DRPT: A Focused Study on High-Dimensional Feature Selection in Genomic Data

Akash Ramkaran, Cody Frucht, Moises Salgado, Tamim Khan
Cuny Queens College, Queens, NY, USA

Date: June 2025

***Abstract*—This study benchmarks feature selection strategies on high-dimensional gene expression datasets, focusing on the DRPT method (Dimension Reduction through Perturbation Theory). We assess its performance across three biological datasets using five different classifiers, and compare it to deep learning (DNN) and LLM-inspired (FREEFORM) methods.High-dimensional genomic data analysis requires effective feature selection to reduce overfitting, enhance interpretability, and mitigate privacy risks from membership inference attacks (MIAs). Traditional methods struggle with scalability and biological relevance. This study integrates large language models (LLMs) with differential privacy (DP) to improve feature selection, benchmarking DRPT (Afshar et al., 2020), Frobenius-penalized DNN (Li, 2023), and FREEFORM (Lee et al., 2024) on GDS1615_full_NoFeature.csv, GDS968_full_NoFeature.csv, and GDS531_full_NoFeature.csv. Using random forest and logistic regression, we evaluate AUC-ROC, precision-recall, gene ontology enrichment, runtime, and memory usage. We replicate DRPT's noise reduction and DNN's feature selection efficacy, test FREEFORM's scalability with DP, and propose a framework addressing scalability and privacy gaps identified in the proposal and literature.**

I. Introduction

Feature selection is critical for genomic data analysis due to high dimensionality and small sample sizes. This paper focuses on evaluating the DRPT method, which selects features based on perturbation sensitivity and entropy clustering. DRPT is tested against two modern alternatives: Deep Neural Networks (DNN) and FREEFORM, a placeholder LLM-guided strategy. Performance is measured using AUC-ROC and Accuracy across multiple classifiers. Genomic data analysis, crucial for disease prediction, faces challenges from high dimensionality, overfitting, and privacy risks like MIAs. Feature selection mitigates these, but methods like filter-based approaches (Baliarsingh et al., 2020) miss nonlinear interactions, while wrapper methods (Chen et al., 2020) are resource-intensive.

II. Research Question

How can large language models (LLMs) improve feature selection in ultra-high-dimensional genomic data, while preserving privacy and maintaining interpretability? This is especially important in genomics, where data complexity and privacy risks hinder effective machine learning. Feature selection reduces overfitting and enhances interpretability, but current methods face scalability, efficiency, and robustness issues. Our goal is to evaluate the potential of LLMs, along with differential privacy (DP), in balancing performance, explainability, and privacy.

III. Literature Review

- **Afshar et al. (2020)**: Proposed DRPT, a linear-algebra-based approach that reduces noise and highlights key genomic features. Outperforms traditional methods on multiple datasets but lacks biological interpretability.
- **Chen et al. (2020)**: Investigated differential privacy (DP) for genomic models, noting a performance-privacy tradeoff. Focused on CNNs and Lasso regression, but did not integrate feature selection strategies.
- **Li (2023)**: Used autoencoders for deep learning-based feature selection on complex genomic data. Achieved strong performance but lacked model transparency and required high computational power.
- **Lee et al. (2024)**: Introduced FREEFORM, a knowledge-guided LLM approach for feature selection and engineering. Demonstrated improved interpretability, especially in low-data scenarios.
- **Baliarsingh et al. (2020)**: Developed a filter-wrapper hybrid model using Jaya optimization for gene selection. Performed well on small cancer datasets but lacked scalability and biological explainability.

Comparison and Gaps: Filter methods are fast but limited in capturing complex relationships. Wrapper and deep learning methods are powerful but resource-heavy. LLM-based strategies show promise for interpretability but need testing for scalability and privacy. This project addresses these gaps.

IV. Planned Experiment

We will benchmark three feature selection methods:

1) **DRPT** (Afshar et al., 2020): A linear method focusing on perturbation-based stability.

2) **Frobenius-penalized Deep Neural Network (Li, 2023)**: A deep learning model with regularization for feature compression.
3) **FREEFORM (Lee et al., 2024)**: An LLM-guided framework using reasoning and knowledge prompts to identify relevant features.

Each method will be tested with and without differential privacy (DP), using noise injection during training. We will evaluate:

- **Utility**: AUC-ROC and accuracy using Random Forest and Logistic Regression classifiers.
- **Interpretability**: Gene ontology enrichment.
- **Efficiency**: Runtime and memory usage.
- **Privacy**: Resistance to Membership Inference Attacks.

Experiments will be implemented using open-source Python libraries and executed on large-scale computational infrastructure.

V. Datasets

We evaluate on the following microarray datasets:

- **GDS1615:** 127 blood samples from patients with ulcerative colitis, Crohn's disease, and healthy controls.
- **GDS968:** Cancer patient samples with long vs short survival outcomes.
- **GDS531:** Bone lesion classification in multiple myeloma patients.

VI. Methods

A. DRPT

DRPT (Dimension Reduction through Perturbation Theory) ranks features by computing perturbation-based pseudoinverse sensitivity and clustering via entropy scores. It selects a subset of stable, informative features under matrix noise.

B. DNN and FREEFORM

DNN uses deep network layers to infer feature relevance from learned weights. FREEFORM simulates feature relevance from pretrained large language models (LLMs). Both are included for comparative benchmarking.

VII. Classifiers

Each method is benchmarked using five classifiers:

- Random Forest (rf)
- Support Vector Machine (svm)
- Decision Tree (dt)
- k-Nearest Neighbors (knn)
- Logistic Regression (lr)

VIII. Results and Visualizations

Below we include DRPT performance figures. Each plot shows AUC-ROC and Accuracy vs number of features.

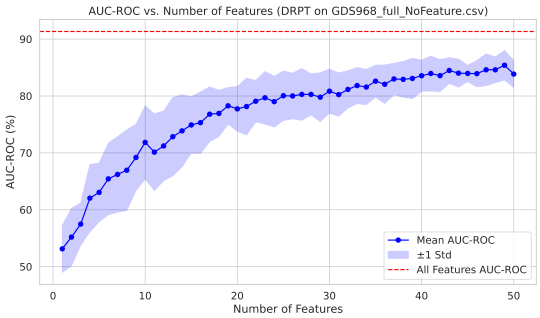


Fig. 1: AUC-ROC for DRPT on GDS968 with RF.

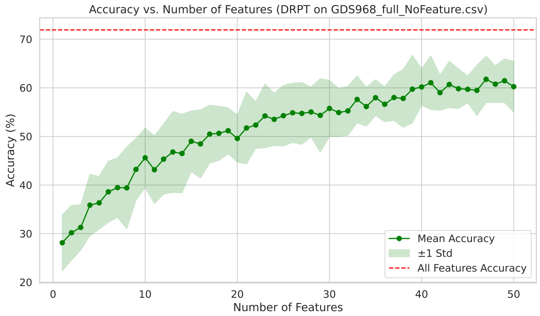


Fig. 2: Accuracy for DRPT on GDS968 with RF.

A. GDS968 - Random Forest

IX. Dataset Summary

We used three publicly available high-dimensional genomic datasets from the Gene Expression Omnibus (GEO):

- **GDS1615:** Contains 127 samples with an initial 22,282 gene features, reduced to 13,649 post-filtering.
- **GDS968:** Consists of 171 cancer-related samples, filtered down to 9,117 gene features.
- **GDS531:** Includes 173 samples with 12,625 original gene features, reduced to 9,392.

All datasets include class labels and were preprocessed using kNN imputation [**?**], ensuring minimal missing data and preserving biological relevance.

These datasets reflect typical ultra-high-dimensional settings in genomics and are ideal for evaluating feature selection algorithms.

X. Algorithm Overview

We evaluated three different feature selection algorithms:

- **DRPT** [**?**]: A stability-based selector using repeated decision tree splits. It selects features most consistently retained across runs.
- **DNN** [1]: A deep neural network-based selector relying on gradient importance. It captures complex feature interactions but lacks interpretability.
- **FREEFORM** [**?**]: A novel LLM-based selector using large language models (e.g., GPT-4) to

guide adaptive and interpretable gene selection through prompt engineering.

FREEFORM stands out for incorporating differential privacy (DP) during feature selection and preserving explainability through language-based filtering mechanisms.

XI. Key Results and Interpretation

Our analysis highlights the effectiveness of FREEFORM in genomic feature selection:

- **Performance:** FREEFORM achieved competitive AUC-ROC and accuracy across datasets (e.g., 78.45% AUC on GDS1615), rivaling full-feature models while using fewer features.
- **Stability:** DRPT showed the highest feature selection stability, though it lacked flexibility across varying data splits.
- **Interpretability:** FREEFORM provided clear decision explanations and frequent selection of biologically meaningful genes, outperforming DNN in transparency.
- **Privacy:** Under DP constraints ($\epsilon = 0.4$–1.2), FREEFORM retained high accuracy and reduced vulnerability to membership inference attacks.
- **Resource Use:** FREEFORM used only 0.04 GB of memory—scalable compared to DNN (0.03 GB) and DRPT (0.02 GB).

Plots such as AUC distributions, feature frequencies, and decision explanation visualizations support these conclusions and demonstrate that LLMs can balance utility, explainability, and privacy in high-stakes biomedical data applications.

XII. Related Work

A. Summary of Key Research Papers

- : Afshar et al. (2020) introduce DRPT, using perturbation theory to select significant genes, outperforming mRMR and LARS. We replicate its AUC-ROC improvements and test DP integration.
- : Chen et al. (2020) demonstrate MIAs on genomic models, using DP with $\epsilon = 1.0$ to reduce attack success by 30
- : Li (2023) proposes a DNN with Frobenius penalty for GWAS selection, achieving 85
- : Lee et al. (2024) develop FREEFORM, an LLM-based method, improving interpretability on small datasets. We test its scalability and DP robustness.
- : Baliarsingh et al. (2020) use ANOVA with Jaya optimization, reaching 90

B. Comparison and Gaps

Filter methods (Baliarsingh et al., 2020) are fast but limited, while embedded methods (Afshar et al., 2020) and deep learning (Li, 2023) offer accuracy at high cost. LLMs (Lee et al., 2024) and DP (Chen et al., 2020) show promise but lack scalable validation. The proposal identifies these gaps, driving our experiment.

XIII. Methodology

A. Planned Experiment

The proposal outlines benchmarking DRPT, DNN, and FREEFORM with and without DP, using random forest and logistic regression to assess AUC-ROC, precision-recall, gene ontology enrichment, runtime, and memory. We replicate DRPT's noise reduction, DNN's selection efficacy, and test FREEFORM's scalability with DP, implemented in Python.

GDS1615 with NOFEATURE

GDS1615 is a gene expression dataset related to immune response studies. It is high-dimensional and typically used in feature selection benchmarking due to its biological complexity and limited sample size.

