Beyond the Black Box: Building Trustworthy AI for Wearable Sleep Trackers

Janet Chen[1], Eric Ortega Rodriguez[1], Rishika Randev[1], Jenny Wu[1], Brinnae Bent, PhD[2]

[1] Interdisciplinary Data Science, Duke University [2] Pratt School of Engineering, Duke University

Duke MIDS
Master in Interdisciplinary Data Science

2026 Duke MIDS Capstone Symposium

Motivation: Current AI models used for sleep stage classification remain largely opaque black boxes that healthcare providers cannot trust or validate. These systems are potentially vulnerable to adversarial attacks that could compromise patient safety or be used to manipulate health data.

① Utilize DREAMT dataset

(Wang, Ke, et al.) [3]

Technician-annotated sleep stage labels

Wrist-worn wearable signals

② Pre-process data

Divide data into 30 second epochs

Statistical feature extraction

Signal processing techniques

Gaussian Filters

③ Structure input data

Pre-extracted features

Raw data

④ Select models

Support Vector Machine

(Logacjov, Aleksej, et al.) [4]

Baseline for evaluating robustness and explanation stability

LightGBM + Long Short-Term Memory

(Wang, Ke, et al.) [3]

Captures population-level patterns and subject-specific temporal dependencies

Convolutional Neural Network + Recurrent Neural Network

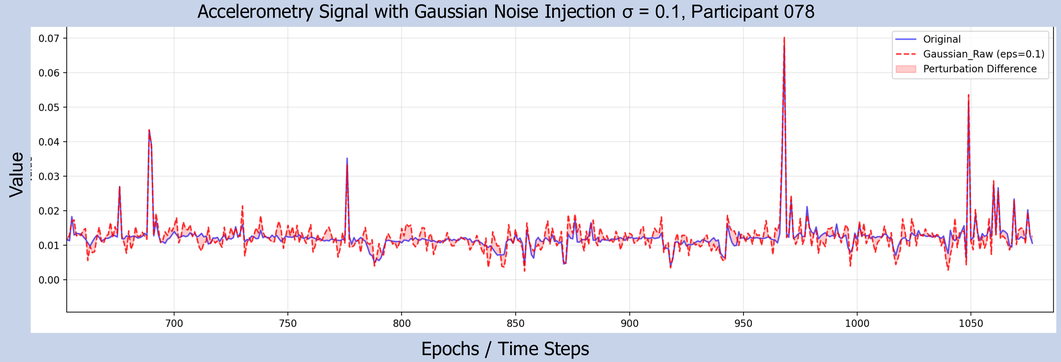
(Korkalainen, H, et al.) [5]

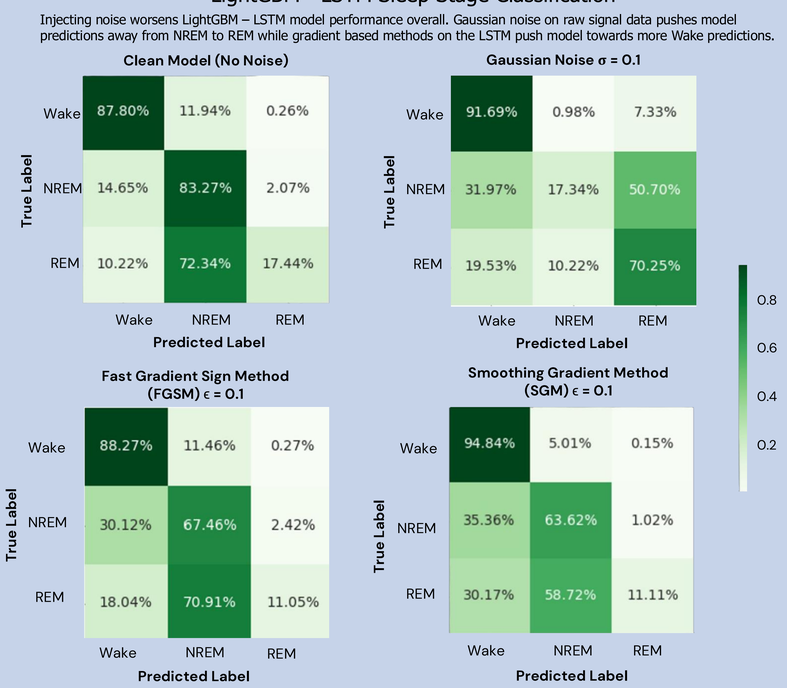
Models complex temporal and cross-sensor feature interactions

