Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00115-w

Original Article

Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems

Kartik Mehta [1], Soharab Hossain [1], Ayushman Raghav [1], Himanshu Mohanty [1], Priyansh Arora [1]

1. *Computer Science, BML Munjal University, Gurgaon, IND*

Received: February 14, 2026 | Review began: February 24, 2026 | Review ended: June 03, 2026 | Published: June 10, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Deep learning-based CAPTCHA recognition systems are increasingly used in modern authentication frameworks; however, their robustness against adversarial attacks remains insufficiently explored, particularly in sequence-based recognition architectures. This study evaluates targeted adversarial vulnerabilities in a Convolutional Recurrent Neural Network (CRNN) optimized using Connectionist Temporal Classification (CTC) loss on a publicly available real-world dataset of 1,070 CAPTCHA images. Unlike prior studies that primarily focus on adversarial robustness in image classification tasks, this work investigates the impact of targeted adversarial attacks on CAPTCHA sequence recognition systems and assesses the effectiveness of a simple preprocessing-based defense.

To assess robustness, white-box Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks are applied. Under clean conditions, the CRNN model achieved a sequence-level recognition accuracy of 98.73%. FGSM achieved an attack success rate of 75.66%, reducing recognition accuracy to 23.07%, while PGD achieved an attack success rate of 34.00%, with post-attack accuracy of 64.73%.

A Gaussian blue preprocessing defense was evaluated at inference time to mitigate adversarial perturbations. The defense reduces FGSM attack success rate from 75.66% to 51.45% and partially restored recognition performance, but provided no measurable improvement against PGD-generated adversarial examples.

The results demonstrate that CRNN-based CAPTCHA recognition systems remain highly vulnerable to targeted adversarial manipulation and that simple preprocessing defenses offer limited protection against adaptive gradient-based attacks. These findings highlight the need for more robust defense mechanisms to improve the security and reliability of AI-driven authentication systems.

Categories: Security and Privacy, The Adversarial AI, Deep Learning

Keywords: adversarial attacks, captcha recognition, deep learning security, gradient based attacks, crnn, projected gradient descent, fast gradient sign method, adversarial robustness

Introduction

CAPTCHAs have been extensively used as protective measures against various types of automated attacks such as spamming, password-guessing attacks, and denial-of-service attacks. In light of recent developments in machine learning techniques, modern-day CAPTCHA recognition systems make extensive use of deep learning architectures for generating and solving more advanced CAPTCHA problems [1-3]. Although deep learning-based CAPTCHA recognition systems exhibit great performance accuracy under usual conditions, their resistance to adversarial attacks is a major issue.

Deep learning models are inherently vulnerable to adversarial attacks, where carefully crafted and visually imperceptible perturbations can significantly alter model predictions [4,5]. A comprehensive body of research has systematically analyzed both adversarial attack strategies and corresponding defense mechanisms in deep learning systems [6]. Prior research has demonstrated that even extremely minimal perturbations, such as the one-pixel attack, can successfully mislead deep neural networks, highlighting the fragility of learned representations [7]. In CAPTCHA-based authentication frameworks, such perturbations introduce a dual risk. On the one hand, adversarially modified inputs may enable automated systems to bypass verification mechanisms. On the other hand, imperceptible perturbations can cause legitimate users to experience authentication failures, as the CAPTCHA appears unchanged to humans while the underlying model prediction is altered. This disconnect between human perception and model interpretation exposes fundamental weaknesses in AI-driven authentication systems. The availability of gradient-based attack techniques, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), further amplifies this vulnerability under white-box threat settings [4,8].

Text-based CAPTCHA recognition systems commonly employ hybrid deep learning architectures in which convolutional layers extract spatial features and recurrent layers model sequential character dependencies. These systems are typically optimized using Connectionist Temporal Classification (CTC) loss to enable sequence prediction without explicit character segmentation [1,9]. While Convolutional Recurrent Neural Network (CRNN) architectures demonstrate strong recognition performance on clean datasets, their end-to-end differentiable structure and reliance on gradient-based optimization render them particularly susceptible to gradient-driven adversarial attacks [8,10]. Despite this known vulnerability in deep learning models, comprehensive empirical evaluations of adversarial robustness in practical CAPTCHA deployment scenarios remain limited.

Although adversarial attacks have been widely studied in image classification and computer vision, their impact on CAPTCHA recognition systems has received comparatively less attention. Most prior work has focused either on improving CAPTCHA recognition performance or on developing adversarial attack techniques in broader machine learning settings. In contrast, fewer studies have examined how targeted adversarial attacks affect CRNN-CTC-based CAPTCHA recognition models or evaluated the effectiveness of simple preprocessing defenses in this context. As a result, the adversarial robustness of CAPTCHA recognition systems remains an open research challenge, particularly for authentication systems that rely on deep learning-based recognition pipelines.

