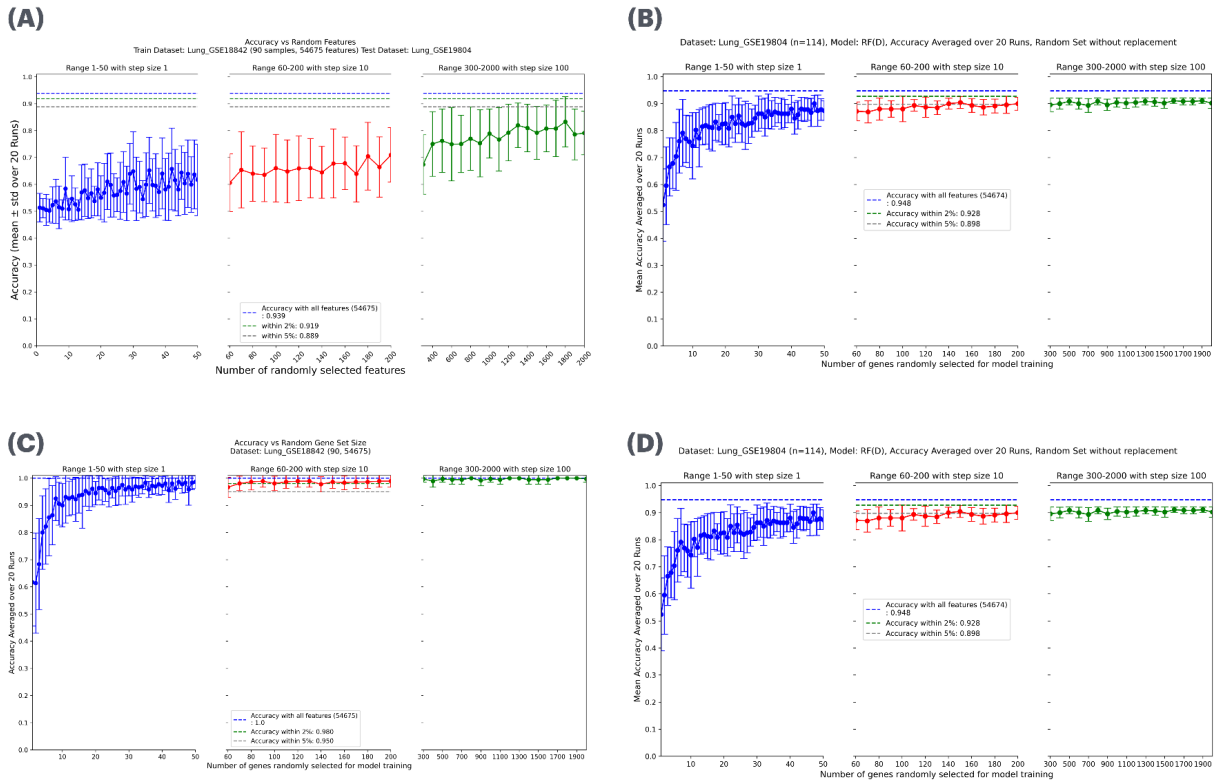
On the (In)Significance of Feature Selection in High-Dimensional Datasets

Supplementary material

S1: Accuracy plots for datasets for which AUC results were presented in the paper

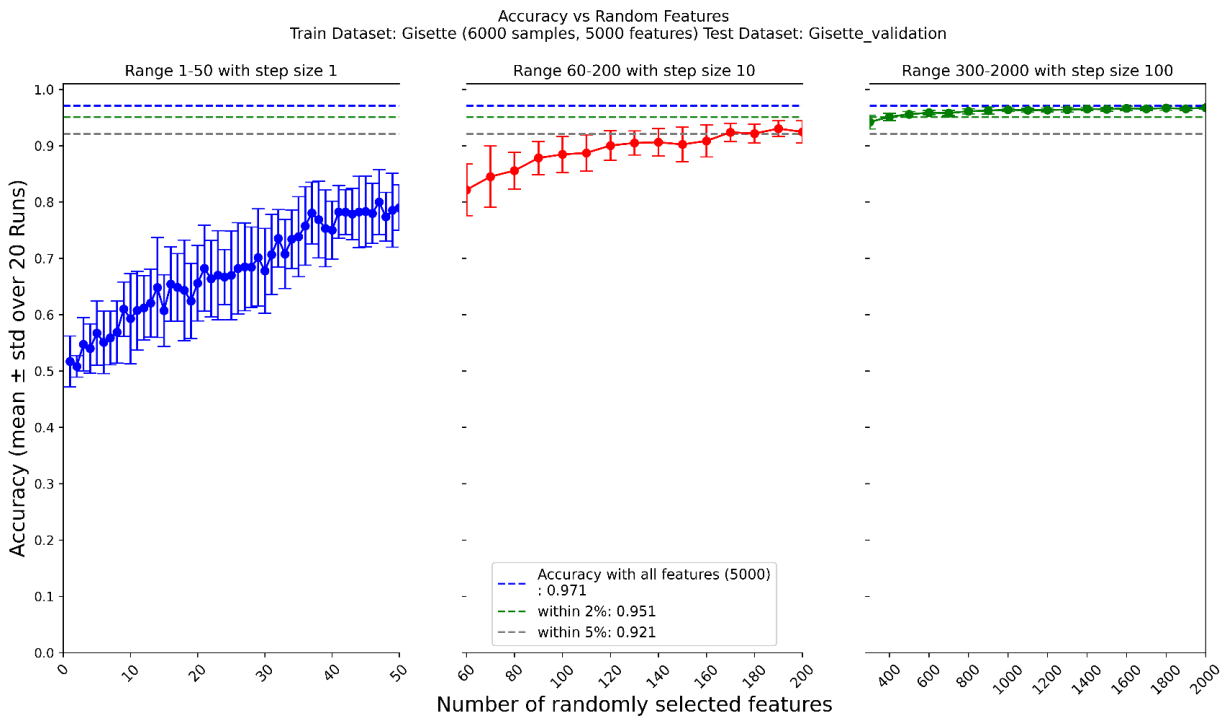


S1 Figure 1: Random Forest performance with **lung microarray** dataset pairs (mean and standard deviation are reported over 20 runs)

(A) RF models trained on GSE18842 and tested on GSE19804 show that randomly selected subsets never achieve accuracy comparable to using all features.

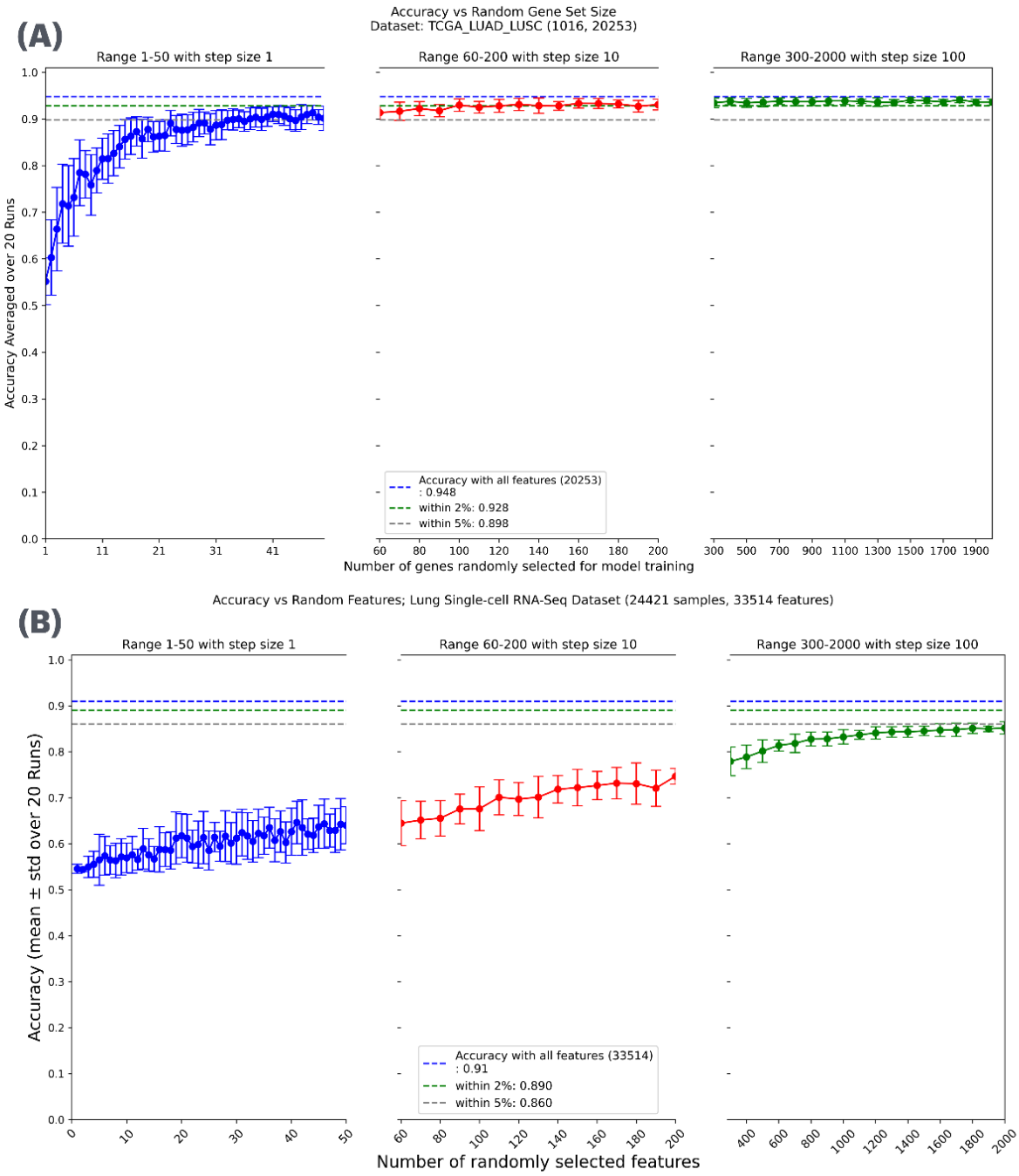
(B) RF models trained on GSE19804 and tested on GSE18842 show that 200 randomly selected features (~0.4% of all features) perform comparable to all features.

(C) Model performance with an 80:20 train-test split using randomly selected feature subsets. For GSE18842, just 50 randomly selected features are sufficient to match the accuracy comparable to all features. Similarly, for GSE19804, 200 features suffice to match accuracy with all features.



S1 Figure 2: Random Forest performance with Gisette **image** dataset (mean and standard deviation are reported over 20 runs)

The task of GISETTE is to discriminate between two confusable handwritten digits: the four and the nine.
(A): Model trained on Gisette train dataset and tested on Gisette validation dataset shows that randomly selected subsets of just 200 achieve accuracy comparable to using all features.

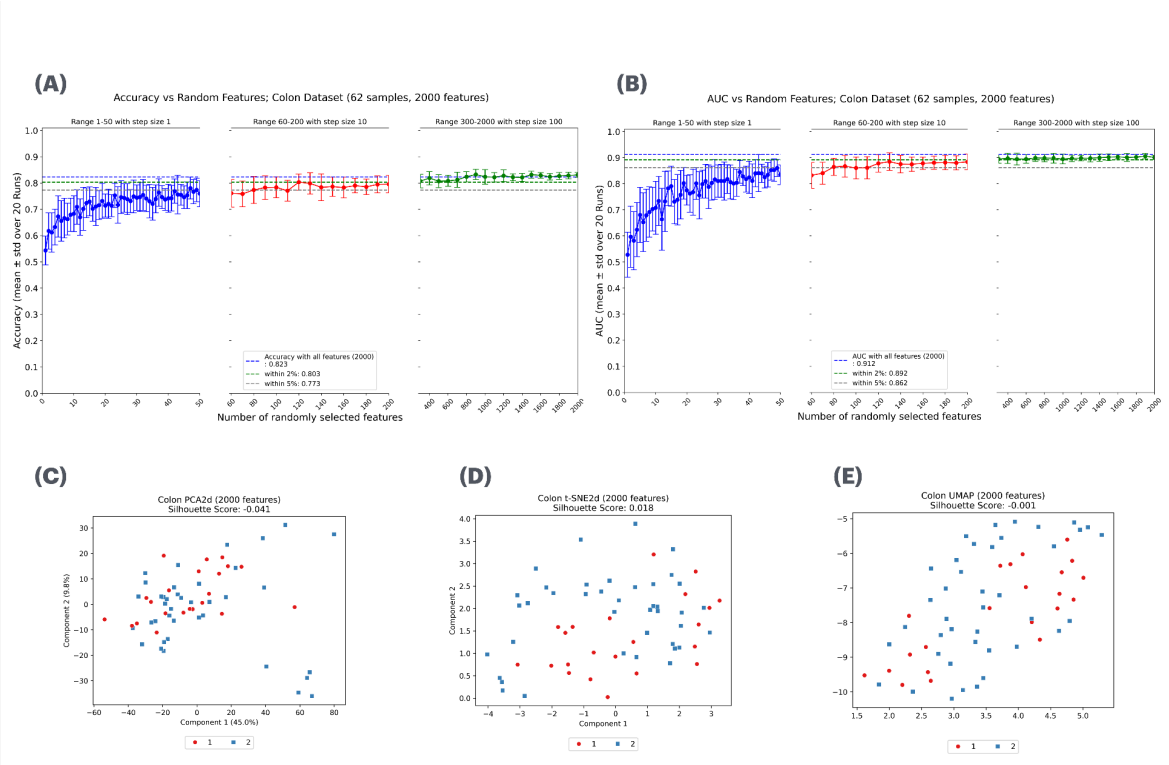


S1 Figure 3: Random Forest performance with **bulk RNA-Seq** and **Single-cell RNA-Seq** datasets (mean and standard deviation are reported over 20 runs)

(A) models trained and tested on TCGA-LUAD-LUSC bulk RNA-Seq dataset (80:20 split) shows that a random subset of size 50 (<0.3%) is able to match within-5% accuracy of all features.

