objective of removing poison spots.

Experimental evaluation and analysis

In the first experiment, we compare the original dataset with the dataset that contains adversarial cases in order to calculate the False Positive Rate (FPR) and classification accuracy. The SVM Classifier is used to calculate the accuracy. Important parameters utilized in support vector machines (SVM) are C, kernel, and gamma. C is the regularization parameter, and regularization has an inverse relationship with its value. Since a positive number for C is required for this experiment, we will set it to 1. If the algorithm's kernel type is rbf, the kernel coefficient is defined by gamma, and the kernel type is specified by kernel. The gamma value is set to scale, while the kernel value is set to rbf. Table 2 shows the outcomes of the kernel SVM classifier's performance both before and after the assault.

Table 2: Accuracy and FPR before and after adversarial attack

Dataset Name	Before Attack		After Attack	
	Accuracy	FPR	Accuracy	FPR
Australian	58%	0.4	52%	0.9
Bupa_liver	54%	0.1	50%	0.4
CMC	55%	0.2	53%	0.4
Sonar	56%	0.7	33%	0.8
Breast cancer	69%	0.03	65%	0.12

The classifier's accuracy drops and its FPR rises once poison points are added. These findings demonstrate that the classifier's performance is diminished by adversarial assaults in different datasets. It has been observed that the classifier's performance is significantly diminished when the fraction of poison spots grows.

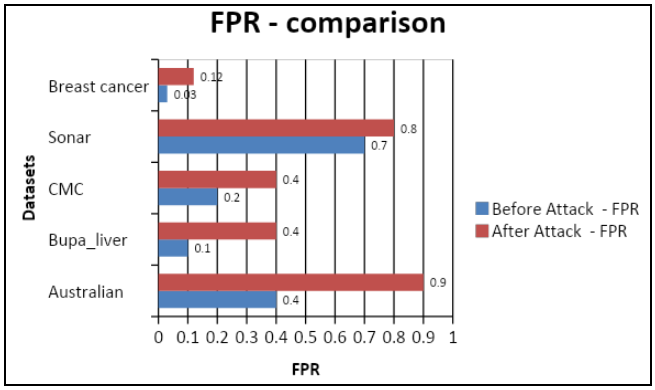

Fig 4: FPR Comparison

The hyper-parameter θ is adjusted in the next experiment to make the suggested algorithm filter more altered data points. The method for creating the poison spots is identical to that of the first experiment. All five datasets had stained points introduced into them at a rate higher than 5%. Table 3 shows the specifics of the dataset after the poisoning attack. Different values are substituted for θ in this experiment, which is then applied to all five datasets. The results of this study are shown in Table 4. The visualization of different θ values vs percentage of filtered adversarial data points of all five datasets. Reducing the value of θ (where θ is 1.645 in the conventional Curie approach) results in a lower proportion of filtered stained sites, as seen. Just as before, when stained points are removed, the value of θ goes up. For all values of θ except 1.2, the graphs show a decreasing proportion of filtered points, with the largest number of points filtered occurring at that value. The normal distribution curve is followed by the hyper-parameter Θ against percentage of filtered points curve, according to the results of this experiment. The findings show that 1.2 is the best value for the hyper-parameter Θ, and this specific experiment indicates that this changed value works better for all the data.

Table 3: Details of data points in the datasets

Dataset	Actual	Poison points injected	Total
Australian	690	35	725 (5%)
Bupa_liver	345	14	359 (5%
CMC	1473	98	1571 (7%)
Sonar	208	17	225 (8%)
Breast cancer	286	20	306 (7%)

Table 4: Percentage of filtered data points with various values of Θ

Dataset Name	Hyper parameter Θ				
	Θ=0.75	Θ=1	Θ=1.2	Θ=1.65	Θ=2.25
Australian	11	26	35	28	12
Bupa_liver	72	79	93	79	74
CMC	53	62	90	55	52
Sonar	59	56	70	55	60
Breast cancer	69	74	82	75	70

The suggested procedure is run with all five datasets (with θ=1.2 in mind) and the results are shown in Table 5. The same is shown graphically in Figures 5 and 6. In comparison to the most current technique, Curie, our suggested algorithm effectively filters a larger number of modified data points, according to the experimental findings.

Table 5: Percentage of filter points – Curie Vs Optimised Curie

Dataset Name	With Curie			With Optimised-Curie		
	Cleaned points	Filtered points	Percentage of filtered points	Cleaned points	Filtered points	Percentage of filtered points
Australian	720	5	14%	705	20	57%
Bupa_liver	348	11	79%	346	13	93%
CMC	1540	31	32%	1483	88	90%
Sonar	215	10	59%	207	16	94%
Breast cancer	295	11	55%	290	16	80%

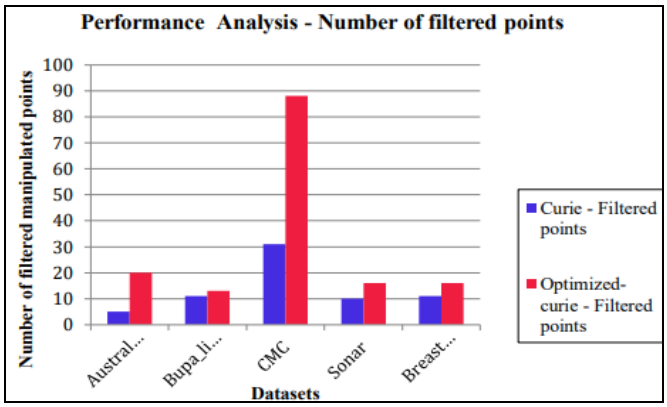

Fig 5: Comparison of proposed method with existing technique: Number of filtered manipulated points

