Assessing Vulnerabilities to Adversarial Perturbations in EEG-Based Pathology Detection Systems

Hira Masood, Maham Jahangir, Muhammad Athar, Muhammad Imran Malik, **Faisal Shafait**, and Hassan Aqeel Khan

National University of Sciences and Technology (NUST), Islamabad, Pakistan

Overview

01 Background

02 Problem

03 Key Idea

04 Methodology

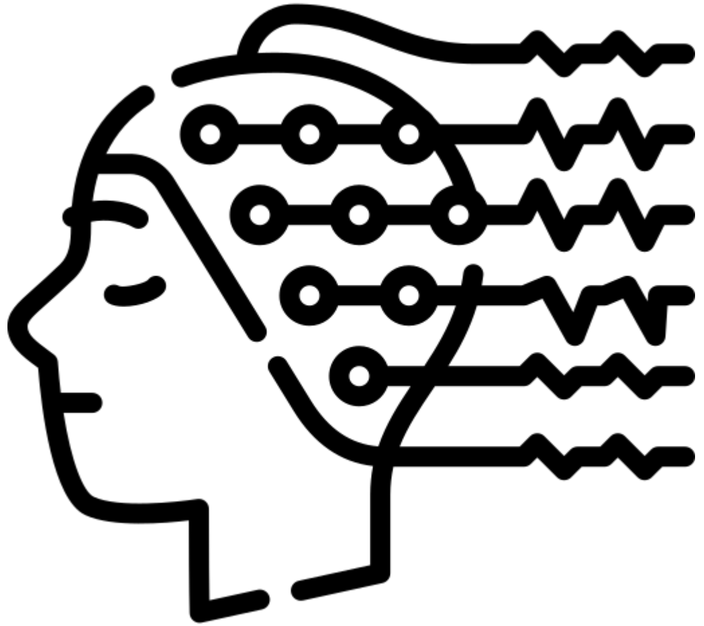
05 Results

06 Analysis

07 Conclusion

Background

- **EEG** provides a non-invasive way to monitor brain activity and support diagnosis of neurological disorders.



- **Deep learning** enables automated EEG analysis, reducing reliance on time consuming and subjective expert interpretation.

- However, DNN-based EEG systems are vulnerable to **adversarial attacks**, raising critical concerns about their security in clinical applications.

Problem & Challenges

Research Scope

- ❑ Adversarial robustness of **clinical EEG models** remains largely unexplored.
- ❑ Existing studies primarily focus on **controlled BCI tasks** rather than pathological EEG.
- ❑ Clinical EEG involves **long-duration, noisy, and highly variable** recordings.

Real-World Relevance

- ❑ Clinical EEG captures **diverse pathological patterns** across patients and temporal/spatial scales.
- ❑ Findings from controlled BCI environments may **not generalize to real-world diagnosis**.
- ❑ Robustness evaluation is essential for **safety-critical clinical deployment**.

Threat Model/ Security

- ❑ Most existing studies assume **white-box access** to model internals.
- ❑ Real-world systems are more likely to face **black-box attacks** with limited model access.
- ❑ This mismatch may lead to a **false sense of security** in deployed EEG systems.

Key Idea/ Motivation

Comprehensive Evaluation

Assess EEG models under white- and black-box threats.