(B) On the lung cancer SIngle-cell RNA-Seq dataset (80:20 split), randomly selected subsets of size 2000 achieve accuracy within 5% of the full-feature model.

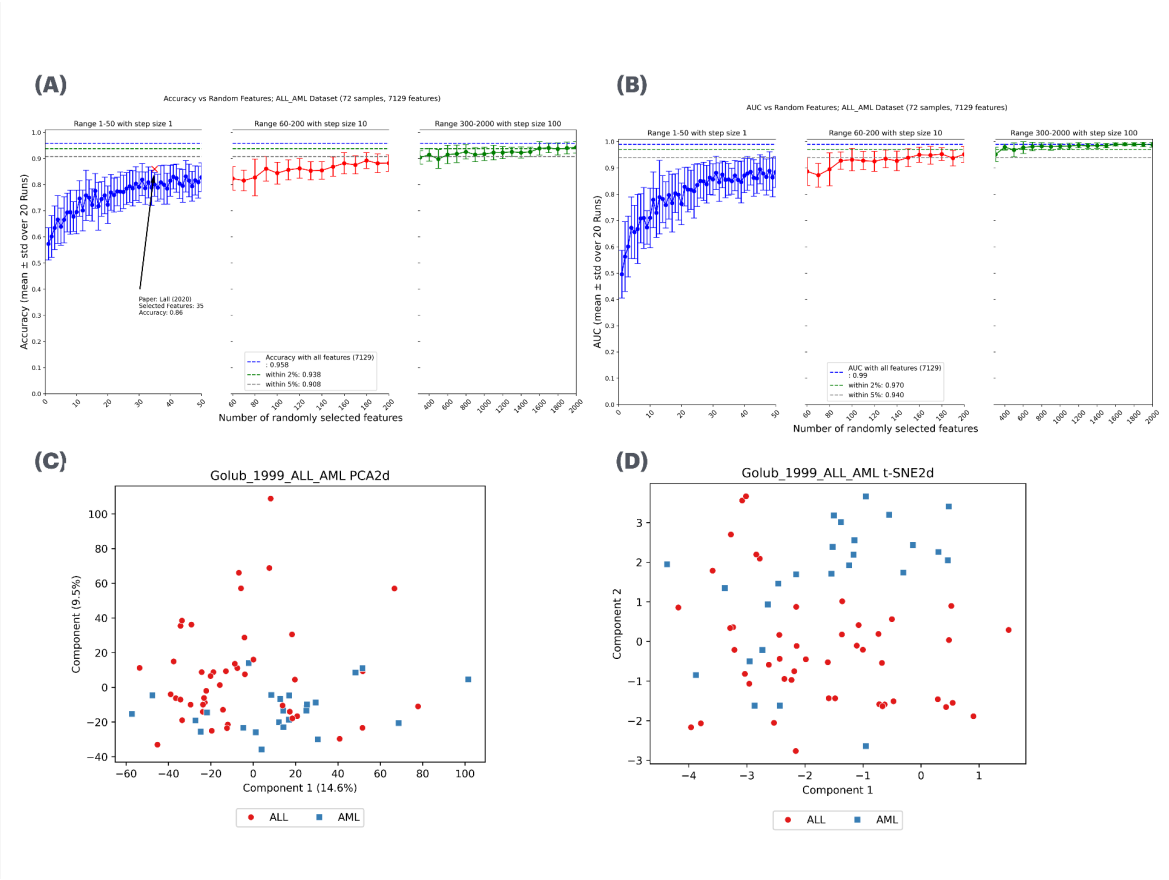
S2: Combined AUC and accuracy plots for all remaining datasets



S2 Figure 1: Random Forest performance with **Colon microarray** dataset (mean and standard deviation are reported over 20 runs)

(A) & (B) models trained and tested on 80:20 split shows that a random subset of size ~100 is able to match accuracy and AUC with all features, respectively.

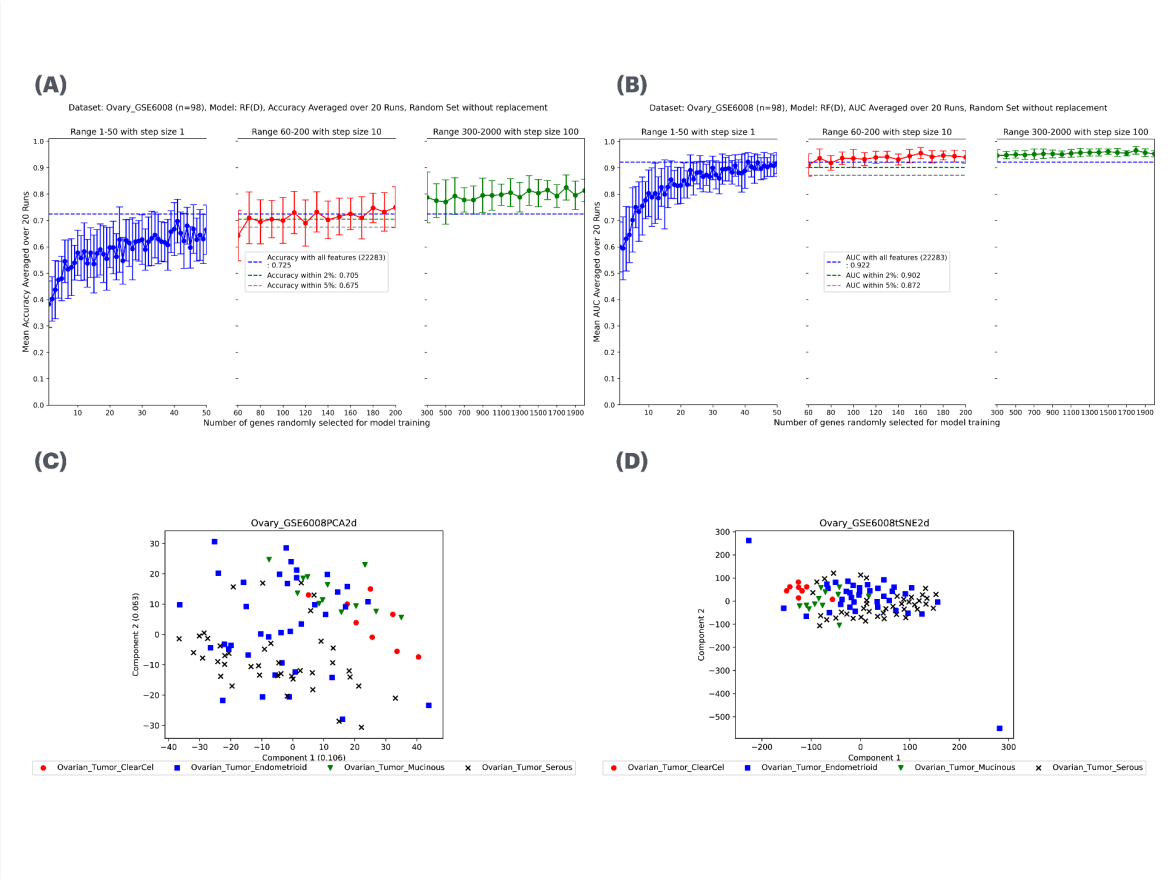
(C) & (D) & (E) PCA, t-SNE and UMAP plots showing class separation.



S2 Figure 2: Random Forest performance with **ALL/AML Leukemia microarray** dataset (mean and standard deviation are reported over 20 runs)

(A) & (B) models trained and tested on 80:20 split shows that a random subset of size ~200 is able to match accuracy and AUC with all features, respectively.

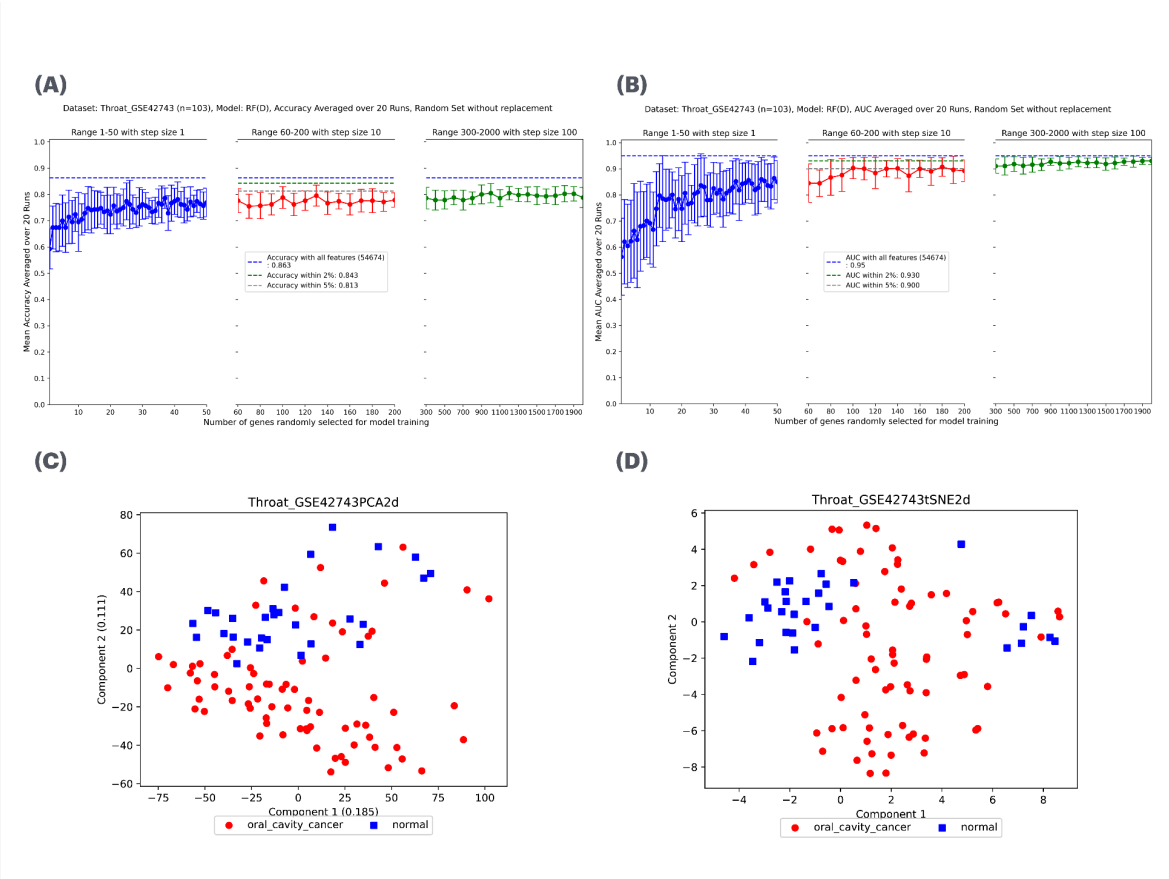
(C) & (D) PCA, t-SNE plots showing class separation.



S2 Figure 3: Random Forest performance with **Ovary (GSE6008) microarray** dataset (mean and standard deviation are reported over 20 runs)

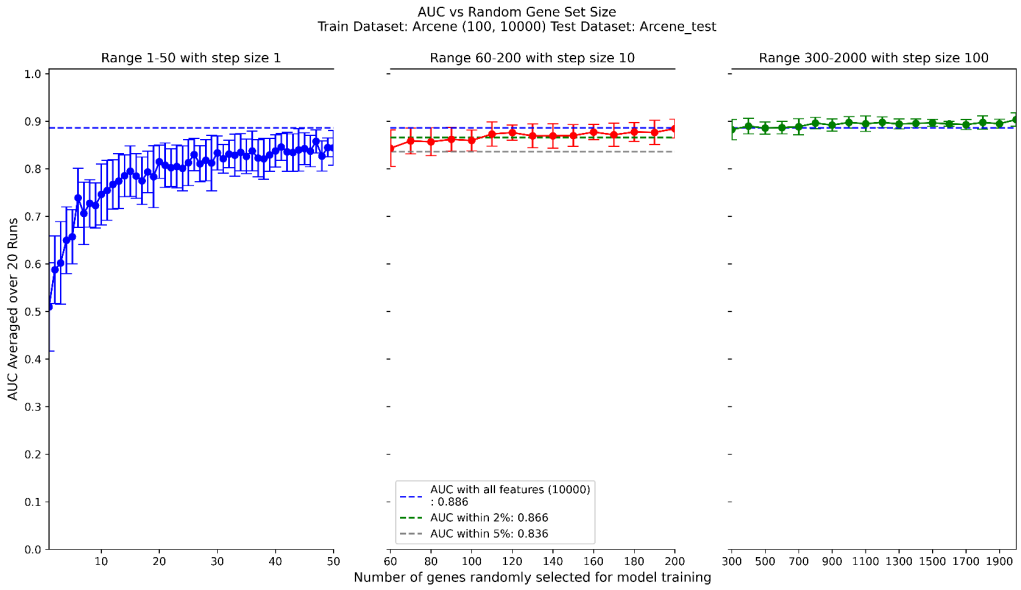
(A) & (B) models trained and tested on 80:20 split shows that a random subset is able to match accuracy and AUC with all features, respectively.