This study systematically evaluates targeted adversarial attacks on a CRNN-based CAPTCHA recognition system using FGSM and PGD under a white-box threat model. Alongside the examination of gradient-based attacks' effects, the study explores the efficacy of a defensive technique based on preprocessing through Gaussian blurring. Experiments conducted on a publicly available real-world CAPTCHA dataset measure sequence-level recognition degradation and adversarial success rates as well as limitations in using input transformations as a means of protection against attacks. The findings provide a focused assessment of adversarial risks in CAPTCHA recognition systems and highlight the need for more robust defense mechanisms to improve the security and reliability of AI-driven authentication frameworks.

The primary contributions of this study are summarized as follows:

1. Evaluation of targeted FGSM and PGD adversarial attacks against a CRNN-CTC CAPTCHA recognition framework under a white-box threat model.

2. Quantitative analysis of attack success rates and sequence-level recognition degradation caused by adversarial perturbations.

3. Assessment of Gaussian blur as a preprocessing-based defense mechanism and evaluation of its effectiveness against both single-step and iterative adversarial attacks.

4. Identification of adversarial security limitations in deep learning-based CAPTCHA recognition systems and discussion of their implications for AI-driven authentication frameworks.

How to cite this article:
Mehta K, Hossain S, Raghav A, et al. (June 10, 2026) Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems. Cureus J Comput Sci 3 : es44389-026-00115-w. DOI https://doi.org/10.7759/s44389-026-00115-w

Materials And Methods

Dataset description

The experimental evaluation is conducted on a publicly available real-world text-based CAPTCHA dataset obtained from Kaggle. The dataset consists of 1,070 grayscale CAPTCHA images containing alphanumeric character sequences of variable lengths. Each image represents an individual CAPTCHA instance, with ground-truth labels encoded in the corresponding file names.

Due to varying character lengths, CAPTCHA images differ in width while maintaining a consistent height. To ensure compatibility with the CRNN architecture, all images are resized to a uniform height of 32 pixels while preserving their original aspect ratios, following standard practices in sequence-based text recognition models [1,9]. Pixel intensities are normalized prior to training and evaluation to improve numerical stability and convergence behavior. This preprocessing pipeline reflects realistic CAPTCHA deployment conditions and provides a practical setting for assessing adversarial attacks and defense mechanisms in authentication systems.

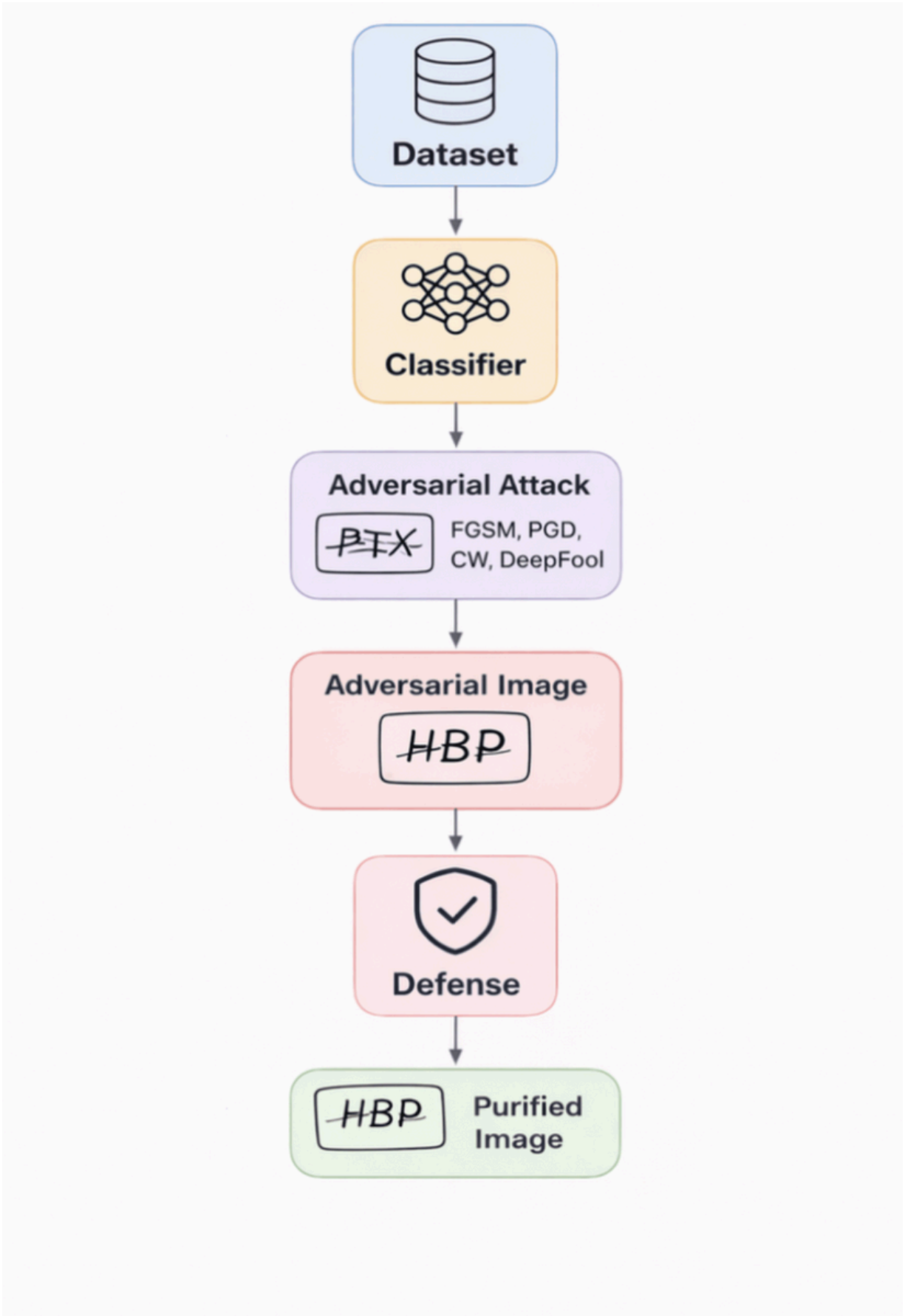
Although the dataset size is modest, it provides a practical benchmark for controlled adversarial robustness evaluation and has been used to assess attack effectiveness under realistic CAPTCHA recognition conditions.

