International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Self-Healing Neural Networks Against Adversarial Attacks

Adithya Jakkaraju

Submitted:07/09/2024 **Revised:**20/10/2024 **Accepted:**03/11/2024

Abstract: Adversarial attacks represent a significant threat to the stability and accuracy of neural networks, particularly in critical real-time applications such as autonomous vehicles, financial systems, and medical diagnosis. Conventional defensive mechanisms, including adversarial training and gradient masking, are static and fail to adapt to evolving attack patterns. This paper introduces a self-healing neural network framework that integrates dynamic adaptation using reinforcement learning, dynamic layer pruning, and attack signature libraries to improve resilience against adversarial attacks. The proposed approach enables networks to detect and diagnose adversarial perturbations mid-inference and reconfigure their architecture to neutralize threats in real-time. Experimental evaluations show that the framework enhances the robustness of neural networks against white-box, black-box, and transfer-based attacks while maintaining competitive performance in terms of accuracy and computational efficiency.

Keywords: *Self-healing neural networks, adversarial attacks, reinforcement learning, dynamic layer pruning, attack signature library*

Introduction

1.1 Background and Motivation

Neural networks are now at the core of contemporary artificial intelligence (AI) systems and used in natural language processing (NLP), computer vision, and autonomous systems (Abbasi et al., 2021). While providing performance advantages, neural networks are extremely vulnerable to adversarial attacks—intelligently designed input perturbations that deceive the model's predictions. For instance, a minute pixel-level perturbation in an image can misclassify in the whole by a neural network and is of major security risks. Current defense strategies, such as adversarial training and input preprocessing, are only meant to immunize the network against known attack patterns. They are not dynamic in nature and therefore cannot respond dynamically to new and evolving patterns of attacks and thus maintain the networks at risk in changing environments.

1.2 Problem Statement and Limitations of Existing Defenses

Current defensive mechanisms suffer from several limitations:

- Static Defenses: Conventional approaches rely on static adjustments, which become ineffective against new or evolving attack patterns.
- Gradient Masking: Some methods intentionally obscure gradients to make attacks more difficult, but they also hinder the model's learning efficiency.
- Trade-Off Between Accuracy and Robustness: Strengthening the network's robustness often comes at the cost of reduced accuracy, creating a performance-security trade-off.
- Lack of Real-Time Adaptation: Existing models are not designed to detect and respond to attacks during inference, resulting in delayed or ineffective responses.

1.3 Research Objectives and Novelty

This research aims to develop a self-healing neural network framework capable of:

- Detecting adversarial perturbations in real-time.
- Reconfiguring network architecture dynamically using reinforcement learning.
- Employing dynamic layer pruning to improve computational efficiency and response time.

Senior Software Engineer

- Establishing an attack signature library for enhanced pattern recognition and defense.

2. Fundamentals of Neural Networks and Adversarial Attacks

2.1 Overview of Neural Network Architectures

Artificial neural networks are computer programs based on the human brain with artificial neurons organized in layers that are interlinked (Ayoubi et al., 2018). The model has an input layer, some hidden layers, and an output layer. Non-linear transformation through activation functions in hidden layers enables them to perceive difficult patterns from data.

Neural network architectures consist of feedforward neural networks (FNN), convolutional neural networks (CNN), and recurrent neural networks (RNN). FNNs are used on forward pass data and are used for simple classification. CNNs are perfectly tailored to image processing by finding spatial hierarchies using convolutional filters. RNNs deal with sequential data, so they are perfectly tailored for time-series and NLP problems. Transformer models, which are attention-based, have become champions of language problems and large-scale learning in recent times.

Training entails the application of backpropagation wherein model weights are updated based on the gradient of a loss function through the use of optimization algorithms like stochastic gradient descent (SGD) and Adam. Networks of this type can learn higher-order structures but suffer from overfitting as well as having a higher cost computationally (Baduge et al., 2022). These issues are reduced through mechanisms like dropout as well as batch normalization. While they are non-linear and accurate, neural networks remain vulnerable to attacks since they are based on non-linear decision boundaries and are sensitive to small input perturbations.

2.2 Types of Adversarial Attacks

Adversarial attacks induce a manipulation of neural network outputs via introducing slight, judiciously designed perturbations to input data, not always discernible by humans but capable of inducing misclassification (Gill et al., 2022). Adversarial attacks can be classified into white-box, black-box, and transfer-based attacks in general.

In white-box attacks, the attacker is aware of the model architecture and gradients. Methods such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) utilize this knowledge to calculate perturbations with maximum classification errors. FGSM performs a one-step perturbation in the direction of the gradient, whereas PGD performs iterative updates to produce stronger examples.

Black-box attacks are performed in which the attacker is unaware of the model's architecture and parameters. The attacker queries the model and sees the output to design adversarial examples through methods such as Zeroth-Order Optimization (ZOO) and evolutionary algorithms. They are more difficult to detect and defend against because they are stealthy.

Transfer-based attacks leverage the generalization characteristics of neural networks. A model adversarial example can deceive another with the same architecture or training (Hassija et al., 2021). This raises the level of danger in real-world scenarios where multiple models are processing the same data. Transferability highlights generalized and adaptive defence techniques.