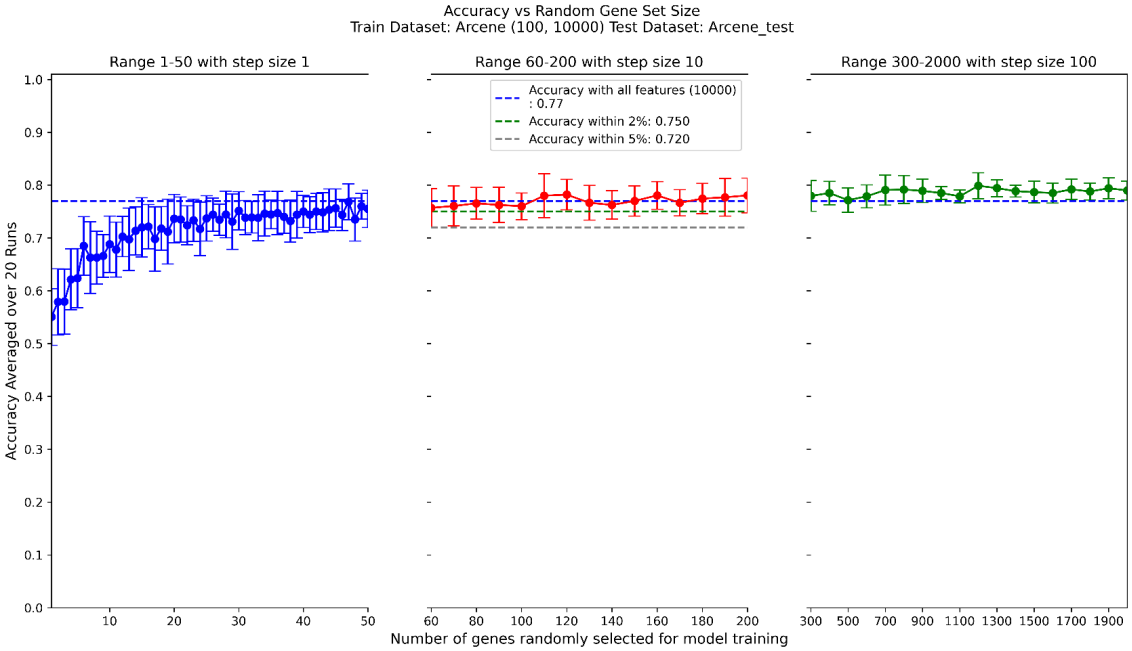
(C) & (D) PCA, t-SNE plots showing class separation.

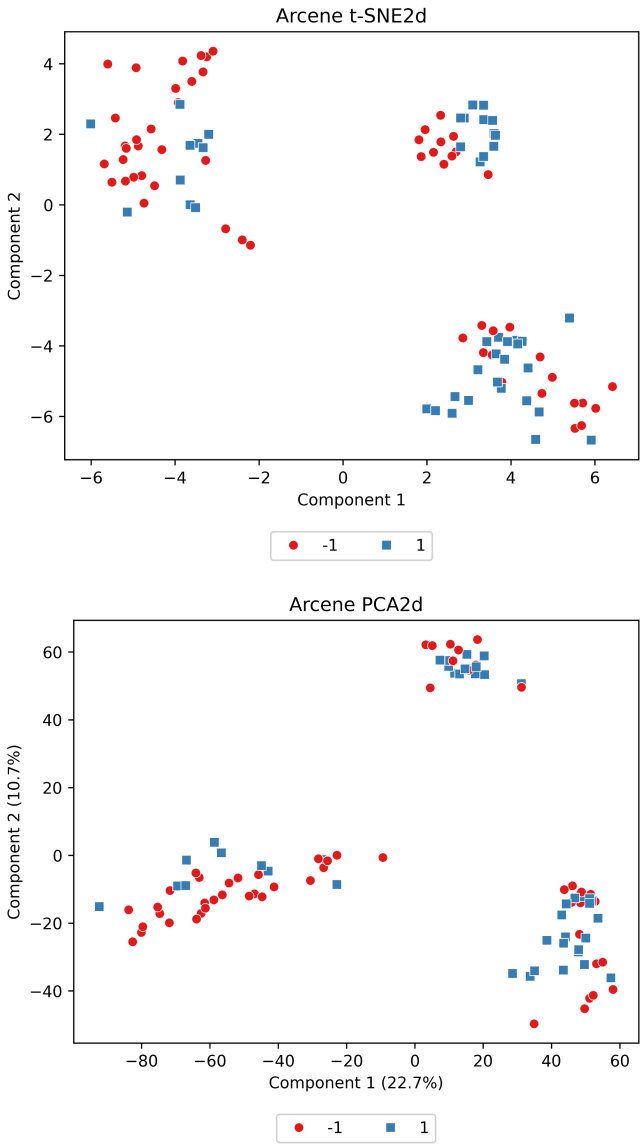


S2 Figure 4: Random Forest performance with **Oral/Throat** (GSE42743)) **microarray** dataset (mean and standard deviation are reported over 20 runs)
(A) & (B) models trained and tested on 80:20 split shows that a random subset is able to match within-5% accuracy and AUC with all features, respectively.
(C) & (D) PCA, t-SNE plots showing class separation.

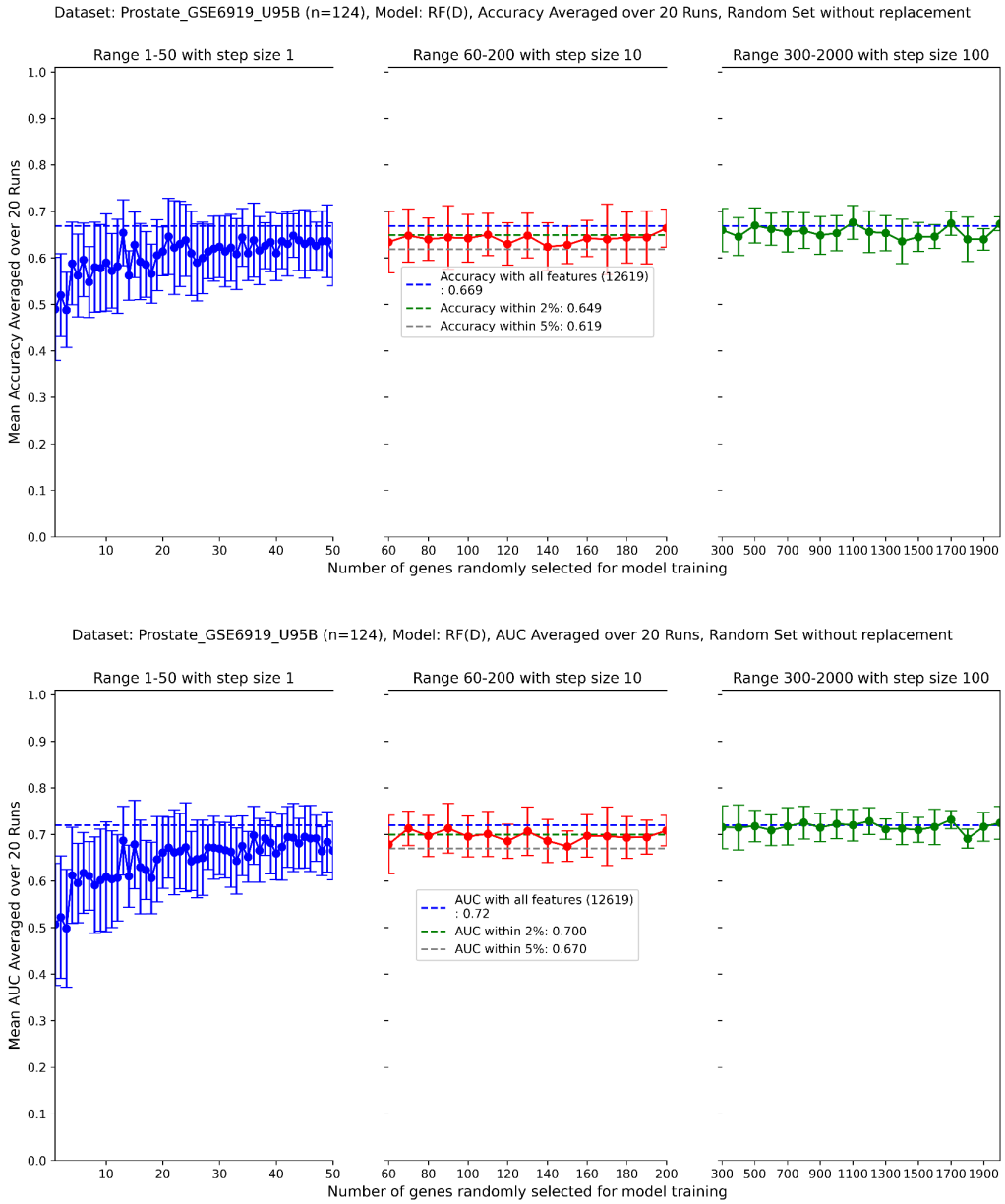
(A)


AUC vs Random Gene Set Size
Train Dataset: Arcene (100, 10000) Test Dataset: Arcene_test
Range 1-50 with step size 1
Range 60-200 with step size 10
Range 300-2000 with step size 100
AUC Averaged over 20 Runs
AUC with all features (10000) : 0.886
AUC within 2%: 0.866
AUC within 5%: 0.836
Number of genes randomly selected for model training

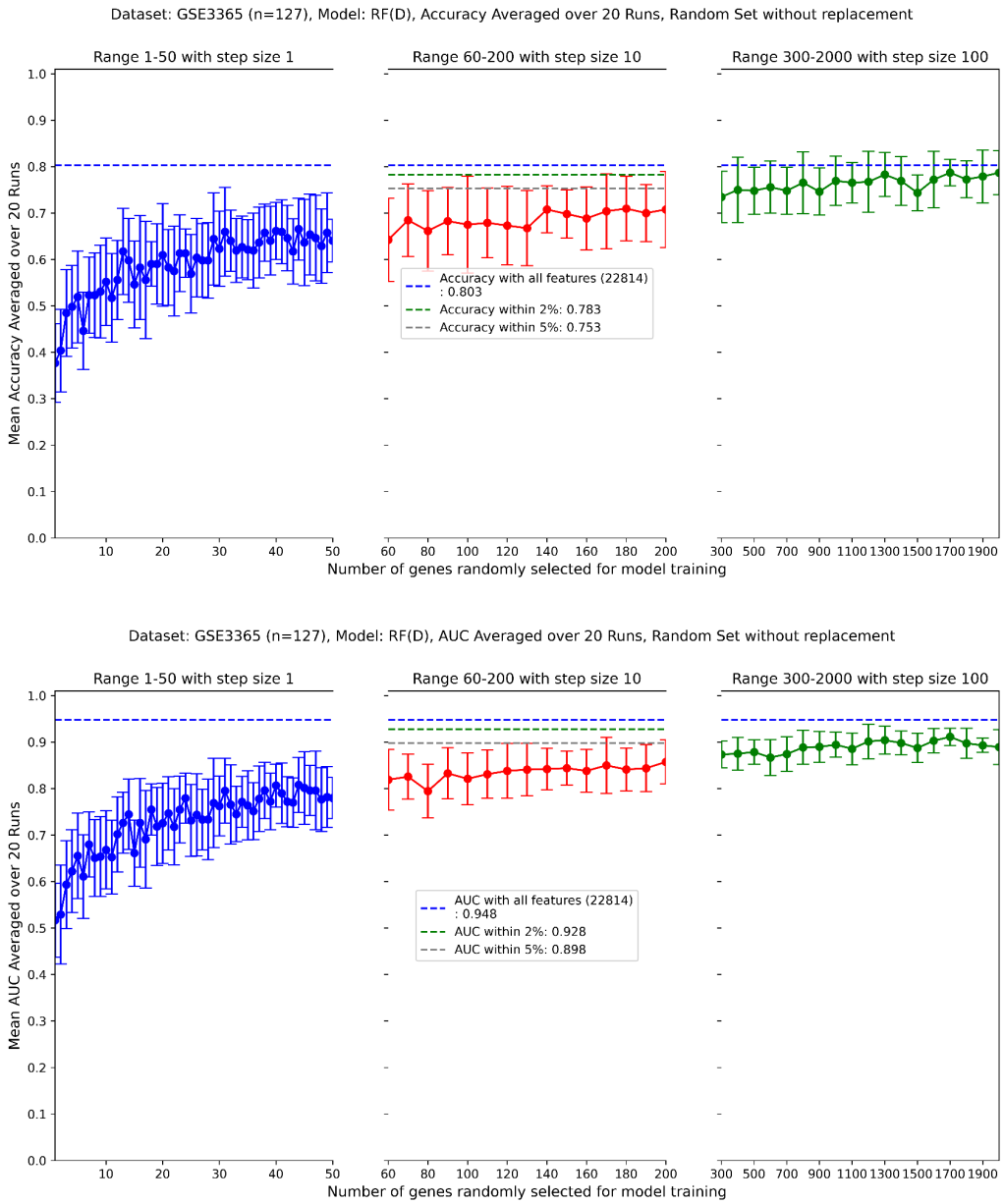

Accuracy vs Random Gene Set Size
Train Dataset: Arcene (100, 10000) Test Dataset: Arcene_test
Range 1-50 with step size 1
Range 60-200 with step size 10
Range 300-2000 with step size 100
Accuracy Averaged over 20 Runs
Accuracy with all features (10000) : 0.77
Accuracy within 2%: 0.750
Accuracy within 5%: 0.720
Number of genes randomly selected for model training