Model architecture

The CAPTCHA recognition framework is based on a CRNN architecture that integrates spatial feature extraction with sequential modeling [1]. The overall architecture of the proposed CRNN-based CAPTCHA recognition model is illustrated in Figure *1*. Convolutional layers first process the input CAPTCHA images to learn discriminative visual representations, which are subsequently transformed into sequential feature vectors suitable for temporal modeling.

The sequential representations are processed by a bidirectional recurrent layer to capture contextual dependencies between preceding and succeeding character positions. This design enables effective interpretation of variable-length CAPTCHA sequences without requiring explicit character segmentation.

How to cite this article:
Mehta K, Hossain S, Raghav A, et al. (June 10, 2026) Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems. Cureus J Comput Sci 3 : es44389-026-00115-w. DOI https://doi.org/10.7759/s44389-026-00115-w



How to cite this article:
Mehta K, Hossain S, Raghav A, et al. (June 10, 2026) Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems. Cureus J Comput Sci 3 : es44389-026-00115-w. DOI https://doi.org/10.7759/s44389-026-00115-w

FIGURE 1: Block diagram of the CRNN-based CAPTCHA recognition model with adversarial attack and defense stages.

CRNN, Convolutional Recurrent Neural Network

CRNN With CTC-Based Sequence Learning

In the proposed CRNN model, convolutional layers extract hierarchical spatial features while suppressing background distortions commonly present in CAPTCHA images. The resulting feature maps are reshaped into a sequence representation, where each column corresponds to a time step, enabling sequential processing without manual segmentation [1].

The sequential features are passed to a bidirectional recurrent layer to model inter-character dependencies [11]. To enable end-to-end sequence learning without pre-segmented labels, the model is optimized using CTC loss [9]. CTC introduces a blank symbol and computes optimal alignment between predicted character probabilities and the target sequence. During inference, repeated characters and blank tokens are collapsed to generate the final predicted CAPTCHA string. An example of a clean CAPTCHA image correctly recognized by the CRNN-based CTC model under non-adversarial conditions is shown in Figure *2*.

