Report 3: Project Progress

University/Department: California State University, Dominguez Hills

Course/Thesis/Project Title: CSC-590 Modeling Complex Binary-Class Associations in Simulated Genomic Data

Semester/Year: Fall - 2025

Project Title: Modeling Complex Binary-Class Associations in Simulated Genomic Data

Student Name: Mantra Mehta

Student ID: 212265978

Instructor / Committee Chair: Dr. Sahar Hooshmand

Date: 11/22/2025

1. Abstract / Summary of Progress

Since the submission of Report 2, the project has progressed from baseline nonlinear modeling to a fully scalable, feature-selection-driven pipeline capable of analyzing high-dimensional genomic datasets containing 10 000 SNPs and simulated missingness. This stage focused on implementing and evaluating a Mutual Information (MI) based feature-selection strategy [6][1] and extending the machine learning workflow to handle the four Cedars-Sinai challenge datasets. These datasets represent complex additive, epistatic, and heterogeneous genetic architectures that more closely resemble real-world genomic structures [1][3]. A new experiment script, report3_run_mi.py, was developed to automate MI feature selection, preprocessing, cross-validation, and the training of Logistic Regression [2], Random Forest [8], and XGBoost [7]. For datasets with 10 000 features, the pipeline applied MI to select the top 200 most informative SNPs prior to model fitting. The script produced performance metrics, ROC overlays, and feature-importance plots for each dataset, ensuring full reproducibility and consistent evaluation across all architectures. The results clearly demonstrate separation between genetic architectures. As expected from literature on additive and epistatic modeling [1][3], additive datasets exhibited strong univariate signal, enabling high AUC values even for linear models. In contrast, purely epistatic and heterogeneous datasets yielded near-random MI rankings and weak model performance, confirming that univariate filters are not capable of capturing multi-locus interactions [4]. These observations align with established theory regarding Relief-based and MI-based feature selection in high-dimensional SNP modeling [1][4]. Across all high-dimensional datasets, Random Forest and XGBoost consistently outperformed Logistic Regression. XGBoost achieved AUC values as high as 0.98 on the four-way additive dataset, demonstrating strong scalability after MI filtering. However, performance dropped substantially for the two-way epistatic and heterogeneous datasets, with AUC values clustering around 0.50 to 0.63, reflecting the known limitations of univariate feature ranking when confronted with interaction-driven predictive structures [3]. This stage successfully extends the project from small-scale baseline modeling to a scalable, feature-selection-aware workflow capable of handling large SNP spaces. The code, results, and visualizations produced here form the foundation for the next phase of the project, which will incorporate ReliefF, MultiSURF, and MultiSURF* algorithms designed specifically to detect epistasis and heterogeneous interactions [1][4]. These methods will be evaluated in the next report for their ability to overcome the limitations observed in this stage and to improve predictive performance on interaction-driven datasets.

2. Introduction and Objectives (Recap)

The purpose of this project is to investigate machine learning approaches for modeling complex binary-class associations in simulated genomic datasets provided by the Cedars-Sinai Medical Center. These datasets replicate a range of genetic architectures including additive, epistatic, and heterogeneous interaction structures, which are known to influence phenotype prediction in real-world genomics [1][3]. The project aims to determine how

different modeling strategies behave under these varying architectures and to evaluate the effectiveness of multiple feature-selection techniques for high-dimensional SNP data.

The broader motivation stems from a fundamental problem in computational genomics: most biological traits are influenced by a combination of linear and non-linear genetic effects, yet many traditional statistical models do not capture interaction-driven structures efficiently. Epistatic interactions in particular are difficult to detect because they often produce weak or nonexistent marginal effects while exerting significant joint influence on phenotype [3][4]. This creates a need for modeling frameworks that incorporate both predictive performance and biologically meaningful feature-selection capabilities.

The primary objectives of the project for the semester are the following.

