PREDICTIVE PERFORMANCE IS OFTEN INSENSITIVE TO FEATURE SELECTION IN HIGH-DIMENSIONAL BIOLOGICAL CLASSIFICATION

Bhavesh Neekhra & Debayan Gupta
Department of Computer Science
Ashoka University
Sonipat, India
{bhavesh.neekhra_phd18,debayan.gupta}@ashoka.edu.in

Partha Pratim Chakrabarti
Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur, India
{ppchak}@cse.iitkgp.ac.in

ABSTRACT

Feature selection (FS) is commonly assumed to improve predictive performance and identify meaningful features in high-dimensional data. We systematically evaluated this assumption across 30 classification benchmarks, primarily drawn from computational genomics, including microarray, bulk RNA-Seq, mass spectrometry, and imaging datasets. Across these datasets, we observe that small random subsets of features (0.02–1.0% of available features) frequently achieve predictive performance that is comparable to, and in some cases statistically indistinguishable from, models trained on full feature set. Notably, performance variability across random subsets of a given size is often low, suggesting substantial redundancy in the predictive signal. Together, these results suggest that, for many widely used high-dimensional biological benchmarks, predictive accuracy alone is not sufficient to justify claims about the importance of specific selected features. They also underscore the need for rigorous validation before interpreting selected features as biologically meaningful or actionable, particularly in computational genomics.

1 INTRODUCTION

Feature selection (FS) is considered a critical part of nearly all machine learning research on high-dimensional datasets. Beyond just computational efficiency, FS is often used to identify features deemed "important" to a given outcome, especially in fields such as computational genomics. Some highly cited, influential studies which do this include Golub et al. (1999); Guyon et al. (2002; 2004); Li et al. (2017); Lall & Bandyopadhyay (2019); Cilia et al. (2019); Chen & Dhahbi (2021); Zanella et al. (2022). As of January 2026, a Google Scholar search for *"feature selection on high dimensional datasets"* lists about 3, 010, 000 results - it is beyond doubt that large resources are being invested in this area.

The aim of such work is typically to identify a small subset of features which, if used to train a machine-learning model, results in high accuracy (sometimes surpassing that of the full feature set). These cleverly-selected features are also often deemed to be important to the underlying task. Strikingly, we found that a critical baseline is almost always missing: a simple null hypothesis, comparing the results against a random subset of features of the same size. Without this comparison, it is unclear whether sophisticated FS methods truly outperform chance.

In this work, we systematically evaluate the null hypothesis of random feature selection across datasets varying in sample size, feature dimensionality, number of classes, and data modality. Across

many settings, models trained on all features achieve performance that is matched by small (as low as $0.02\% - 1.0\%$) random feature subsets. Consequently, attempts to reject the null hypothesis of random feature selection frequently fail. We find that in many high-dimensional biological datasets, predictive performance is largely insensitive to the specific features chosen, with low variance across random subsets. In practice, small random subsets often match, and in some cases outperform, all features or features selected by established methods.