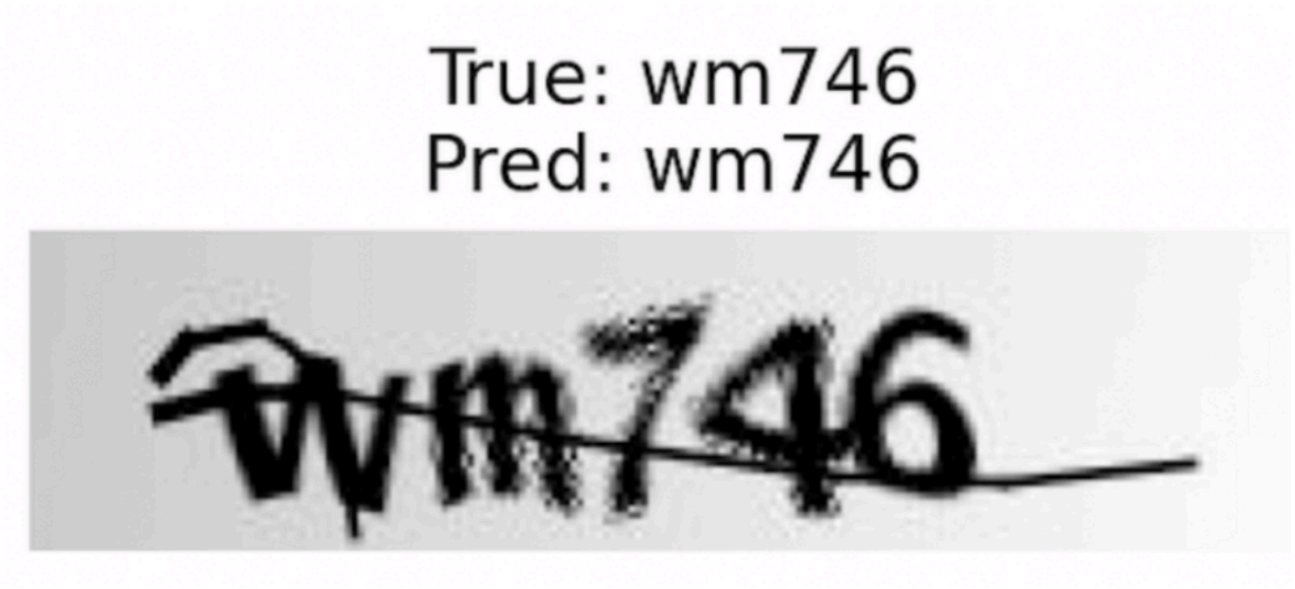


FIGURE 2: Example of clean CAPTCHA image correctly recognized by the CRNN-based CTC model under non-adversarial conditions.

CRNN, Convolutional Recurrent Neural Network; CTC, Connectionist Temporal Classification

Threat model

The present study considers a white-box adversarial threat model in which the attacker is assumed to possess complete knowledge of the target model, including its architecture, parameters, and training methodology [4,8]. This assumption represents a worst-case security scenario and is widely adopted in adversarial machine learning research for evaluating model robustness under highly capable attack conditions.

All attack scenarios examined were targeted in nature, where the adversary aims to manipulate the model into predicting a specific incorrect target label rather than merely inducing general misclassification. The adversary introduces visually imperceptible perturbations into input images that do not alter human perception but significantly impact model

predictions, consistent with established adversarial attack formulations [4].

For targeted attack generation, adversarial perturbations were optimized with respect to incorrect target CAPTCHA sequences that differed from the corresponding ground-truth labels. The objective of the attack was to steer model predictions toward a predefined target sequence while maintaining visually imperceptible perturbations.

Adversarial attacks

Fast Sign Gradient Method

FGSM is a single-step adversarial attack that generates perturbed inputs by leveraging the gradient of the loss function with respect to the input image [4]. Rather than modifying model parameters, FGSM perturbs the input in the direction that maximizes the loss.

For an input image *x*, ground truth label *y*, and model parameters θ, the adversarial example x' is computed as:

$$x' = x - \epsilon \cdot \text{sign}\left(\nabla_x L(x, y_{target}; \theta)\right)$$

where ϵ controls the perturbation magnitude. Due to its computational efficiency, FGSM is commonly used as a baseline approach for evaluating adversarial robustness in deep learning models [4]. A visual comparison between a clean CAPTCHA image and its FGSM-based adversarial counterpart is presented in Figure *3*.

FIGURE 3: Visual comparison of clean CAPTCHA and FGSM-based adversarial CAPTCHA image.

FGSM, Fast Gradient Sign Method

Projected Gradient Descent

PGD is an iterative adversarial attack that extends the FGSM by applying multiple gradient-based updates [8]. Starting from the original input, the attack performs repeated perturbation steps with step size α, followed by projection into an ϵ -bounded region around the clean sample.

Formally, for an input x, ground truth label y, and model parameters θ, the adversarial example at iteration $t+1$ is computed as:

$$x_{t+1} = \Pi_{\mathcal{B}_\epsilon(x)}\left(x_t - \alpha \cdot \text{sign}\left(\nabla_x L(x_t, y_{target}; \theta)\right)\right)$$

where α denotes the step size, ϵ defines the perturbation bound, and $\Pi_{\mathcal{B}_\epsilon(x)}$ represents the projection operator that constrains the adversarial sample within an ϵ-ball centered around the original input. The effect of the iterative PGD attack on clean CAPTCHA images is illustrated in Figure *4*.