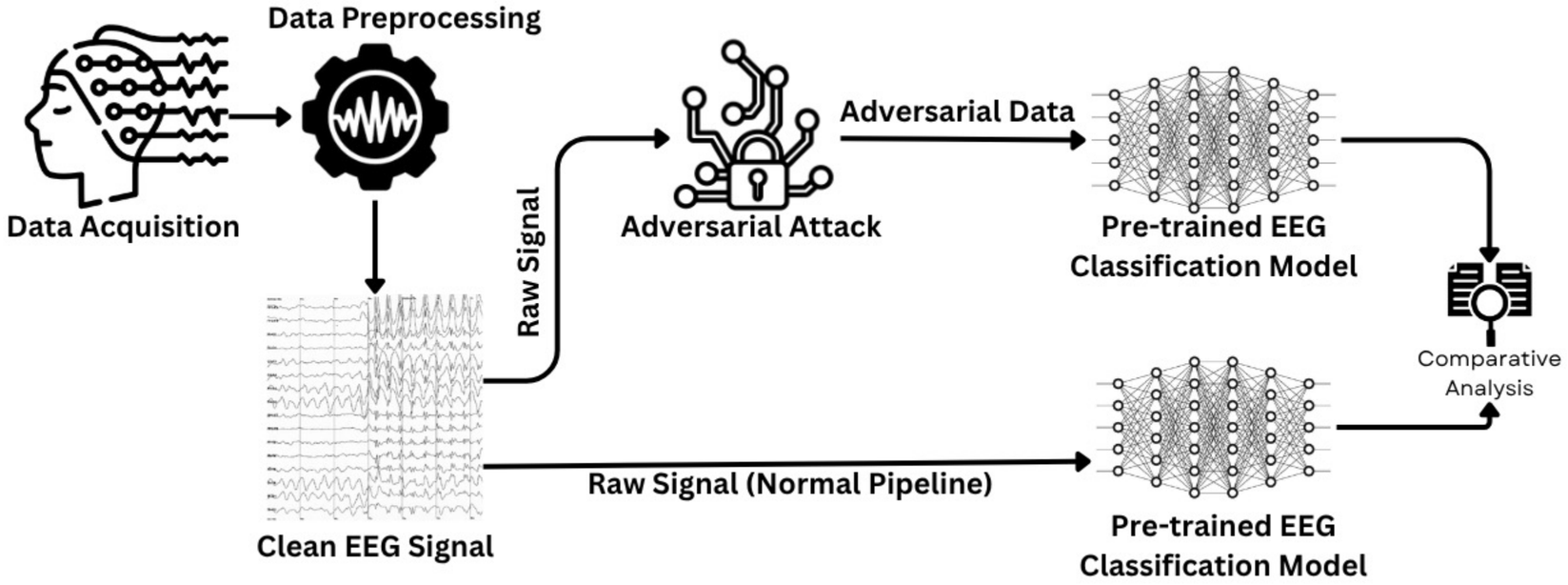
PCA-based Black Box Attack

Generate adversarial EEG signals without model internals.

Clinical Security

Reveal vulnerabilities and motivate robust EEG diagnostic systems.

Methodology: Threat Model



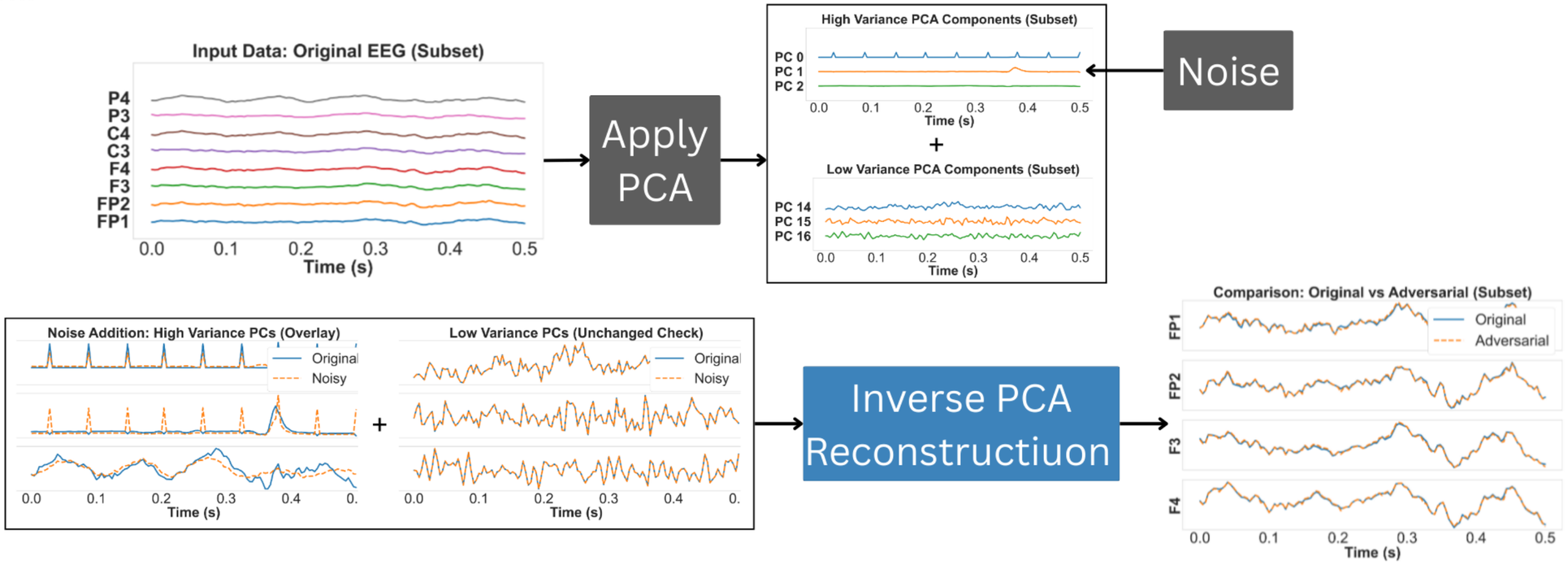
Methodology: Evaluation Pipeline

- **White-box Attacks:**
  - Fast Gradient Sign Method (FGSM)
  - Basic Iterative Method (BIM)
  - Projected Gradient Descent (PGD)
  - Proposed Structured White-box Attack: FGSM-PCA

- **Proposed Black-box Attacks**
  - Rotation-PCA Attack
  - Laplacian-PCA Attack
  - Frequency-domain PCA Attack (FFT-PCA)

Methodology: Black-box PCA-based Attack.