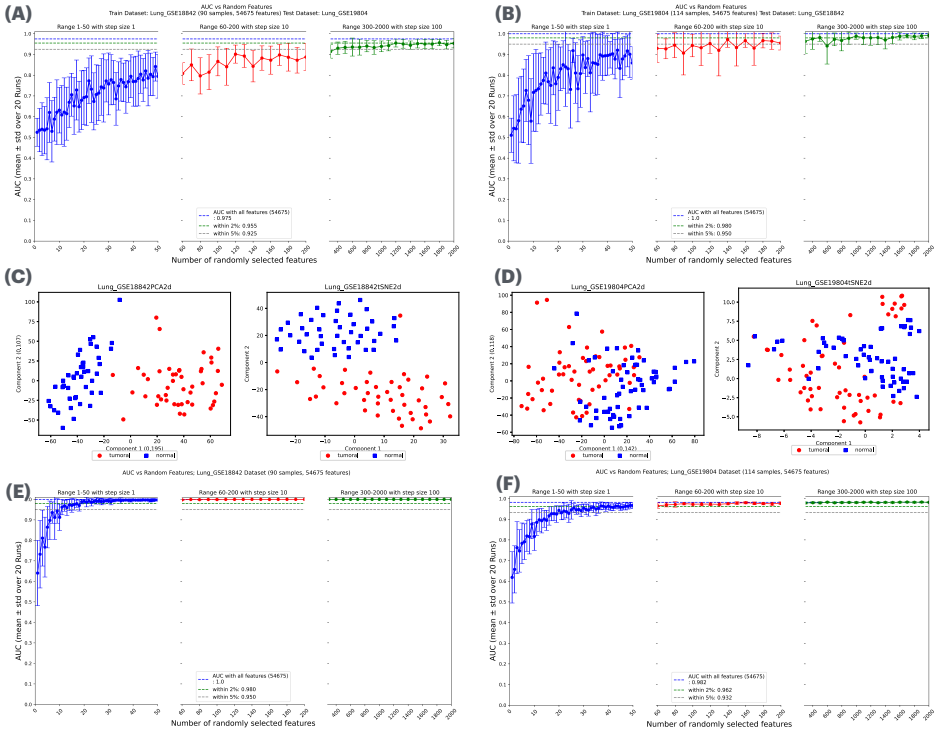


Figure 1: RF results on lung cancer **microarray** datasets (A) Training on GSE18842 and testing on GSE19804 shows that random subsets of only 400 features ($\sim$0.8% of all features) achieve AUC comparable to using all features. (B) Training on GSE19804 and testing on GSE18842 shows similar performance with just 200 ($\sim$0.4%) randomly selected features. (C–D) PCA and t-SNE visualizations illustrating class separation in GSE18842 and GSE19804, respectively. (E–F) With an 80:20 train-test split, subsets of only 20 features ($\sim$0.04%) for GSE18842 and 100 features ($\sim$0.2%) for GSE19804 match the performance of the full feature set.

Across 27 out of 30 diverse datasets that we tested (dataset inclusion criteria specified in Section 3.2), including microarray, RNA-Seq (bulk and single-cell), mass spectrometry, and imaging, we find that extremely small, random subsets of features ($0.02 - 1\%$ of all features) match or outperform the predictive performance of full feature sets. The only three cases where full feature set does better than chance are in journalistic text categorization, drug activity prediction, and synthetic data. Results for all 30 datasets are shown in Table 1.

We also find that, for a given dataset, performance (accuracy or AUC) obtained from randomly sampled feature subsets rapidly stabilizes as subset size increases, exhibiting low variance beyond a dataset-specific scale. Sampling with or without replacement seems to make little difference. To formalize this behavior, we introduce Minimum Sufficient Random Sample Size (MSRSS) (Section 3.4.2), defined as the minimum number of randomly sampled features such that increasing the subset size yields less than ϵ relative improvement in mean performance across random draws, where $\epsilon \in [0.02, 0.05]$, relative to the asymptotic performance.

In our analysis, for each dataset, we compare model performance when trained on:

1. all features.

2. randomly selected subsets of features
    - of the same size as in the published feature selection (FS) studies, for all datasets where FS results are available,
    - across various subset sizes ranging from 1 to 2000 features
3. features selected using standard feature selection methods

This emphasizes the importance of rigorous validation before interpreting selected features as biologically or functionally significant. The work carries broad implications across genomics, neuroscience, imaging, and other fields that rely on high-dimensional data for discovery and prediction.

We realize that this is a very strong claim to make [1] ; as such, all of our code and sample datasets are provided at the following anonymous GitHub link: `https://anonymous.4open.science/r/Feature_Selection_HD-D853/`. All datasets used in our experiments are publicly available and their original sources are cited in the appendix A.