How to cite this article:
Mehta K, Hossain S, Raghav A, et al. (June 10, 2026) Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems. Cureus J Comput Sci 3 : es44389-026-00115-w. DOI https://doi.org/10.7759/s44389-026-00115-w

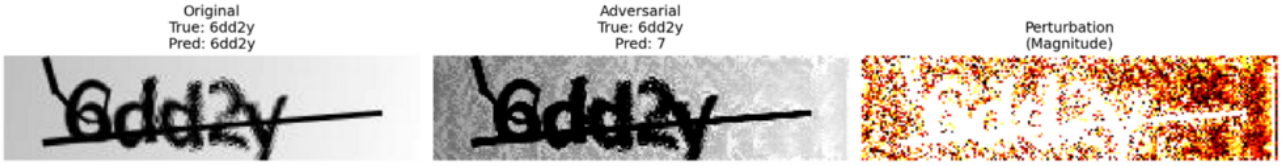


FIGURE 4: Effect of iterative PGD attack on clean CAPTCHA images.

PGD, Projected Gradient Descent

Defense mechanism

To mitigate adversarial perturbations, a preprocessing-based defense strategy was evaluated. Specifically, Gaussian blur was selected as a representative preprocessing-based defense because of its simplicity, low computational overhead, and its established use as a baseline adversarial mitigation technique in image-based recognition systems [12]. Such preprocessing-based defenses are particularly relevant in practical deployment scenarios where attackers may operate under limited model knowledge, including black-box threat settings [13].

Gaussian blur functions as a low-pass filter and has been explored as a simple baseline defense in image-based adversarial settings [12]. The defense is applied exclusively at inference time without modifying model parameters or retraining the network.

Defense effectiveness is evaluated by comparing recognition accuracy under clean, adversarial, and defended conditions. While Gaussian blur partially restores performance under single-step FGSM attacks, it does not provide meaningful robustness against stronger iterative PGD attacks, consistent with observations in prior adversarial defense research [10,14]. Prior work on input transformation-based defenses has further examined techniques such as image resizing, compression, and filtering to suppress adversarial perturbations before inference [15]. Examples of FGSM-based adversarial CAPTCHA images and the effect of Gaussian blur defense, including both successful and failed recovery cases, are presented in Figure *5*. In contrast, Figure *6* illustrates the failure of the Gaussian blur defense against PGD-based adversarial CAPTCHA images, demonstrating that correct recognition is not restored under stronger iterative attack conditions.

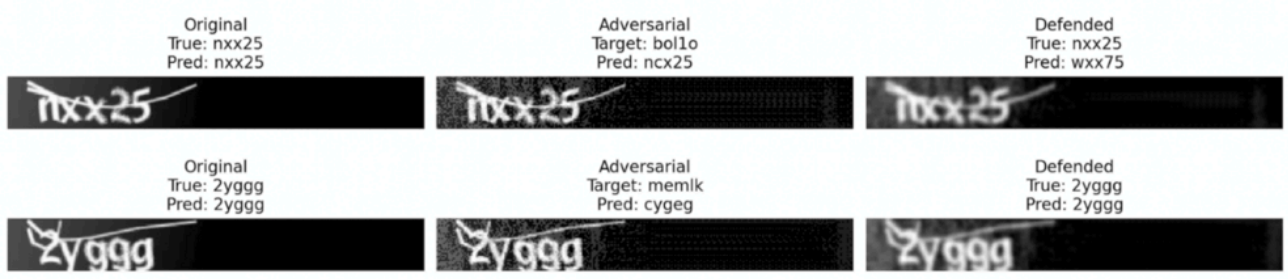


FIGURE 5: Visual examples of FGSM-based adversarial CAPTCHA images and the effect of Gaussian Blur defense, showing both successful and failed recovery cases.

FGSM, Fast Gradient Sign Method

FIGURE 6: Failure of Gaussian Blur defense against PGD-based adversarial CAPTCHA images, showing that correct recognition is not restored after preprocessing.

PGD, Projected Gradient Descent

Results And Discussion

Experimental setup

All experiments were conducted using a real-world text-based CAPTCHA dataset consisting of 1,070 grayscale samples. To evaluate model generalization, the dataset was divided into training and testing subsets using a standard train-test split protocol. Each CAPTCHA image was resized to a uniform height of 32 pixels while preserving its original aspect ratio, resulting in variable width inputs compatible with CRNN-based processing [1,9]. Pixel intensities were normalized to improve numerical stability during training and inference.

