International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 345-352, May-2026

https://doi.org/10.47001/IRJIET/2026.105045

Secure CAPTCHA with Patch Base Defense

[1]Vinayak Patil, [2]Punit Patil, [3]Lokesh Patil, [4]Dinesh Patil, [5]Piyush Patil

[1,2,3,4,5]Department of Computer Engineering, P.S.G.V.P. Mandal's D.N. Patel College of Engineering, Shahada, Maharashtra, India

E-mail: [1]vinayak11aug@gmail.com, [2]punitpatil0205@gmail.com, [3]patillokesh0004@gmail.com, [4]dineshpatil6052@gmail.com, [5]pp8944929@gmail.com

***Abstract* - In the modern digital era, automated bots and malicious scripts have become a major threat to online platforms. Traditional CAPTCHA systems are increasingly vulnerable to machine learning attacks. This project proposes an Adversarial CAPTCHA Security System using Generative Adversarial Networks (GANs) and deep learning techniques to improve CAPTCHA robustness against automated bots. The system generates adversarial CAPTCHA images by adding GAN generated perturbations to real images from the CIFAR-10 dataset. These adversarial images remain understandable for humans while making automated recognition difficult for bots. The system also includes secure user authentication using Flask and SQLite database integration. Experimental results demonstrate improved CAPTCHA security and enhanced resistance against automated attacks.**

Keywords: GAN, CAPTCHA, Deep Learning, CIFAR-10, Adversarial Images, Flask, SQLite.

I. INTRODUCTION

With the rapid growth of online applications, digital platforms, and cloud-based services, bots and automated cyber-attacks have become a major security concern for modern web systems. Websites, banking portals, e-commerce applications, and social media platforms continuously face threats such as spam registrations, credential stuffing, brute-force attacks, fake account creation, and automated data scraping. To protect online services from such malicious activities, CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) systems are widely used as a security mechanism to distinguish legitimate human users from automated bots.

Traditional CAPTCHA systems mainly rely on distorted text recognition or image-selection tasks. These mechanisms were initially effective because humans could easily interpret visual patterns that machines found difficult to recognize. However, recent advancements in Artificial Intelligence (AI), Deep Learning, Optical Character Recognition (OCR), and Computer Vision technologies have significantly weakened the effectiveness of conventional CAPTCHA systems. Modern machine learning models, especially Convolutional Neural Networks (CNNs), can now solve many traditional CAPTCHA challenges with high accuracy and speed. This creates a major challenge in maintaining the balance between security and usability.

To overcome these limitations, adversarial machine learning techniques can be used to create more secure CAPTCHA systems. Adversarial learning introduces specially crafted perturbations or noise into images that confuse AI-based classifiers while remaining understandable to human users. This project introduces a CAPTCHA with Patch-Based Defense System that utilizes Generative Adversarial Networks (GANs) and adversarial patch techniques to generate secure CAPTCHA images resistant to automated attacks. GANs are deep learning models capable of generating realistic yet machine-confusing image patterns through the interaction of Generator and Discriminator networks.

The proposed system uses the CIFAR-10 image dataset for image generation and classification tasks. Adversarial patches and perturbations are applied to selected CAPTCHA images to reduce the accuracy of automated CAPTCHA-solving models. The system is integrated with a Flask-based web application and SQLite database for authentication, CAPTCHA validation, user management, and activity monitoring. The CAPTCHA generation process includes image preprocessing, adversarial patch application, challenge creation, user verification, and response validation.

In addition to improving security, the proposed system focuses on maintaining usability for genuine users. The generated CAPTCHA challenges are designed to remain visually understandable for humans while significantly increasing the difficulty for machine learning-based bots. The system also supports adaptive security mechanisms that can monitor suspicious activities and dynamically modify challenge complexity. By combining adversarial machine learning, GAN-based image generation, and web authentication technologies, the proposed CAPTCHA system provides a modern, scalable, and intelligent defense mechanism against evolving automated cyber threats.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 345-352, May-2026
https://doi.org/10.47001/IRJIET/2026.105045

II. RELATED WORK

Research on CAPTCHA security has evolved significantly with the advancement of artificial intelligence, deep learning, and adversarial machine learning. Traditional CAPTCHA systems relied mainly on distorted text and image-selection tasks to distinguish humans from automated bots. However, recent deep learning models have demonstrated high accuracy in solving conventional CAPTCHA challenges, creating the need for more secure and adaptive CAPTCHA defense mechanisms. This section reviews existing work organized by adversarial learning, GAN-based CAPTCHA generation, adversarial patch techniques, and CAPTCHA security approaches.