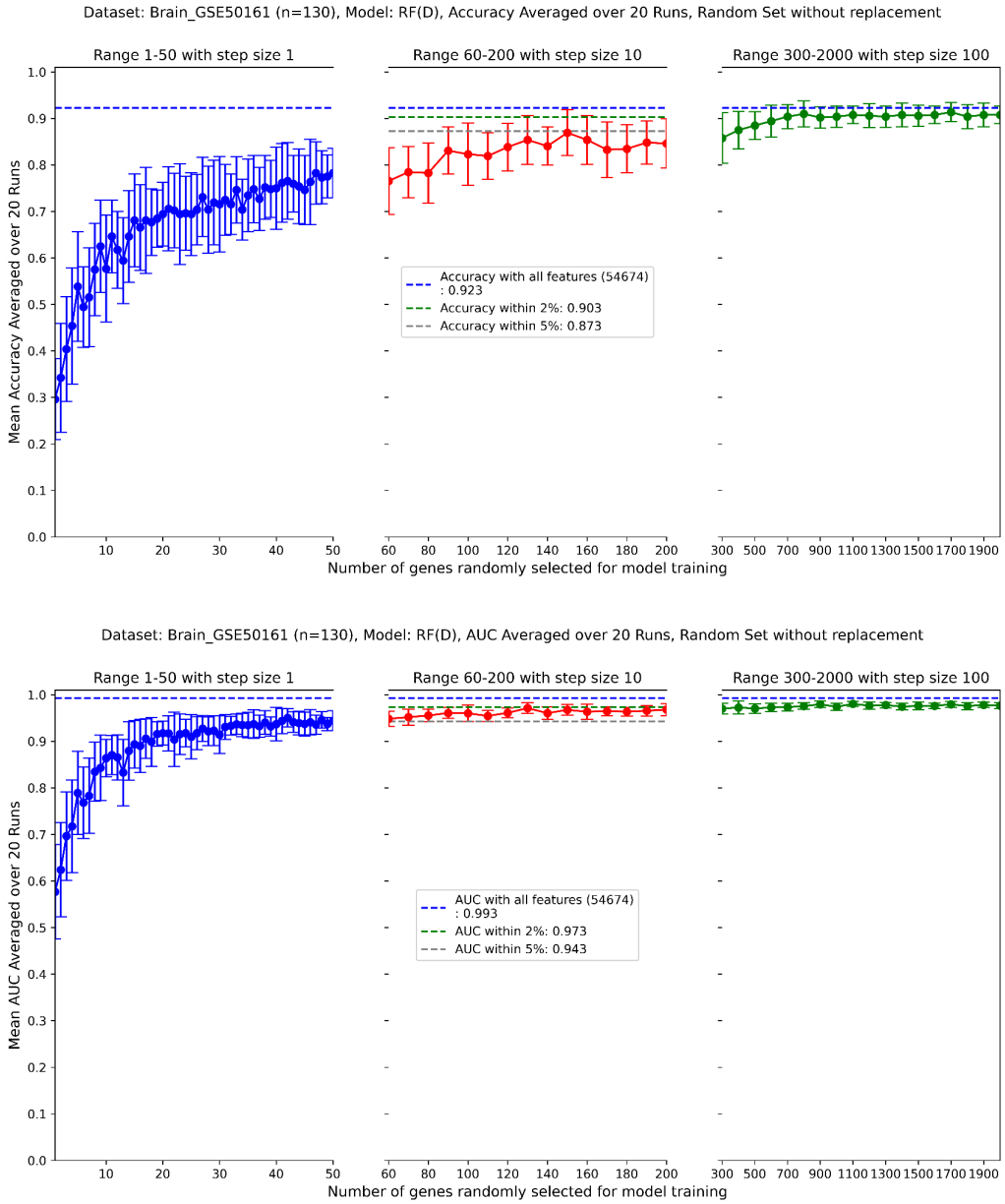
S2 Figure 5: Random Forest performance with **Arcene mass-spectrometry** dataset (mean and standard deviation are reported over 20 runs)
The task of ARCENE is to distinguish cancer versus normal patterns from mass-spectrometric data.
(A) Models trained and tested on 80:20 split shows that a random subset of size ~50 (0.5% of all features) is able to match within-5% Accuracy and AUC of all features.
(B), (C) PCA, t-SNE plots showing class separation.



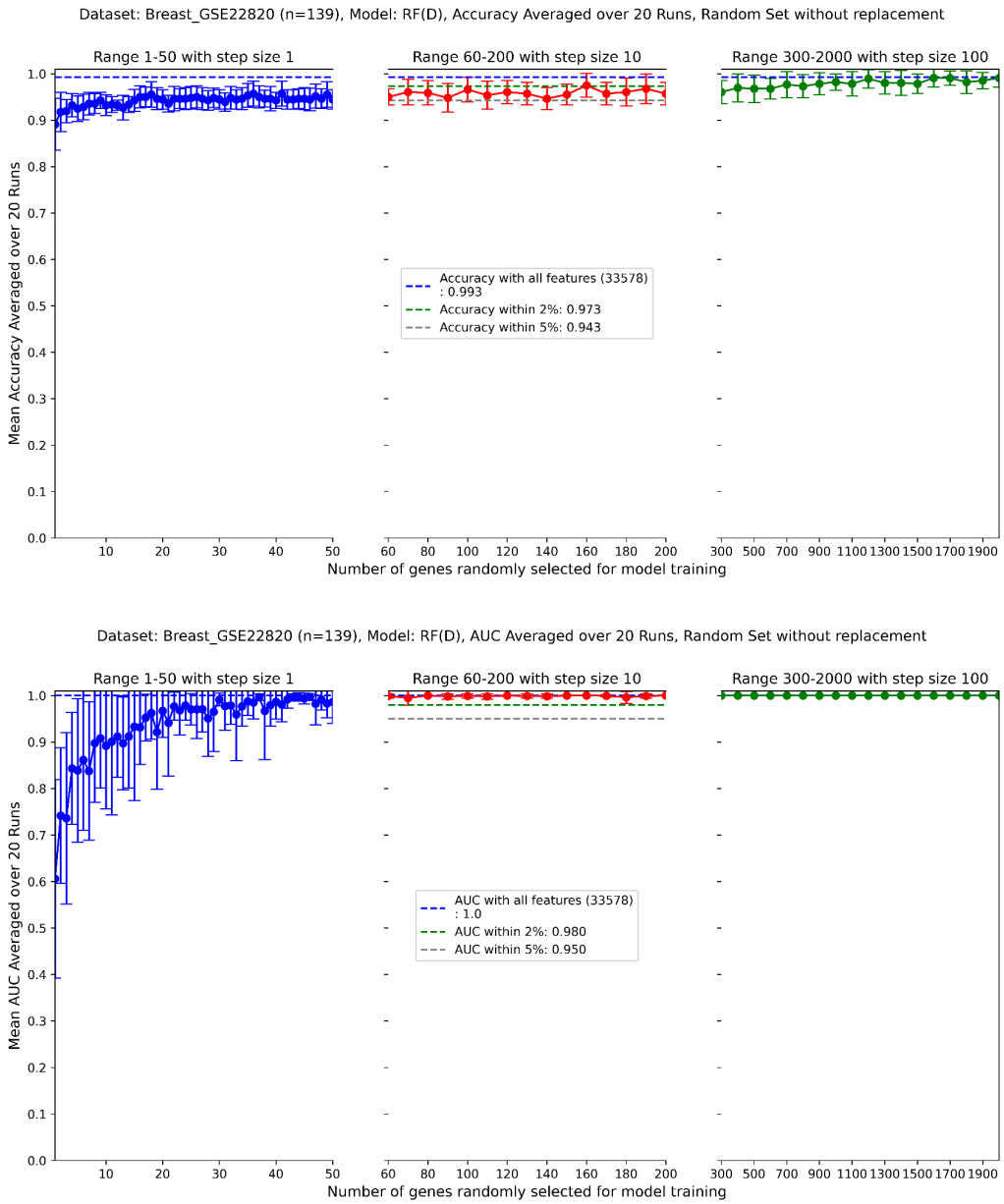
S2 Figure 6: Random Forest performance with Prostate (GSE6919_U95B) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset of size ~50 (0.4% of all features) is able to match within-5% Accuracy and AUC of all features.



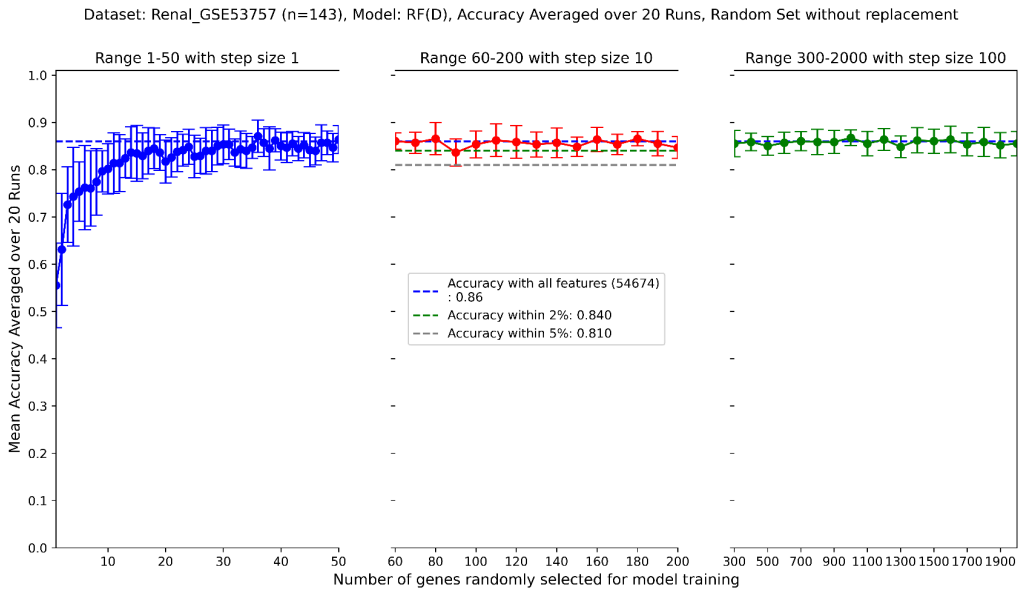
S2 Figure 7: Random Forest performance with Bowl (GSE3365) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset of size ~500 (~2.2% of all features) is able to match within-5% Accuracy and AUC of all features.

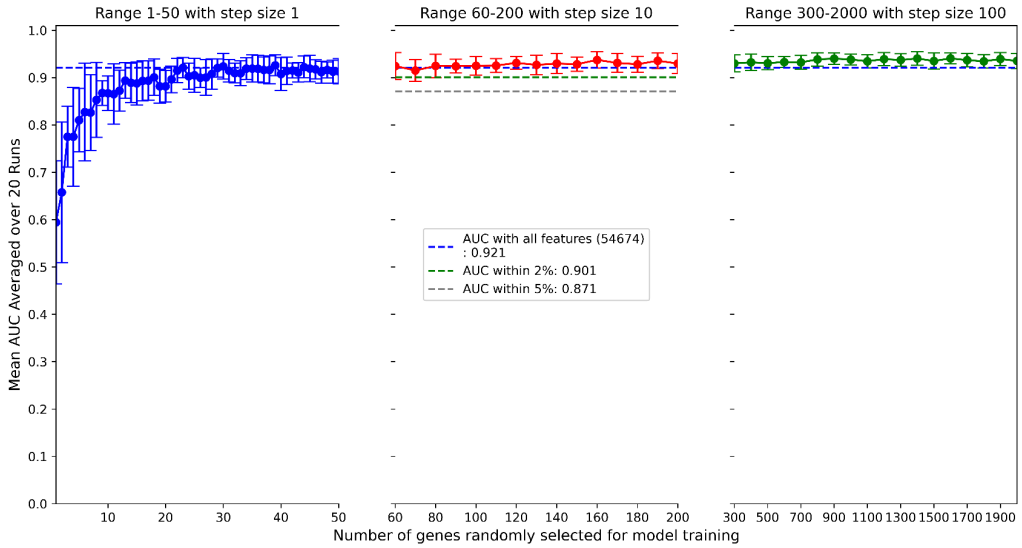


S2 Figure 8: Random Forest performance with Brain (GSE50161) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset of size ~50 (~0.09% of all features) is able to match within-5% Accuracy and AUC of all features.