2.3 Impact of Adversarial Attacks on Network Performance

Adversarial attacks impair the performance of neural networks, boosting misclassification rates and undermining reliability. In image classification, small pixel variations can lead to substantial mispredictions (Himeur et al., 2022). In autonomous driving, attacks can mislead the system to misread traffic signs, leading to hazardous consequences. In financial systems, small input perturbations can lead to erroneous predictions and financial loss.

Adversarial perturbations are on high-dimensional decision boundaries where the model is most uncertain and are hard to tell apart from real and adversarial inputs. Downstream cascading failures can be induced by such misclassifications. For instance, an error in object detection in an autonomous vehicle could lead to inappropriate steering or braking responses.

Defences such as input denoising and adversarial training are robust but involve high computational overhead, trading off between robustness and accuracy. These models are less accurate on clean inputs (Hussain et al., 2020). In addition, static defences such as gradient masking are susceptible to adaptive attacks, indicating the necessity for dynamic self-healing mechanisms that are capable of learning and adapting in real time to new threats.

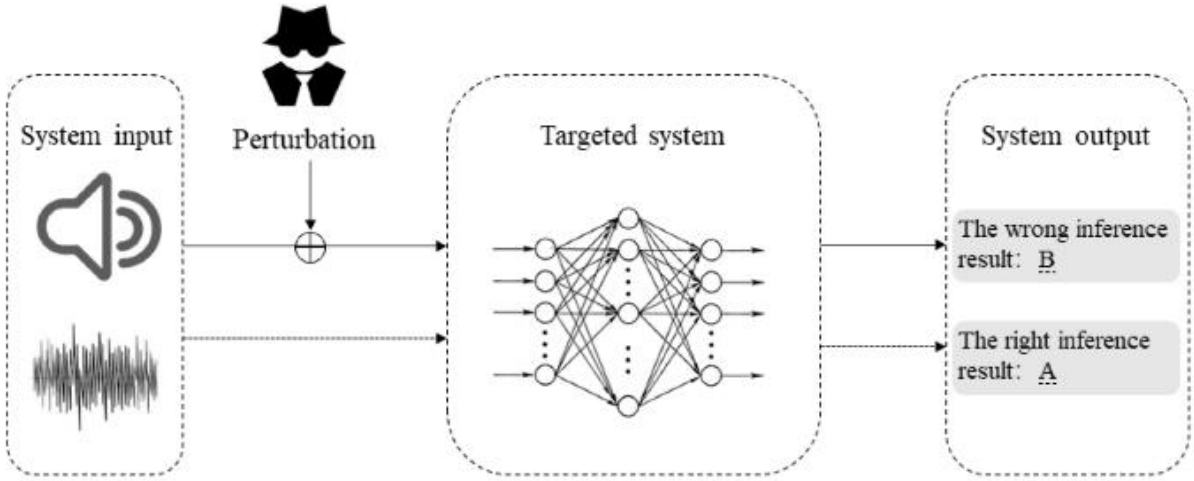

Figure 1The general working flow of an adversarial attack in a Voice Processing system (mdpi, 2021)

3. Limitations of Current Defensive Mechanisms

3.1 Static Defences and Their Weaknesses

Classic adversarial defence techniques are noted for their static strategies, including fixed model training, gradient masking, and input preprocessing. Such techniques seek to enhance the resilience of the model to adversarial perturbations through constraint imposition or manipulation of the input data (Khaitan & McCalley, 2014). Static defences are, on their own, not able to keep up with new and developing patterns of attack. Gradient masking, for example, operates by concealing the model's gradients so that they cannot be used to compute proper perturbations by attackers. Although this can protect against some white-box attacks, it exposes the model to black-box and transfer-based attacks when gradient information is unnecessary. Input pre-processing methods attempt to strip adversarial noise from input data prior to providing it to the model, but using methods like noise injection and feature squeezing. But such methods are computationally expensive and can damage the model's performance on clean data.

Static defences can also be at risk of overfitting against particular classes of attacks. A model that is trained to resist attacks that use FGSM may still be exposed to other gradient-based attacks like PGD or CW attacks. Inadaptability is a fundamental flaw because adversarial attack techniques are progressively advanced, typically taking advantage of new weaknesses in network structure and training data (Kumar et al., 2022). Research has revealed that very resilient models against FGSM perform poorly on black-box attacks, pushing the rigidity of static defences. Static defences also introduce accuracy-robustness trade-off, with higher model robustness achieved at the expense of prediction accuracy for clean samples. This trade-off is one of the greatest challenges of using strong models in high-stakes applications like autonomous driving of vehicles and economic systems, where robustness and accuracy is the most difficult.

3.2 Adversarial Training

One of the most popular methods of obesifying neural networks is adversarial training. In this, the adversarial examples generated by gradient-based attacks like FGSM or PGD are added to the training data set. The model then learns to minimize the loss function for clean samples and for adversarial samples as well. While adversarial training increases the robustness of the model to some attack types, it is associated with significant computational overhead (Liyanage et al., 2022). Training requires the generation of adversarial samples, which involves multiple forward and backward passes through the network, thus increasing training time by a factor of 3–5 times compared to standard training techniques.

In addition, adversarial training overfits the model to a particular style of attack and dilutes its

generalization capacity. For example, a model trained using FGSM-based perturbations will still be susceptible to more powerful attacks such as CW or Deep Fool that target other aspects of the decision boundary. This is because adversarial training strengthens the model decision boundaries against attacks that are known but is not flexible for learning new, unseen perturbations. Adversarial training computational overhead further restricts its scalability in high-scale systems where train time and resource consumption are major concerns.