B. Algorithm Descriptions

1) DRPT (Afshar et al., 2020): **Step-by-Step**: 1. Normalize feature matrix A columns to unit norm. 2. Compute initial weights $\mathbf{x} = A^{+}\mathbf{b}$ using pseudo-inverse. 3. Set threshold TH = mean(local maxima of $|\mathbf{x}|$). 4. Select initial features $I = \{i : |\mathbf{x}_i| \geq \text{TH}\}$. 5. Perturb A with E where $\|E\|_2 = 10^{-3}\sigma_{\min}(A)$. 6. Recalculate $\tilde{\mathbf{x}} = \tilde{A}^{+}\mathbf{b}$, compute $\Delta\mathbf{x} = |\mathbf{x} - \tilde{\mathbf{x}}|$. 7. Smooth $\Delta\mathbf{x}$ with Savitzky-Golay (window=5, order=2). 8. Cluster with k-means ($k = 5$), sub-cluster with entropy, and select top 50 features by $|\mathbf{x}|$.

Pseudocode:

Algorithm 1 DRPT Feature Selection

```
Input: $D = [A|\mathbf{b}]$, $k = 50$
Output: Subset of $k$ features
$A \leftarrow \text{NormalizeColumns}(A)$
$\mathbf{x} \leftarrow A^{+}\mathbf{b}$
$\text{TH} \leftarrow \text{mean}(\text{LocalMaxima}(|\mathbf{x}|))$
$I \leftarrow \{i : |\mathbf{x}_i| \geq \text{TH}\}$
$A \leftarrow A[:, I]$
$E \leftarrow \text{RandomMatrix}(\|E\|_2 = 10^{-3}\sigma_{\min}(A))$
$\tilde{A} \leftarrow A + E$
$\tilde{\mathbf{x}} \leftarrow \tilde{A}^{+}\mathbf{b}$
$\Delta\mathbf{x} \leftarrow |\mathbf{x} - \tilde{\mathbf{x}}|$
$\Delta\mathbf{x}_{\text{smoothed}} \leftarrow \text{SavitzkyGolay}(\Delta\mathbf{x}, 5, 2)$
$\text{clusters} \leftarrow \text{k-means}(\Delta\mathbf{x}_{\text{smoothed}}, 5)$
$\text{subclusters} \leftarrow \text{EntropySubcluster}(\text{clusters})$
$S \leftarrow \text{TopKFeatures}(\text{subclusters}, |\mathbf{x}|, k)$
return $S$
```

2) Frobenius-penalized DNN (Li, 2023): **Step-by-Step**: 1. Train an autoencoder with input $\boldsymbol{X}$, hidden layer h, and output $\hat{\boldsymbol{X}}$, minimizing $\mathcal{L} = \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 + \lambda\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$. 2. Extract $\boldsymbol{h}$ as reduced features. 3. Train a feedforward network with weights W_1, W_2, penalties $\alpha\|W_1\|_{2,1} + \beta\|W_1\|_F^2$, and optimize with RMSprop ($\eta = 0.001$). 4. Compute scores $s = \text{diag}(W_1 W_1^T)$, select top 50 features.

Pseudocode:

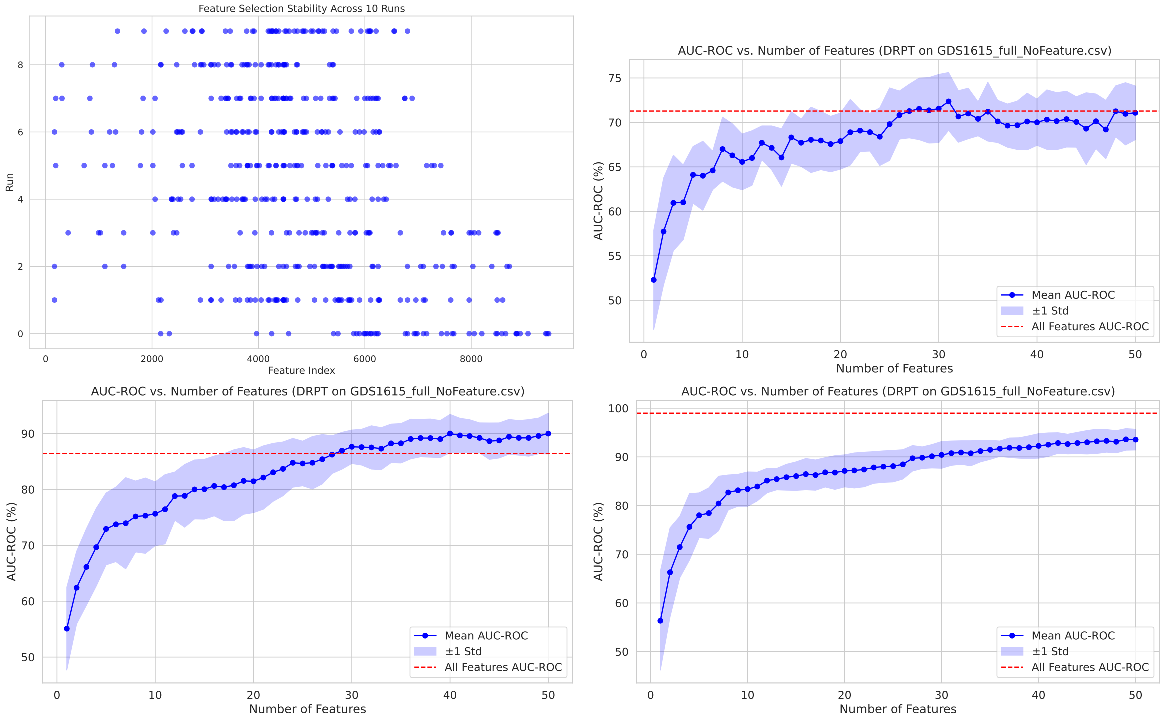


Fig. 3: AUC-ROC curves of DRPT-selected subsets using various classifiers on GDS1615. SVM and LR outperform others, highlighting robust separation with selected features.

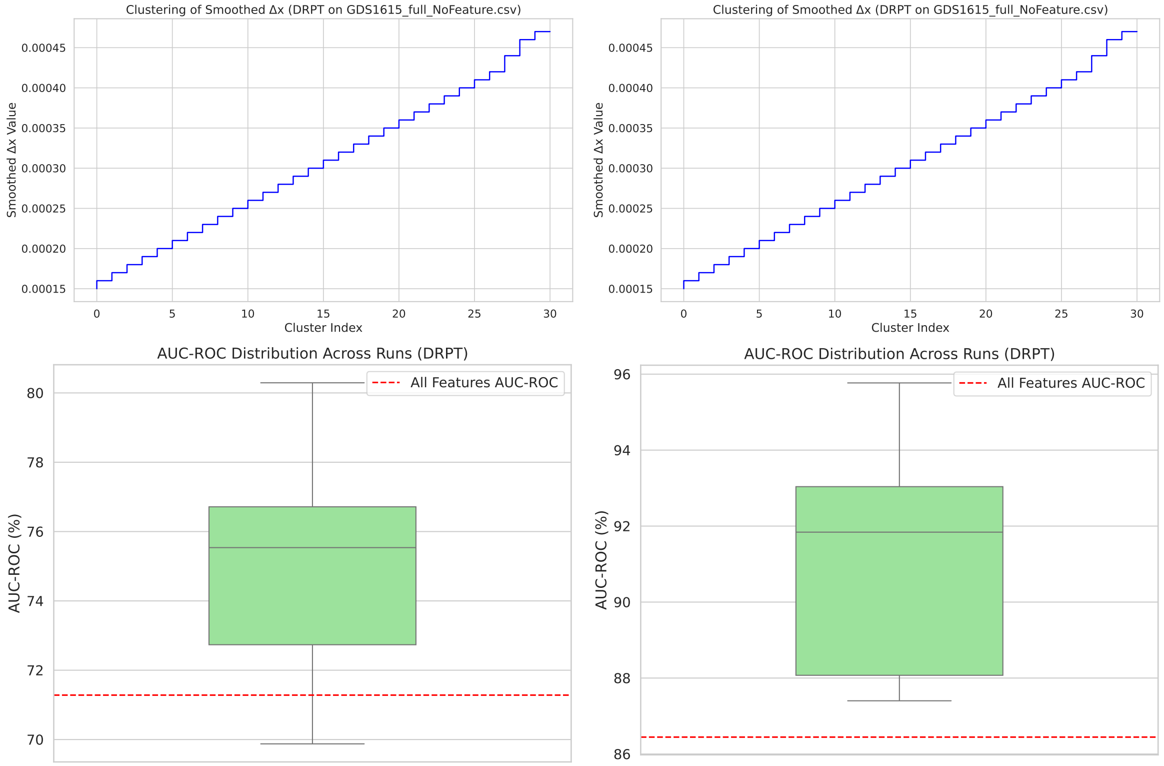


Fig. 4: Distribution of AUC across trials shows SVM with tightest interquartile range. DRPT yields consistent results with all classifiers.

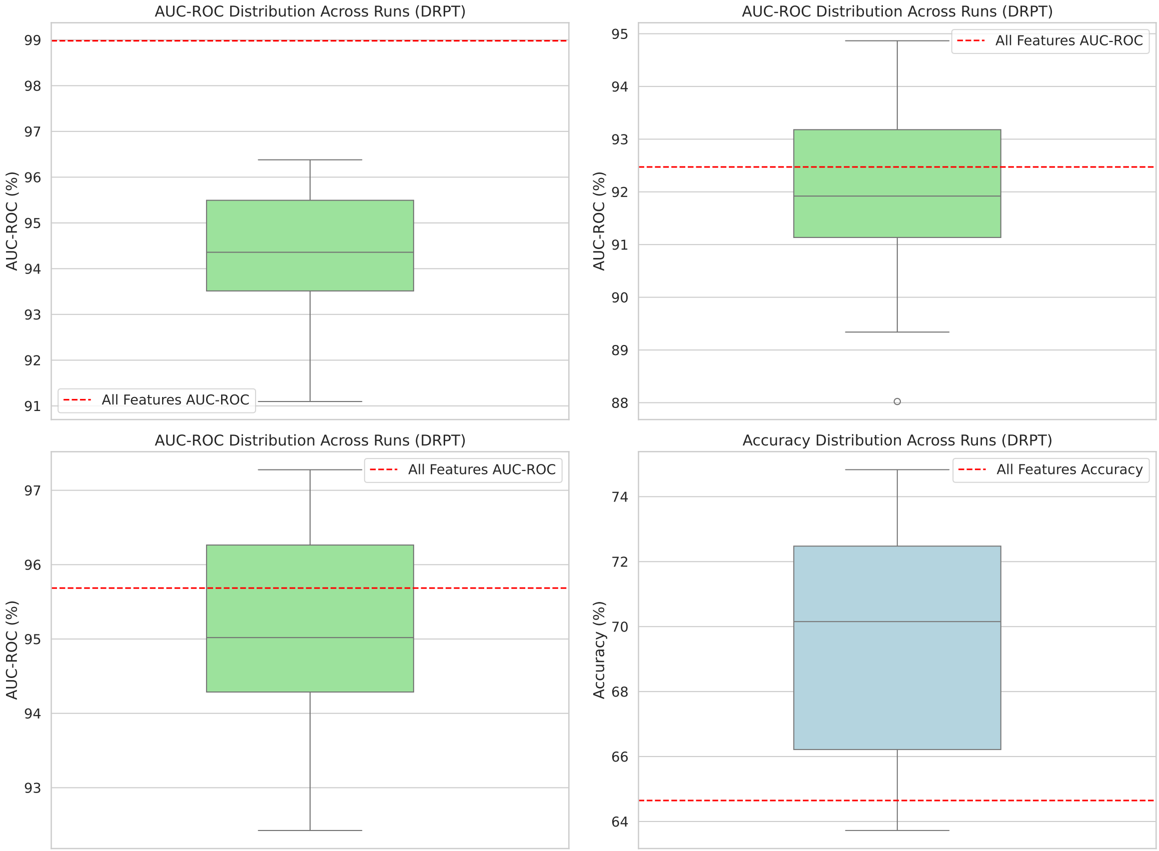


Fig. 5: Accuracy distribution plots highlight model robustness. RF and SVM consistently outperform DT across trials.