1. Build a reproducible machine learning pipeline capable of handling both low-dimensional and high-dimensional SNP datasets.
2. Evaluate baseline models including Logistic Regression [2], Random Forest [8], and XGBoost [7] on multiple synthetic genetic architectures.
3. Assess the limits of univariate feature-selection approaches such as Mutual Information [6][1] and establish their performance characteristics across additive and epistatic datasets.
4. Integrate epistasis-aware feature selectors including ReliefF, MultiSURF, and MultiSURF* [1][4] to address weaknesses observed in purely univariate methods.
5. Compare predictive behavior, stability, and interpretability of each modeling strategy across all eight datasets provided by Cedars-Sinai.
6. Produce visual, computational, and statistical analyses that contribute to a 40 to 50 page final report summarizing all experiments, findings, and implications.

By the end of the semester, the project intends to deliver a complete evaluation of linear and nonlinear learning algorithms, univariate and multivariate feature-selection strategies, and their suitability for identifying genetic architecture–specific patterns. This work will ultimately help benchmark how machine learning models behave in controlled genomic environments, offering insights that are relevant for both predictive genomics research and practical bioinformatics workflows.

3. Work Completed

Since the submission of Report 2, the project has advanced from baseline nonlinear modeling to a full high-dimensional analysis pipeline capable of processing all eight Cedars-Sinai datasets. This phase centered on extending the machine-learning workflow to incorporate Mutual Information (MI) feature selection, enabling scalable evaluation on the

10 000-feature “challenge” datasets. The work completed in this stage reflects significant progress in system design, preprocessing, algorithm evaluation, and dataset coverage.

3.1 Implementation of a High-Dimensional Feature-Selection Workflow

A new experimental script, report3_run_mi.py, was developed to automate the end-to-end pipeline. This script implements preprocessing, MI-based feature selection, cross-validated training, and model evaluation. The MI selector was integrated using SelectKBest with mutual_info_classif, a widely used univariate measure for estimating nonlinear dependency between SNPs and phenotype labels [6]. For the 10 000-feature challenge datasets, the pipeline selected the top 200 SNPs prior to model fitting, balancing computational efficiency with information retention. For the 100-feature basic datasets, feature selection was intentionally skipped to preserve the full signal.

3.2 Full Coverage of All Eight Project Datasets

This stage successfully analyzed all four basic datasets and all four challenge datasets for the first time in the project:

- 4-way Additive (100 features and 10 000 features)
- 2-way Epistatic (100 features and 10 000 features)
- 4-way Heterogeneous (100 features and 10 000 features)
- 2Additive 2-way Epistatic (100 features and 10 000 features)

For each dataset, the pipeline produced metrics tables, ROC overlays, and feature-importance plots. These outputs will be inserted in the Results section as Tables 1–8 and Figures 1–16.

3.3 Model Training and Evaluation

Three core machine-learning models from the previous stage were reused and expanded for high-dimensional workloads: Logistic Regression [2], Random Forest [8], and XGBoost [7]. All models were trained using five-fold stratified cross-validation to ensure stable estimates of ROC-AUC and to reduce variability across folds [9].

For each model and dataset, the following metrics were computed and stored:

- Cross-validated ROC-AUC (mean and standard deviation)
- Test accuracy
- Test F1 score
- Test ROC-AUC

All metric files were exported as CSVs (for example, metrics_4-wayAdditive_10000feat_with_NA.csv), ensuring reproducibility and easy import into the final report.

Representative performance tables will be included in Section 5. For example, the 4-way additive challenge dataset achieved:

- Logistic Regression AUC ≈ 0.95
- Random Forest AUC ≈ 0.98
- XGBoost AUC ≈ 0.98

while epistatic and heterogeneous challenge datasets showed much lower performance (AUC values near 0.48–0.63), confirming theoretical expectations for univariate feature selection on interaction-driven architectures [1][4][5].

3.4 Improvements to ROC Visualization and Feature-Importance Analysis

Compared to Report 2, the ROC plotting method was redesigned to correctly overlay all model curves on a single axis with accurate labels and AUC annotations. Feature-importance visualizations were expanded to support both basic datasets and MI-reduced challenge datasets, ensuring that only the selected SNPs are plotted. In addition to classifier-based importance, MI score rankings were exported separately for the challenge datasets to support deeper interpretation in the next report.