The CAPTCHA recognition framework is based on CRNN optimized using CTC loss, enabling sequence prediction without explicit character segmentation [1,9]. The model was trained exclusively on clean, nonadversarial samples until convergence and achieved high recognition accuracy under benign conditions. Recognition accuracy was computed at the sequence level. A prediction was considered correct only when the complete predicted CAPTCHA string exactly matched the corresponding ground-truth sequence. Partial character matches were not considered correct predictions.

The experiments were implemented using the PyTorch deep learning framework and executed on a system utilizing GPU acceleration when available. The dataset was divided using a 90:10 training-validation split. The model was trained for 200 epochs with a batch size of 32. A bidirectional LSTM layer with 256 hidden units was employed. Optimization was performed using the Adam optimizer with a learning rate of 1×10^{-4}. CTC loss was used as the training objective, and gradient clipping with a maximum norm of 5.0 was applied to stabilize training.

To evaluate adversarial robustness, the trained model was assessed in a white-box threat setting using FGSM and PGD attacks [4,8]. As a preprocessing-based defense, Gaussian Blur was applied at inference time to suppress high-frequency adversarial perturbations [12,14].

Attack Configuration

Adversarial attacks were configured as follows:

FGSM: FGSM generates adversarial examples in a single gradient step. Perturbations are constrained by a predefined noise bound ϵ, which limits the magnitude of pixel level modifications. Due to its one step formulation, FGSM is computationally efficient but typically less destructive than iterative methods. In the present study, the perturbation magnitude was set to $\epsilon = 0.3$ under a targeted attack formulation.

PGD: PGD is an iterative extension of FGSM that applies multiple gradient updates with step size α while constraining the perturbation within the same ϵ-bound region around the original input sample. This iterative optimization makes PGD significantly stronger and more capable of misleading the model. For iterative adversarial evaluation, PGD was configured with $\epsilon = 0.1$, step size$\alpha = 0.01$, and 20 attack iterations.

The attack parameters used in this study were selected based on commonly adopted configurations reported in the adversarial machine learning literature. These values provide a practical balance between attack effectiveness and perturbation imperceptibility while enabling consistency with established adversarial robustness evaluation methodologies.

All attacks were implemented in a targeted setting, where adversarial perturbations were optimized to steer the model toward a predefined incorrect CAPTCHA string. In this formulation, the attack objective minimizes the loss with respect to a chosen target label rather than merely increasing overall prediction error. Although complete convergence to the exact target label was not consistently achieved across all samples, the targeted optimization process significantly altered model outputs and resulted in effective misclassification. These findings indicate that even partial target-driven perturbations are sufficient to compromise recognition reliability in CRNN-based CAPTCHA systems.

Results and Analysis

Table *1* presents the recognition accuracy and attack success rate of the CRNN-based CAPTCHA recognition system under clean conditions, targeted adversarial attacks, and preprocessing-based defense.

How to cite this article:
Mehta K, Hossain S, Raghav A, et al. (June 10, 2026) Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems. Cureus J Comput Sci 3 : es44389-026-00115-w. DOI https://doi.org/10.7759/s44389-026-00115-w

Condition	Accuracy (%)	Attack Success Rate (%)
Clean	98.73	-
FGSM	23.07	75.66
FGSM + Gaussian Blur	47.28	51.45
PGD	64.73	34.00
PGD + Blur	64.73	34.00

TABLE 1: CAPTCHA recognition accuracy under adversarial attacks and preprocessing-based defense.

FGSM, Fast Gradient Sign Method; PGD, Projected Gradient Descent

Under clean conditions, the model achieves a recognition accuracy of 98.73%, confirming its effectiveness in sequence modeling and character level prediction in the absence of adversarial interference.

When subjected to targeted FGSM attack, the attack success rate reaches 75.66%, indicating that adversarial perturbations successfully altered the predicted CAPTCHA outputs in the majority of samples. Consequently, recognition accuracy drops substantially to 23.07%, demonstrating the high vulnerability of the model to single-step gradient-based targeted attacks.

Application of Gaussian Blur partially mitigates the effect of FGSM. Approximately 32% of the successfully perturbed samples are restored to correct prediction, reducing the attack success rate to 51.45%. As a result, recognition accuracy improves to 47.28%. This behavior suggests that FGSM generated perturbations contain high-frequency components that can be partially suppressed through simple low-pass filtering.