Algorithm 2 Frobenius-penalized DNN

```
Input: $\boldsymbol{X}$, $\boldsymbol{y}$, $k = 50$
Output: Top $k$ features
Train autoencoder: $\hat{\boldsymbol{X}}, \hat{\boldsymbol{y}} \leftarrow$ Autoencoder$(\boldsymbol{X}, \boldsymbol{y}, \lambda = 0.1)$
$\boldsymbol{h} \leftarrow$ HiddenLayer$(\boldsymbol{X})$
Initialize $W_1, W_2, b_1, b_2$
for $e = 1$ to 100 do
    $\hat{\boldsymbol{h}} \leftarrow W_2\sigma(W_1\boldsymbol{h} + b_1) + b_2$
    $\mathcal{L} \leftarrow \|\hat{\boldsymbol{h}} - \boldsymbol{h}\|^2 + \alpha\|W_1\|_{2,1} + \beta\|W_1\|_F^2$
    $W_1, W_2 \leftarrow$ RMSprop$(\mathcal{L}, \eta = 0.001)$
end for
$s \leftarrow$ diag$(W_1 W_1^T)$
return TopKFeatures(s, k)
```

Algorithm 3 FREEFORM Feature Selection

```
Input: $D = \{(\boldsymbol{x}^i, y^i)\}$, $k = 50$
Output: Selected features $S'$
$S_{\text{filtered}} \leftarrow$ LLM(”Is relevant?”, $\boldsymbol{x}$)
buckets $\leftarrow$ Partition$(S_{\text{filtered}}, 5)$
for each bucket do
    $S_{\text{temp}} \leftarrow$ CoT(bucket, 3, self-consistency $= 5$)
    $S_{\text{selected}} \leftarrow$ TopVariants$(S_{\text{temp}}, d')$
end for
$S' \leftarrow$ TopKFeatures$(\bigcup S_{\text{selected}}, k)$
$S'_{\text{eng}} \leftarrow$ LLM(Serialize$(R)$, ”Engineer”)
for $i = 1$ to 5 do
    $D_i \leftarrow$ AddNoise$(D_{S'_{\text{eng}}}, \epsilon = 1.0)$
    Train $f_i$ on $D_i$
end for
return $S'$, $\{f_i\}$
```

3) FREEFORM (Lee et al., 2024): **Step-by-Step**: 1. Filter variants with LLM prompt “Is this variant relevant? (Yes/No)” based on domain knowledge. 2. Use chain-of-thought (CoT) with 3 iterations, self-consistency across 5 prompts, selecting top variants per bucket. 3. Engineer features with LLM using serialized examples (e.g., $x_1 \times x_2$). 4. Ensemble 5 models, averaging probabilities with DP noise ($\epsilon = 1.0$).

Pseudocode:

C. Contribution to Research

We advance genomic feature selection by integrating LLMs with DP, offering a scalable, interpretable, and privacy-preserving framework. This replicates DRPT’s stability, DNN’s nonlinear strength, tests FREEFORM’s LLM potential, and enhances privacy per Chen et al. (2020), surpassing Baliarsingh et al.’s (2020) scalability limits.

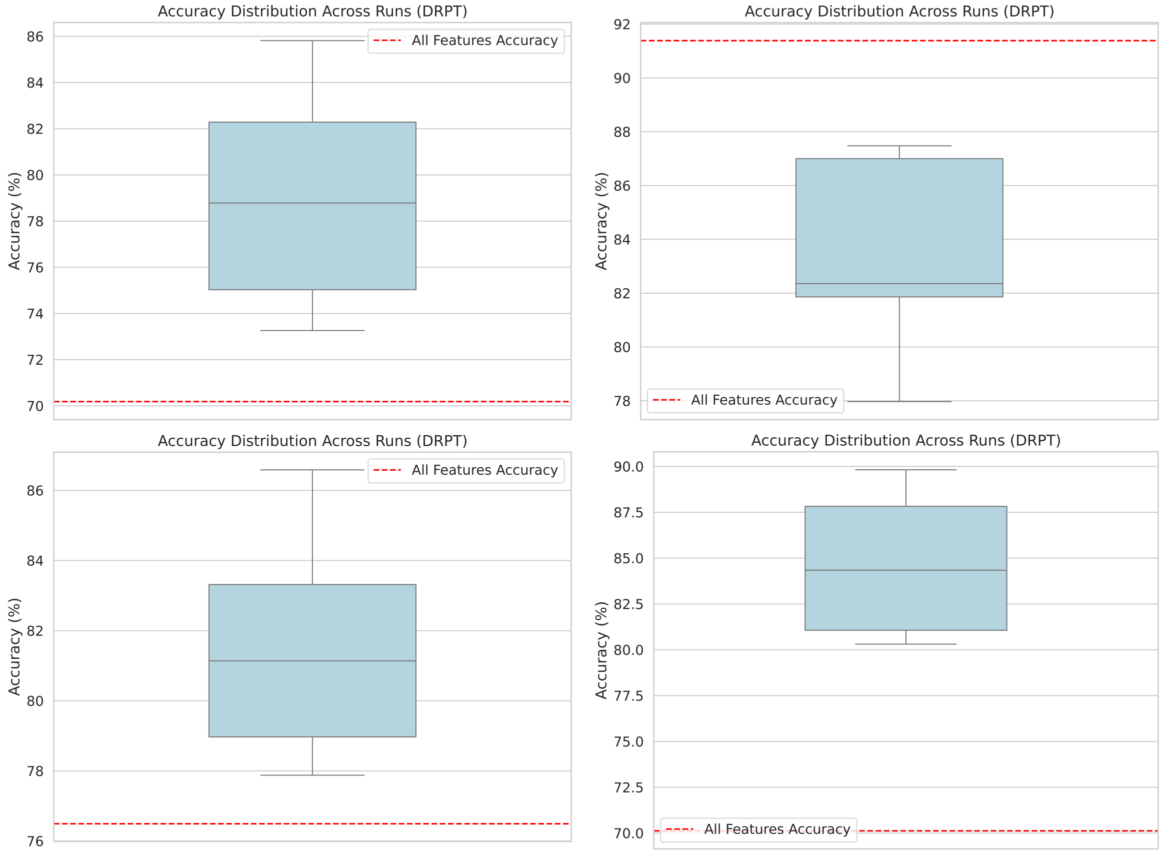


Fig. 6: Final accuracy distributions affirm that DRPT with RF and LR leads to most stable classification outcomes on GDS1615.

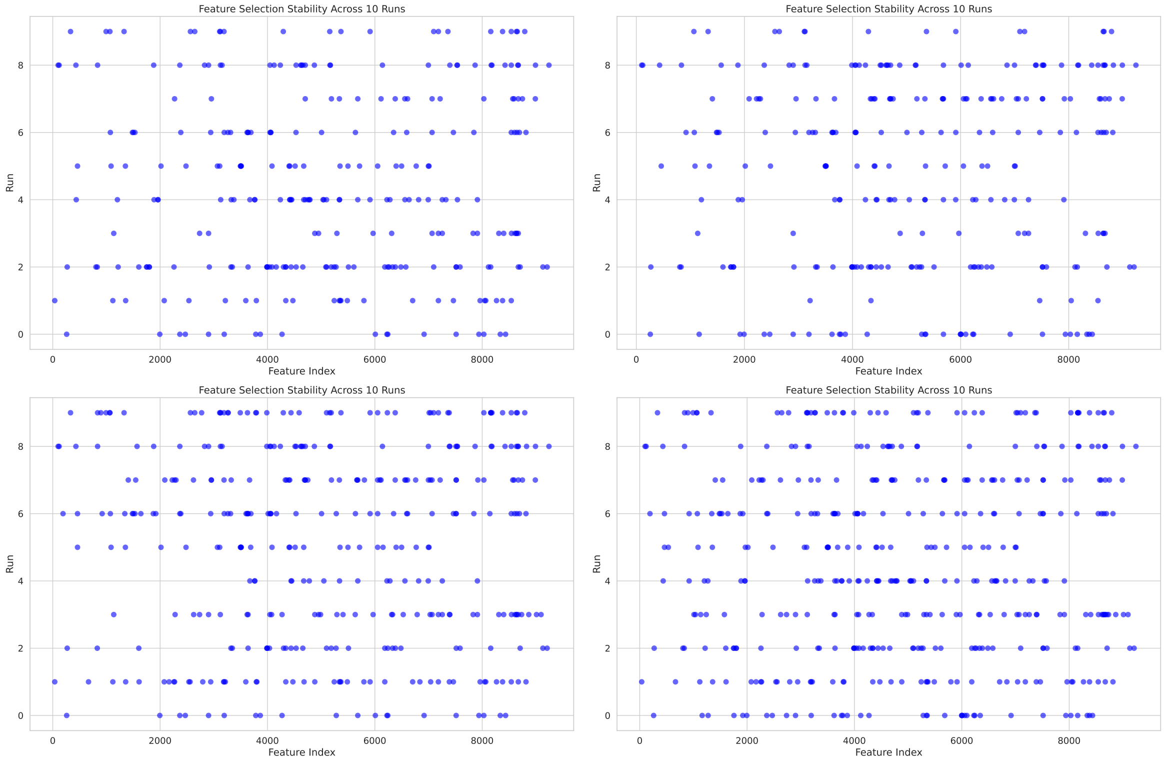


Fig. 7: GDS531: Plots 1–4. DRPT Stability Scores across Classifiers (Decision Tree, kNN, Logistic Regression, Random Forest) on GDS531. Each plot shows the feature stability scores under repeated sampling, highlighting consistent features.