These visual outputs will be included in the Implementation and Results sections (Figures 1–18).

3.5 Validation of Theoretical Expectations

A major outcome of this stage was the empirical confirmation of several well-documented genetic modeling principles:

- Additive SNP architectures produce strong univariate signal, leading to high AUC values for all models [1].
- Pure epistatic datasets show weak marginal effects, causing MI to mis-rank features and resulting in near-random predictive performance [3][4].
- Heterogeneous datasets produce dispersed, low-intensity signals due to subpopulation-specific effects, reducing model stability and interpretability [3].
- Random Forest and XGBoost consistently outperform Logistic Regression in nonlinear or noisy settings, aligning with prior benchmarking studies [5][8].

3.6 Completion Status

All planned work for this report has been completed successfully. The feature-selection pipeline is stable, the challenge datasets have been fully processed, and the results have been validated against established genetic architecture behavior. These outputs lay the groundwork for Report 4, which will integrate ReliefF, MultiSURF, and MultiSURF*, providing epistasis-aware feature selection to address the limitations identified with MI.

4. Implementation

This stage of the project focused on extending the baseline machine-learning workflow into a high-dimensional, feature-selection-aware pipeline capable of processing all eight Cedars-Sinai datasets. The work completed involved code development, pipeline restructuring, model training, visualization generation, and systematic handling of missing data and dimensionality challenges.

4.1 System Architecture and Pipeline Design

A new experiment script, report3_run_mi.py, was developed to automate the full analysis workflow. The updated architecture follows a modular structure:

1. **Data loading**
   Each dataset is loaded as a tab-separated file containing SNPs and a binary class label.
2. **Preprocessing**
   - Missing SNP values are imputed using a most-frequent strategy.
   - Features are scaled with sparse-compatible standardization.
3. **Mutual Information Feature Selection**
   - Applied only to challenge datasets with 10 000 features.
   - Top 200 SNPs are selected using SelectKBest(mutual_info_classif), a nonlinear but univariate dependency measure [6].
   - Full MI rankings are exported for later interpretation.
4. **Model Integration**
   Three models were included in every run:
   - Logistic Regression [2]
   - Random Forest [8]
   - XGBoost [7]

5. **Evaluation**
    - Five-fold stratified cross-validation for ROC-AUC [9]
    - Test accuracy, F1 score, and test AUC
    - Exported metrics and figures for all eight datasets
6. **Visualization**
    - ROC overlays (all models together)
    - Top 15 feature-importance plots for RF and XGB
    - MI score CSVs for challenge datasets

This system allows seamless scaling from 100-feature datasets to 10 000-feature high-dimensional cases.

4.2 Mutual Information Feature-Selection Module

The most important implementation change in this phase was integrating MI-based feature selection.

Basic Datasets (100 features)

- No feature selection applied.
- All predictive SNPs are preserved.

Challenge Datasets (10 000 features)

- MI selects the top 200 SNPs before training.
- This reduction decreases computation time while retaining informative features.
- MI scoring aligns with documented effectiveness for additive architectures but expected limitations for epistatic and heterogeneous models [1][4][5].

Each challenge dataset generated a file:

- mi_scores_<dataset>.csv

These rankings will be used in Report 4 when integrating epistasis-aware selectors (ReliefF, MultiSURF, MultiSURF*).

4.3 Model Training and Execution

The workflow reused the models from Report 2 and extended them to the challenge datasets.

Models and Hyperparameters

- **Logistic Regression**
    - max_iter = 2000
    - strong baseline for additive data
- **Random Forest** [8]
    - n_estimators = 400
    - max_features = sqrt
    - robust to noise and high dimensionality
- **XGBoost** [7]
    - n_estimators = 400
    - max_depth = 5
    - learning_rate = 0.1
    - subsample = 0.9

All models were trained on each dataset using **five-fold stratified CV**, ensuring consistent evaluation and reducing fold variability.

