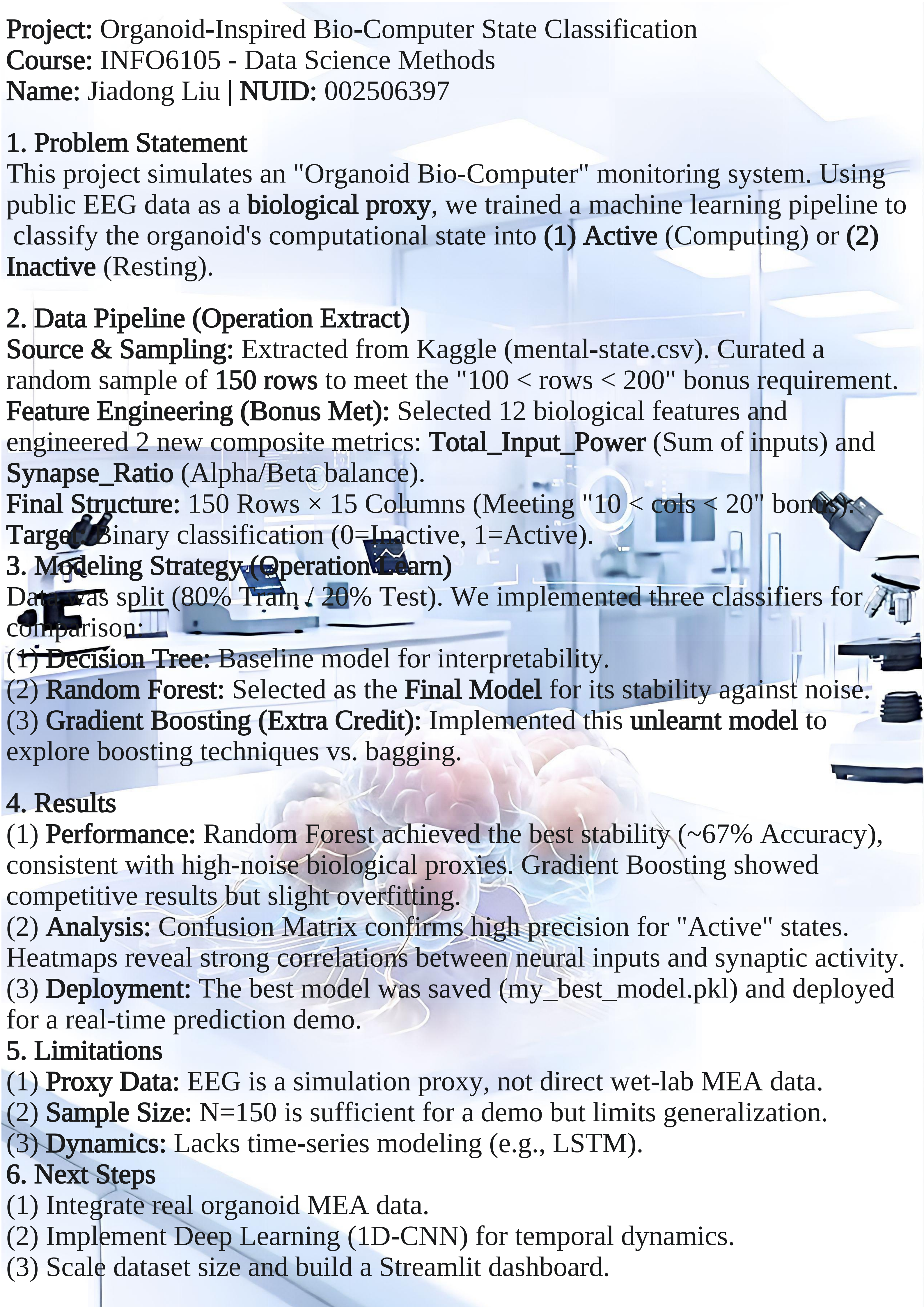

Project: Organoid-Inspired Bio-Computer State Classification
Course: INFO6105 - Data Science Methods
Name: Jiadong Liu | **NUID:** 002506397

1. Problem Statement

This project simulates an "Organoid Bio-Computer" monitoring system. Using public EEG data as a **biological proxy**, we trained a machine learning pipeline to classify the organoid's computational state into **(1) Active** (Computing) or **(2) Inactive** (Resting).

2. Data Pipeline (Operation Extract)

Source & Sampling: Extracted from Kaggle (mental-state.csv). Curated a random sample of **150 rows** to meet the "100 < rows < 200" bonus requirement.
Feature Engineering (Bonus Met): Selected 12 biological features and engineered 2 new composite metrics: **Total_Input_Power** (Sum of inputs) and **Synapse_Ratio** (Alpha/Beta balance).
Final Structure: 150 Rows × 15 Columns (Meeting "10 < cols < 20" bonus).
Target: Binary classification (0=Inactive, 1=Active).

3. Modeling Strategy (Operation Learn)

Data was split (80% Train / 20% Test). We implemented three classifiers for comparison:
(1) **Decision Tree:** Baseline model for interpretability.
(2) **Random Forest:** Selected as the **Final Model** for its stability against noise.
(3) **Gradient Boosting (Extra Credit):** Implemented this **unlearnt model** to explore boosting techniques vs. bagging.

4. Results

(1) **Performance:** Random Forest achieved the best stability (~67% Accuracy), consistent with high-noise biological proxies. Gradient Boosting showed competitive results but slight overfitting.
(2) **Analysis:** Confusion Matrix confirms high precision for "Active" states. Heatmaps reveal strong correlations between neural inputs and synaptic activity.
(3) **Deployment:** The best model was saved (my_best_model.pkl) and deployed for a real-time prediction demo.

5. Limitations

(1) **Proxy Data:** EEG is a simulation proxy, not direct wet-lab MEA data.
(2) **Sample Size:** N=150 is sufficient for a demo but limits generalization.
(3) **Dynamics:** Lacks time-series modeling (e.g., LSTM).

6. Next Steps

(1) Integrate real organoid MEA data.
(2) Implement Deep Learning (1D-CNN) for temporal dynamics.
(3) Scale dataset size and build a Streamlit dashboard.