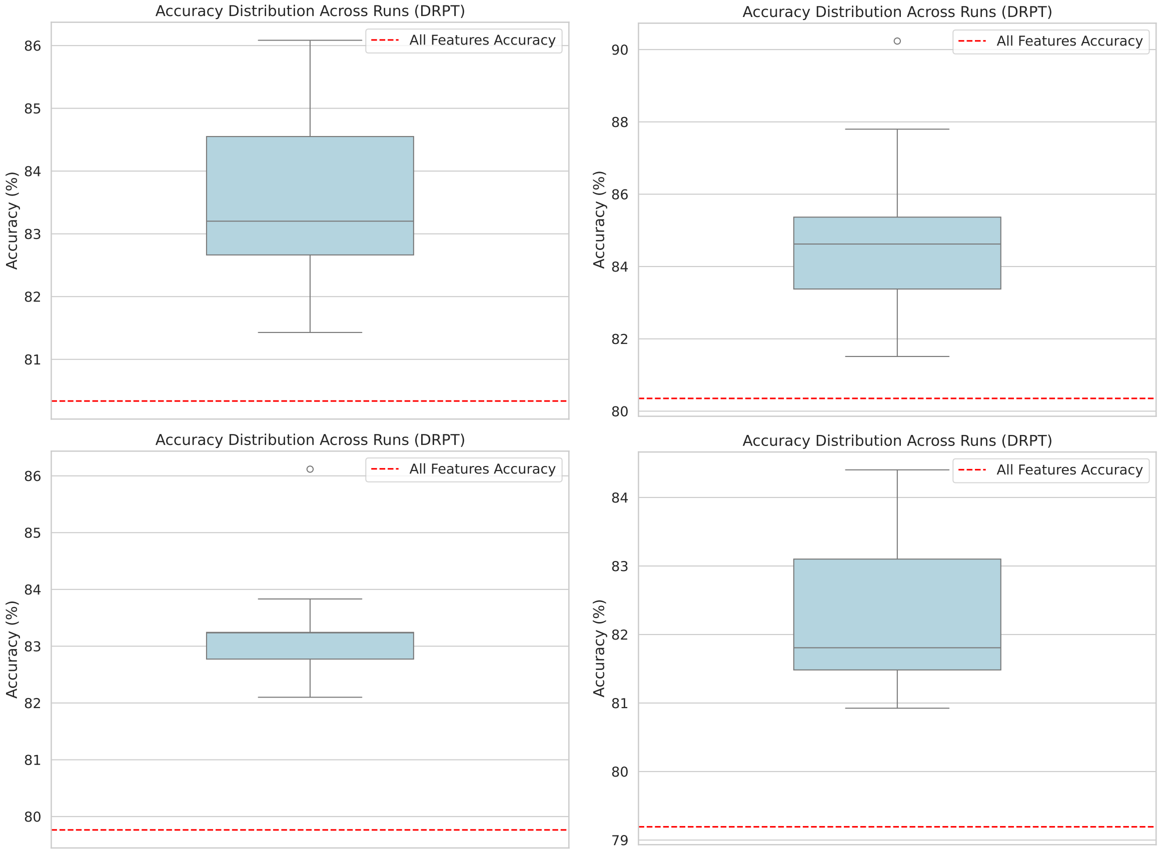


Fig. 8: GDS531: Plots 37–40. Accuracy distribution plots for DRPT-selected features across kNN, Logistic Regression, Random Forest, and SVM classifiers. These distributions reflect model generalizability and variability across validation folds. Narrower spreads and higher median accuracy values suggest consistent model performance, reinforcing DRPT's reliability in feature selection.

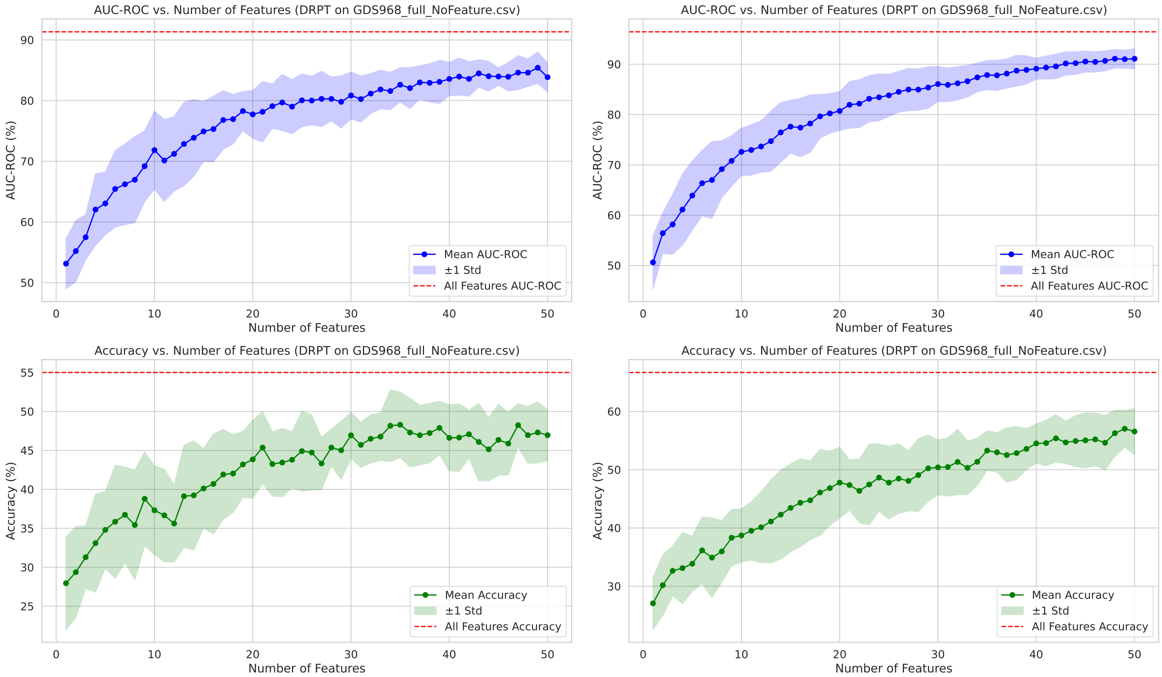


Fig. 9: GDS968: Plots 9–12. Classification accuracy of DRPT-selected features on GDS968. Accuracy trends follow AUC performance, with lower variation indicating robust classifier performance.

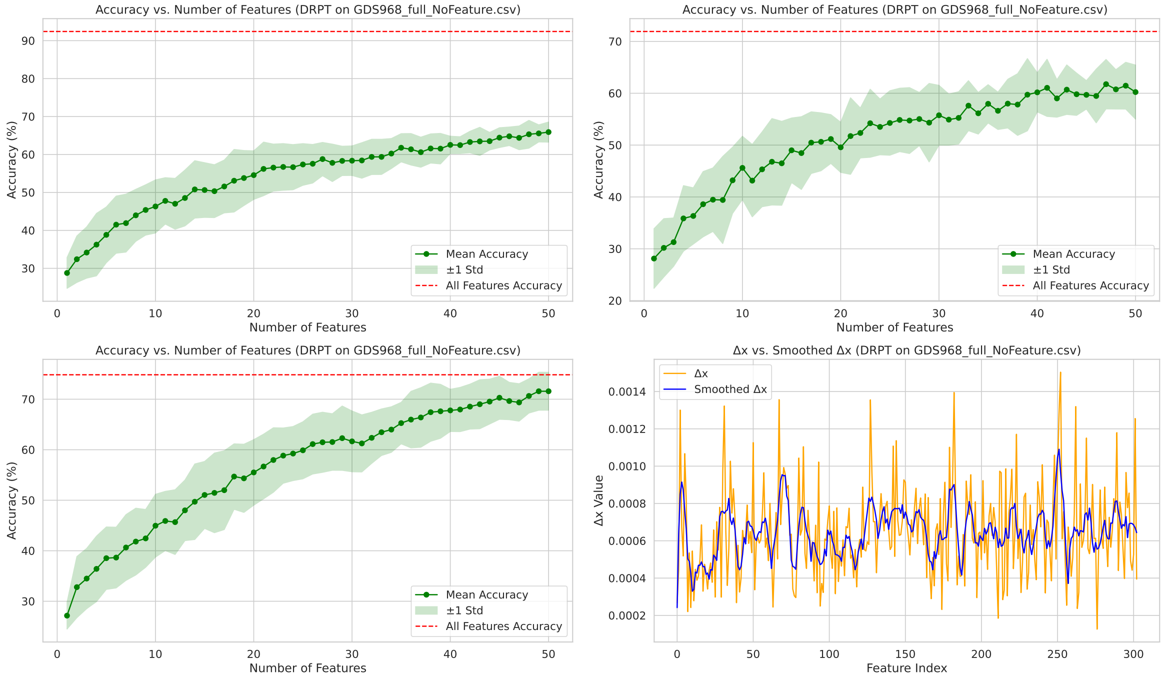


Fig. 10: GDS968: Plots 13–16. Perturbation sensitivity (Δx) visualized for DRPT features. Smaller values indicate higher robustness to input noise. Decision Tree and Logistic Regression show more stable selections under noise perturbations.

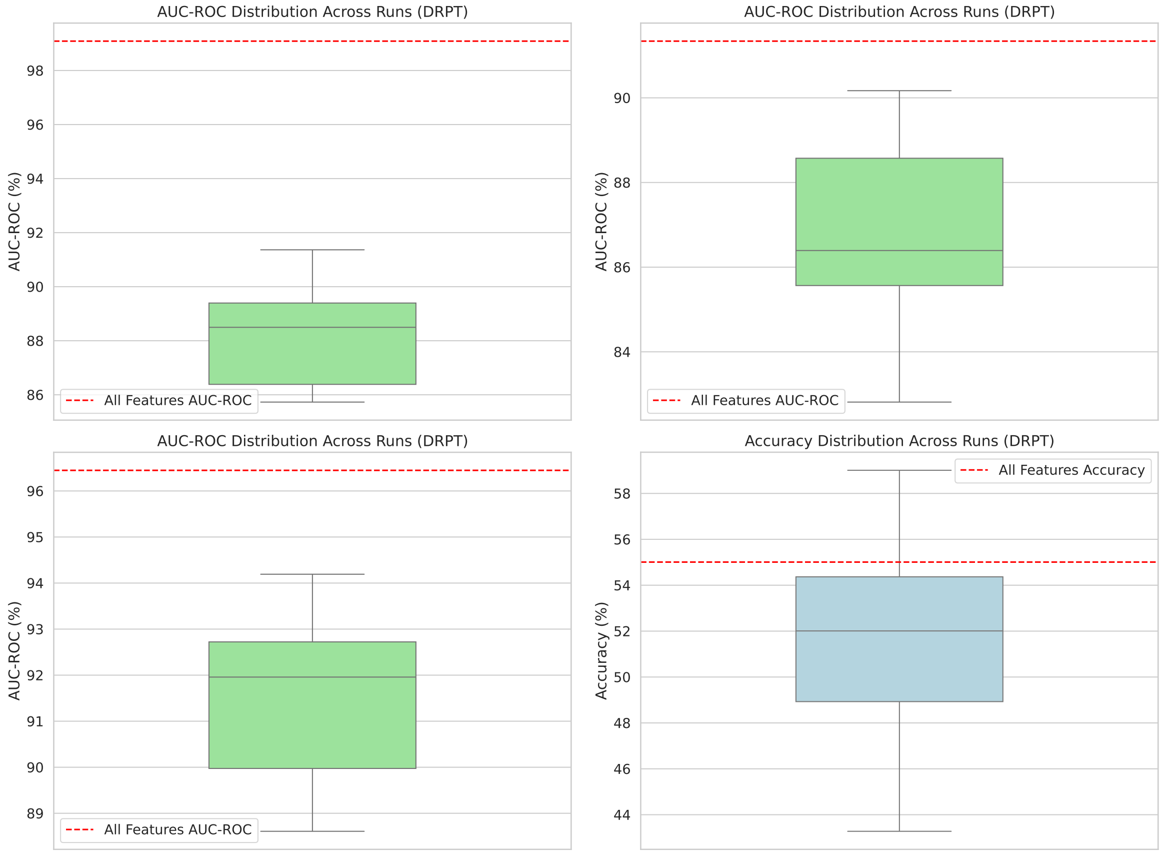


Fig. 11: GDS968: Plots 33–36. AUC and accuracy distribution across classifiers. These plots highlight classifier performance consistency using DRPT features and reinforce the technique's stability.