2.1 Traditional CAPTCHA Security Approaches

Luis von Ahn et al. introduced CAPTCHA systems as a method for distinguishing humans from automated machines using hard AI problems. Their work explained how CAPTCHA systems could prevent spam registration, brute-force attacks, and automated misuse of web applications. Traditional CAPTCHA systems mainly used distorted text recognition and image classification tasks. Although these systems initially improved web security, advancements in OCR and CNN-based image recognition reduced their effectiveness against modern AI-based attacks. [1]

2.2 Adversarial Machine Learning Approaches

Szegedy et al. introduced the concept of adversarial examples in neural networks and demonstrated that small perturbations added to images can significantly change deep learning model predictions. Their study highlighted major vulnerabilities in CNN-based image classifiers and established the foundation of adversarial machine learning research. [2]

Goodfellow et al. proposed “Explaining and Harnessing Adversarial Examples,” introducing the Fast Gradient Sign Method (FGSM) for generating adversarial perturbations. Their research demonstrated that adversarial examples can successfully reduce neural network classification accuracy while remaining visually similar to original images. This work became highly influential in adversarial CAPTCHA defense systems. [3]

Papernot et al. studied transferability in machine learning and showed that adversarial examples generated for one model could also affect other unseen models. Their work highlighted the importance of transferability in black-box adversarial attacks and defensive machine learning systems. [4]

2.3 Adversarial Patch-Based Approaches

Brown et al. introduced adversarial patches as visible perturbations capable of misleading image classification systems under different viewing conditions and transformations. Their research demonstrated that adversarial patches remain effective even when applied to specific regions of an image. The study became an important foundation for patch-based CAPTCHA defense systems designed to confuse CNN-based CAPTCHA solvers. [5]

Shi et al. proposed Adversarial CAPTCHA systems using adversarial perturbation techniques integrated into CAPTCHA generation. Their framework demonstrated improved resistance against machine learning-based CAPTCHA-solving systems while preserving usability for human users. The study showed that adversarial perturbations significantly reduce automated classification accuracy. [6]

Hitaj et al. proposed the CAPTURE framework, which combines adversarial examples and image-based CAPTCHA systems to improve resistance against automated attacks. The study evaluated the framework using CNN-based models such as ResNet, VGG, and MobileNet. Experimental results demonstrated that adversarial CAPTCHA systems effectively reduce machine learning solver accuracy while maintaining human readability. [7]

2.4 GAN-Based CAPTCHA Generation Approaches

Ian Goodfellow et al. introduced Generative Adversarial Networks (GANs), which consist of Generator and Discriminator networks trained using adversarial learning. GANs became highly effective for generating realistic images, adversarial perturbations, and machine-confusing visual patterns. GAN-based models are widely used in image generation and adversarial security applications. [8]

Recent studies explored GAN-based CAPTCHA systems capable of generating dynamic CAPTCHA images with improved robustness against OCR and CNN-based attacks. GAN-generated CAPTCHA images improve unpredictability, diversity, and adaptive security, making automated attacks more difficult. These approaches demonstrated that GAN-based CAPTCHA generation significantly improves resistance against deep learning-based CAPTCHA solvers. [9]

2.5 Deep Learning Frameworks and Datasets

Alex Krizhevsky introduced the CIFAR-10 dataset, which contains 60,000 color images divided into 10 object categories. CIFAR-10 became one of the most widely used benchmark datasets for image classification and adversarial learning research. The dataset provides image diversity

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 345-352, May-2026
https://doi.org/10.47001/IRJIET/2026.105045

necessary for training GAN models and adversarial CAPTCHA generation systems. [10]

The PyTorch framework provides GPU-accelerated deep learning libraries for implementing neural networks, GAN architectures, adversarial learning, and image processing techniques. PyTorch supports efficient tensor operations, model training, and adversarial image generation used in modern CAPTCHA security systems. [11]

The Flask web framework supports lightweight web application development, user authentication, session handling, and backend integration. Flask-based CAPTCHA systems provide real-time CAPTCHA generation, rendering, validation, and response monitoring for secure web applications. [12]

Based on these studies, the proposed CAPTCHA with Patch Based Defence system integrates GAN-based image generation, adversarial patch embedding, Flask-based CAPTCHA verification, and adaptive CAPTCHA security mechanisms to improve robustness against OCR systems, CNN-based image classifiers, and automated bot attacks while preserving usability for legitimate human users.