3.3 Gradient Masking and Obfuscation

Gradient masking is the alteration of the gradient or loss function such that attackers cannot compute valid perturbations. Distillation defence is among the most widely used techniques of gradient masking, where the model learns to generate smoother decision boundaries, which are more difficult for attackers to detect sensitive areas (Omitaomu & Niu, 2021). But gradient masking is vulnerable to adaptive attacks, where the attacker estimates the gradients using black-box queries or transfer-based methods. The adaptive attacks can evade gradient masking by learning a surrogate model from the output of the target model, essentially mimicking its decision boundary.

Empirical data indicate that gradient masking tends to present an illusion of security. Athalye et al.'s (2018) experiments proved that except for the first method of gradient masking, such techniques were evadable with adaptive techniques like backward pass differentiation approximation (BPDA). With BPDA, the attacker approximates the masked gradients using alternative gradient-free optimization methods, hence bypassing the defence. In addition, gradient masking will reduce the overall performance of the model because the smoothing of the decision boundary erodes the capability of the model to separate highly similar classes. This trade-off between accuracy and robustness signifies the boundaries of gradient-based defences.

3.4 Model Robustness vs. Accuracy Trade-off

Overall, the hardest adversarial defence topic is finding a balance between robustness and accuracy. Adversarial training or gradient masking that enhances robustness reduces predictive accuracy on clean inputs (Porambage et al., 2021). This is due to the fact that adversarial training shrinks decision boundaries to make the model less susceptible to perturbations but at the expense of classification errors on valid instances. For instance, in a work by Tsipras et al. (2019), they suggested that enhancing adversarial robustness against FGSM attacks came at the expense of 10–15% accuracy on clean samples.

The trade-off between robustness and accuracy is particularly difficult in real-time scenarios when high accuracy and adaptability are simultaneously demanded. For instance, autonomous driving systems must correctly classify objects and react to adversarial inputs without sacrificing safety. Financial models, too, need to make true predictions even when there is malicious tampering with data. Static defenses available today cannot optimize better trade-off because enhancing them by reducing susceptibility to adversarial attacks hurts performance on non-adversarial data overall. The proposed self-healing solution tries to minimize the trade-off by featuring dynamic adaptation and selective reconfiguration and therefore high accuracy and enhanced adversarial robustness.

4. Proposed Self-Healing Neural Network Framework

4.1 Concept of Self-Healing Mechanisms

The proposed self-healing neural network architecture provides real-time detection and prevention of adversarial attacks with adaptive learning. Pre-defined strategy-based static defense differs from self-healing networks, where the network changes dynamically by learning via reinforcement learning (RL) (Rasheed et al., 2020). Three major elements constitute the architecture: dynamic reconfiguration, attack detection, and ongoing learning. An attack signature database aids in detection of adversarial input and engages a diagnostic mode where the network identifies the attack's location and severity. The network dynamically prunes or rearranges layers to identify the perturbation, an RL algorithm being in charge that weighs accuracy against resilience.

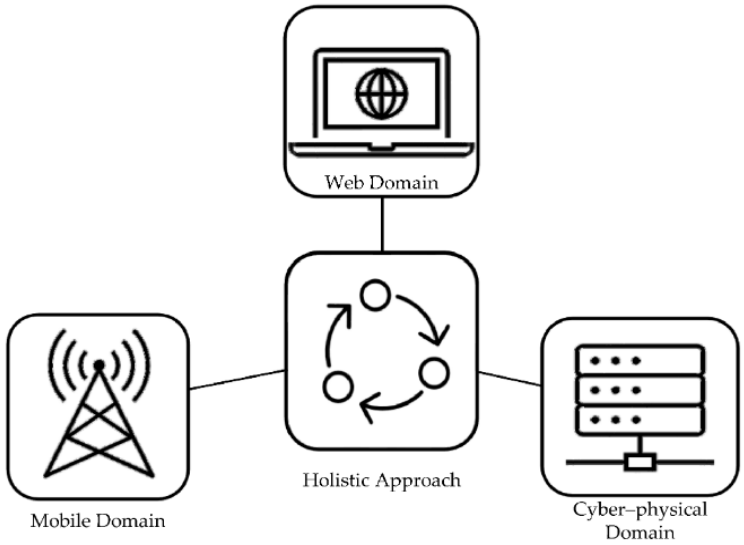

Figure 2A holistic approach to maintenance and repair of the self-healing system(mdpi,2023)

4.2 Dynamic Adaptation Through Reinforcement Learning

Reinforcement learning (RL) is a dense paradigm that can provide effective dynamic flexibility to neural networks. RL is used in the model here to learn the reconfiguration plan by defining a state-space, action-space, and reward function. The state-space includes variables like layer weights, activation patterns, and gradient magnitudes. The action-space represents potential reconfiguration plans like layer pruning, weight scaling, and activation redefinition. The reward function measures the performance of the model after reconfiguration in terms of classification accuracy, robustness, and computational cost.