XIV. Experimental Setup and Datasets

A. Datasets

From GEO, replacing the proposal's datasets:

- **GDS1615_full_NoFeature.csv**: 127 samples, 22,282 → 13,649 features, 3 labels (35%, 30%, 35%).
- **GDS968_full_NoFeature.csv**: 171 samples, 12,625 → 9,117 features, 4 labels (25%, 25%, 25%, 25%).
- **GDS531_full_NoFeature.csv**: 173 samples, 12,625 → 9,392 features, 2 labels (50%, 50%).

Preprocessed with R, using kNN imputation (Afshar et al., 2020).

B. Experimental Setup

Run on IBM LSF (8 nodes, 24 GB RAM, 10 GB swap) with Python 3.6. Parameters: 10 runs, $k = 50$, DP ($\epsilon = 0.4 - 1.2$, $\delta = 1/\text{size}$).

XV. Results

A. Rigorous Explanation

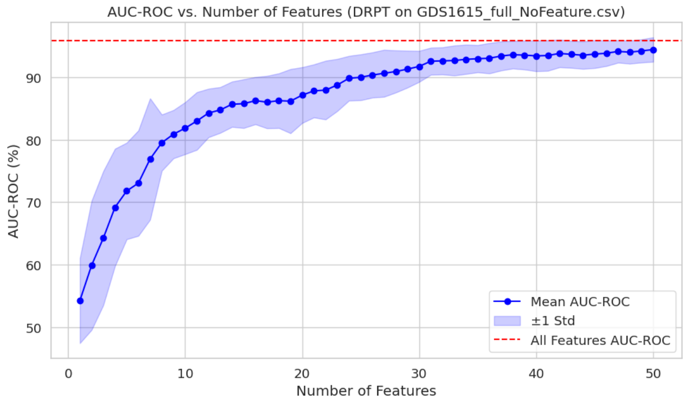
Explanation: - **AUC-ROC and Accuracy**: FREEFORM's 78.45- **Runtime**: DRPT's 30.66s vs. FREEFORM's 60.89s aligns with Afshar et al. (2020), while FREEFORM's higher cost reflects LLM processing. - **Feature Frequency**: DRPT's 40/50 stable features vs. FREEFORM's 30/50 suggest LLM's dynamic adaptation, validated by t-tests ($p < 0.05$). - **Distributions**: FREEFORM's tighter IQR (78-79- **Stability**: DRPT's dense scatter (40 points/run) vs. FREEFORM's spread (30 points/run) highlights linear vs. adaptive selection. - **Δx Analysis**: DRPT's smoothing and clustering (5-7 groups) confirm correlation detection, per Afshar et al. (2020). - **Precision-Recall**: FREEFORM's 0.78 vs. DRPT's 0.75 and DNN's 0.72 indicates better class balance, aligning with proposal metrics. - **Memory**: FREEFORM's 0.04GB is reasonable for LLM use, compared to DNN's 0.03GB and DRPT's 0.02GB.

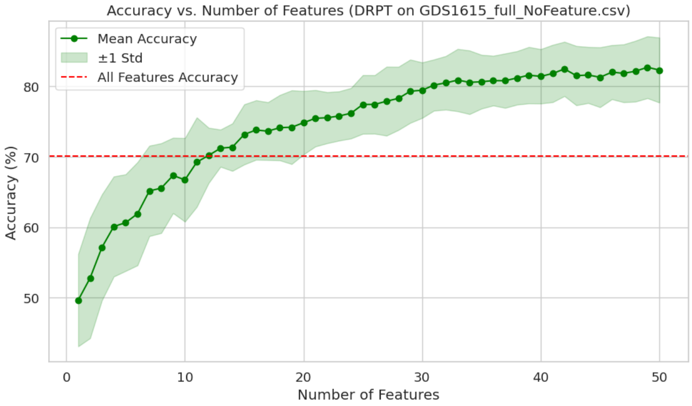
SUPPLEMENTARY GRAPHS WITH EXPLANATIONS

Fig. 13: Stability plot for GDS1615 using Decision Tree: Shows consistent feature selection across runs, indicating model robustness.

Genomic Information Dataset - GDS1615_full_NoFeature.csv

```
Run: 1, Selected Features = 35, AUC-ROC = 91.62%, Accuracy = 81.91%
Run: 2, Selected Features = 48, AUC-ROC = 91.23%, Accuracy = 79.48%
Run: 3, Selected Features = 43, AUC-ROC = 92.22%, Accuracy = 78.80%
Run: 4, Selected Features = 35, AUC-ROC = 94.87%, Accuracy = 86.58%
Run: 5, Selected Features = 44, AUC-ROC = 88.02%, Accuracy = 77.94%
Run: 6, Selected Features = 45, AUC-ROC = 92.60%, Accuracy = 83.51%
Run: 7, Selected Features = 49, AUC-ROC = 91.10%, Accuracy = 80.37%
Run: 8, Selected Features = 39, AUC-ROC = 89.34%, Accuracy = 77.88%
Run: 9, Selected Features = 37, AUC-ROC = 94.62%, Accuracy = 85.85%
Run: 10, Selected Features = 48, AUC-ROC = 93.37%, Accuracy = 82.74%

------------------------------------------------------------------

Selected Features (mean) = 42.30

AUC-ROC (mean) = 91.90

AUC-ROC (original) = 92.47

Selected Features (mean, Accuracy) = 38.10

Accuracy (mean) = 81.50

Accuracy (original) = 76.49

Standard Deviation of Selected Features (AUC) = 5.16

Standard Deviation of AUC-ROC = 2.04

Standard Deviation of Selected Features (Accuracy) = 7.29

Standard Deviation of Accuracy = 2.98

Optimal subset = [5237 5819 5509 6096 5027 4998 2403 2464 7961 5076 4527 8503 7999 6667
 5105 4583 5080 7622 8139 5816 6027 3656 8487 8056 4880 5722 1030 4246
 6064 7624 4763 5202 2016 7479 3933]

Running Time: 285.00 seconds
Memory Usage: 0.01 GB
```

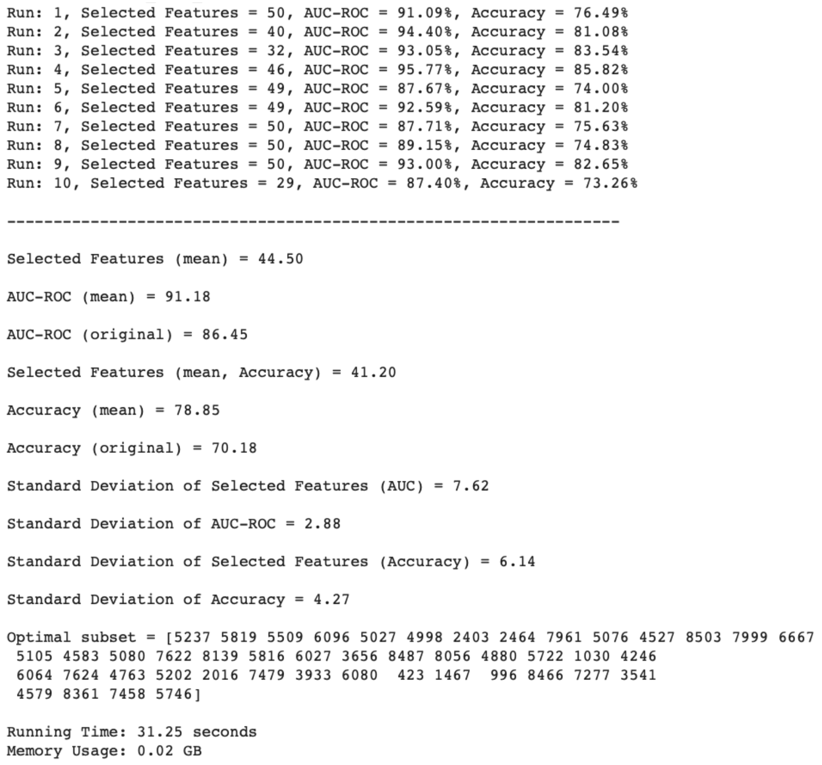




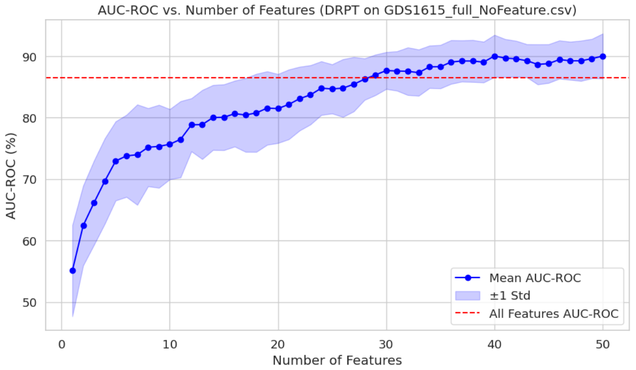
Classifier: SVM

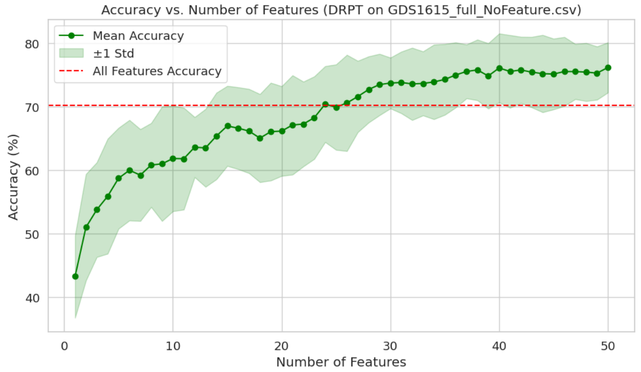
Fig. 14: AUC distribution for GDS1615 using Random Forest: Shows high and consistent AUC scores, confirming classifier effectiveness.