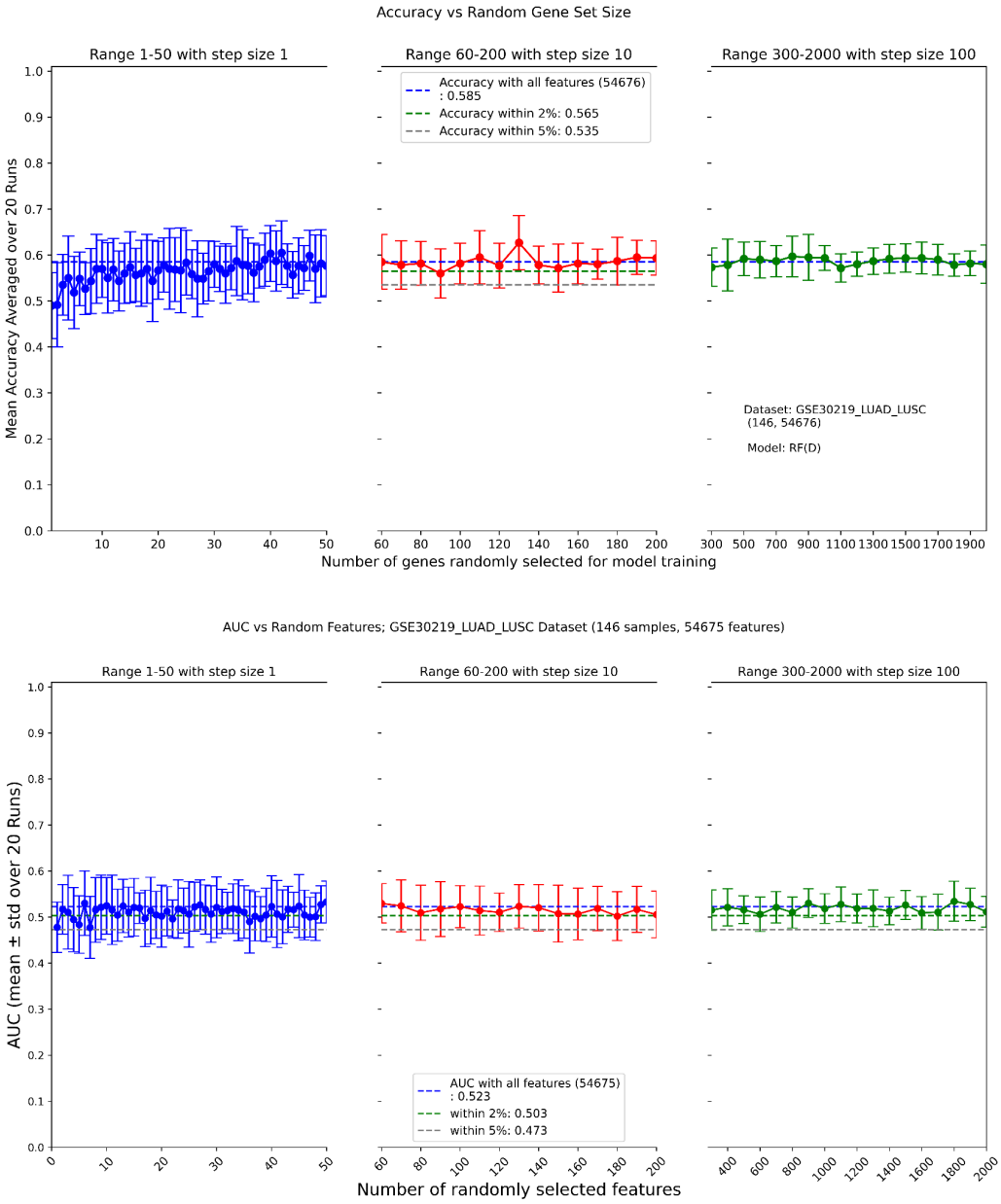


S2 Figure 9: Random Forest performance with Breast (GSE22820) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset of size ~50 (~0.14% of all features) is able to match within-5% Accuracy and full AUC of all features. (The unusually high accuracy with just one feature is because there is a severe class imbalance in this dataset)

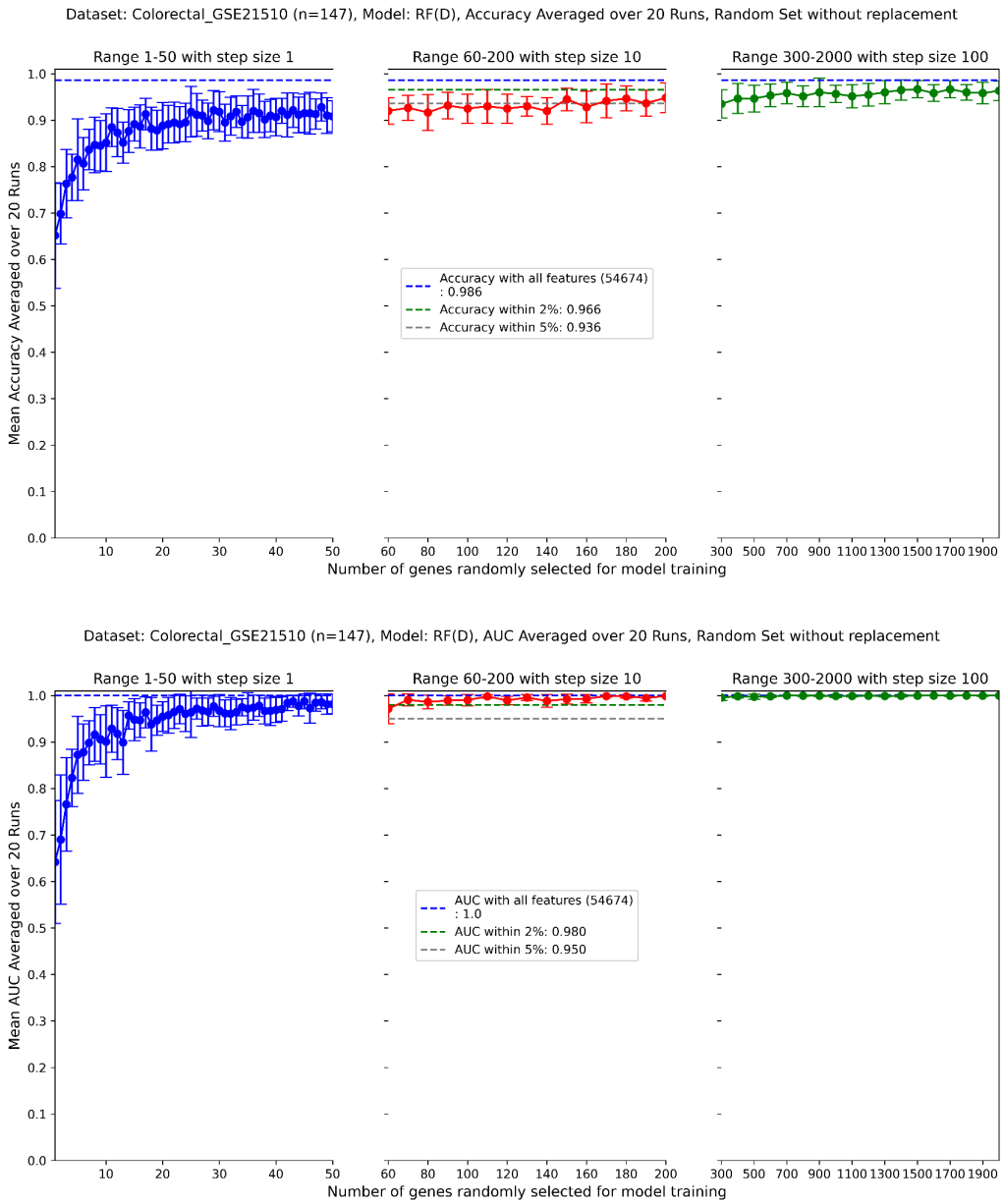




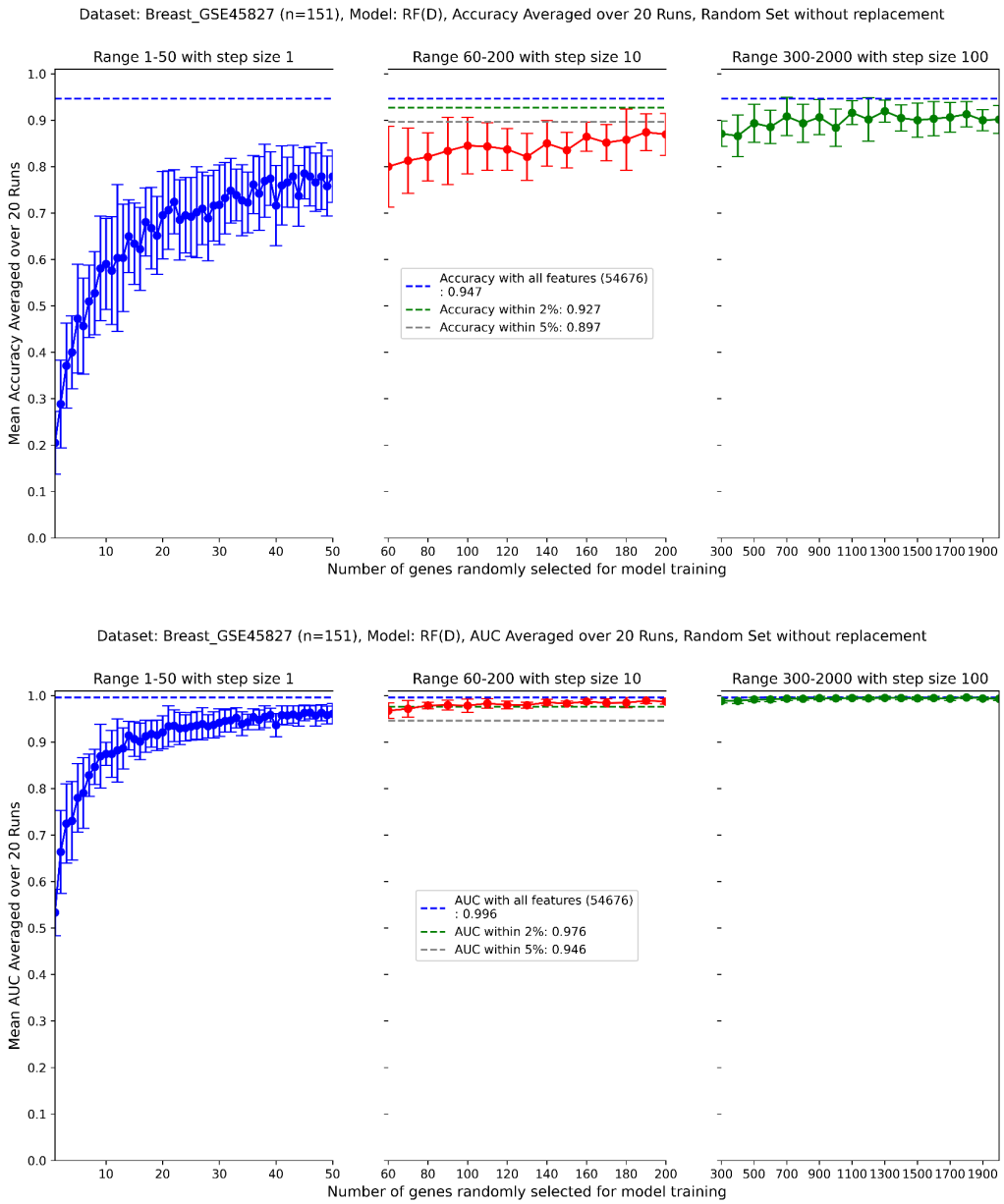
S2 Figure 10: Random Forest performance with Renal (GSE53757) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset of size ~30 (~0.06% of all features) is able to match full Accuracy and full AUC of all features.



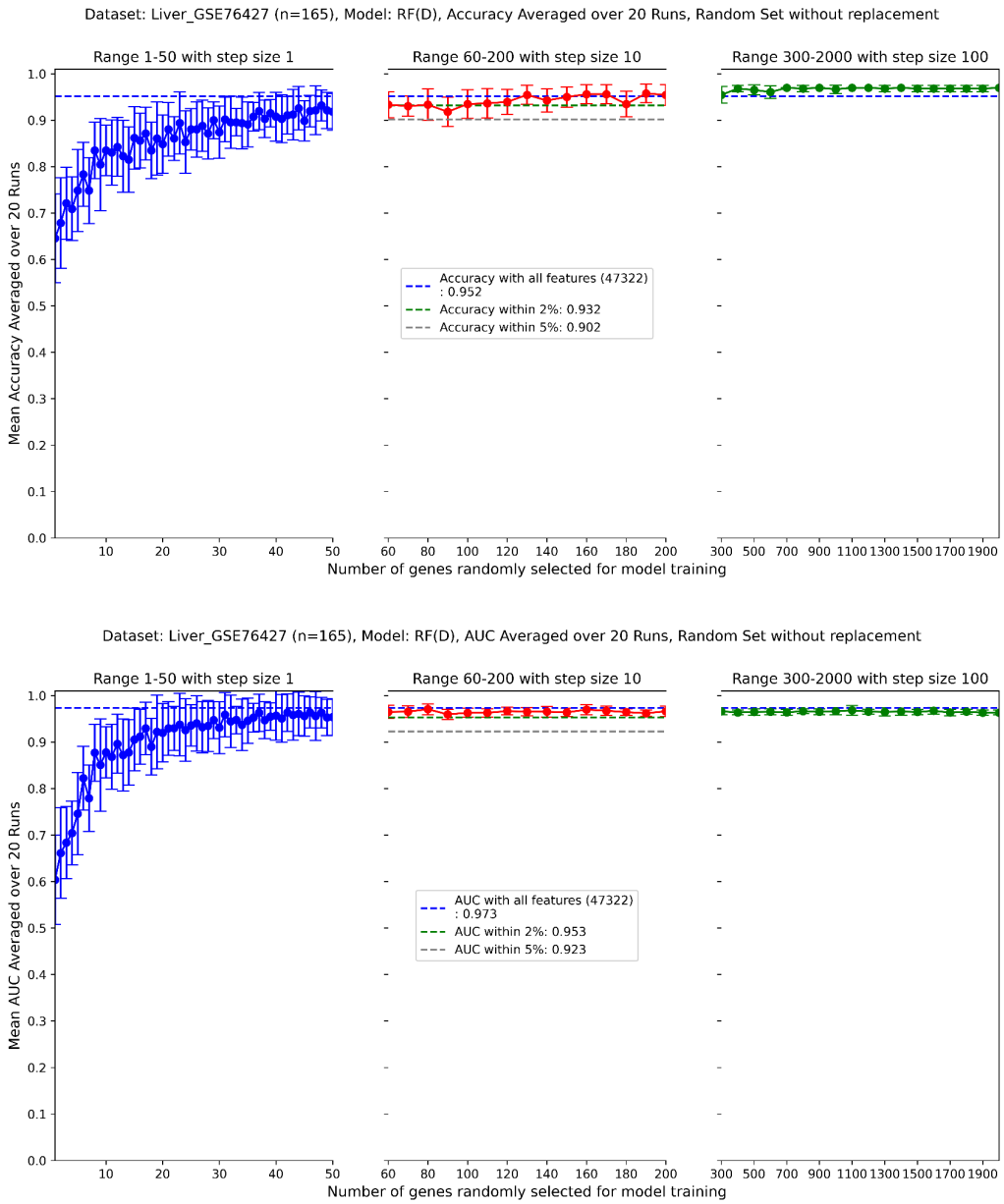
S2 Figure 11: Random Forest performance with Lung Cancer (GSE30219) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match full Accuracy and full AUC of all features.