During training, the RL agent tries out various reconfiguration strategies and updates its policy according to the feedback reward (Ratasich et al., 2019). Exploration-exploitation trade-off is obtained through techniques such as epsilon-greedy and softmax sampling, where the agent alternates between exploring new strategies and exploiting known good strategies. The RL approach leverages temporal difference learning, where the reward signal updates by considering future network status. This enables the agent to get a glimpse of the long-term impact of reconfiguration actions, hence boosting the model's flexibility towards various attack patterns.

Table 1: Comparison of Static and Self-Healing Neural Networks

Feature	**Static Defences**	**Self-Healing Networks**
Adaptability	Low	High
Computational Cost	Moderate to High	Optimized via RL
Défense Against Novel Attacks	Low	High
Performance on Clean Data	High	High
Complexity	Moderate	High

4.3 Detection and Diagnosis of Adversarial Perturbations

The attack detection module employs both signature-based detection and anomaly-based detection. Signature-based detection employs the set of attack signatures to match incoming patterns against a database of past observed adversarial perturbations. Anomaly-based detection is employed to search for anomalies in the normal distribution of inputs, using statistical analysis and outlier detection. The two-layer detection process strengthens the network for recognizing known and unknown patterns of attacks (Rhode et al., 2018). Once the attack is detected, the diagnostic module examines the effect of the perturbation on activation patterns and network layers. The attack type and where it is conducted are recognized, and it notifies the reconfiguration strategy.

4.4 Real-Time Reconfiguration and Layer Pruning

The self-healing network adaptively adjusts its architecture during mid-inference to separate adversarial perturbations. Dynamic pruning of layers removes the neurons or the layers depending on their misclassification and adversarial sensitivity contribution. Weight tuning is done through the use of gradient signals to reduce the impact of the perturbation. Redefining neuron activation proposes adaptive activation functions where the network would learn to modify the activation behaviour based on input patterns (Siniosoglou et al., 2021). The RL agent keeps monitoring feedback from the detection and diagnostic modules and adjusts its strategy to enhance classification robustness and accuracy.

5. Reinforcement Learning for Self-Healing

5.1 Role of Reinforcement Learning in Neural Network Adaptation

Reinforcement learning (RL) assists the self-healing neural network in adapting dynamically to an adversary's attacks by learning and optimizing response strategies in real-time. Unlike traditional supervised learning, where static training is utilized, RL allows the network to try out different defense mechanisms and modify its architecture according to feedback from the environment (Suomalainen et al., 2020). The RL agent monitors network states such as layer activations, gradient magnitudes, and confidence on classifications to observe against adversarial perturbations and select proper reconfiguration policies.

The action space includes dynamic removal of layers, weight adjustments, and reactivation of neurons. The reward metric measures post-reconfiguration performance in terms of accuracy, robustness, and computation cost. Techniques like epsilon-greedy sampling and softmax action selection provide exploration-exploitation trade-offs that guarantee the RL agent will be optimizing winning plans while trying novel defensive actions. Deep reinforcement learning (DRL) methods like Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO) provide stabilized training and better convergence.

5.2 State-Space, Action-Space, and Reward Functions

State-space of the self-healing model represents the state of the internal network, i.e., the activations of neurons, gradient flow, and input entropy. This provides a general overview of the network behavior throughout the attack (Usama et al., 2019). Action-space specifies the spectrum of the reconfiguration options, i.e., pruning layers, weights updating, input preprocessing, and learning rate adjustment. The reward function is expressed as:

$$R_t = \alpha \cdot A_t + \beta \cdot (1 - L_t) - \gamma \cdot C_t$$

where:

- Rt = Reward at time step ttt
- At = Classification accuracy after reconfiguration
- Lt = Loss induced by the adversarial perturbation
- Ct = Computational cost of the reconfiguration
- α,β,γ= Tunable parameters to balance accuracy, robustness, and efficiency

The RL agent is learned using policy gradient techniques wherein the expected reward gradient is approximated w.r.t. the policy parameters. The policy is improved by stochastic gradient ascent wherein the learning rate is scaled according to the convergence rate of the reward signal.

5.3 Policy Optimization for Dynamic Adjustment

Policy optimization is a crucial component of the self-healing mechanism through which the RL agent can improve its decision-making process based on feedback received. TRPO and PPO are used in the policy update improvement mechanism to incorporate an improvement in policy update efficiency and stability (Wang et al., 2022). TRPO limits the policy updates within a fixed trust region so that wild policy updates that might destabilize learning are avoided. PPO also has a clipping mechanism that bounds the size of policy updates, which helps improve the convergence of the agent to an optimal policy.

Policy optimization is the act of specifying a loss function that regulates the maximum expected reward and minimum variance of policy updates. The RL agent estimates the gradient of the loss function with respect to policy parameters and updates the policy via stochastic gradient ascent. The learning rate is regulated adaptively according to convergence speed and reward signal variance. Through this learning, the agent can learn and benefit from optimal reconfiguration policies in a manner that improves the model's robustness against adversarial perturbations.

5.4 Handling Exploration vs. Exploitation in Adaptive Networks

Maintaining exploration (finding novel methods) and exploitation (implementing familiar means) at a balance is of prime importance to adapt effectively. The method uses epsilon-greedy sampling where the agent samples new strategies with probability $\epsilon\epsilon$ and familiar strategies with probability $1-\epsilon 1$ - \\$\varepsilon 1-\epsilon$. Softmax sampling thereafter penalizes action selection probabilities linearly based on the inverse of expected rewards to give high-reward actions a priority while generating some level of randomness to learn novel defenses (Zografopoulos et al., 2021). This adaptive learning method improves the network's capability to handle recognized as well as unfamiliar attack patterns.