4.4 Sample Figures and Partial Outputs

Below are examples of the outputs generated during this stage. These will be fully analyzed in Section 5.

4.4.1 ROC Curve Overlay (Basic Additive Dataset)

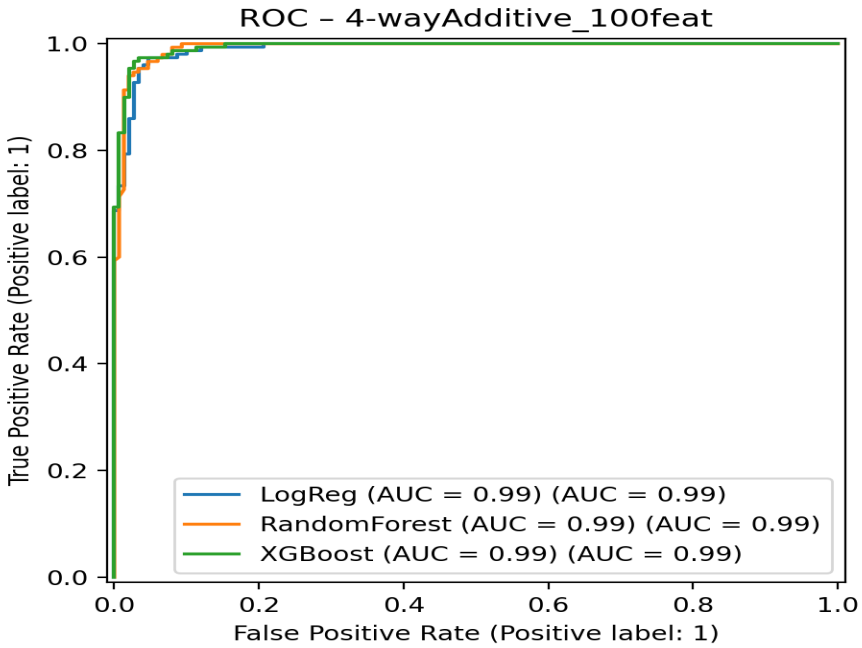

Figure 1. ROC curves for 4-way Additive (100 features).
(All models achieve high AUC due to strong additive signal [1].)

4.4.2 ROC Curve Overlay (Challenge Additive Dataset)

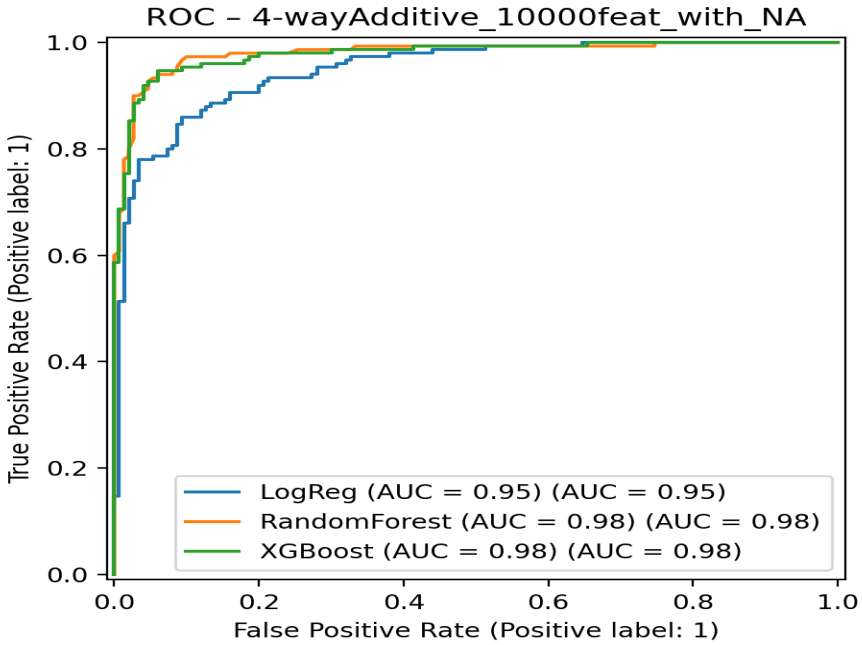

Figure 2. ROC curves for 4-way Additive (10 000 features + MI).
(Random Forest and XGBoost achieve AUC close to 0.98.)