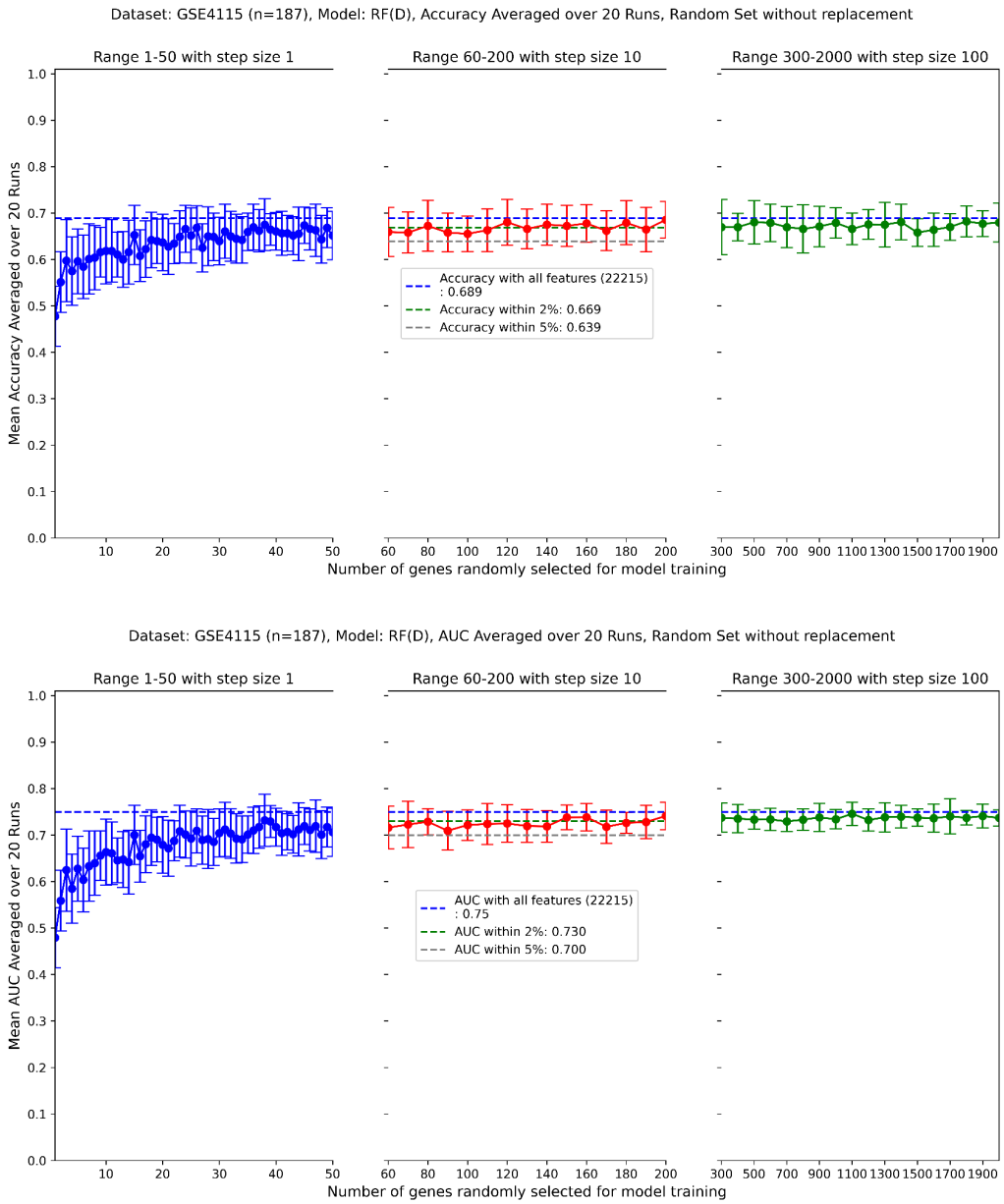
S2 Figure 12: Random Forest performance with Lung Cancer (GSE30219) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match within-2% Accuracy and full AUC of all features



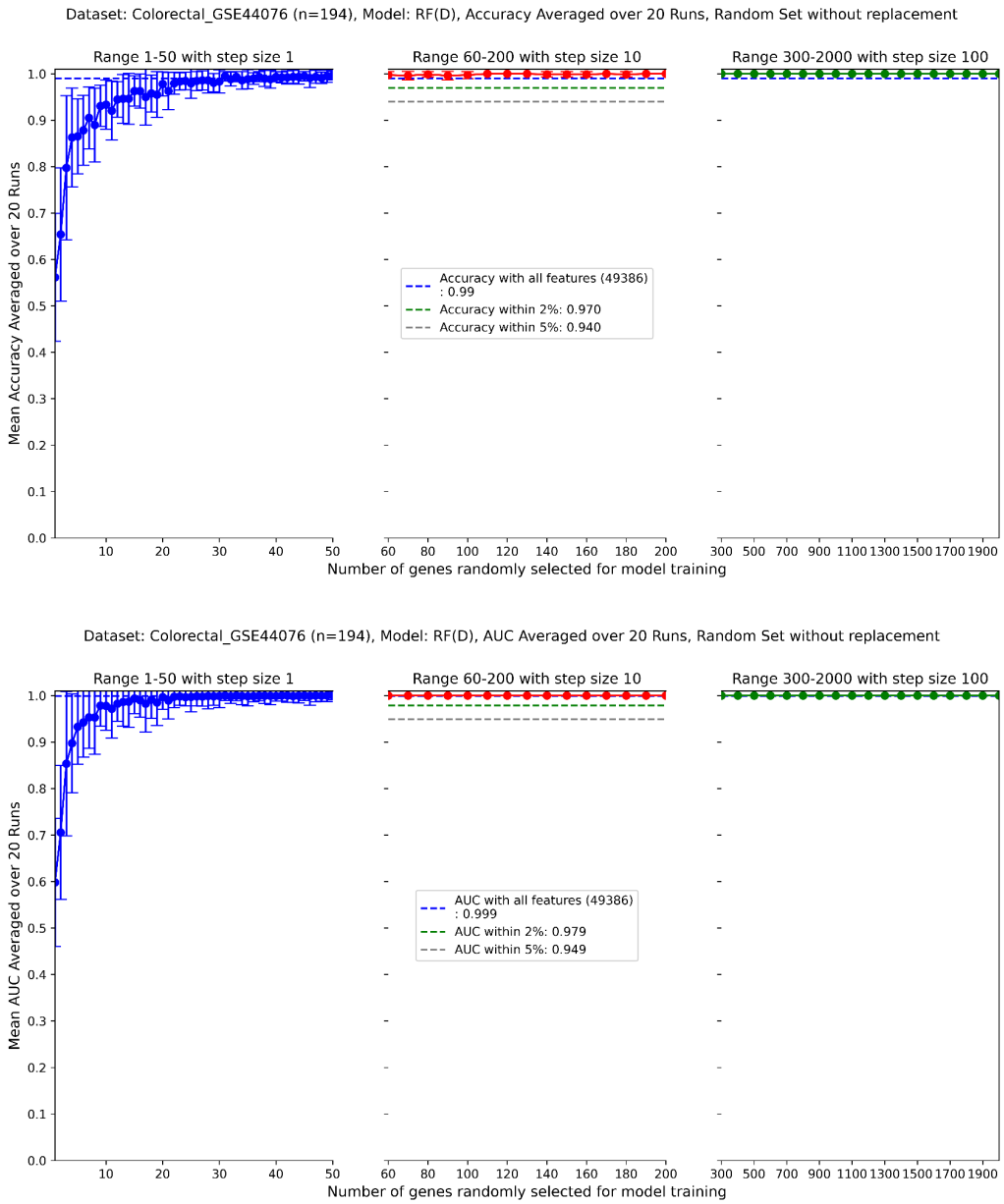
S2 Figure 13: Random Forest performance with Breast Cancer (GSE45827) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match within-5% Accuracy and full AUC of all features



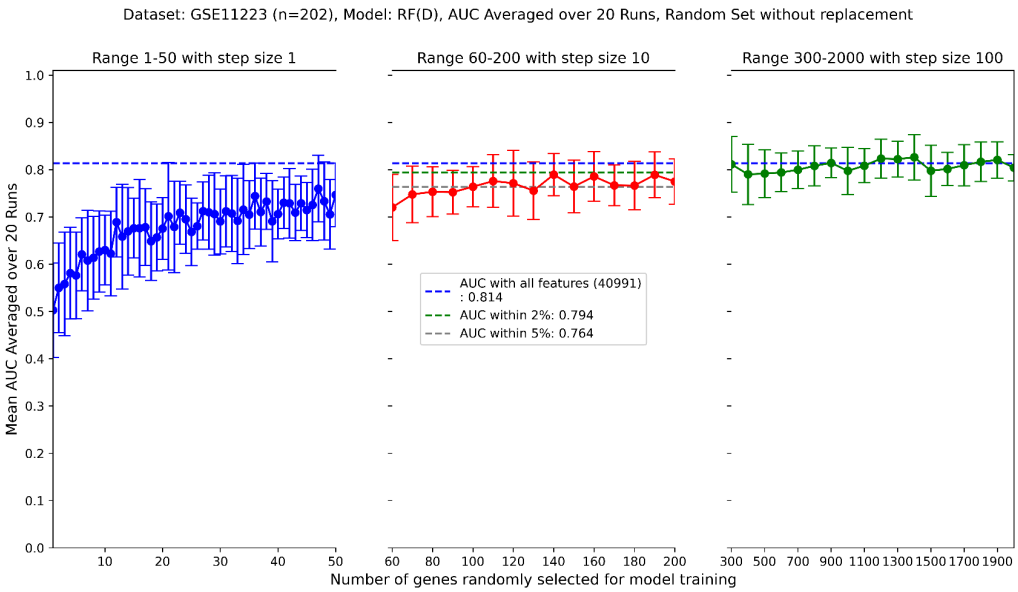
S2 Figure 14: Random Forest performance with Liver Cancer (GSE76427) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match full Accuracy and full AUC of all features



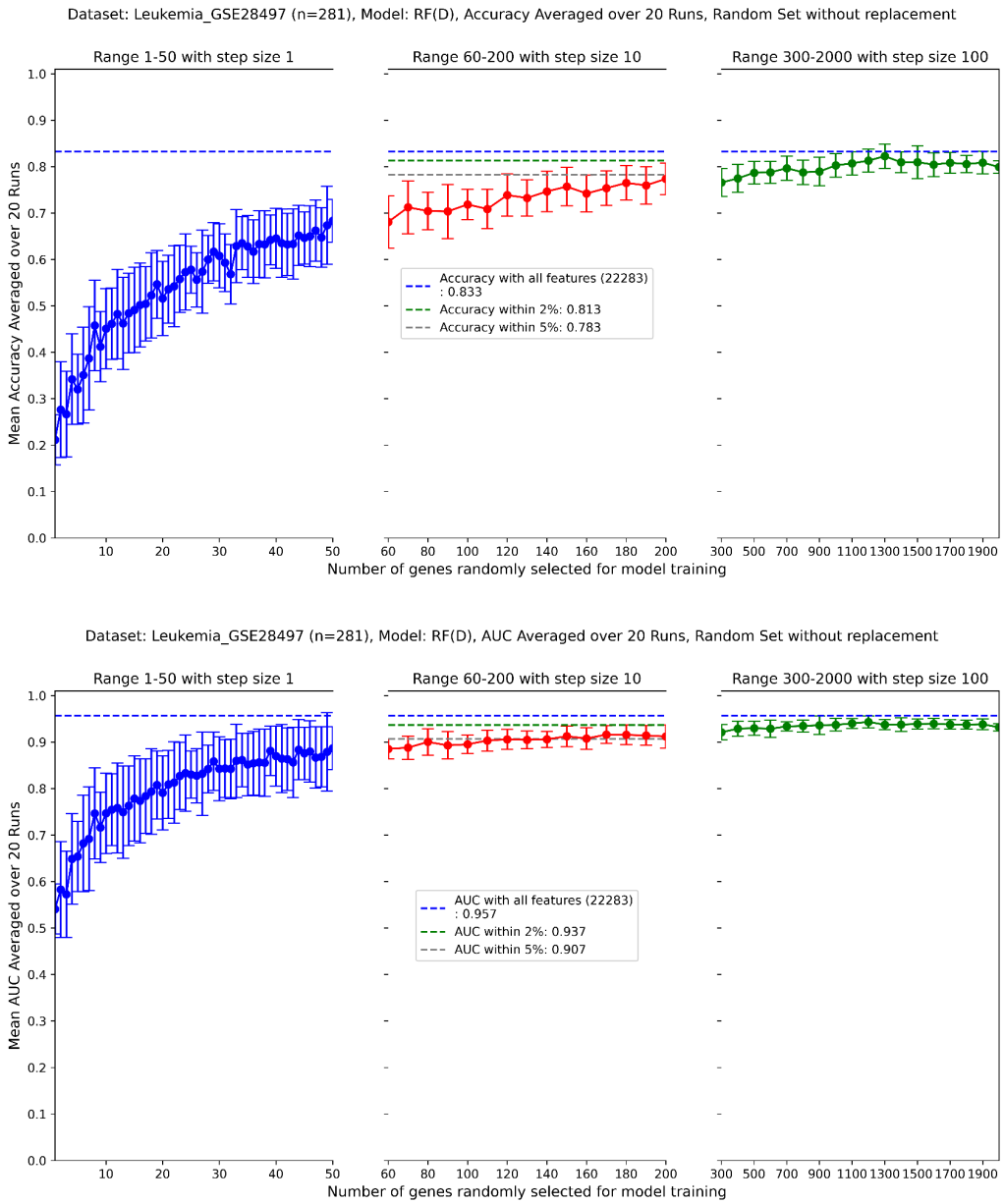
S2 Figure 15: Random Forest performance with Lung Cancer (GSE4115) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match within-2% Accuracy and AUC of all features