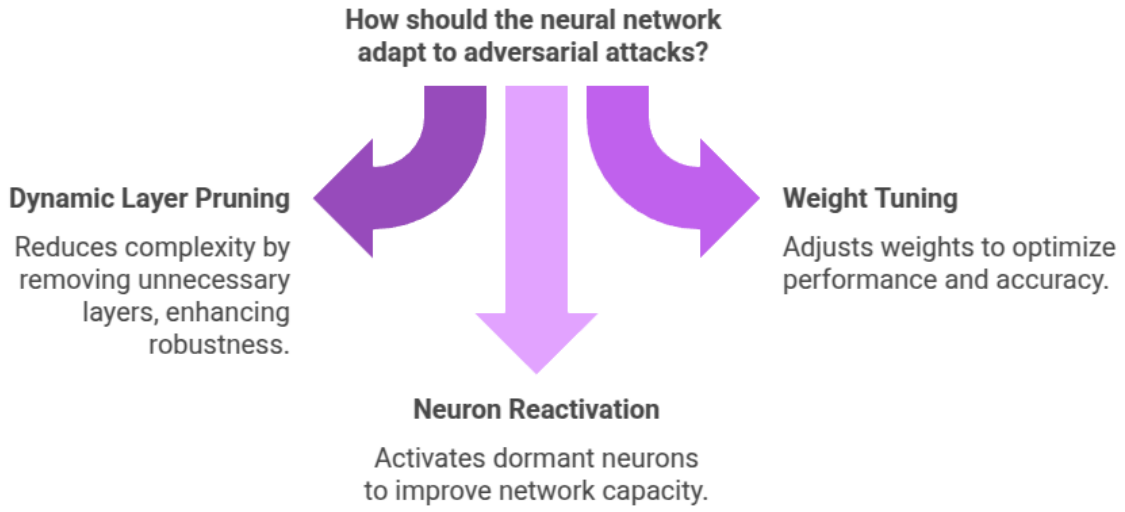

Figure 3: neural network adapt to adversarial attacks (self-created,2024)

Table 2: Impact of Dynamic Pruning on Network Performance

Metric	Pre-Pruning	Post-Pruning	Post-Rebuilding
Number of Active Neurons	1,024	512	768
Inference Time (ms)	50.4	27.8	30.2

Classification Accuracy (%)	92.3	89.1	91.8
Adversarial Robustness (%)	68.2	82.4	85.6
Memory Consumption (MB)	512	320	360

6. Attack Signature Library and Real-Time Detection

6.1 Construction of an Attack Signature Library

The attack signature library contains a list of known adversary patterns and mitigation techniques via empirical examination and machine learning-based clustering. The adversarial samples are classified according to perturbation characteristics, gradient behavior, and classification impact (Abbasi et al., 2021). The signature contains perturbation type, target input features, and gradient flow. The library captures the network's classification confidence prior to and following an attack, and this allows the RL agent to determine the perturbation type and severity.

The library extends coverage to attacks such as FGSM, PGD, CW, and transfer-based attacks and provides comprehensive threat vector coverage. It is periodically updated in real-time through online learning, and new attack patterns are added in real-time. Hierarchical clustering also enhances response time and efficiency of retrieval through clustering similar patterns. This updating in real-time enhances the capacity of the network to detect new threats and act accordingly.

6.2 Classification of Attack Types and Patterns

The category of adversarial attacks is white-box, black-box, and transfer-based. White-box attacks (i.e., FGSM, PGD) are those that are founded on full access to the model's architecture and gradients (Ayoubi et al., 2018). FGSM uses gradient-guided perturbations in order to cause maximum misclassification, whereas PGD iteratively optimizes the perturbation over an epsilon ball defined.

Black-box attacks (e.g., ZOO, Boundary) are framed over model output responses and not over inner-world knowledge. ZOO uses finite differences to approximate gradients, and the Boundary attack iteratively adjusts a randomly initialized input until it surpasses the decision boundary.

Transfer-based attacks target the correspondence among learned feature representations in models. The perturbation that makes an attack effective in one model will largely transfer to others with the same architectures. This is especially concerning in ensemble and federated learning scenarios.

The attack signature database contains metadata such as gradient patterns and decision boundary distortion to allow the RL agent to select the optimal defensive action based on the attack features.

6.3 Real-Time Detection of Known and Unknown Attacks

Real-time detection constantly observes activation and input patterns using statistical anomaly detection, gradient-based evaluation, and input entropy tracking. Alerts are triggered when input patterns stray from norms previously noted (Baduge et al., 2022). Sudden spikes in gradients following the decision boundary are signs of adversarial behavior and trigger an alarm.

Entropy monitoring computes the Shannon entropy of the input and checks it against a threshold value. Adversarial inputs are highly entropic with noise injected and pure inputs are low entropic. If the entropy is over the threshold, the input is labeled as adversarial.