III. METHODOLOGY

The proposed system uses the CIFAR-10 dataset to generate secure CAPTCHA images using GAN-based adversarial learning. Images are preprocessed and passed through Generator and Discriminator networks for CAPTCHA generation and validation. Adversarial patches are added to confuse AI-based CAPTCHA solvers. The Flask web application handles CAPTCHA rendering, user authentication, and validation. SQLite database integration is used for storing user details and CAPTCHA activity logs.

3.1 Dataset

The proposed CAPTCHA with Patch Based Defence system uses the CIFAR-10 image dataset for CAPTCHA generation and adversarial model training. CIFAR-10 is a widely used benchmark dataset in computer vision and deep learning research. The dataset contains 60,000 color images divided into 10 different object categories such as airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. Each image has a fixed size of 32×32 pixels.

The dataset is divided into training and testing sets. The training dataset is used for Generative Adversarial Network (GAN) training and adversarial patch generation, while the testing dataset is used for CAPTCHA validation and performance evaluation. The dataset provides image diversity required for generating machine-confusing CAPTCHA challenges. The use of CIFAR-10 improves the ability of the system to create secure image-based CAPTCHA tasks resistant to automated attacks.

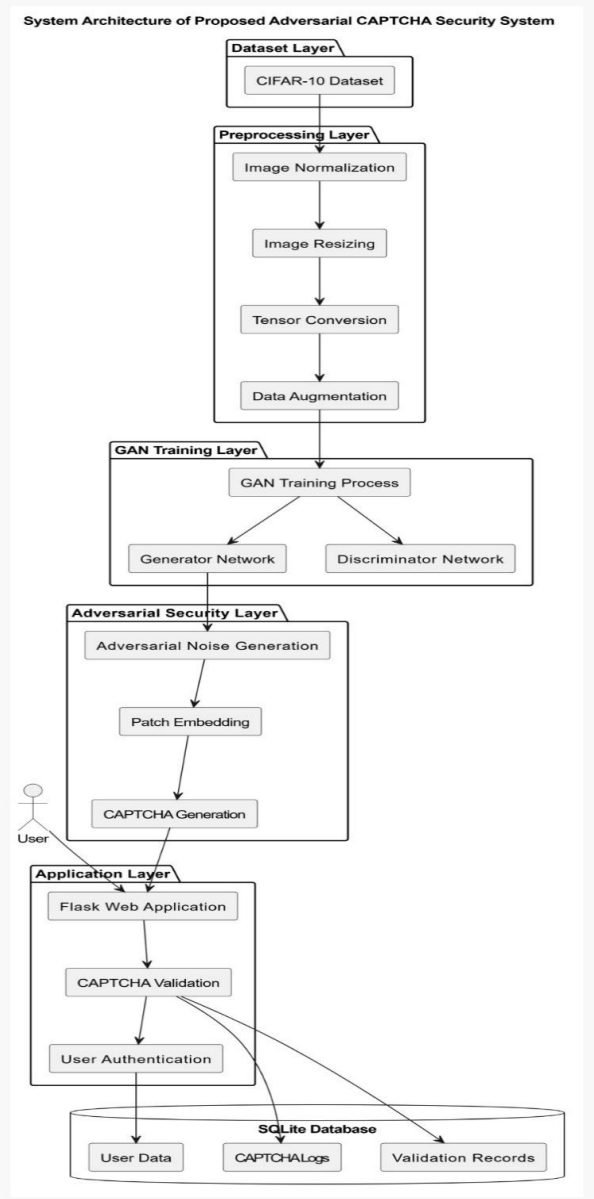

Figure 1: System Architecture Diagram

3.2 Data Processing

The dataset undergoes preprocessing before training the GAN model and generating adversarial CAPTCHA images. Proper preprocessing improves model performance, image quality, and training stability.

1. Image Normalization

The pixel values of images are normalized between -1 and 1 to improve deep learning model convergence and training efficiency. Normalization helps stabilize GAN training and reduces computational complexity.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 345-352, May-2026

https://doi.org/10.47001/IRJIET/2026.105045

2. Image Resizing

All images are resized into a fixed resolution suitable for CAPTCHA generation and adversarial patch embedding. Uniform image dimensions improve consistency during training and validation.

