Overview:
We are building an interpretable machine learning “liquid biopsy” model that uses blood-based gene expression to diagnose osteoarthritis (OA). The model will be trained on a real human dataset (GSE48556, PBMC microarray data) and will output a simple symbolic equation involving a small number of genes that can predict whether someone has OA.

[REFERENCE PAPER INSPIRATION]:
Copy of INTERPRETABLE ML FOR PANCREATIC CANCER

The Problem/Gap:

- Osteoarthritis is extremely common and a leading cause of disability, but it is usually diagnosed late, only after joint damage is visible on X-rays or patients have significant symptoms.
- There is no widely used blood test for early OA detection; current diagnosis relies on imaging and clinical symptoms that appear after irreversible damage.
- Early, non-invasive diagnosis could delay progression, reduce pain, and postpone or prevent joint replacement.

Goal: Develop an early, blood-based diagnostic model for OA using gene expression from immune cells in the blood (PBMCs).

Key Requirements:

- The model must be interpretable, not a black box.
- It should use only a few genes (e.g., 2–3) in a clear mathematical formula.
- It should perform competitively with standard black-box machine learning models.

Dataset: [LINK]

- Dataset ID: GSE48556 (NCBI GEO).
- Samples: 139 total → 106 patients with osteoarthritis + 33 healthy controls
- Tissue: Peripheral Blood Mononuclear Cells (PBMCs) → a blood-based “liquid biopsy” signal.
- Technology: Illumina HumanHT-12 v3 microarray (~48,000 probes).

Overall Timeline:

1. Environment and data setup
2. Data cleaning and labeling
3. Feature selection (t-tests on genes)
4. Train–test split
5. Train black-box baseline models (Random Forest, Logistic Regression, Gradient Boosting)
6. Train QLattice interpretable model
7. Extract an interpretable equation and structure
8. Evaluate the QLattice model
9. Compare QLattice vs black-box models
10. Understand biological interpretations of resulting gene outcomes
11. Visualize selected genes and decision logic

Final Deliverables:

1. Interpretable equation using 2–3 genes (your key result).
2. Semantic model diagram from QLattice.
3. Performance tables (baselines vs QLattice).
4. Graphs: confusion matrices, ROC curves, gene scatter/decision plots.
5. Biological interpretation of the selected genes as potential OA blood biomarkers.