In contrast, PGD attack achieves an attack success rate of 34%, resulting in a post attack recognition accuracy of 64.73%. Although the attack success rate is lower than that of FGSM in this configuration, PGD remains effective in inducing targeted misclassification. Importantly, the application of Gaussian Blur does not produce any measurable reduction in attack success rate under PGD. Both the attack success rate (34%) and recognition accuracy (64.73%) remain unchanged after preprocessing, indicating that iterative adversarial optimization produces perturbations that are robust to basic input-level filtering [8,10,14].

Discussion

This study presents a structured evaluation of targeted adversarial vulnerabilities in CRNN-based CAPTCHA recognition systems under both single-step and iterative attack settings. The experimental results demonstrate that adversarial optimization significantly compromises recognition reliability, even when perturbations remain visually imperceptible to human observers.

Under targeted FGSM attack, the model exhibits a high attack success rate of 75.66%, resulting in a substantial reduction in recognition accuracy. Although FGSM is a single-step gradient-based method, its targeted formulation effectively shifts model predictions in the majority of samples. The partial recovery observed after applying Gaussian Blur indicates that FGSM-generated perturbations contain high-frequency components that can be partially attenuated through low-pass filtering. The reduction of attack success rate to 51.45% following preprocessing suggests limited defensive capability against single-step adversarial manipulation.

In contrast, PGD-generated adversarial examples demonstrate greater resilience against preprocessing-based defenses. Despite achieving a lower attack success rate of 34% in this configuration, PGD generated adversarial examples remain unaffected by Gaussian blur. Both attack success rate and recognition accuracy remain unchanged after preprocessing, indicating that iterative adversarial optimization produces perturbations aligned with the model's decision boundaries rather than merely introducing high-frequency noise. This behavior reflects the adaptive nature of iterative attacks and their resilience to basic input-level filtering mechanisms [10,12,14].

The failure of Gaussian Blur against PGD highlights a fundamental limitation of preprocessing-based defense strategies. While such techniques may provide partial mitigation against weaker, single-step perturbations, they do not offer reliable protection in adversarial environments where attackers possess model knowledge and optimize perturbations iteratively. These findings underscore that input-level denoising alone is insufficient for securing deep learning-based CAPTCHA recognition pipelines under targeted adversarial pressure.

From a computational perspective, FGSM requires only a single gradient computation and therefore incurs relatively low computational overhead. In contrast, PGD performs multiple iterative optimization steps, resulting in increased computational cost proportional to the number of attack iterations. This trade-off between computational efficiency and attack sophistication is consistent with observations reported in adversarial machine learning literature.

Several limitations should be acknowledged. First, the evaluation was conducted using a publicly available CAPTCHA dataset consisting of 1,070 samples, which may limit the generalizability of the findings to larger and more diverse CAPTCHA datasets. Second, the study focused on a single preprocessing-based defense mechanism using Gaussian blur. Although this approach provides useful insight into the effectiveness of simple input-transformation defenses, more advanced defense strategies such as adversarial training, randomized smoothing, and ensemble-based methods were beyond the scope of the present work. Finally, the analysis was performed using a representative CRNN-CTC architecture and did not include comparisons across multiple CAPTCHA recognition models.

Overall, the results confirm that CRNN-based CAPTCHA solvers remain vulnerable to gradient-driven adversarial manipulation, particularly when attacks are formulated with targeted objectives. Beyond robustness-oriented training, adversarial detection mechanisms have also been explored to identify manipulated inputs prior to inference, providing an additional defensive layer in adversarial settings [16]. In addition to detection-based approaches, adversarial training methods, including ensemble adversarial training, have been proposed to improve resilience against adaptive gradient-based attacks [17]. The study emphasizes the necessity of adopting more robust and adaptive defense strategies, including adversarial training, robustness-aware optimization, randomized smoothing, and detection mechanisms specifically designed for sequence recognition architectures. Strengthening adversarial resilience is essential to maintaining the reliability of AI-driven authentication systems in real-world deployment scenarios.

Conclusions

This study experimentally evaluated targeted adversarial vulnerabilities in a CRNN-based CAPTCHA recognition system optimized using CTC loss on a publicly available real-world CAPTCHA dataset. Unlike prior studies that primarily focus on adversarial robustness in image classification tasks, the present work specifically examined sequence-based CAPTCHA recognition under targeted gradient-based attacks and evaluated the effectiveness of a preprocessing-based defense mechanism. The experimental results demonstrated that although the model achieved high recognition accuracy (98.73%) under clean conditions, it remained highly susceptible to adversarial manipulation. The targeted FGSM attack achieved an attack success rate of 75.66%, reducing recognition accuracy to 23.07%, while the iterative PGD attack achieved an attack success rate of 34.00% and remained unaffected by Gaussian blur preprocessing. These findings indicate that simple input-transformation defenses may provide limited protection against single-step attacks but are insufficient against stronger iterative adversarial optimization strategies.