After PCA decomposition, dominant components are selectively perturbed using Rotation-PCA or Laplacian-PCA before inverse reconstruction.



Experiments & Results

Experimental Setup

Dataset: TUAB clinical EEG corpus for **normal vs. abnormal** classification.

Evaluation: Focus on **abnormal EEG samples correctly classified** under clean inference.

Models: Evaluate two state-of-the-art models: **WaveNet–LSTM** and **SCNet**.

Implementation

Input Protocol: 1-minute segments for WaveNet–LSTM and 7-minute segments for SCNet.

PyTorch-based experiments on an **NVIDIA RTX 3090 GPU**

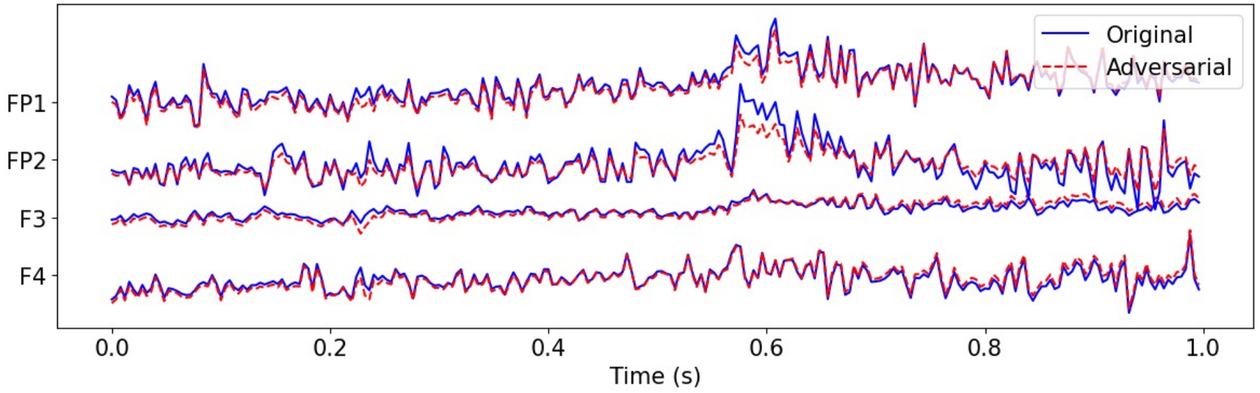
White-box robustness comparison on TUAB. Miscls.: number of abnormal samples flipped under attack.

Attack	ε	SCNet			WaveNet–LSTM		
		Miscls.	ASR (%)	SNR	Miscls.	ASR (%)	SNR
FGSM	0.1	6	5.94	43.22	7	7.07	37.52
	0.5	24	23.76	29.24	53	53.54	23.54
	1.0	42	41.58	23.22	70	70.71	17.52
BIM	0.1	6	5.94	46.77	7	7.07	39.61
	0.5	22	21.78	32.80	47	47.47	25.89
	1.0	42	41.58	26.86	69	69.70	20.06
PGD	0.1	3	2.97	45.40	5	5.05	40.19
	0.5	16	15.84	31.46	30	30.30	26.19
	1.0	28	27.72	25.48	61	61.62	20.11
FGSM-PCA	0.1	6	5.94	42.87	7	7.07	35.13
	0.5	26	25.74	27.11	54	54.55	22.78
	1.0	46	45.54	21.16	72	72.73	16.40