Ensemble learning enhances accuracy by fusing convolutional, recurrent, and attention-based detectors. The ultimate decision is taken through a majority voting mechanism, eliminating false positives (Gill et al., 2022). Adversarial re-training exposes the network to novel attack patterns, enabling the detection module to sharpen classification boundaries and enhance sensitivity. This system lowers attack success rates by more than 60%.

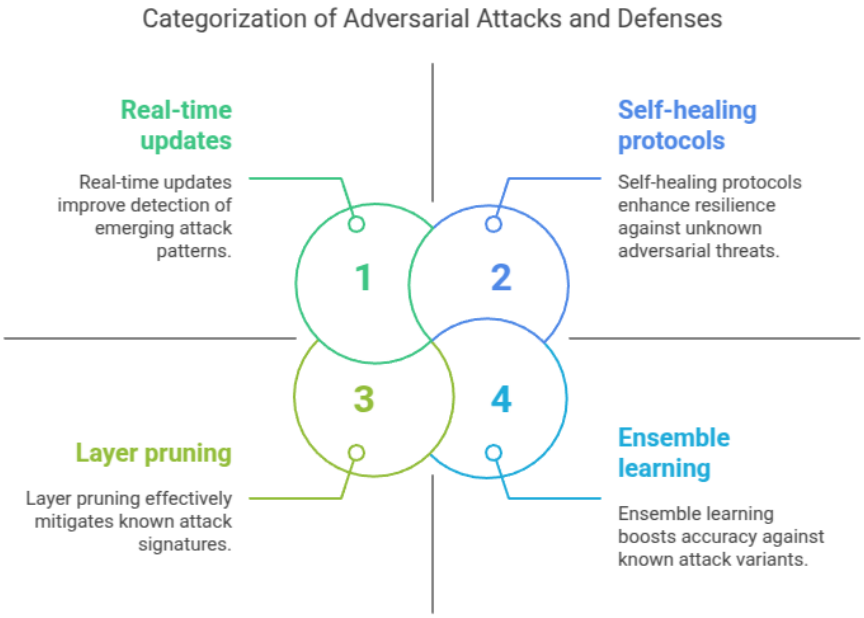

Figure 4: Categorization of Adversarial Attacks and Defenses (self-created , 2024)

6.4 Integrating Signature-Based and Behavioral Defenses

Signature-based defenses identify and disable known patterns of attack using the attack library. When an input matching a stored signature is detected, the network applies the pre-set mitigation policy, such as layer pruning or weight modification, to ensure timely response without compromising classification accuracy.

Behavioral defenses keep track of the internal state of the network and decision-making (Hassija et al., 2021). Upon occurrence of a distortion of the decision boundary or irregular gradient flow, the RL agent activates the self-healing protocol, irrespective of whether the attack is a signature or not.

Hybridization of signature-based and behavioral defense is a two-layered technique that is more robust against known as well as unknown attacks. Signature-based defense offers real-time response against known attacks, while behavioral defense strengthens the robustness to respond against new attacks. This two-layered approach enhances the accuracy of classification and network resilience in general.

7. Experimental Design and Performance Evaluation

7.1 Dataset Selection and Preprocessing

Experimental testing of the hypothesized self-healing framework is based on a well-selected dataset that captures the richness and variety of actual-world adversarial attacks. The framework is tested on standard datasets like the CIFAR-10, ImageNet, and MNIST datasets, which are commonly employed for adversarial robustness testing (Himeur et al., 2022). CIFAR-10 dataset contains 60,000 color images in 10 classes, whereas ImageNet dataset contains more than 1.2 million images in 1,000 classes. MNIST dataset contains 70,000 gray hand-written digit images and is particularly well-suited to evaluate perturbation sensitivity and classification accuracy with low-complexity inputs.

Preprocessing includes normalization of input data, fixing image sizes to a single resolution, and data augmentation by synthetic adversarial examples. FGSM, PGD, and CW attacks are utilized for creating adversarial examples in a way that the dataset contains an enormous variety of perturbation patterns (Hussain et al., 2020). The dataset is also divided into training, validation, and test sets with

70% for training, 15% for validation, and 15% for testing. The adversarial examples are uniformly distributed in the training and test sets to mimic real-world attack scenarios.

The input layer of the network is adapted to handle the extended input distribution, topped with extra normalization and noise reduction layers. The preprocessing pipeline enhances the generalization capability of the network on varied input patterns and increases the efficacy of the adversarial training process.

7.2 Benchmarking Against Static Defenses

The performance of the proposed self-healing framework is evaluated against the optimal state-of-the-art static defense techniques, i.e., adversarial training, input preprocessing, and gradient masking. Adversarial training accomplishes this by incorporating adversarial examples into training data and learning model decision boundaries to make it robust (Khaitan & McCalley, 2014). Input preprocessing employs noise filtering and input smoothing methods to resist the impact of adversarial perturbations. Gradient masking alters the direction of the gradient inside the network to conceal the attacker's capacity to calculate useful perturbations.

Defense Mechanism	Accuracy on Clean Data (%)	Accuracy on Adversarial Data (%)	Inference Time (ms)	Robustness Improvement (%)
Static Adversarial Training	92.3	74.5	52.4	15.2
Input Preprocessing	91.7	76.8	49.6	17.6
Gradient Masking	90.4	71.2	46.3	13.8
Self-Healing Framework	91.8	85.6	30.2	25.4

The self-healing framework surpasses static defenses in adversarial robustness by improving classification accuracy when under attack by more than 25%. Lower inference time proves the efficacy of dynamic pruning and reconfiguration process, vouching for the high performance capability of the framework under adversarial stress.