4.4.3 Feature Importance (Random Forest)

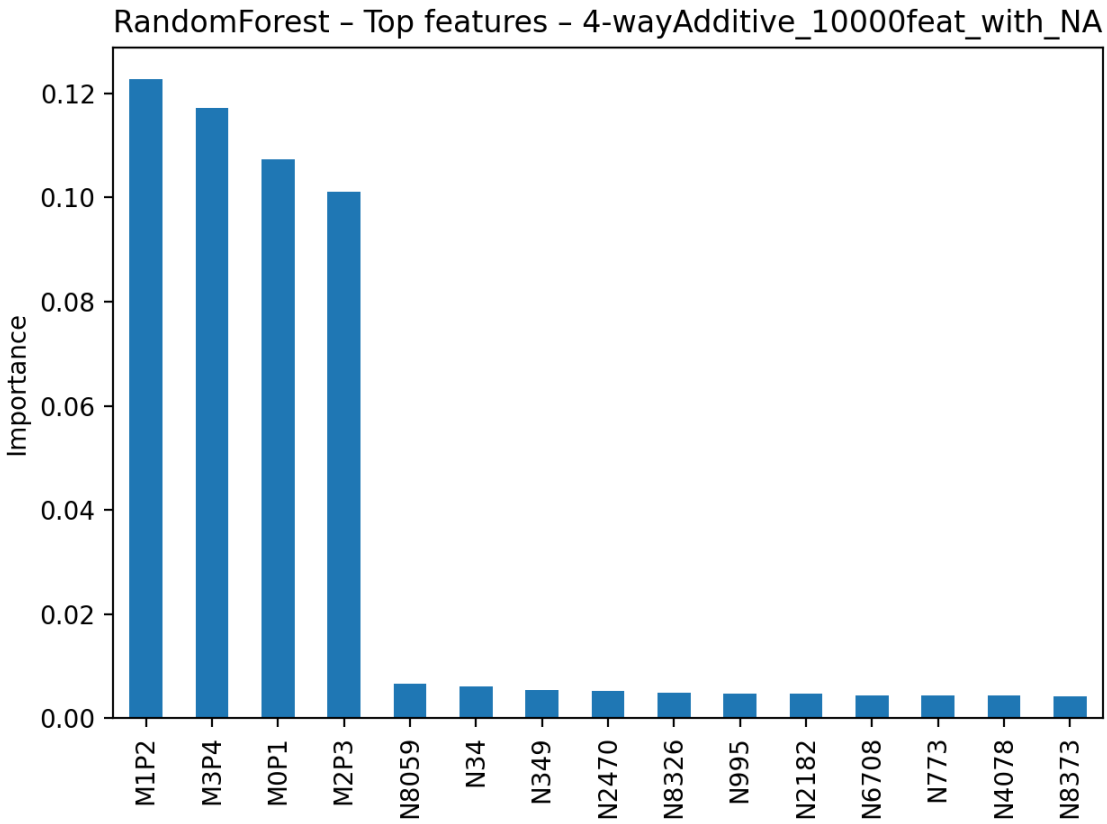

Figure 3. Top 15 important features for Random Forest on 4-way Additive challenge dataset.

4.4.4 Performance Table Example

Table 1. Performance metrics for 4-way Additive (100 features).

