Project Guidelines

Project Theme

Students may choose **one of the following application domains**:

1. **Human Action Recognition (HAR)**
2. **Cancer Identification** using medical imaging data
3. **Industrial Anomaly Detection** - this dataset is proposed by **Renault Group and eVantage company** (script for model evaluation)

Note: For HAR and cancer identification you can find a set of datasets at the end of this document.

Both domains are highly relevant for the course topic *AI for Trustworthy Decision Making* as they raise crucial questions about **data privacy**, **model fairness**, **robustness**, and **interpretability**.
In **HAR**, data originates from personal sensors, making **privacy preservation and fairness across users** critical.
In **medical imaging**, models influence clinical decisions, where **transparency, accountability, and bias mitigation** are essential for trustworthy deployment.

The project should demonstrate how trustworthy AI principles can be applied in one of these domains — ideally through techniques such as **Federated Learning (FL)**, **Differential Privacy (DP)**, or **Explainable AI (XAI)**.

Project Format and Stages

Students can work individually or in a team of three.
Projects are structured in **two graded stages**:

Stage	Description	Weight
Stage 1 – Baseline Development and Minimal Federated Setup	Implementation of a functional baseline and minimal federated setup; initial results and analysis	**2 points**

Stage 2 – Trustworthiness Enhancements and Evaluation	Integration of trust mechanisms, complete evaluation, report, and presentation	**4 points**

STAGE 1 (2 points) – Baseline Development and Minimal Federated Setup

Objectives

- Understand the problem and formulate research questions.
- Select and preprocess a dataset suitable for the chosen domain (HAR or medical).
- Implement a **working baseline model** (centralized or federated).
- Demonstrate at least one **minimal federated learning experiment** using an existing framework (e.g., Flower, TensorFlow Federated, FedML).
- Perform initial evaluation with standard metrics and brief analysis.

Deliverables

1. **Technical Report (3–4 pages)** including:
    - Motivation and problem definition.
    - Related work (3–5 references).
    - Dataset description and preprocessing pipeline.
    - Baseline model architecture.
    - Initial federated setup description (number of clients, partitioning, aggregation method).
    - Preliminary results (accuracy, F1, or AUC).
    - Short discussion on observed limitations and possible trust dimensions to address in Stage 2.
2. **Code Deliverables**
    - **Baseline implementation** (centralized training).
    - **Minimal federated experiment** using any open-source model or framework.
    - Configuration files or notebooks reproducible in Colab or local environment.
3. **Short Presentation**
    - Dataset, model, preliminary results, and identified next steps for improving trustworthiness.

Expected Technical Work

- Implement or adapt an existing model (e.g., from PyTorch or TensorFlow examples).
- Simulate 3–5 federated clients (even on a single machine) with non-IID data splits.
- Train and evaluate a simple FL setup (e.g., using FedAvg).

- Compare performance between centralized and federated versions.
    - Students working individually should compare their federated baseline against a state-of-the-art (SOTA) model from the literature.
    - Teams must coordinate so that one or two members train models using centralized setups (data aggregated) and other members train models in a federated configuration.

STAGE 2 (4 points) – Trustworthiness Enhancements and Evaluation

Objectives

- Integrate **trust-enhancing mechanisms** such as:
    - Differential Privacy (DP-SGD, noise injection).
    - Fairness-aware training or evaluation across subgroups.
    - Robustness to noise or data poisoning.
    - Interpretability (Grad-CAM, LIME, SHAP) for understanding model behavior.
- Analyze the trade-offs between accuracy and trustworthiness.
- Document, evaluate, and present findings comprehensively.

Deliverables

1. **Final Report**
    - 6–8 pages (individual) / 18–20 pages (team).
    - Clear separation between baseline (Stage 1) and enhancements (Stage 2).
    - Quantitative evaluation of both performance and trust-related metrics.
    - Ethical discussion and future work suggestions.
    - The report may be **generated directly from the project's notebook** (e.g., Jupyter/Colab), provided that:
        - the notebook includes **structured narrative sections** (titles, descriptions, markdown explanations);
        - **code cells are clearly commented** and accompanied by interpretations of the results;
        - all figures, tables, and results are properly labeled and discussed;
        - the notebook is **self-contained and readable as a standalone document**, suitable for evaluation without external context.
    -
2. **Complete Code Repository**
    - With modular organization and README for reproducibility.
    - Each script or notebook clearly linked to the corresponding report section.
3. **Final Presentation**
    - **Individual projects**
    - **Team projects**: each member must present part of the work

Individual Project Requirements

Each individual project must include:

1. A baseline model and a federated version tested on a small dataset.
2. At least one implemented trust mechanism (privacy, interpretability, fairness, or robustness).
3. Quantitative analysis comparing centralized vs. federated vs. trust-enhanced versions.
4. A reflection on ethical and societal implications.