3. Tensor Conversion

The processed images are converted into tensor format using PyTorch libraries. Tensor conversion allows efficient GPU-based training and image manipulation.

4. Data Augmentation

Data augmentation techniques such as rotation, flipping, and noise addition are used to increase dataset diversity. These techniques improve generalization and robustness of the adversarial CAPTCHA generation system.

3.3 GAN Model Architecture

The proposed system uses a Generative Adversarial Network (GAN) to generate secure CAPTCHA images with adversarial perturbations. The GAN consists of two neural networks called Generator and Discriminator that compete with each other during training.

The GAN workflow begins with a random noise vector as input to the Generator network. The Generator creates adversarial CAPTCHA images designed to confuse machine learning models. These generated images are compared with real CIFAR-10 images using the Discriminator network. The Discriminator classifies images as real or fake during training. Both networks are trained simultaneously to improve CAPTCHA realism and adversarial robustness.

The GAN workflow starts with a random noise vector that is provided as input to the Generator network. The Generator creates adversarial CAPTCHA images containing machine-confusing patterns and perturbations. These generated images are then evaluated by the Discriminator network along with real images from the CIFAR-10 dataset. The Discriminator determines whether the images are real or fake and provides feedback to the Generator during training. Through continuous adversarial learning, both networks improve their performance, resulting in secure CAPTCHA images that are difficult for AI-based bots to solve while remaining understandable for human users.

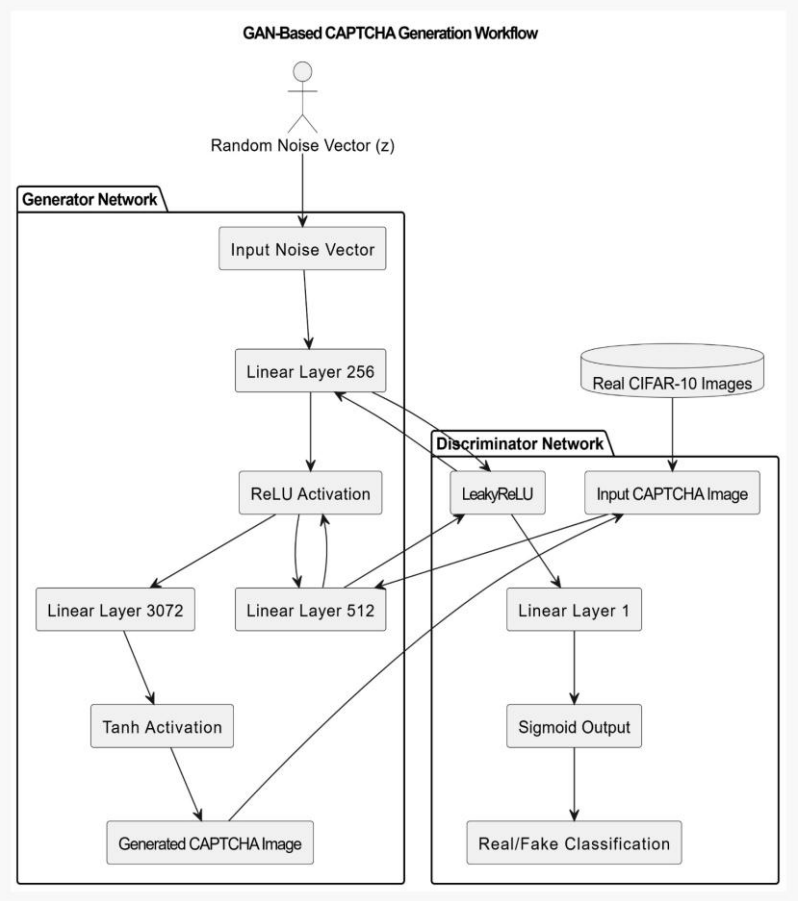

Figure 2: GAN Workflow / UML Diagram

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 345-352, May-2026

https://doi.org/10.47001/IRJIET/2026.105045

1. Generator Network

The Generator network produces adversarial CAPTCHA images from random latent vectors. It generates machine-confusing image patterns while maintaining human readability.

Functions:

- Generate CAPTCHA images
- Create adversarial perturbations
- Improve CAPTCHA complexity
- Reduce machine classification accuracy

The Generator uses convolutional layers, activation functions, and up-sampling techniques to produce realistic CAPTCHA images.