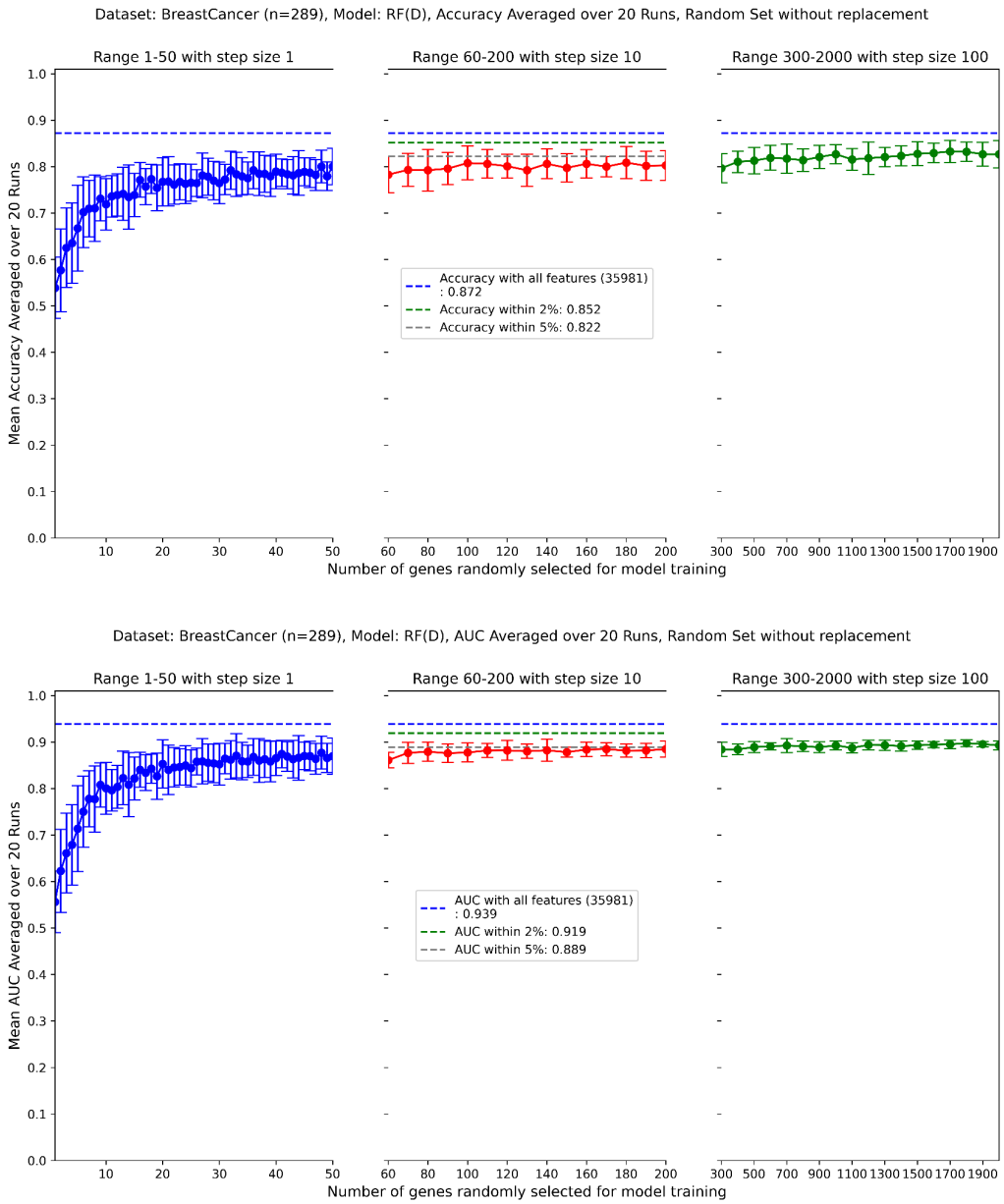
S2 Figure 16: Random Forest performance with Colorectal Cancer (GSE44076) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match full Accuracy and full AUC of all features



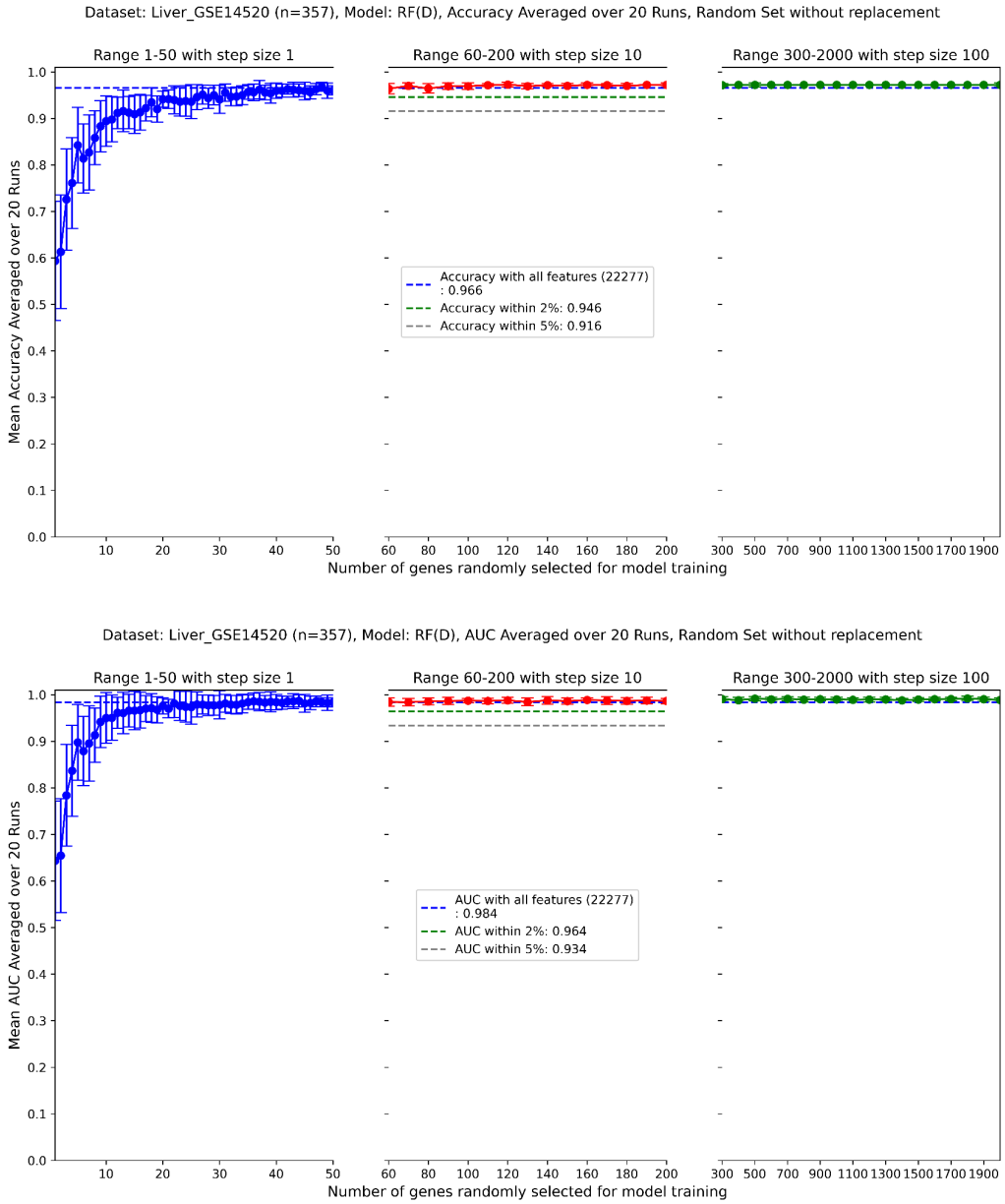
S2 Figure 17: Random Forest performance with Colon Cancer (GSE11223) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match full AUC of all features



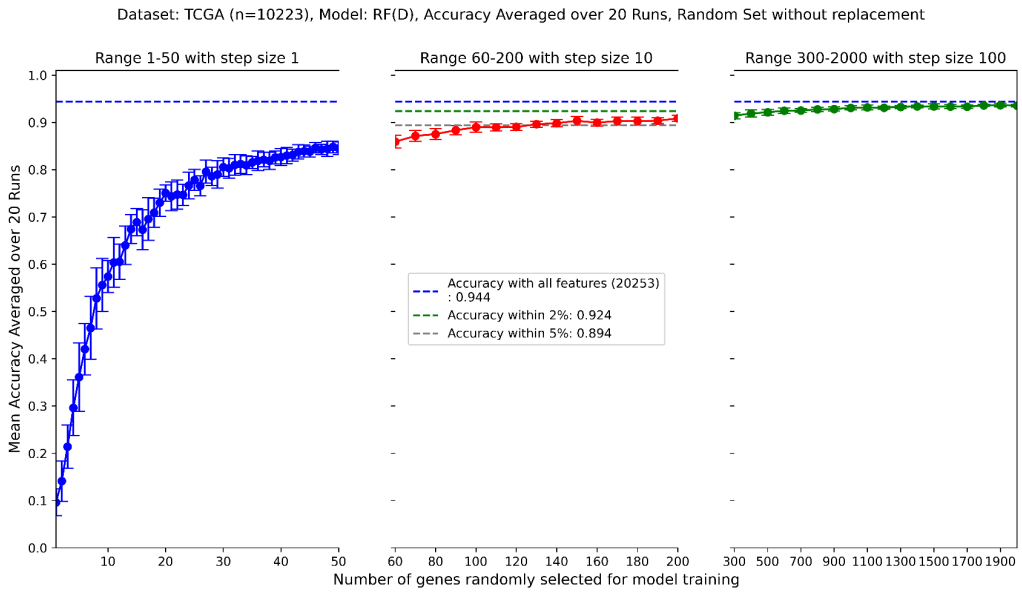
S2 Figure 18: Random Forest performance with Leumkemia (GSE28497) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match within-2% Accuracy and AUC of all features.

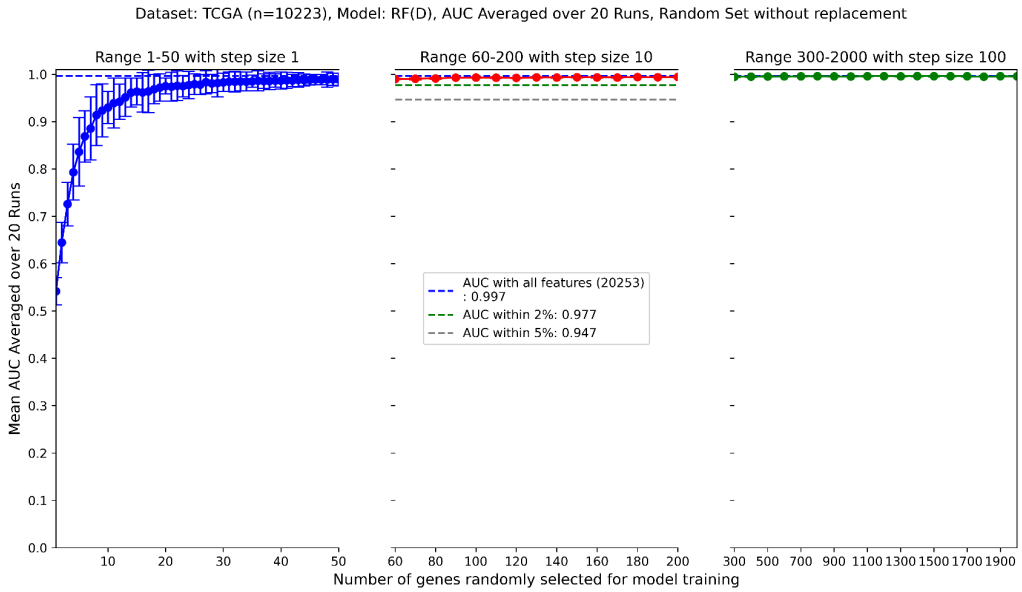


S2 Figure 19: Random Forest performance with Breast Cancer (GSE70947) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a random subset is able to match within-5% Accuracy and AUC of all features