Several limitations should be acknowledged. The study was conducted using a single CAPTCHA dataset and evaluated only one preprocessing-based defense mechanism. Furthermore, the analysis focused on a representative CRNN-CTC architecture and did not compare multiple CAPTCHA recognition models or advanced defense strategies. Overall, the

How to cite this article:
Mehta K, Hossain S, Raghav A, et al. (June 10, 2026) Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems. Cureus J Comput Sci 3 : es44389-026-00115-w. DOI https://doi.org/10.7759/s44389-026-00115-w

findings provide a focused empirical assessment of adversarial risks in deep learning-based CAPTCHA recognition systems and highlight the need for more robust, attack-aware defense mechanisms in AI-driven authentication frameworks. Future research may investigate larger and more diverse datasets, adversarial training techniques, randomized smoothing methods, ensemble-based defenses, and hybrid machine learning-deep learning approaches to further improve adversarial robustness in CAPTCHA recognition systems.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Kartik Mehta, Soharab Hossain

Acquisition, analysis, or interpretation of data: Kartik Mehta, Ayushman Raghav, Himanshu Mohanty, Priyansh Arora

Drafting of the manuscript: Kartik Mehta

Critical review of the manuscript for important intellectual content: Kartik Mehta, Ayushman Raghav, Himanshu Mohanty, Priyansh Arora, Soharab Hossain

Supervision: Soharab Hossain

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data supporting this study are openly available in Kaggle at https://www.kaggle.com/datasets/fournierp/captcha-version-2-images

Acknowledgements

The CAPTCHA dataset used in this study was obtained from the publicly available Kaggle repository: https://www.kaggle.com/datasets/fournierp/captcha-version-2-images. The authors acknowledge the contributors of this dataset.

References

1. Shi B, Bai X, Yao C: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017, 39:2298-2304. 10.1109/tpami.2016.2646371
2. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016, 770-778. 10.1109/CVPR.2016.90
3. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature. 2015, 521:436-444. 10.1038/nature14539
4. Akhtar N, Mian A: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access. 2018, 6:14410-14430. 10.1109/access.2018.2807385

5. Biggio B, Roli F: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition. 2018, 84:317-331. 10.1016/j.patcog.2018.07.023

6. Yuan X, He P, Zhu Q, Li X: Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems. 2019, 30:2805-2824. 10.1109/tnnls.2018.2886017

7. Su J, Vargas DV, Sakurai S: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation. 2019, 23:828-841. 10.1109/tevc.2019.2890858

8. Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D: A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology. 2021, 6:25-45. 10.1049/cit2.12028

9. Graves A, Fernández S, Gomez F, Schmidhuber J: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning (ICML). 2006, 369-376. 10.1145/1143844.1143891

10. Athalye A, Carlini N, Wagner D: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. Proceedings of the 35th International Conference on Machine Learning (ICML). 2018, 1-10.

11. Hochreiter S, Schmidhuber J: Long short-term memory. Neural Computation. 1997, 9:1735-1780. 10.1162/neco.1997.9.8.1735

12. Xu W, Evans D, Qi Y: Feature squeezing: Detecting adversarial examples in deep neural networks. Network and Distributed System Security Symposium (NDSS). 2018, 1-15. 10.14722/ndss.2018.23198

13. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A: Practical black-box attacks against machine learning. Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (AsiaCCS). 2017, 506-519. 10.1145/3052973.3053009

14. Carlini N, Wagner D: Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy. 2017, 39-57. 10.1109/SP.2017.49

15. Chang CL, Hung JL, Tien CW, Tien CW, Kuo SY: Evaluating Robustness of AI Models against Adversarial Attacks. Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence (co-located with AsiaCCS 2020). 2020, 47-54. 10.1145/3385003.3410920

16. Roth K, Kilcher Y, Hofmann T: The odds are odd: A statistical test for detecting adversarial examples. Proceedings of the 36th International Conference on Machine Learning (ICML). 2019, 97:5498-5507.

17. Pang T, Xu K, Du C, Chen N, Zhu J: Improving adversarial robustness via promoting ensemble diversity. Proceedings of the 36th International Conference on Machine Learning (ICML). 2019, 97:4970-4979.

How to cite this article:
Mehta K, Hossain S, Raghav A, et al. (June 10, 2026) Adversarial Attacks and Defense Evaluation in Deep Learning-Based CAPTCHA Recognition Systems. Cureus J Comput Sci 3 : es44389-026-00115-w. DOI https://doi.org/10.7759/s44389-026-00115-w