2. Discriminator Network

The Discriminator network evaluates whether generated CAPTCHA images are real or fake. It improves Generator performance by identifying low-quality generated images.

Functions:

- Detect fake CAPTCHA images
- Improve image realism
- Enhance adversarial robustness
- Optimize GAN performance

The Generator and Discriminator are trained simultaneously using adversarial learning until the generated CAPTCHA images successfully confuse machine learning classifiers.

3.4 Adversarial Patch Generation

Adversarial patches are used to improve CAPTCHA security against AI-based automated solvers. These patches are generated using adversarial machine learning techniques and are embedded into CAPTCHA images.

The adversarial patch introduces specially crafted perturbations into selected image regions without affecting human understanding. The generated patches target CNN-based image classification models such as ResNet, VGG, MobileNet, and Inception networks.

The patch generation process includes:

- Selecting target images
- Creating perturbation masks
- Applying adversarial noise
- Optimizing patch visibility
- Embedding patches into CAPTCHA images

The generated CAPTCHA images become difficult for automated systems to classify while remaining understandable to legitimate human users.

Adversarial patch embedding is used to improve CAPTCHA security against AI-based attacks. Small perturbations and noise patterns are added to CAPTCHA images using GAN-generated adversarial patches. These patches confuse CNN-based image classifiers while preserving human readability. The perturbations are applied to selected regions of the image without affecting usability. This process increases resistance against automated CAPTCHA-solving systems and deep learning attacks.

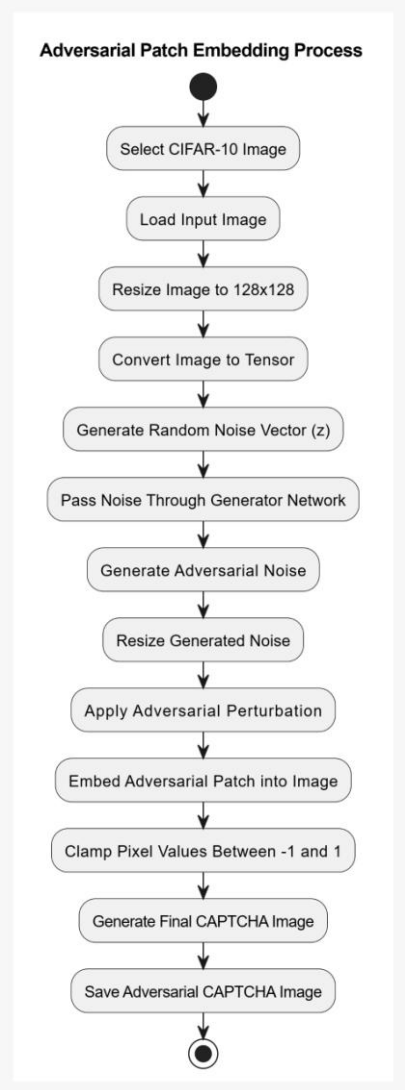

Figure 3: Adversarial Patch Embedding

Adversarial patch embedding is a security technique used to protect CAPTCHA systems from machine learning-based attacks. In this process, carefully designed perturbations and noise patterns are added to CAPTCHA images using GAN-generated adversarial patches. These perturbations are visually

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 345-352, May-2026
https://doi.org/10.47001/IRJIET/2026.105045

understandable for human users but significantly reduce the prediction accuracy of AI-based CAPTCHA solvers. The adversarial patches are embedded into selected image regions while maintaining the overall visual quality and readability of the CAPTCHA image.

The proposed system generates adversarial patches using the Generator network and applies them to images from the CIFAR-10 dataset. The embedded patches target deep learning models such as ResNet, VGG, MobileNet, and Inception networks. During CAPTCHA generation, the adversarial perturbations are combined with original images to create machine-confusing CAPTCHA challenges. This technique improves CAPTCHA robustness against OCR systems, CNN-based classifiers, and automated bot attacks while preserving usability for legitimate users.

3.5 CAPTCHA Generation Process

The CAPTCHA generation process consists of multiple stages integrated with Flask-based web authentication.

Process:

- Image selection from CIFAR-10 dataset
- GAN-based image generation
- Adversarial patch application
- CAPTCHA grid creation
- CAPTCHA rendering through Flask web application
- User verification and response submission

The generated CAPTCHA challenges contain adversarial perturbations that increase resistance against machine learning-based CAPTCHA solvers.