This work makes the following contributions:

1. We present a large-scale empirical study across 30 high-dimensional datasets to evaluate the effectiveness of randomly selected feature subsets for classification tasks, comparing full feature set results against random subsets.
2. We find that randomly selected subsets, sometimes comprising as little as $0.02\% - 1\%$ of features, can match or exceed the performance of models trained on all features.
3. We discuss the broader implications of these findings, including discussing possible metrics (MSRSS) for evaluating datasets and FS methods, understanding variance in model performance and interpreting computational results in genomics and other high-dimensional domains.

2 Related Work

Early work on feature selection in genomics laid the foundation for much of the field. Seminal studies showed how gene expression data could support molecular cancer classification (Golub et al., 1999), how SVM-based gene selection could be applied to high-dimensional biology (Guyon et al., 2002). Gene expression profiles were also applied to predict breast cancer outcomes (West et al., 2001), underscoring the clinical promise of FS. More recent methods include manifold-preserving FS for detecting rare cell types in single-cell data (Liang et al., 2021) and Wx, a neural network–based FS algorithm for transcriptomic datasets (Park et al., 2019). Applied studies have used overlapping FS to distinguish lung cancer subtypes (Chen & Dhahbi, 2021) and ℓ_1-norm copula-based FS for microarray data (Lall & Bandyopadhyay, 2019), and pan-cancer classification of TCGA expression data (Li et al., 2017).

Additionally, comparative evaluations have sought to benchmark FS methods. For example, Guyon et al. (2004) reported the results of the NIPS 2003 FS Challenge, providing one of the earliest systematic benchmarks of FS methods across diverse datasets, Zanella et al. (2022) evaluated FS algorithms for cancer phenotype classification, and Cilia et al. (2019) provided a systematic evaluation of FS and classification methods on microarray datasets. Yet, performance has rarely been compared against a simple null baseline of randomly chosen feature subsets, leaving open a critical question of whether complex FS methods genuinely outperform chance.

Additionally, while signature multiplicity theory (Statnikov & Aliferis, 2010; Borboudakis & Tsamardinos, 2021; Statnikov et al., 2013) shows that several distinct but causally meaningful gene sets may achieve similar predictive performance, our results go even further: the accuracy of any random feature subset is comparable to selected features across many datasets. This indicates that predictive signal is highly redundant and distributed - to the point where feature selection may provide little practical advantage in high-dimensional biological classification tasks. Thus our results are stronger than presented in these papers. To the best of our knowledge, previously no one has ever shown a universal random subset is good enough - there is no cleverness involved (which is the whole point, making the FS results fail in terms of significance vs the null hypothesis.).

[1] We asked multiple other teams to independently rewrite all code and re-test everything, with similar results.

Table 1: Summary of datasets and random subset size (percentage of full features) required to match full feature performance within 5%. For 27 out of 30 datasets, randomly selected feature subsets can match or even outperform models trained on the full feature set.