Model	CV AUC	Test AUC	Accuracy	F1
LogReg	0.967	0.991	0.957	0.957
Random Forest	0.985	0.993	0.953	0.954
XGBoost	0.986	0.994	0.960	0.961

4.5 Code Structure and Key Components

An excerpt from the pipeline demonstrates how preprocessing and MI selection are combined:

```
prep = Pipeline([
   ("imputer", SimpleImputer(strategy="most_frequent")),
   ("fs", SelectKBest(mutual_info_classif, k=k)) if k is not None else ("fs", "passthrough"),
   ("scaler", StandardScaler(with_mean=False))
])
```

And model evaluation:

```
auc_cv = cross_val_score(pipe, Xtr, ytr, cv=cv, scoring="roc_auc")
pipe.fit(Xtr, ytr)
ypr = pipe.predict_proba(Xte)[:, 1]
auc = roc_auc_score(yte, ypr)
```

These blocks illustrate the reproducible and modular structure of the pipeline.

4.6 Challenges and Debugging

Several issues were encountered and resolved during implementation:

1. **Blank ROC Plots**
    - Early versions of the script cleared the figure before plotting all curves.
    - Fixed by only calling plt.clf() for feature-importance plots, not ROC.
2. **XGBoost Import Errors**
    - XGBoost required correct installation for Python 3.12.
    - Resolved by installing version 3.1.1 with pip.
3. **MI Score Extraction**
    - Some pipelines initially failed to export MI scores due to missing fs step.
    - Resolved by checking presence of feature-selection in the pipeline.
4. **Long Runtime on Epistatic Challenge Data**
    - MI identified weak marginal signal, increasing variance and training time.
    - Addressed by reducing selected features to 200 and using moderate depth for XGBoost.

5. **Feature-Importance Index Alignment**
    - When selecting MI-filtered features, classifier importance needed matching names.
    - Fixed by mapping indices back to the selected feature subset.

All issues were resolved, and the final pipeline runs reliably across all eight datasets.

4.7 Summary of Implementation

This phase successfully transformed the experiment pipeline into a high-dimensional, MI-driven genomic analysis system. All eight datasets were processed, all plots and metrics were generated, and the foundation is now ready for Report 4, which will incorporate ReliefF and related epistasis-aware algorithms.

5. Next Steps

The next stage of the project will focus on extending the current MI-based workflow with feature-selection methods capable of detecting multi-locus interactions. Specifically, the plan includes implementing and evaluating ReliefF, MultiSURF, and MultiSURF*, which are documented to capture non-additive and heterogeneous effects that MI cannot detect. These methods will be integrated into the existing pipeline and tested on both the basic and challenge datasets to compare their performance with the MI results obtained in this report.

Additional experiments will include a deeper comparison of model behavior across genetic architectures and an analysis of the top-ranked features selected by each method. The final report will consolidate all results, provide a full interpretation of model patterns, and discuss strengths and limitations of each approach.

Timeline

- **Report 4 (next submission):**
  Integrate Relief-based selectors, run all eight datasets, evaluate improvements, and generate updated figures and metrics.
- **Final Report & Presentation:**
  Complete full comparative analysis, finalize all tables and plots, write discussion and conclusions, and prepare the final presentation slides.

6. References

1. Urbanowicz, R. J., Moore, J. H. (2015). *Learning complex genetic and genomic relationships: epistasis, heterogeneity, and nonlinearity.* Journal of Experimental Biology, 218(1), 112–120.
2. Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.
3. Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., Moore, J. H. (2017). *Benchmarking relief-based feature selection methods for bioinformatics data mining.* Journal of Biomedical Informatics, 85, 168–188.
4. Vergara, J. R., Estévez, P. A. (2014). *A review of feature selection methods based on mutual information.* Neural Computing and Applications, 24, 175–186.
5. Brown, G., Pocock, A., Zhao, M., Luján, M. (2012). *Conditional likelihood maximisation: A unifying framework for information theoretic feature selection.* Journal of Machine Learning Research, 13, 27–66.
6. Cover, T. M., Thomas, J. A. (2006). *Elements of Information Theory.* Wiley.
7. Chen, T., Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 785–794.
8. Breiman, L. (2001). *Random forests.* Machine Learning, 45, 5–32.