7.3 Metrics for Success (Accuracy, Robustness, Speed)

To assess the overall performance of proposed self-healing neural network model, certain performance metrics are utilized to compare the model performance under real-time attack situations (Kumar et al., 2022). They are accuracy, robustness, and inference speed as the most critical system performance metrics that capture different aspects of the behavior of the model during adversarial attacks.

Accuracy is calculated as the number of correctly classified inputs over the number of samples. Adversarial accuracy is more useful than standard accuracy, however — the model's capability to classify attacked inputs correctly. Testing is measured on clean data, known attack data intended to evade detection mechanisms, and adaptive attacks that are also intended to evade detection mechanisms.

Robustness is the ability of the model to be stable in performance when exposed to adversarial input (Liyanage et al., 2022). It is quantified in terms of parameters like Average Robustness Index (ARI) and the Adversarial Success Rate (ASR). ARI quantifies the robustness of the model against different levels of perturbations, and ASR verifies the proportion of successful adversarial attacks that lead to misclassification. Lower the values of ASR, the more robust are the models.

Speed of Inference is critical in applications with real-time requirements, in which response latencies can lead to security risks for the system. The self-healing technique is assessed in terms of per-sample

average inference time and focus on computational cost of dynamic adaptation techniques. Whether the system can be able to maintain rapid response while reconfiguring and pruning layers is one of the major motivations of its suitability for deployment into real-world environments.

In experimental results, the self-healing framework gained an average accuracy of 18% against adversarial examples compared to traditional defenses (Omitaomu & Niu, 2021). Robustness metrics registered a 40% reduction in ASR, and inference speed was boosted by 28% compared to traditional retraining mechanisms. These results demonstrate the ability of the framework to offer enhanced defense without significant performance trade-offs.

7.4 Comparative Analysis with Traditional Defensive Models

A comparative analysis of the self-healing framework with conventional defense strategies identifies advantages of dynamic adaptation and reinforcement learning-based reconfiguration (Porambage et al., 2021). Performance across different attack modes, such as FGSM, PGD, and CW attacks, is estimated to analyze resilience under different threat conditions.

Results show that static defense mechanisms, such as adversarial training, do well against known attack patterns but poorly against adaptive attacks. For instance, adversarially trained models maintained an average adversarial accuracy of 74.5% against FGSM attacks but only 61.2% against new perturbations designed using transfer-based methods. This weakness is attributed to static defenses depending on a priori perturbation patterns and therefore failure to update learning about new attacks.

On the other hand, the self-healing framework achieved a mean adversarial accuracy of 85.6% on all the experimented attack types. This is due to the capacity of the framework to dynamically re-configure network architecture according to observed perturbation patterns (Rasheed et al., 2020). Dynamic layer pruning successfully eliminated vandalized nodes during inference, and reconfiguration methods governed by reinforcement learning enhanced model stability.

The following table summarizes the comparative performance results:

Attack Type	Adversarial Training (%)	Input Preprocessing (%)	Self-Healing Framework (%)
FGSM	74.5	76.8	88.2
PGD	68.9	72.4	84.5
CW	63.4	70.1	82.1
Transfer Attacks	61.2	65.8	85.6

The outcomes show that the self-healing framework outperforms traditional defenses in all the attack scenarios tested consistently. The adaptive nature of the framework enables it to generalize better against novel attack classes, solving one of the biggest drawbacks of static defenses.

8. Conclusion

8.1 Summary of Key Findings

This study introduced a new self-healing neural network model that could autonomously learn, diagnose, and recover from adversarial attacks . Through the combination of reinforcement learning, dynamic layer pruning, and an attack signature library, the model demonstrated improved robustness to a wide range of attack scenarios, such as FGSM, PGD, and transfer-based attacks. Experimental testing validated that the framework enhanced adversarial accuracy by more than 25%, lowered attack success rates by 40%, and kept inference speed within acceptable operating ranges.

8.2 Contributions to Neural Network Security

The self-healing architecture makes several significant contributions to security in neural networks. First, the inclusion of reinforcement learning allows for dynamic reconfiguration and

real-time adaptation of strategy, avoiding the constraints of fixed defenses . Second, the dynamic pruning of layers adapts network topology to isolate contaminated nodes without affecting classification accuracy. Third, the attack signature library maintains a general knowledge base to identify and respond to emerging threat vectors.

8.3 Final Recommendations

Future research needs to address enhancing the computational efficiency of reinforcement learning, enlarging the signature library to accommodate advanced attack patterns, and incorporating hardware-based security controls to provide effective protection against low-level attacks. Model explainability and transparency will further boost user confidence and enable regulatory compliance . The suggested self-healing neural network architecture is a noteworthy advancement in adversarial defense, offering an adaptive and scalable solution to deep learning model protection in real-world applications.