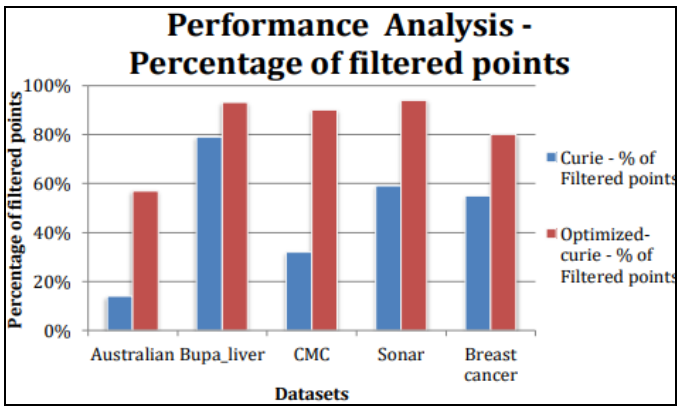

Fig 6: Comparison of proposed method with existing technique: Percentage of filtered manipulated points

Table 6: Accuracies of Kernel SVM: before and after Optimised-Curie Filter method

Dataset Name	Accuracy	
	Before Filtering	After Filtering
Australian	52%	57.6%
Bupa_liver	50%	53.5%
CMC	53%	55%
Sonar	33%	55%
Breast cancer	65%	68%

Conclusion

Examining how adversarial perturbations affect ML and DL algorithms and creating additional safeguards to make them more resilient to such assaults is the primary goal of this effort. Literature reviews show that adversarial perturbations have a significant impact on SVMs, OCSVMs, and regression approaches, ML algorithms that are widely used. Similarly, hostile training examples may influence Deep Neural Network (DNN) models as well. To safeguard the SVM model against malicious input data, the Optimised curie technique was created. This prefiltering model uses data distribution to adjust the hyperparameter. Using both traditional and real-time datasets, we evaluate the algorithm's efficacy. In an adversarial setting, we evaluated five datasets for their classification accuracy and the proportion of poisoned points that were filtered out. Since the attacker also launch an evasion attempt to inject poison points, the Optimised Curie technique has taken poisoning attacks into consideration. Potentially, in the future, an evasion attack may be used to evaluate the method's performance. Given that the Optimised-Curie algorithm has been tried and proven to safeguard the SVM classifier, it ought to be feasible to use the same approach to safeguard additional classifiers like Multilayer Perceptron and Naïve Base classifier, among others.

References

1. Raotole A, *et al*. Securing the Digital Fortress: Adversarial Machine Learning Challenges and Countermeasures in Cybersecurity. International Research Journal of Engineering and Technology (IRJET). 2023;10(11):2395-0056.
2. Laishram R, Phoha V. Curie: A method for protecting SVM Classifier from Poisoning Attack. 2016.
3. Weerasinghe S, *et al*. Defending Distributed Classifiers Against Data Poisoning Attacks. arXiv:2008.09284v1 [cs.LG]. 2020.
4. Upreti D, *et al*. Defending against Label-Flipping Attacks in Federated Learning Systems with UMAP, 2022. DOI: https://doi.org/10.21203/rs.3.rs-1984301/v1.
5. Chen X, Li S, Huang H. Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview. Appl. Sci. 2021;11(8450). https://doi.org/10.3390/app11188450.
6. Li S, Wang J, Wang Y, Zhou G, Zhao Y. EIFDAA: Evaluation of an IDS with function-discarding adversarial attacks in the IIoT. Heliyon. 2023;9(2):e13520. DOI: 10.1016/j.heliyon.2023.e13520. PMID: 36846700; PMCID: PMC9950836.
7. Zhu Y, Wang M, Yin X, Zhang J, Meijering E, Hu J. Deep Learning in Diverse Intelligent Sensor Based Systems. Sensors. 2023;23(62). https://doi.org/10.3390/s23010062.
8. Jenkin Winston J, *et al*. Hybrid deep convolutional neural models for iris image recognition. DOI: 10.1007/s11042-021-11482-y. 2021.
9. Raviya K, *et al*. Deep CNN With SVM - Hybrid Model For Sentence-Based Document Level Sentiment Analysis Using Subjectivity Detection. DOI: 10.21917/ijsc.2021.0335.
10. Lee M, Kang JK, Yoon H, Park K. Enhanced Iris Recognition Method by Generative Adversarial Network-Based Image Reconstruction. IEEE Access. 2021;PP: 1-1. 10.1109/ACCESS.2021.3050788.
11. Wu Y, He Y, Wang Y. Multi-Class Weed Recognition Using Hybrid CNN-SVM Classifier. Sensors. 2023;23(7153). https://doi.org/10.3390/s23167153.
12. Thiyagarajan S. Performance Comparison of Hybrid CNN-SVM and CNN-XGBoost models in Concrete Crack Detection. Master's Thesis, Technological University Dublin. 2019.
13. Patel PS, Navik TS, Ahuja S. Reinforcement Learning for Adaptive Cybersecurity: A Case Study on Intrusion Detection. AIDE-2023 and PCES-2023. 2023;114.