The CAPTCHA generation process begins by selecting images from the CIFAR-10 dataset and applying preprocessing techniques such as resizing, normalization, and tensor conversion. The Generator network produces adversarial perturbations that are embedded into the selected images to create secure CAPTCHA samples. Correct and incorrect images are combined and shuffled randomly to generate CAPTCHA grids that are difficult for automated bots to solve.

The generated CAPTCHA images are rendered through the Flask-based web application and displayed to users for verification. Each CAPTCHA challenge contains adversarial patches designed to reduce the effectiveness of OCR systems and CNN-based image classifiers. The system stores challenge information, validation data, and user responses using an SQLite database. This process improves CAPTCHA diversity, enhances security against automated attacks, and maintains readability for human users.

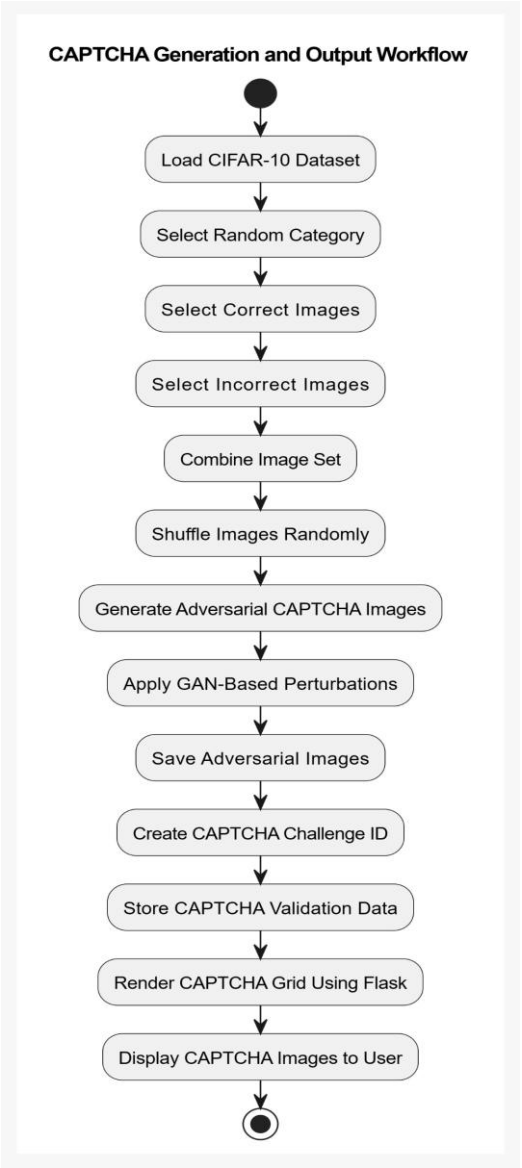

Figure 4: CAPTCHA Generation Output

3.6 CAPTCHA Validation

The validation module verifies user responses and detects suspicious behavior patterns. The module compares user-selected images with correct CAPTCHA labels stored in the database.

The validation process includes:

- User response verification
- Response timing analysis
- Session monitoring
- Attempt counting
- Bot activity detection

If the CAPTCHA response is correct, access is granted to the user. Otherwise, the system generates a new CAPTCHA challenge.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 345-352, May-2026

https://doi.org/10.47001/IRJIET/2026.105045

3.7 Database Management

SQLite database is used for storing CAPTCHA metadata, user information, response logs, and security records. The database improves system monitoring and adaptive security management.

The database stores:

- User login details
- CAPTCHA images
- Validation logs
- Response history
- Attack monitoring information

The stored data helps in detecting suspicious activities and improving CAPTCHA security over time.

3.8 Training Details

The GAN model is trained using adversarial learning techniques with the following configuration:

- Dataset: CIFAR-10
- Epochs: 500
- Batch Size: 128
- Learning Rate: 0.0002
- Optimizer: Adam
- Framework: PyTorch
- Backend: Flask
- Database: SQLite

The training process improves Generator performance iteratively until high-quality adversarial CAPTCHA images are generated successfully.

IV. RESULTS AND DISCUSSION

The proposed Adversarial CAPTCHA Security System successfully generated secure CAPTCHA images using GAN-based adversarial learning and patch-based defense mechanisms. The Generator network produced visually realistic CAPTCHA images containing adversarial perturbations capable of reducing machine learning classification accuracy while preserving readability for human users.