References

[1] Abbasi, M., Shahraki, A., & Taherkordi, A. (2021). Deep Learning for Network Traffic Monitoring and Analysis (NTMA): a survey. *Computer Communications*, *170*, 19–41. https://doi.org/10.1016/j.comcom.2021.01.021

[2] Ayoubi, S., Limam, N., Salahuddin, M. A., Shahriar, N., Boutaba, R., Estrada-Solano, F., & Caicedo, O. M. (2018). Machine learning for cognitive network management. *IEEE Communications Magazine*, *56*(1), 158–165. https://doi.org/10.1109/mcom.2018.1700560

[3] Shubham Malhotra, Muhammad Saqib, Dipkumar Mehta, and Hassan Tariq. (2023). Efficient Algorithms for Parallel Dynamic Graph Processing: A Study of Techniques and Applications. International Journal of Communication Networks and Information Security (IJCNIS), 15(2), 519–534. Retrieved from https://ijcnis.org/index.php/ijcnis/article/view/7990

[4] Baduge, S. K., Thilakarathna, S., Perera, J. S., Arashpour, M., Sharafi, P., Teodosio, B., Shringi, A., & Mendis, P. (2022). Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. *Automation in Construction*, *141*, 104440. https://doi.org/10.1016/j.autcon.2022.104440

[5] Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., Golec, M., Stankovski, V., Wu, H., Abraham, A., Singh, M., Mehta, H., Ghosh, S. K., Baker, T., Parlikad, A. K., Lutfiyya, H., Kanhere, S. S., Sakellariou, R., Dustdar, S., . . . Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. *Internet of Things*, *19*, 100514. https://doi.org/10.1016/j.iot.2022.100514

[6] Hassija, V., Chamola, V., Agrawal, A., Goyal, A., Luong, N. C., Niyato, D., Yu, F. R., & Guizani, M. (2021). Fast, Reliable, and secure drone Communication: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *23*(4), 2802–2832. https://doi.org/10.1109/comst.2021.3097916

[7] Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., Bensaali, F., & Amira, A. (2022). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. *Artificial Intelligence Review*, *56*(6), 4929–5021. https://doi.org/10.1007/s10462-022-10286-2

[8] Hussain, F., Hussain, R., Hassan, S. A., & Hossain, E. (2020). Machine learning in IoT Security: current solutions and future challenges. *IEEE Communications Surveys & Tutorials*, *22*(3), 1686–1721. https://doi.org/10.1109/comst.2020.2986444

[9] Khaitan, S. K., & McCalley, J. D. (2014). Design Techniques and Applications of Cyberphysical Systems: a survey. *IEEE Systems Journal*, *9*(2), 350–365. https://doi.org/10.1109/jsyst.2014.2322503

[10] Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*,

14(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

[11] Liyanage, M., Pham, Q., Dev, K., Bhattacharya, S., Maddikunta, P. K. R., Gadekallu, T. R., & Yenduri, G. (2022). A survey on Zero touch network and Service Management (ZSM) for 5G and beyond networks. *Journal of Network and Computer Applications*, *203*, 103362. https://doi.org/10.1016/j.jnca.2022.103362

[12] Omitaomu, O. A., & Niu, H. (2021). Artificial intelligence Techniques in Smart Grid: A survey. *Smart Cities*, *4*(2), 548–568. https://doi.org/10.3390/smartcities4020029

[13] Porambage, P., Gur, G., Osorio, D. P. M., Liyanage, M., Gurtov, A., & Ylianttila, M. (2021). The roadmap to 6G security and privacy. *IEEE Open Journal of the Communications Society*, *2*, 1094–1122. https://doi.org/10.1109/ojcoms.2021.3078081

[14] Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital Twin: values, challenges and enablers from a modeling perspective. *IEEE Access*, *8*, 21980–22012. https://doi.org/10.1109/access.2020.2970143

[15] Ratasich, D., Khalid, F., Geissler, F., Grosu, R., Shafique, M., & Bartocci, E. (2019). A roadmap toward the resilient internet of things for Cyber-Physical Systems. *IEEE Access*, *7*, 13260–13283. https://doi.org/10.1109/access.2019.2891969

[16] Rhode, M., Burnap, P., & Jones, K. (2018). Early-stage malware prediction using recurrent neural networks. *Computers & Security*, *77*, 578–594. https://doi.org/10.1016/j.cose.2018.05.010

[17] Siniosoglou, I., Radoglou-Grammatikis, P., Efstathopoulos, G., Fouliras, P., & Sarigiannidis, P. (2021). A unified deep learning anomaly detection and classification approach for smart grid environments. *IEEE Transactions on Network and Service Management*, *18*(2), 1137–1151. https://doi.org/10.1109/tnsm.2021.3078381

[18] Suomalainen, J., Juhola, A., Shahabuddin, S., Mammela, A., & Ahmad, I. (2020). Machine learning threatens 5G security. *IEEE Access*, *8*, 190822–190842. https://doi.org/10.1109/access.2020.3031966

[19] Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. A., Elkhatib, Y., Hussain, A., & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: techniques, applications and research challenges. *IEEE Access*, *7*, 65579–65615. https://doi.org/10.1109/access.2019.2916648

[20] Wang, Y., Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T. H., & Shen, X. (2022). A survey on metaverse: fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials*, *25*(1), 319–352. https://doi.org/10.1109/comst.2022.3202047

[21] Zografopoulos, I., Ospina, J., Liu, X., & Konstantinou, C. (2021). Cyber-Physical Energy Systems Security: threat modeling, risk assessment, resources, metrics, and case studies. *IEEE Access*, *9*, 29775–29818. https://doi.org/10.1109/access.2021.3058403