Team Project Requirements (3-4 Members)

General Guidelines

- The team must address **at least two trust dimensions** (e.g., privacy + interpretability, or fairness + robustness).
- All members should contribute to **data handling, model training, and evaluation**, ensuring comparable workloads.
- Teams must coordinate their work through **shared experiments and cross-evaluation**, not isolated modules.
- The project must clearly document how collaboration and mutual validation were achieved.

Collaborative Workflow and Role Options

To promote balanced work and deeper insight into trustworthy learning, each team should organize around **three complementary roles**, but with shared responsibilities in training and testing.

Role	Title	Main Responsibilities
Member 1	*Data & Experiment Design Lead*	- Selects or defines one dataset variant (e.g., subset, modality, or client distribution). - Prepares preprocessing pipeline and federated setup (client simulation, partitioning). - Conducts initial training on their dataset split. - Shares model weights for cross-evaluation by teammates.

Member 2	*Modeling & Fairness/Privacy Lead*	- Implements the main model architecture (e.g., CNN/LSTM/FedAvg baseline). - Integrates one trust-enhancing mechanism (e.g., Differential Privacy, fairness constraint). - Trains the model on a different dataset variant or client group. - Exchanges trained models with others for testing and comparison.
Member 3	*Evaluation & Robustness/Interpretability Lead*	- Performs interpretability (Grad-CAM, SHAP, etc.) or robustness testing (noise, data shifts). - Trains an alternative version of the model or runs fine-tuning using another data partition. - Evaluates all models (own + teammates') using consistent trust metrics and reports comparative results.

If the project is conducted by a team of four, the additional member must focus on cross-domain generalization and data distribution analysis. Specifically, this member will curate or select a new dataset and evaluate how models trained by the other team members perform on this unseen data. Their task is to analyze domain shift effects, quantify performance degradation, and discuss trust implications (e.g., robustness, fairness, or reliability across sources).

Cross-Evaluation and Collaboration Options

Teams are required to design **collaborative experimental protocols**, for instance:

1. **Cross-Model Evaluation:**
   - Each member trains one model; all models are tested by the other two members on their local data splits.
   - This simulates real federated or multi-center validation.
2. **Data Diversity:**
   - Each member uses a distinct dataset subset or source.
   - For medical tasks, they may use **different label sources** (e.g., labels from multiple radiologists) or simulate **annotation uncertainty**.
   - For HAR, they may simulate **different users or devices**.
3. **Trust Dimension Diversity:**
   - Each member focuses on one trust aspect (privacy, fairness, robustness).
   - Final integration compares trade-offs and discusses compatibility of the mechanisms.
4. **Model Exchange and Integration:**
   - Members periodically exchange model weights to fine-tune on each other's data.
   - Optionally, aggregate results via FedAvg or a simple averaging protocol to simulate distributed consensus.

Deliverables (Team Project)

1. **Joint Report (18–20 pages)**
   - Unified structure, but with **clearly attributed contributions** (each member's dataset, model, and evaluation).
   - Include cross-evaluation tables and inter-member comparisons.
   - Discuss observed discrepancies (e.g., model bias due to different labels or sensors).
2. **Shared Repository (GitHub or similar)**
   - Each member maintains a subfolder or branch for their experiments.
   - Include scripts for aggregation and comparative evaluation.
   - Commit history must show parallel and intersecting contributions.
3. **Group Presentation**
   - Joint introduction and conclusions.
   - Each member presents their specific dataset/model and how cross-validation or model sharing was performed.

Suggested Datasets for Human Activity Recognition (HAR)

- WISDM–Actitracker Dataset (Wireless Sensor Data Mining)
- NTU RGB+D 60 / 120
- OPPORTUNITY Activity Recognition
- PRECIS HAR

Suggested Datasets for Medical Projects

- LIDC-IDRI (Lung Image Database Consortium)
- ISIC Challenge Datasets
- COVIDx CXR-4
- Medical Segmentation Decathlon (MSD)
- INBreast Dataset
- CBIS-DDSM: Breast Cancer Image Dataset
- DMID Breast Cancer Mammography Dataset