Step 1 — Project Plan & Validation

Project Title

Interactive RNA-Seq Classifier and DEG Explorer for Heart Disease Using the Magnetique Dataset

---

1. Scientific Question & Dataset (Combined)

Rough Idea and Brainstorming

Using GenAI brainstorming (ChatGPT), several ideas were explored around integrating transcriptomic data for heart disease prediction and gene prioritization. Simpler RNA-only strategies were selected for feasibility. Among possible directions (multi-omics integration, network inference, or eQTL mapping), a two-group RNA-seq approach emerged as both realistic and scientifically meaningful.

Refined Question

Scientific question:
Can machine-learning models trained on normalized RNA-seq count data distinguish cardiac disease from control samples and identify the most important genes driving this distinction through interpretable methods (e.g., SHAP values) while also confirming differential expression through statistical analysis?

Dataset Selection

Dataset: Magnetique (heart failure study).
Source: https://zenodo.org/records/7547318 (described in *Sci Data* 2023, DOI 10.1038/s41597-023-01934-5).
Description: RNA-seq data from left-ventricular tissue of patients with dilated cardiomyopathy, ischemic cardiomyopathy, and non-failing donor hearts.
Size: ~180 samples × ~20 000 genes (counts).
License/Ethics: Publicly available under Creative Commons BY 4.0; fully de-identified.
Subset used: Dilated cardiomyopathy (DCM) vs. non-failing donor (NF) for a two-group comparison.

Scientific Relevance

Heart failure is a major global health burden. Understanding expression signatures distinguishing diseased from healthy tissue can reveal candidate genes and pathways. A lightweight, interpretable RNA-seq classifier allows fast hypothesis testing and educational use without accessing raw sequencing reads.

Task Type

- **Differential expression (statistical)**
- **Supervised classification (ML)**
- **Model interpretability (SHAP feature importance)**

---

2. Exploration Goals

Before modeling, the following exploratory analyses will be performed:

1. **Data distribution & normalization:** Examine library sizes, variance stabilization, and log-transformed counts to confirm normality assumptions.
2. **Group balance & sample quality:** Check the number of DCM vs. NF samples, remove outliers using PCA.
3. **Gene filtering:** Investigate expression ranges, proportion of low-count genes, and their effect on downstream performance.

Optional exploratory checks include correlation of top features and enrichment of top DEGs.

---

3. Proposed Model & Evaluation

3.1 Simple ML models

Baseline Model

- Logistic Regression on normalized expression of top variable genes.

Additional Models (Benchmarked)

- Random Forest
- Decision Tree
- Support Vector Machine (linear kernel)

All models will use 5-fold cross-validation.

Evaluation Metrics

- Accuracy, Precision, Recall, F1-score, ROC AUC.
- Model interpretability via SHAP values (top 50 genes).
- Concordance of SHAP-important genes with DEGs (from DESeq2) as a biological validation step.

Good performance definition:
AUC > 0.85 with biologically meaningful top SHAP features overlapping known cardiac genes.

3.2 Extension with Advanced AI Methods

To strengthen the project's methodological scope and align with the course expectations, an **advanced AI component** will be added alongside the baseline models.

Deep Neural Network (DNN) Model

A simple fully connected **feed-forward neural network** will be implemented using TensorFlow/Keras.
Architecture:

- Input: normalized expression matrix (top 1000 most variable genes).
- Layers:
  - Dense(512) → ReLU → Dropout(0.3)
  - Dense(128) → ReLU → Dropout(0.2)
  - Dense(1) → Sigmoid
- Optimization: Adam optimizer, binary cross-entropy loss.
- Training: 70/30 train–test split, early stopping to prevent overfitting.

This design ensures the model remains computationally light and can run within Google Colab in under 30 minutes.

Explainability and Interpretation

To maintain interpretability and biological insight:

- **SHAP values** will be used to quantify the contribution of each gene to the neural network's predictions.
- **Integrated Gradients (IG)** will be applied as a complementary interpretability approach to validate top predictive genes.
- The overlap between DNN-explained genes and DEGs will be visualized as a Venn diagram and compared with the traditional ML results.

Evaluation

The DNN's performance will be benchmarked against baseline models (logistic regression, random forest, SVM).
Metrics: ROC-AUC, accuracy, and F1-score.
Good performance will be defined as AUC > 0.85 and consistent biological interpretability (high SHAP/IG importance for known cardiac genes).

SHAP values will be applied to both traditional ML models (via TreeSHAP or Linear SHAP) and the deep neural network (via Deep SHAP) to identify the most influential genes driving disease–control classification.

Rationale and Feasibility

The DNN extension introduces a meaningful deep-learning component without exceeding hardware or time limits. The neural model enhances feature abstraction and non-linear pattern

recognition in RNA-seq data, providing a stronger predictive layer while preserving transparency through explainability methods.

---

4. Accessibility Plan

Chosen Option: A – Web Application

The final workflow will be deployed as a **Shiny** (or Streamlit) web app with the following modules:

1. **Upload data:** Users upload an RNA-seq count matrix (genes × samples) and a metadata file defining two groups.
2. **Run analysis:** Performs normalization, DEG analysis (DESeq2), and trains selected classifiers.
3. **Visualize results:** Displays:
    - Volcano plot of DEGs
    - Confusion matrix and ROC curve
    - SHAP summary and feature importance plots
4. **Download results:** Summary tables and plots downloadable; user data not stored.

The public GitHub repository and demo web app will be accessible through the course website.

---

5. Feasibility Check

Aspect	Description
Timeframe	Each analysis (DESeq2 + ML + SHAP) expected < 1 hour on a laptop/Colab.
Hardware	CPU runtime only, < 8 GB RAM.
Risks	(1) Imbalanced sample groups may bias models; mitigated by cross-validation. (2) Noisy genes or low counts; mitigated by filtering.
Deliverables	(i) RNA count processing pipeline notebook (ii) trained model benchmark report (iii) interactive web app with upload support.

Conclusion:
The project is fully feasible within 2–3 weeks, educationally valuable, and relevant to precision transcriptomics.

Appendix — AI Deep Research Transcript (Summary)

Completed chat history: https://chatgpt.com/share/68e4326c-c598-800b-82cc-9daa64a90f58

User Input Examples

I have RNA-seq count data (CHD / healthy); I need a feasible two-group design for a course project with a web interface.

AI Outputs (ChatGPT brainstorm)

- Suggested limiting to RNA-only for feasibility.
- Proposed Magnetique dataset for public reproducibility.
- Recommended combining DE analysis + classification + SHAP interpretation for interpretability.
- Outlined technical stack: Python (pandas, scikit-learn, SHAP) + R (DESeq2) + Streamlit/Shiny.

Decision Outcome
Adopted disease vs. control two-group design with DEGs + ML + web app workflow.