5a Apply adversarial attacking techniques on all models to test robustness

F1 Scores on Selected Models and Attacking Techniques

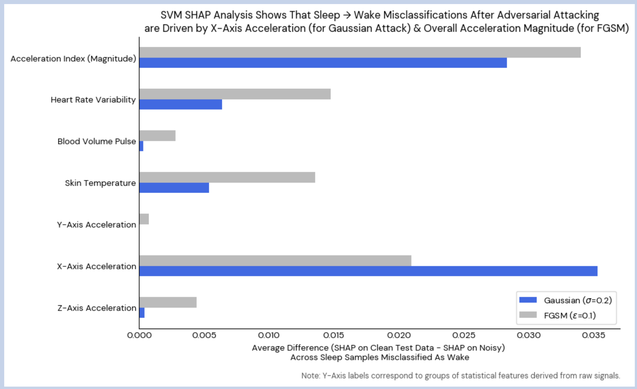
	Baseline	Gaussian Noise (σ = 0.05)	Gaussian Noise (σ = 0.1)	Gaussian Noise (σ = 0.2)	FGSM (ϵ = 0.1)	Smoothing (ϵ = 0.1)
SVM (2 Stage)	0.81	0.81	0.81	0.80	0.75	-
LightGBM + biLSTM (3 Stage)	0.64	0.47	0.45	0.44	0.55	0.55
CNN + RNN (3 Stage)	0.38	0.36	0.36	0.36	-	-

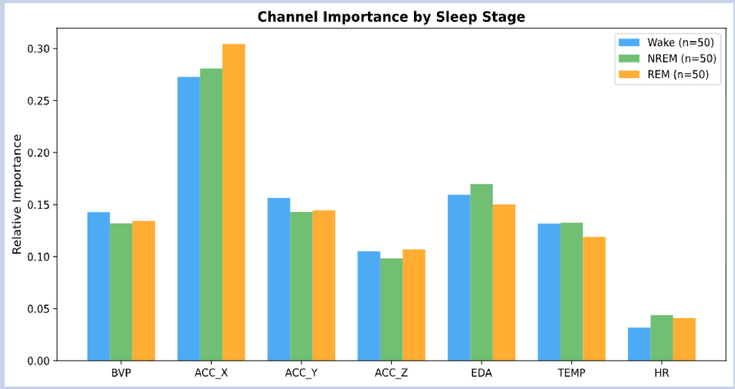




5b Use explainability techniques to understand how feature attributions change with noise injection

Differencing SHAP values for samples that are misclassified by the SVM model after an adversarial attack helps us see what feature groups push less towards a sleep classification (which is the true classification) after the attack, and thus what groups are most impacted by noise and drive misclassification.





Grad-CAM was used on the CNN-RNN sleep staging model to evaluate feature importance by channel/signal. Results were averaged across 50 balanced samples per class. X-Axis acceleration (ACC_X) emerges as the dominant feature across all three sleep stages. Its relative importance is consistent with the two other models' results.

[3] Wang, Ke, et al. "DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology" (version 2.1.0). PhysioNet (2025). RRID:SCR_007345. https://doi.org/10.13026/7r9r-7r24
[4] Logacjov, Aleksej et al. "A Machine Learning Model for Predicting Sleep and Wakefulness Based on Accelerometry, Skin Temperature and Contextual Information." Nature and science of sleep vol. 16 699-710. 6 Jun. 2024, doi:10.2147/NSS.S452799
[5] Korkalainen, Henri, et al. "Deep Learning Enables Sleep Staging from Photoplethysmogram for Patients with Suspected Sleep Apnea." Sleep, vol. 43, no. 11, 21 May 2020, https://doi.org/10.1093/sleep/zsaa098.