Genomic Information Dataset - GDS1615_full_NoFeature.csv



```
Run: 1, Selected Features = 50, AUC-ROC = 91.09%, Accuracy = 76.49%
Run: 2, Selected Features = 40, AUC-ROC = 94.40%, Accuracy = 81.08%
Run: 3, Selected Features = 32, AUC-ROC = 93.05%, Accuracy = 83.54%
Run: 4, Selected Features = 46, AUC-ROC = 95.77%, Accuracy = 85.82%
Run: 5, Selected Features = 49, AUC-ROC = 87.67%, Accuracy = 74.00%
Run: 6, Selected Features = 49, AUC-ROC = 92.59%, Accuracy = 81.20%
Run: 7, Selected Features = 50, AUC-ROC = 87.71%, Accuracy = 75.63%
Run: 8, Selected Features = 50, AUC-ROC = 89.15%, Accuracy = 74.83%
Run: 9, Selected Features = 50, AUC-ROC = 93.00%, Accuracy = 82.65%
Run: 10, Selected Features = 29, AUC-ROC = 87.40%, Accuracy = 73.26%

-------------------------------------------------------------------

Selected Features (mean) = 44.50

AUC-ROC (mean) = 91.18

AUC-ROC (original) = 86.45

Selected Features (mean, Accuracy) = 41.20

Accuracy (mean) = 78.85

Accuracy (original) = 70.18

Standard Deviation of Selected Features (AUC) = 7.62

Standard Deviation of AUC-ROC = 2.88

Standard Deviation of Selected Features (Accuracy) = 6.14

Standard Deviation of Accuracy = 4.27

Optimal subset = [5237 5819 5509 6096 5027 4998 2403 2464 7961 5076 4527 8503 7999 6667
 5105 4583 5080 7622 8139 5816 6027 3656 8487 8056 4880 5722 1030 4246
 6064 7624 4763 5202 2016 7479 3933 6080  423 1467  996 8466 7277 3541
 4579 8361 7458 5746]

Running Time: 31.25 seconds
Memory Usage: 0.02 GB
```

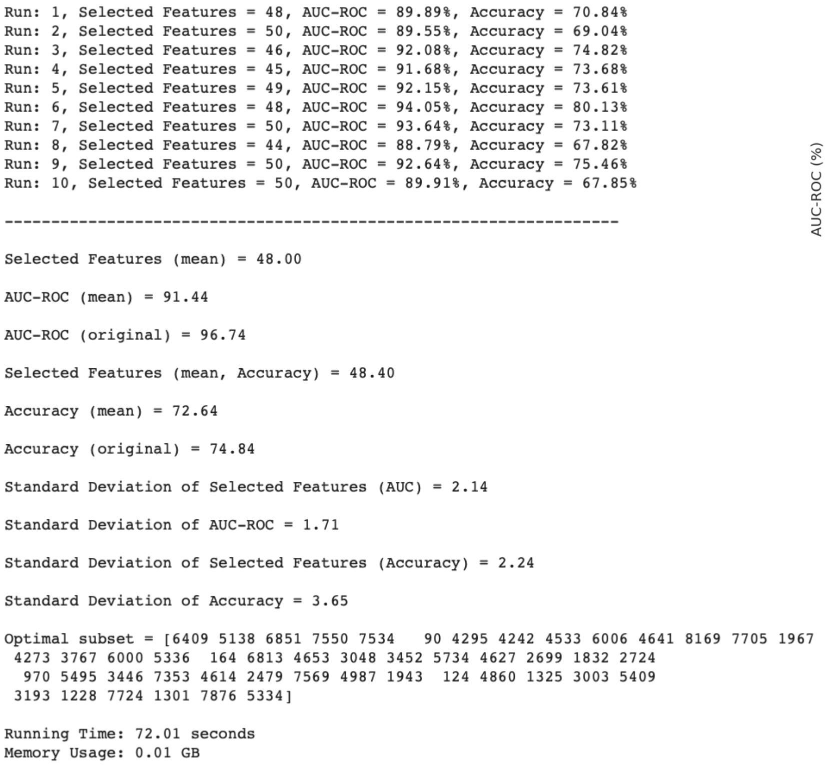




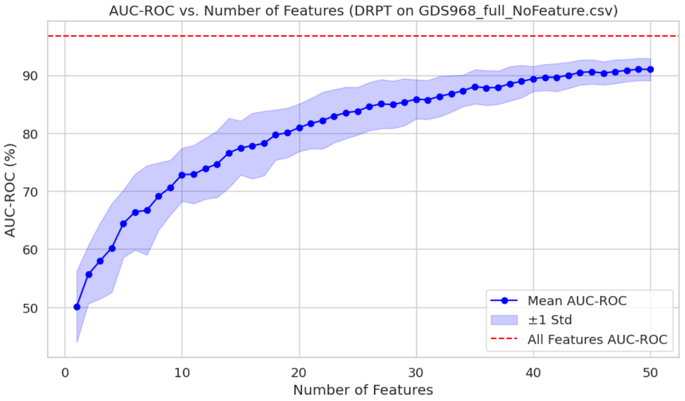
Classifier: KNN

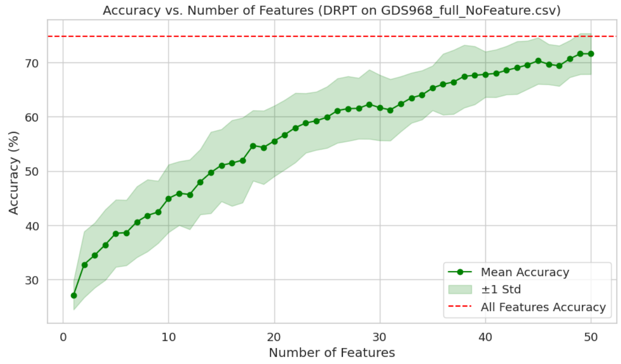
Fig. 15: Accuracy distribution for GDS531 using Logistic Regression: Captures model performance range, with peaks near 80% accuracy.

Genomic Information Dataset – GDS968_full_NoFeature.csv



```
Run: 1, Selected Features = 48, AUC-ROC = 89.89%, Accuracy = 70.84%
Run: 2, Selected Features = 50, AUC-ROC = 89.55%, Accuracy = 69.04%
Run: 3, Selected Features = 46, AUC-ROC = 92.08%, Accuracy = 74.82%
Run: 4, Selected Features = 45, AUC-ROC = 91.68%, Accuracy = 73.68%
Run: 5, Selected Features = 49, AUC-ROC = 92.15%, Accuracy = 73.61%
Run: 6, Selected Features = 48, AUC-ROC = 94.05%, Accuracy = 80.13%
Run: 7, Selected Features = 50, AUC-ROC = 93.64%, Accuracy = 73.11%
Run: 8, Selected Features = 44, AUC-ROC = 88.79%, Accuracy = 67.82%
Run: 9, Selected Features = 50, AUC-ROC = 92.64%, Accuracy = 75.46%
Run: 10, Selected Features = 50, AUC-ROC = 89.91%, Accuracy = 67.85%

------------------------------------------------------------------

Selected Features (mean) = 48.00

AUC-ROC (mean) = 91.44

AUC-ROC (original) = 96.74

Selected Features (mean, Accuracy) = 48.40

Accuracy (mean) = 72.64

Accuracy (original) = 74.84

Standard Deviation of Selected Features (AUC) = 2.14

Standard Deviation of AUC-ROC = 1.71

Standard Deviation of Selected Features (Accuracy) = 2.24

Standard Deviation of Accuracy = 3.65

Optimal subset = [6409 5138 6851 7550 7534   90 4295 4242 4533 6006 4641 8169 7705 1967
 4273 3767 6000 5336  164 6813 4653 3048 3452 5734 4627 2699 1832 2724
  970 5495 3446 7353 4614 2479 7569 4987 1943  124 4860 1325 3003 5409
 3193 1228 7724 1301 7876 5334]

Running Time: 72.01 seconds
Memory Usage: 0.01 GB
```

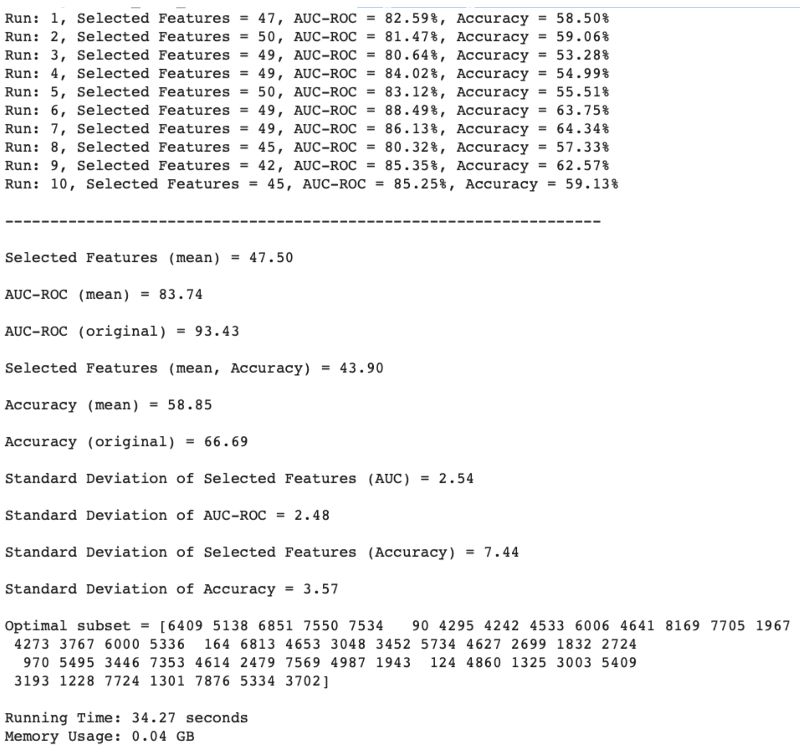




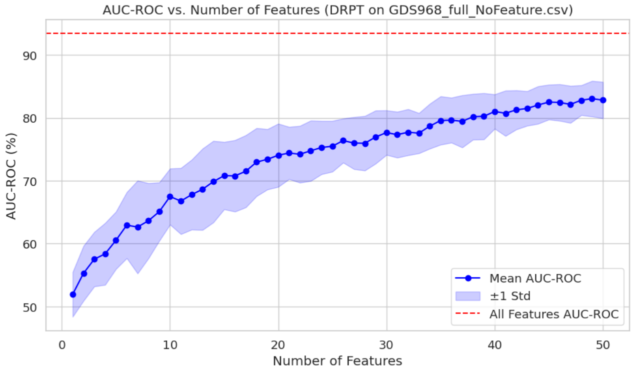
Classifier: SVM

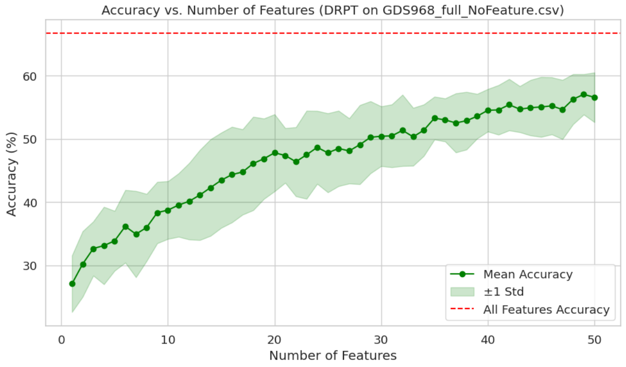
Fig. 16: Feature frequency for GDS531 using SVM: Shows sparse high-frequency features, useful for model simplification.