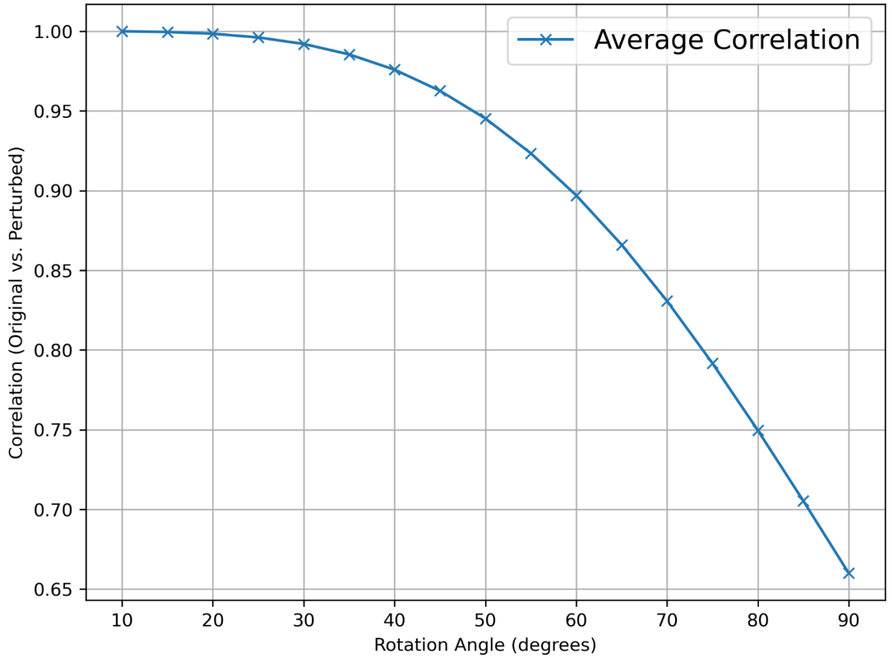
Experiments & Results

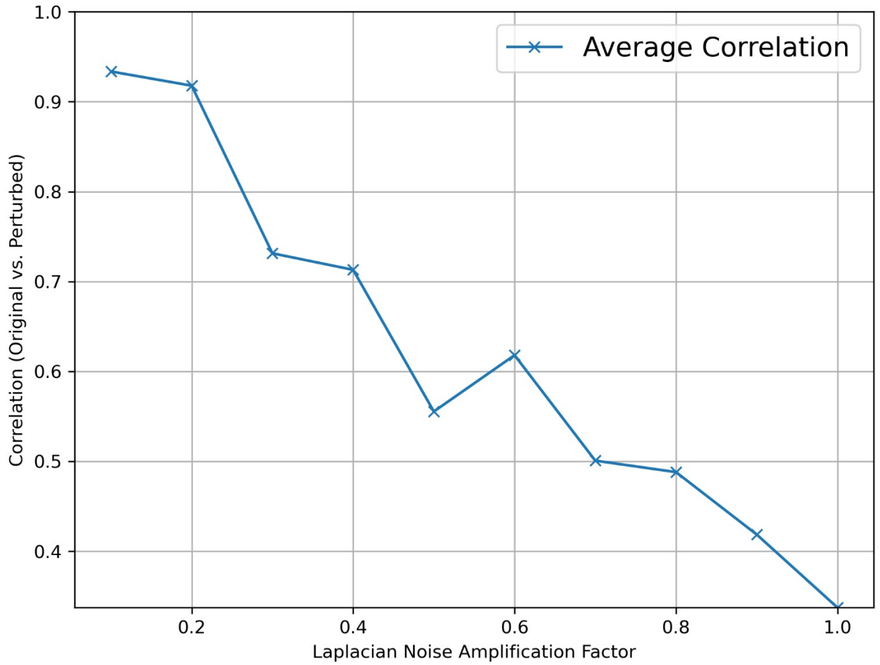
Black-box robustness on TUAB under FFT-PCA, Rotation-PCA (60◦),and Laplacian-PCA (amplitude 0.2)

Attack	PCA	SCNet			WaveNet–LSTM		
		Miscls.	ASR (%)	SNR	Miscls.	ASR (%)	SNR
FFT-PCA	1	70	69.31	0.71	38	38.38	0.64
	10	28	27.72	6.70	9	9.09	6.29
	20	13	12.87	12.24	3	3.03	11.93
Rotation-PCA (60°)	1	26	25.74	1.24	70	70.71	1.19
	10	10	9.90	4.04	31	31.31	5.64
	20	5	4.95	5.11	26	26.26	8.94
Laplacian-PCA (0.2)	1	41	40.59	1.86	55	55.56	1.57
	10	9	8.91	6.82	3	3.03	10.32
	20	1	0.99	9.61	0	0.00	24.87



Experiments & Results





Mean correlation between clean and perturbed EEG signals versus perturbation strength for Rotation-PCA (left) and Laplacian-PCA (right).

Analysis/ Key Findings

- **Clinical EEG classifiers are highly vulnerable** to both white-box and black-box attacks.
- **WaveNet–LSTM is more susceptible** than SCNet across most attack scenarios.
- **SCNet shows greater resilience**, suggesting that explicit spatial modeling improves robustness.
- **PCA-based black-box attacks** can cause substantial misclassification without model gradients.
- **Frequency-domain attacks (FFT-PCA)** effectively expose vulnerabilities, particularly in SCNet.

Conclusion

- **Clean accuracy does not guarantee adversarial security** in EEG diagnostic systems.
- Realistic black-box attacks pose a **significant risk to clinical EEG deployment**.
- Adversarial robustness should be considered alongside **classification performance** in medical AI.
- Future work should explore **adversarial training, robustness-aware feature learning, and certified defenses**.

Thank You