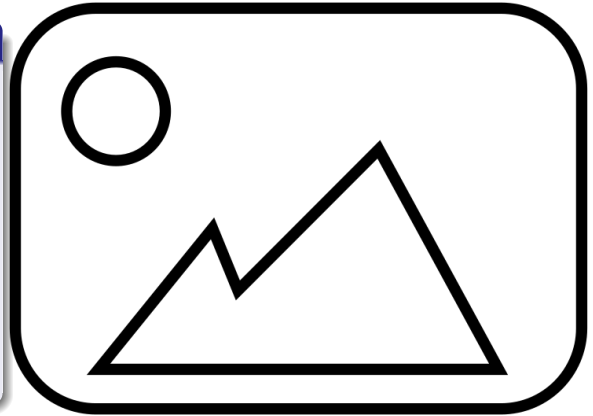
Step 4: Downstream Classification

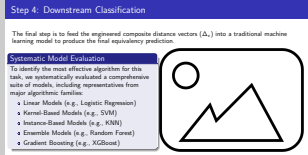
The final step is to feed the engineered composite distance vectors (Δ_c) into a traditional machine learning model to produce the final equivalency prediction.

Systematic Model Evaluation

To identify the most effective algorithm for this task, we systematically evaluated a comprehensive suite of models, including representatives from major algorithmic families:

- Linear Models (e.g., Logistic Regression)
- Kernel-Based Models (e.g., SVM)
- Instance-Based Models (e.g., KNN)
- Ensemble Models (e.g., Random Forest)
- Gradient Boosting (e.g., XGBoost)





1. The goal here was to find the best possible classifier for our specific feature vectors. By testing models from different families, we could probe the data for things like linear separability, complex non-linear decision boundaries, and feature interactions.
2. The graphic on the right shows the simple version of this process: our feature vector goes in, the classifier makes a judgment, and a binary prediction comes out.
3. The specific results of this competitive evaluation—that is, which models performed best—will be discussed in the upcoming Results section.

Finding 2: Final Classifier Performance

The final evaluation, conducted on the held-out test data, measured the efficacy and efficiency of the top-performing classifiers.

Exceptional Performance Across the Board

All finalist models (SVM, RF, XGBoost, KNN) achieved exceptionally high and stable performance, with mean F_1-scores approaching or exceeding **0.97**.

The Accuracy vs. Efficiency Trade-Off

Our analysis revealed a classic performance trade-off, leading to context-dependent recommendations:

- **For Maximum Accuracy:** The **Support Vector Machine (SVM)** was the statistical winner, proving to be the most accurate and consistent classifier.
- **For Optimal Efficiency: Random Forest (RF) and XGBoost** were nearly as accurate but an order of magnitude faster and more predictable at inference time, making them practical choices for a scalable, low-latency system.