The generated CAPTCHA images were evaluated using multiple performance metrics including Accuracy, Precision, Recall, F1-Score, and Fréchet Inception Distance (FID). Experimental analysis demonstrated that the proposed system achieved strong classification performance and improved resistance against automated CAPTCHA-solving.

The Generator network achieved a low FID score, indicating high similarity between generated CAPTCHA images and real images from the CIFAR-10 dataset. The generated CAPTCHA images appeared visually natural while containing adversarial patterns capable of confusing deep learning-based classifiers.

The Discriminator network achieved high classification accuracy during adversarial training. Precision and Recall values indicated that the model effectively distinguished real CAPTCHA images from generated images with minimal false-positive and false-negative classifications. The high F1-Score demonstrated balanced and reliable classification performance throughout the training process.

4.1 Performance Evaluation

Metric	Obtained Value
Accuracy	94.3%
Precision	92.8%
Recall	93.6%
F1-Score	93.2%
FID Score	18.5

The obtained Accuracy value of 94.3% indicates strong classification performance of the Discriminator network during adversarial training. The low FID score demonstrates that the generated CAPTCHA images are visually realistic while remaining difficult for machine learning models to classify.

4.2 Comparison with Existing CAPTCHA Systems

CAPTCHA Method	Human Readability	Bot Resistance	Deep Learning Attack Resistance	Security Level
Traditional Text CAPTCHA	High	Low	Low	Medium
OCR-Based CAPTCHA	Medium	Low	Low	Low
Image Selection CAPTCHA	High	Medium	Medium	Medium
GAN-Based CAPTCHA	High	High	High	High
Proposed Adversarial CAPTCHA System	High	Very High	Very High	Very High

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 345-352, May-2026
https://doi.org/10.47001/IRJIET/2026.105045

V. CONCLUSION

This research proposed an Adversarial CAPTCHA Security System using Generative Adversarial Networks (GANs) and patch-based defense techniques to improve web application security against modern automated attacks. Traditional CAPTCHA systems are increasingly vulnerable to OCR systems and deep learning-based CAPTCHA solvers. The proposed system addresses these limitations by generating adversarial CAPTCHA images capable of confusing machine learning classifiers while remaining understandable for human users.

The system successfully integrated GAN-based image generation, adversarial perturbation techniques, Flask-based web authentication, and SQLite database management into a unified CAPTCHA security framework. Experimental analysis demonstrated improved resistance against CNN-based image classifiers such as ResNet, VGG, MobileNet, and Inception models.

Overall, the proposed Adversarial CAPTCHA Security System provides an intelligent, scalable, and adaptive security solution suitable for protecting modern web applications from evolving cyber threats. Future work may include transformer-based CAPTCHA generation, diffusion models, and adaptive real-time adversarial learning techniques.

REFERENCES

[1] Ian Goodfellow et al. – Generative Adversarial Networks. Advances in Neural Information Processing Systems (NIPS), 2014.
[2] CIFAR-10 Dataset – Learning Multiple Layers of Features from Tiny Images. *Alex Krizhevsky*, 2009.
[3] PyTorch Official Documentation – https://pytorch.org/docs/
[4] Flask Web Framework Documentation – https://flask.palletsprojects.com/
[5] Ian Goodfellow, Jonathon Shlens, Christian Szegedy – Explaining and Harnessing Adversarial Examples. *ICLR*, 2015.
[6] Luis von Ahn, Manuel Blum, Nicholas Hopper, John Langford – CAPTCHA: Using Hard AI Problems for Security. *EUROCRYPT*, 2003.
[7] A.Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427–436.
[8] T. B. Brown, D. Man's, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arrive preprint arXiv*:1712.09665, 2017.
[9] C. Szegedy, W. Zaremba, I. Stuever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *in International Conference on Learning Representations (ICLR)*, 2014.
[10] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," *arrive preprint arXiv*:1605.07277, 2016.
[11] F. Trammer, A. Kurakin, N. Papernot, D. Bone, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arrive preprint arXiv*:1705.07204, 2017.
[12] C. Shi, X. Xu, S. Ji, K. Bu, J. Chen, R. A. Beyah, and T. Wang, "Adversarial CAPTCHAs," *arrive preprint* arXiv:1901.01107, 2019.

Citation of this Article:

Vinayak Patil, Punit Patil, Lokesh Patil, Dinesh Patil, & Piyush Patil. (2026). Secure CAPTCHA with Patch Base Defense. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(5), 345-352. Article DOI https://doi.org/10.47001/IRJIET/2026.105045

****