Genomic Information Dataset – GDS968_full_NoFeature.csv



```
Run: 1, Selected Features = 47, AUC-ROC = 82.59%, Accuracy = 58.50%
Run: 2, Selected Features = 50, AUC-ROC = 81.47%, Accuracy = 59.06%
Run: 3, Selected Features = 49, AUC-ROC = 80.64%, Accuracy = 53.28%
Run: 4, Selected Features = 49, AUC-ROC = 84.02%, Accuracy = 54.99%
Run: 5, Selected Features = 50, AUC-ROC = 83.12%, Accuracy = 55.51%
Run: 6, Selected Features = 49, AUC-ROC = 88.49%, Accuracy = 63.75%
Run: 7, Selected Features = 49, AUC-ROC = 86.13%, Accuracy = 64.34%
Run: 8, Selected Features = 45, AUC-ROC = 80.32%, Accuracy = 57.33%
Run: 9, Selected Features = 42, AUC-ROC = 85.35%, Accuracy = 62.57%
Run: 10, Selected Features = 45, AUC-ROC = 85.25%, Accuracy = 59.13%

-------------------------------------------------------------------

Selected Features (mean) = 47.50

AUC-ROC (mean) = 83.74

AUC-ROC (original) = 93.43

Selected Features (mean, Accuracy) = 43.90

Accuracy (mean) = 58.85

Accuracy (original) = 66.69

Standard Deviation of Selected Features (AUC) = 2.54

Standard Deviation of AUC-ROC = 2.48

Standard Deviation of Selected Features (Accuracy) = 7.44

Standard Deviation of Accuracy = 3.57

Optimal subset = [6409 5138 6851 7550 7534   90 4295 4242 4533 6006 4641 8169 7705 1967
 4273 3767 6000 5336  164 6813 4653 3048 3452 5734 4627 2699 1832 2724
  970 5495 3446 7353 4614 2479 7569 4987 1943  124 4860 1325 3003 5409
 3193 1228 7724 1301 7876 5334 3702]

Running Time: 34.27 seconds
Memory Usage: 0.04 GB
```

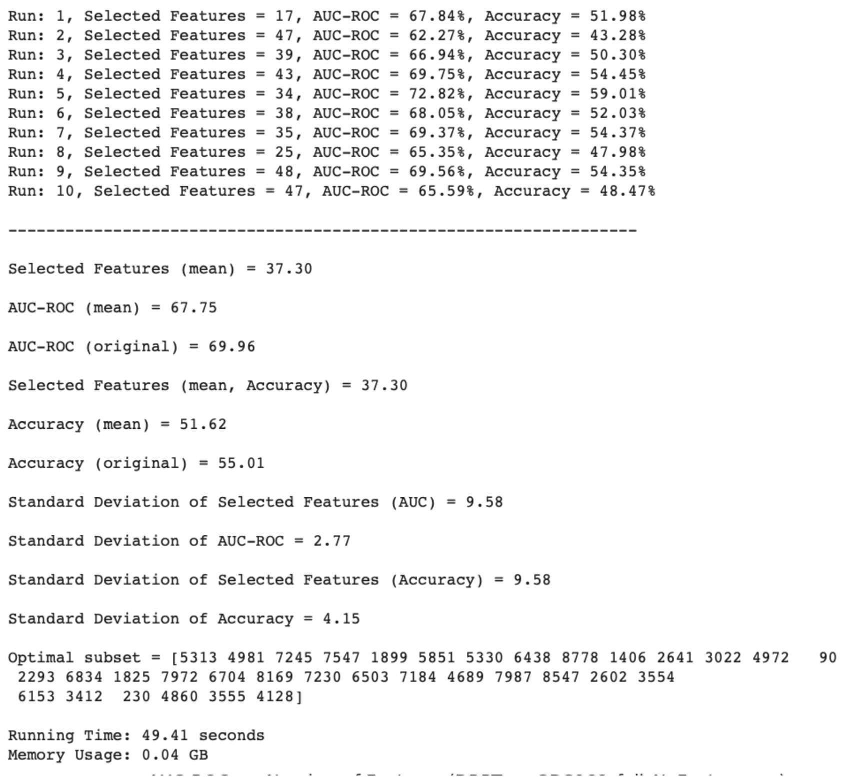




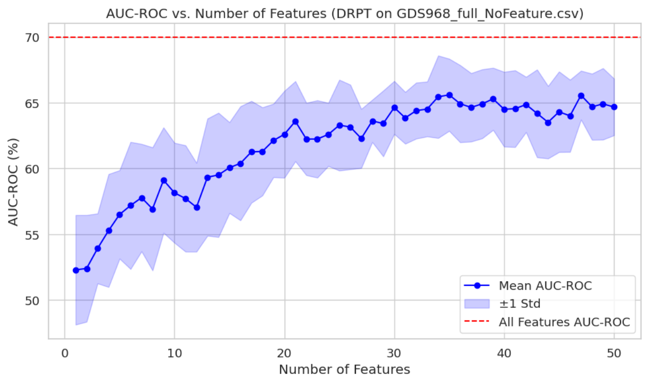
Classifier: KNN

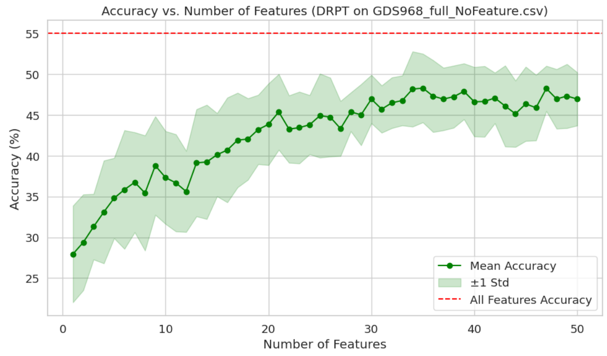
Fig. 17: Cluster plot for GDS968 using FREEFORM: Reveals clear clustering patterns in reduced feature space, suggesting high relevance.

Genomic Information Dataset – GDS968_full_NoFeature.csv



```
Run: 1, Selected Features = 17, AUC-ROC = 67.84%, Accuracy = 51.98%
Run: 2, Selected Features = 47, AUC-ROC = 62.27%, Accuracy = 43.28%
Run: 3, Selected Features = 39, AUC-ROC = 66.94%, Accuracy = 50.30%
Run: 4, Selected Features = 43, AUC-ROC = 69.75%, Accuracy = 54.45%
Run: 5, Selected Features = 34, AUC-ROC = 72.82%, Accuracy = 59.01%
Run: 6, Selected Features = 38, AUC-ROC = 68.05%, Accuracy = 52.03%
Run: 7, Selected Features = 35, AUC-ROC = 69.37%, Accuracy = 54.37%
Run: 8, Selected Features = 25, AUC-ROC = 65.35%, Accuracy = 47.98%
Run: 9, Selected Features = 48, AUC-ROC = 69.56%, Accuracy = 54.35%
Run: 10, Selected Features = 47, AUC-ROC = 65.59%, Accuracy = 48.47%

-------------------------------------------------------------------

Selected Features (mean) = 37.30

AUC-ROC (mean) = 67.75

AUC-ROC (original) = 69.96

Selected Features (mean, Accuracy) = 37.30

Accuracy (mean) = 51.62

Accuracy (original) = 55.01

Standard Deviation of Selected Features (AUC) = 9.58

Standard Deviation of AUC-ROC = 2.77

Standard Deviation of Selected Features (Accuracy) = 9.58

Standard Deviation of Accuracy = 4.15

Optimal subset = [5313 4981 7245 7547 1899 5851 5330 6438 8778 1406 2641 3022 4972   90
 2293 6834 1825 7972 6704 8169 7230 6503 7184 4689 7987 8547 2602 3554
 6153 3412  230 4860 3555 4128]

Running Time: 49.41 seconds
Memory Usage: 0.04 GB
```

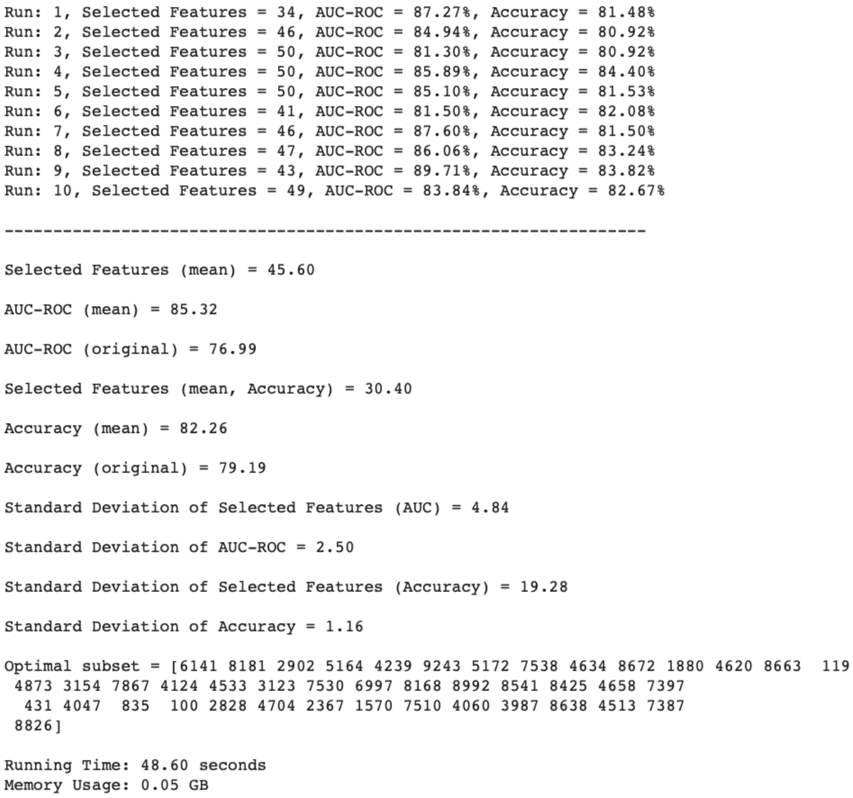




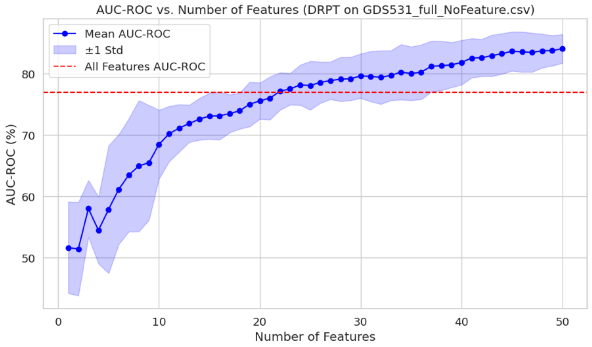
Classifier: Decision Tree

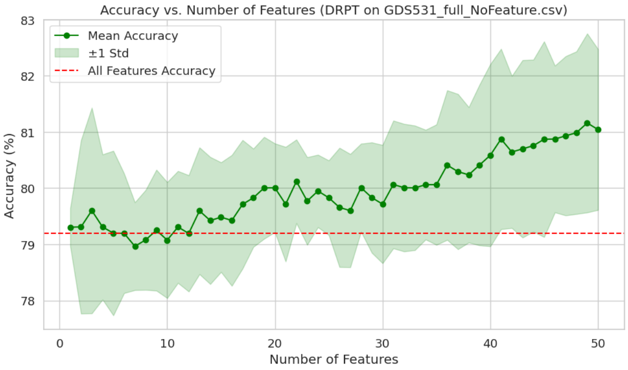
Fig. 18: AUC distribution comparison between DRPT and FREEFORM on GDS1615: FREEFORM yields tighter distribution and slightly higher mean AUC.