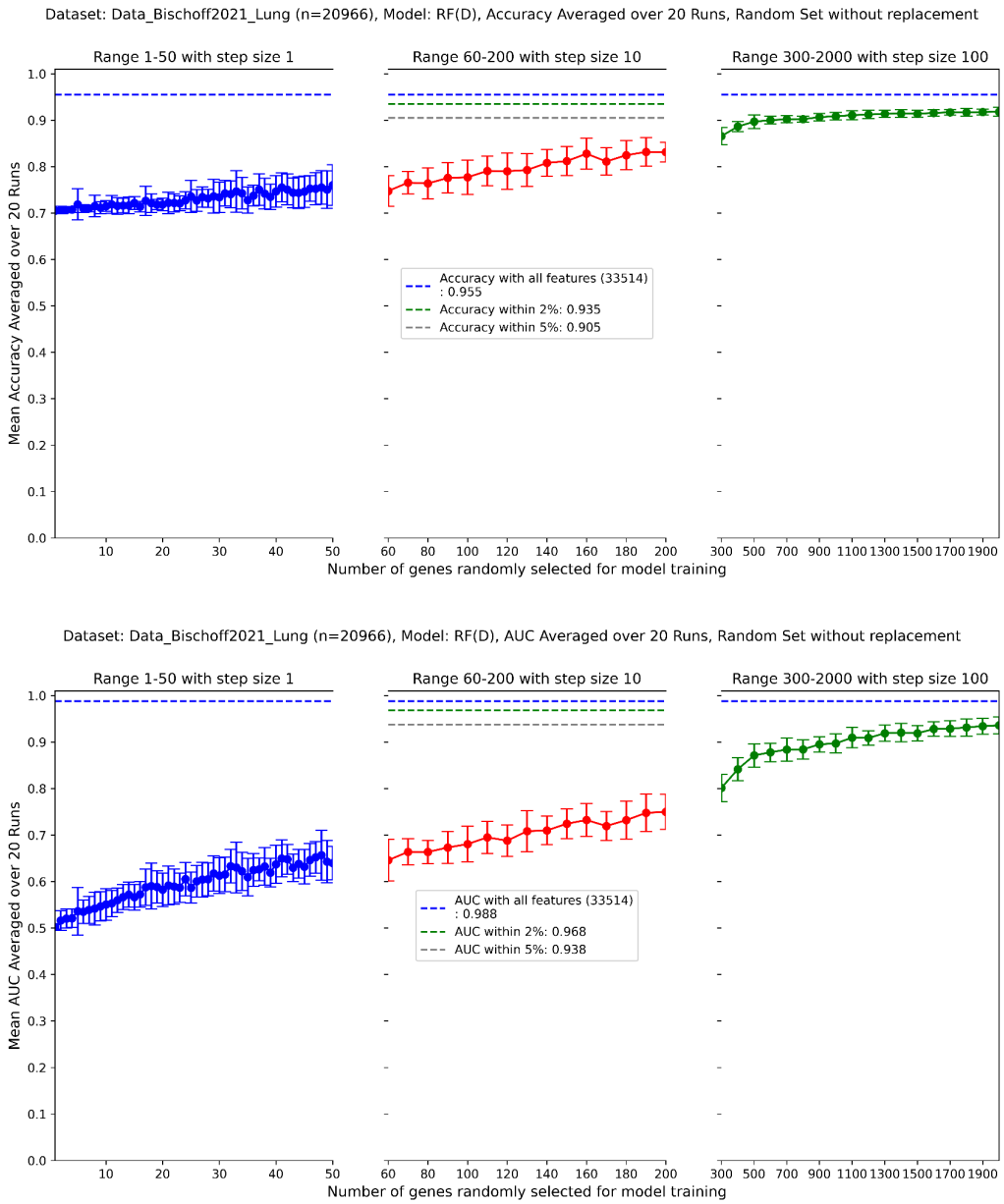


S2 Figure 20: Random Forest performance with Liver Cancer (GSE14520) dataset (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a small random subset is able to match full Accuracy and AUC of all features

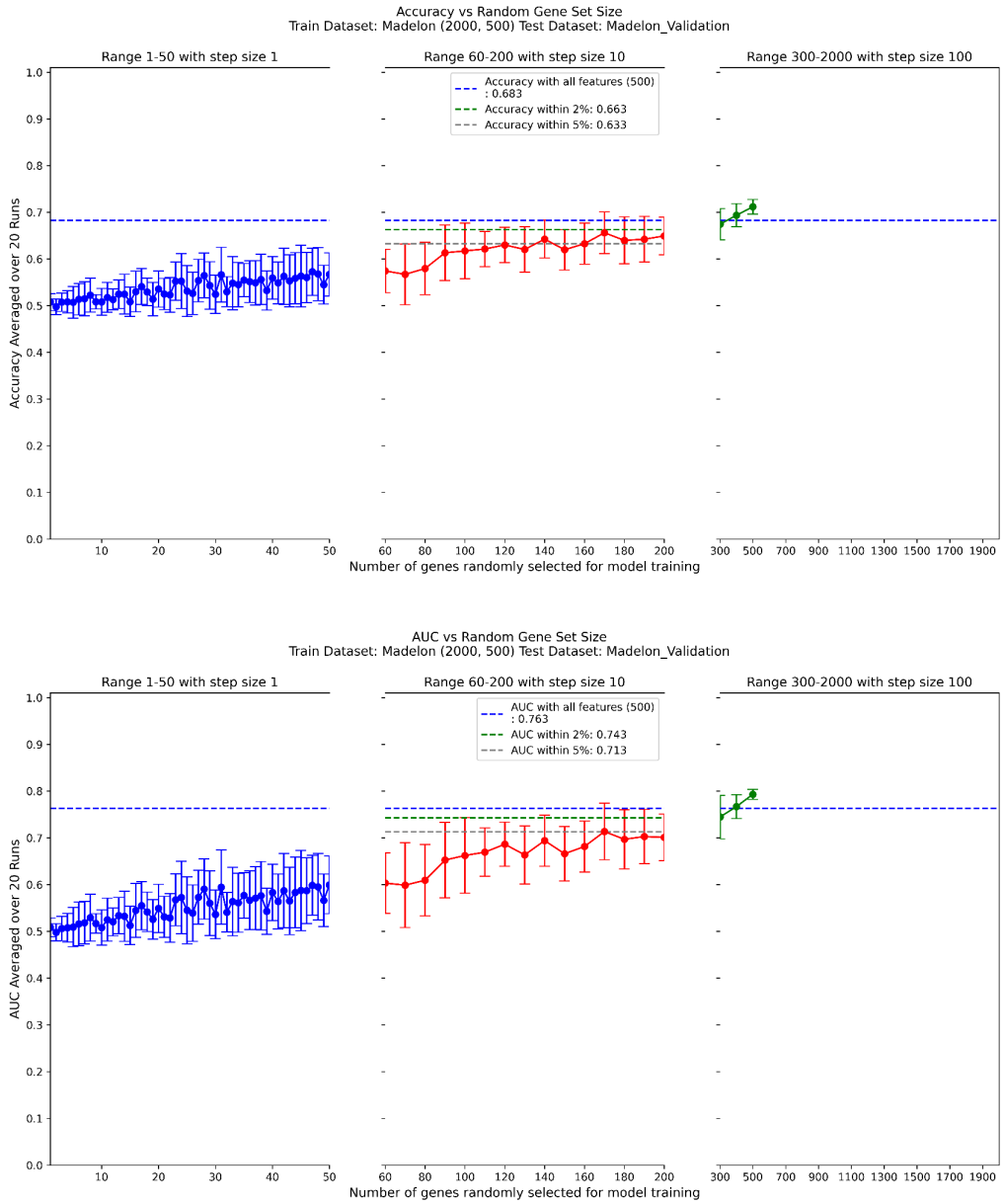




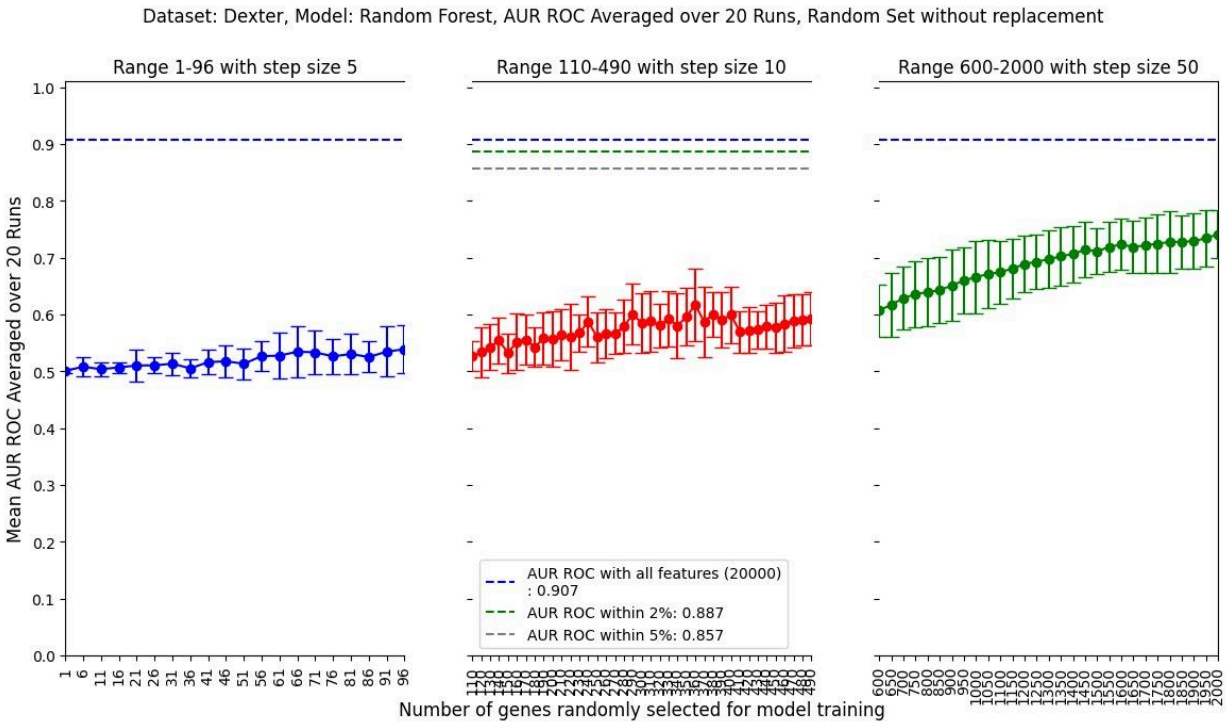
S2 Figure 21: Random Forest performance with bulk RNA-Seq TCGA dataset iwth 33 classes (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a small random subset is able to match full Accuracy and AUC of all features



S2 Figure 22: Random Forest performance with single-cell RNA-Seq Lung dataset with 9 classes (mean and standard deviation are reported over 20 runs)
Models trained and tested on 80:20 split shows that a small random subset is able to match within-5% Accuracy and AUC of all features



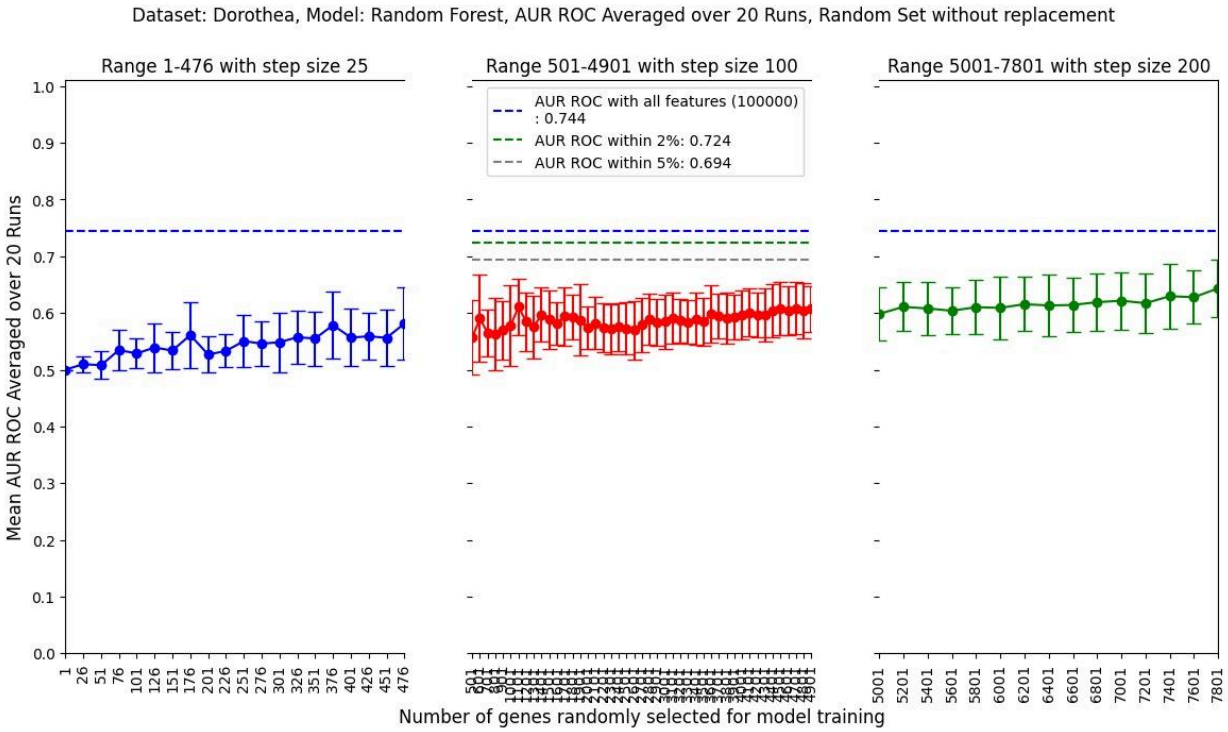
S2 Figure 23: Random Forest performance with Madelon dataset (mean and standard deviation are reported over 20 runs)
The task of MADELON is to classify random data.
Models trained and tested on 80:20 split shows that a small random subset is able to match within-5% Accuracy and AUC of all features. (As there are only 500 features in this dataset, there is no result beyond those many features)



S2 Figure 24: Random Forest performance with Dexter dataset (mean and standard deviation are reported over 20 runs)

The task of DEXTER is to filter texts about "corporate acquisitions".

Models trained and tested on 80:20 split shows that a random subset is NOT able to match AUC of all features.



S2 Figure 25: Random Forest performance with Dorothea dataset (mean and standard deviation are reported over 20 runs)
The task of DOROTHEA is to predict which compounds bind to Thrombin.
Models trained and tested on 80:20 split shows that a random subset is NOT able to match AUC of all features.