Goal: Train baseline classifiers, generate adversarial examples, test defenses, and compare performance under varying attack strengths.

Baseline models:

Train two baseline classifiers:

- Random Forest
- Neural Network (MLP) — required for gradient-based attacks.

Baseline Evaluation

Report the following metrics for each baseline:

- Accuracy
- Precision
- Recall
- F1 score
- AUC (ROC)

Adversarial example generation (against the NN):

- Methods: **FGSM** and **PGD**.
- Vary attack strength across several ε (epsilon) values.
- Measure and record how the NN's performance degrades as ε increases.

Notes for Random Forest:

- Option A — Transfer attack: Attack a substitute NN and test the transferability of those adversarial examples against the Random Forest.
- Option B — Constrained feature perturbation: Apply realistic, domain-specific feature perturbations (simulate plausible changes) and evaluate the Random Forest.

Defenses to implement and compare

Implement the following three defenses and evaluate each against FGSM/PGD attacks at the same ε values used above:

1. Adversarial training
- Augment training data with adversarial examples (FGSM / PGD) and re-train.
2. Noise injection
- Add Gaussian noise to inputs during training (and optionally at inference) to increase robustness.
3. Adversarial detector
- Train a detector (e.g., autoencoder reconstruction error or isolation forest) to flag likely adversarial inputs before they reach the classifier.

Experiments & outputs (required figures/tables)

- Performance vs. ε plots for each model and defense.
- Confusion matrices for key conditions for each defense.
- ROC curves (clean vs. attacked) and AUC values.
- Comparison table summarizing: clean metrics, degraded metrics at selected ε values, and defense-specific results.