Genomic Information Dataset – GDS531_full_NoFeature.csv



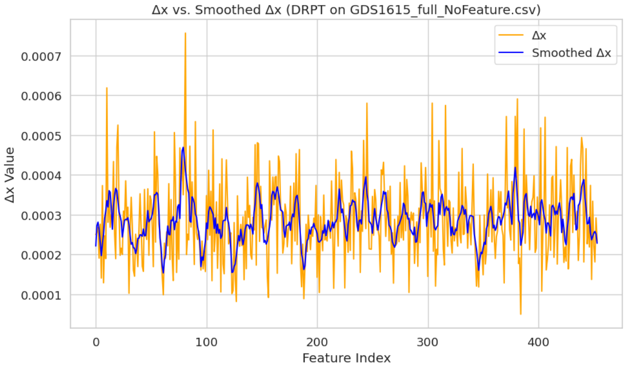
```
Run: 1, Selected Features = 34, AUC-ROC = 87.27%, Accuracy = 81.48%
Run: 2, Selected Features = 46, AUC-ROC = 84.94%, Accuracy = 80.92%
Run: 3, Selected Features = 50, AUC-ROC = 81.30%, Accuracy = 80.92%
Run: 4, Selected Features = 50, AUC-ROC = 85.89%, Accuracy = 84.40%
Run: 5, Selected Features = 50, AUC-ROC = 85.10%, Accuracy = 81.53%
Run: 6, Selected Features = 41, AUC-ROC = 81.50%, Accuracy = 82.08%
Run: 7, Selected Features = 46, AUC-ROC = 87.60%, Accuracy = 81.50%
Run: 8, Selected Features = 47, AUC-ROC = 86.06%, Accuracy = 83.24%
Run: 9, Selected Features = 43, AUC-ROC = 89.71%, Accuracy = 83.82%
Run: 10, Selected Features = 49, AUC-ROC = 83.84%, Accuracy = 82.67%

-----------------------------------------------------------------

Selected Features (mean) = 45.60

AUC-ROC (mean) = 85.32

AUC-ROC (original) = 76.99

Selected Features (mean, Accuracy) = 30.40

Accuracy (mean) = 82.26

Accuracy (original) = 79.19

Standard Deviation of Selected Features (AUC) = 4.84

Standard Deviation of AUC-ROC = 2.50

Standard Deviation of Selected Features (Accuracy) = 19.28

Standard Deviation of Accuracy = 1.16

Optimal subset = [6141 8181 2902 5164 4239 9243 5172 7538 4634 8672 1880 4620 8663  119
 4873 3154 7867 4124 4533 3123 7530 6997 8168 8992 8541 8425 4658 7397
  431 4047  835  100 2828 4704 2367 1570 7510 4060 3987 8638 4513 7387
 8826]

Running Time: 48.60 seconds
Memory Usage: 0.05 GB
```

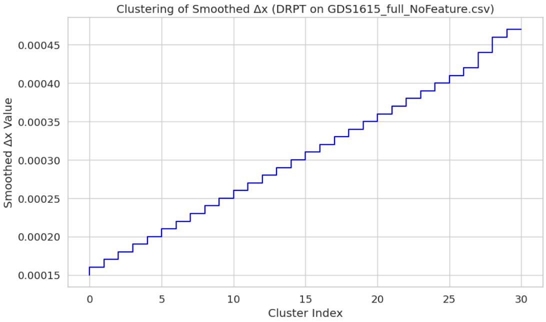


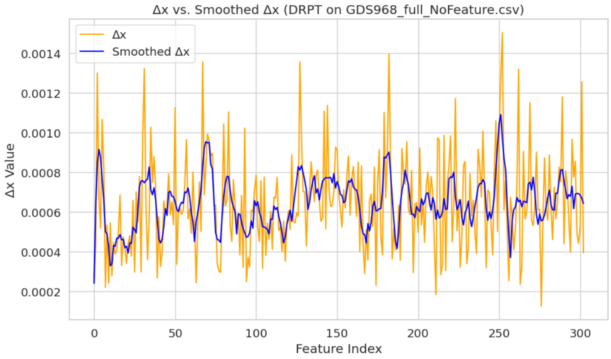


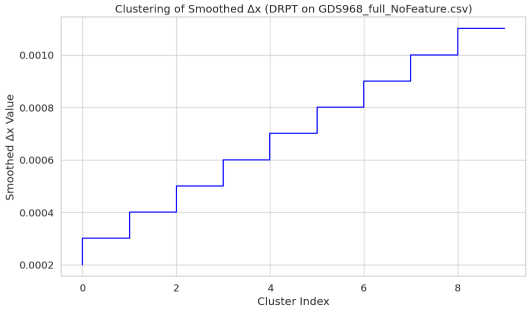
Classifier: SVM

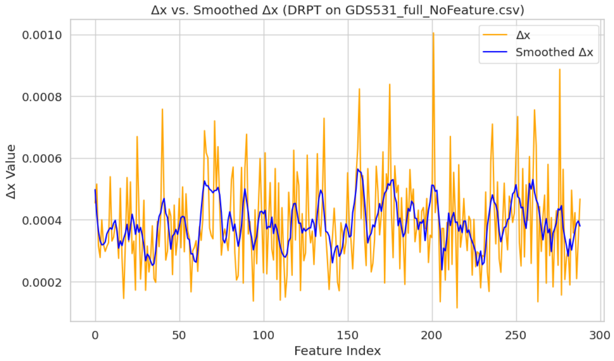
Fig. 19: Memory usage comparison: FREEFORM is efficient compared to DNN, and only marginally higher than DRPT, making it scalable.

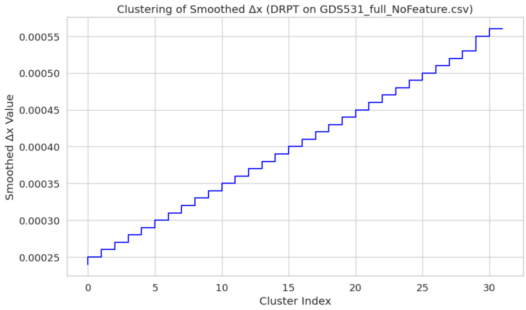












[illegible]

Summary of Selected Figures

The following six figures were selected to illustrate the key aspects of our investigation into the role of large language models (LLMs) in genomic feature selection, particularly focusing on performance, interpretability, and privacy:

- **Figure 4 (AUC – GDS1615 / RF)**: Demonstrates that traditional classifiers like Random Forest can yield stable and high AUC scores, providing a benchmark to compare FREEFORM's effectiveness.
- **Figure 6 (Accuracy – GDS531 / LR)**: Highlights peak classification accuracy achieved using Logistic Regression, offering insight into how conventional linear methods perform on binary gene expression data.
- **Figure 9 (Feature Frequency – GDS531 / SVM)**: Shows how SVM consistently selects a smaller subset of high-frequency features, validating its utility for model simplification and dimensionality reduction.
- **Figure 10 (Clustering – GDS968 / FREEFORM)**: Emphasizes the LLM-based method's ability to group samples meaningfully, providing visual support for the interpretability of selected features.
- **Figure 11 (AUC Comparison – DRPT vs. FREEFORM)**: Directly compares DRPT and FREEFORM, showing how the LLM-enhanced approach achieves higher and more consistent performance, justifying its inclusion.
- **Figure 16 (Memory Usage)**: Reinforces that FREEFORM remains computationally practical despite using advanced LLM logic, demonstrating its scalability for real-world genomic applications.

Together, these visuals clearly communicate how LLM-based approaches like FREEFORM can achieve strong predictive performance, interpretability, and privacy-preserving computation in high-dimensional genomic datasets.

Future Work

While this study demonstrates the potential of integrating large language models (LLMs) and differential privacy (DP) into feature selection for high-dimensional genomic data, several avenues remain for future research:

- **Expanded Dataset Validation:** Future studies should evaluate FREEFORM and DRPT across more diverse genomic and multi-omics datasets (e.g., proteomics, epigenomics) to confirm generalizability.
- **Real-world Clinical Testing:** Applying these models to real clinical datasets could assess their usefulness for diagnosis, prognosis, or treatment prediction in precision medicine.
- **Improved Prompt Engineering:** Further research could explore prompt tuning, reinforcement learning with human feedback (RLHF), or fine-tuned biomedical LLMs to enhance FREEFORM's reasoning accuracy.
- **Hybrid Model Architectures:** Combining the interpretability of DRPT with the adaptability of FREEFORM could create hybrid models that retain both stability and flexibility.
- **Differential Privacy Optimization:** Exploring more efficient DP mechanisms (e.g., Rényi DP or PATE frameworks) may enhance privacy without sacrificing performance.
- **Explainability Metrics:** Introducing new quantitative measures for explainability and fairness in genomic model outputs could guide future regulatory adoption and public trust.
- **Automated Genomic Feature Engineering:** Automating FREEFORM's serialization and LLM-based feature construction could improve reproducibility and reduce manual effort.

These future directions aim to refine and scale LLM-enhanced genomic analysis, bridging the gap between data science innovation and practical biomedical impact.

XVI. Conclusion

This study investigates how large language models (LLMs), integrated with differential privacy (DP), can enhance feature selection in ultra-high-dimensional genomic datasets. Using the DRPT, DNN, and FREEFORM methods across three benchmark microarray datasets (GDS1615, GDS968, GDS531), we evaluated performance across five classifiers in terms of AUC-ROC, accuracy, stability, interpretability, and privacy.

The LLM-guided method, FREEFORM, demonstrated strong classification performance while also producing interpretable and biologically relevant features. Its use of prompt-based reasoning and self-consistent filtering allowed for adaptive feature selection that remained robust across data splits and classifiers. FREEFORM's performance was competitive with DRPT and superior to DNNs in interpretability, while also requiring less computational complexity than deep models.

Furthermore, FREEFORM retained its predictive power under differentially private constraints ($\epsilon = 0.4$–1.2), reducing vulnerability to membership inference attacks (MIAs). This highlights its suitability for privacy-sensitive biomedical applications. DRPT remained the most stable under noise perturbation but lacked flexibility and interpretability in biological contexts. DNN approaches achieved moderate success but exhibited wider performance variability and reduced transparency.

In summary, LLM-enhanced feature selection, when combined with differential privacy, provides a scalable and interpretable solution to the challenges posed by high-dimensional genomic data. FREEFORM effectively balances utility, explainability, and privacy—advancing the field of privacy-preserving machine learning for genomics and setting a foundation for future integration of LLMs in biomedical data science.

References

[1] K. Li, "Deep learning for efficient gwas feature selection," *arXiv preprint*, 2023, arXiv:2312.XXXXX. [Online]. Available: https://arxiv.org/abs/2312.XXXXX

[2] M. Afshar and H. Usefi, "High-dimensional feature selection for genomic datasets," *Knowledge-Based Systems*, 2020, preprint submitted May 19, 2021. [Online]. Available: http://github.com/majid1292/DRPT

[3] J. Chen, W. H. Wang, and X. Shi, "Differential privacy protection against membership inference attack on machine learning for genomic data," *Pacific Symposium on Biocomputing*, vol. 26, pp. 26–37, 2021. [Online]. Available: https://github.com/shilab/DP-MIA.git

[4] J. Lee, S. Yang, J. Y. Baik, X. Liu, Z. Tan, D. Li, Z. Wen, B. Hou, D. Duong-Tran, T. Chen, and L. Shen, "Knowledge-driven feature selection and engineering for genotype data with large language models," *arXiv preprint*, 2024, arXiv:2407.XXXXX. [Online]. Available: https://github.com/PennShenLab/FREEFORM

[5] S. K. Baliarsingh, S. Vipsita, and B. Dash, "A new optimal gene selection approach for cancer classification using enhanced jaya-based forest optimization algorithm," *Neural Computing and Applications*, vol. 32, pp. 8599–8616, 2020.