S No	Name	Classes	Samples	Features	Random Subset Size (matching percentage)	Datatype	Domain
1	Colon (Alon)*	2	62	2000	100 (5%)	microarray	colon cancer
2	ALL/AML*	2	72	7129	200 (3%)	microarray	leukemia
3	GSE6008	4	98	22283	45 (0.2%)	microarray	ovarian cancer
4	GSE18842	2	90	54675	20 (0.04%)	microarray	lung cancer
5	GSE42743	2	103	54675	200 (0.4%)	microarray	oral cavity cancer
6	GSE19804	2	114	54675	100 (0.2%)	microarray	lung cancer
7	GSE6919_U95B	2	124	12620	120 (1%)	microarray	prostate cancer
8	GSE3365	3	127	22814	300 (1.5%)	microarray	bowel disease
9	GSE50161	5	130	54575	60 (0.1%)	microarray	brain tumours
10	GSE22820	2	139	33579	50 (0.15%)	microarray	breast cancer
11	GSE53757	2	143	54675	60 (0.1%)	microarray	kidney cancer
12	GSE30219	2	146	54675	10 (0.02%)	microarray	lung cancer
13	GSE21510	3	147	54675	60 (0.1%)	microarray	colorectal cancer
14	GSE45827	6	151	54675	60 (0.1%)	microarray	breast cancer
15	GSE76427	2	165	47322	100 (0.2%)	microarray	liver cancer
16	GSE4115	2	187	22215	110 (0.5%)	microarray	lung epithelial
17	GSE44076	2	194	49386	50 (0.1%)	microarray	colorectal cancer
18	GSE11223	3	202	40991	200 (0.5%)	microarray	colon inflammation
19	GSE28497	7	281	22284	110 (0.5%)	microarray	pediatric leukemia
20	GSE70947	2	289	35981	110 (0.3%)	microarray	breast cancer
21	GSE14250	2	357	22277	50 (0.2%)	microarray	liver cancer
22	TCGA (LUAD/LUSC)	2	1016	20253	50 (0.3%)	bulk RNASeq	lung cancer
23	TCGA pan-cancer	33	10223	20253	50 (0.25%)	bulk RNASeq	pan-cancer
24	Lung (scRNA-Seq)	9	20966	33514	1600 (5.0%)	scRNA-Seq	lung adenocarcinoma
25	Lung (scRNA-Seq)	2	24421	33514	1200 (3.6%)	scRNA-Seq	lung adenocarcinoma
26	Arcene†	2	200	10000	100 (1.0%)	mass-spectrometry	ovarian vs prostate cancer
27	Dexter†	2	600	20000	**20000 (100%)**	text	corporate acquisition
28	Dorothea†	2	1150	100000	**100000 (100%)**	fingerprint	drug activity prediction
29	Madelon†*	2	2600	500	**200 (40.0%)**	synthetic	cluster classification
30	Gisette†*	2	7000	5000	90 (1.8%)	image	digit (4 vs 9)

† NIPS 2003 FS Challenge datasets; all five datasets contain random probes, ranging from 30% (Arcene) $\sim$ 50% (Dexter, Gisette, Dorothea) to 96% (Madelon).

* indicates subsets sampled with replacement (datasets with $< 20{,}000$ features).

3 METHOD

3.1 DATASETS

We conducted our experiments on a diverse set of 30 high-dimensional datasets, spanning multiple data modalities: 21 microarray gene expression, 4 RNA-Seq (bulk and single-cell), 1 mass spectrometry, 2 image, and 2 other data types(Table 1, sorted by sample size, with NIPS 2003 FS challenge datasets listed at the end (Guyon et al., 2004)). We intentionally chose this heterogeneous collection of datasets spanning multiple cancer types and molecular profiling platforms, varying widely in tissue origin, sample size, feature dimensionality, datatypes and class distribution. In addition, we include five benchmark datasets from the NIPS feature selection challenge, covering domains such as cancer prediction via mass-spectrometry data, handwritten digit recognition, text classification, and molecular activity prediction along with one synthetically generated dataset. Notably, all five of these challenge datasets (marked with † in Table 1) were constructed with random probe features intentionally added as distractors, making them especially relevant for evaluating the robustness of feature selection methods. Most of the microarray dataset are sourced from Feltes et al. (2019), which curated them specifically for ML model training. All the datasets are publicly available. None were imputed before download, and we did not apply an imputation. For all the datasets, the original sources are cited in Table 3 in the Appendix A.

3.2 DATASET INCLUSION CRITERIA

We include datasets only if they satisfy the following:

1. at least 100 samples (three canonical FS benchmarks, ALL/AML, Colon and Madelon are included despite slightly fewer samples due to their widespread use),
2. at least 2,000 features,
3. prior use in feature-selection studies. We have also included datasets from NIPS 2003 FS challenge (Guyon et al., 2004).

3.3 MODELS

We evaluate a representative set of models covering major learning paradigms: linear (Logistic Regression (LR), Ridge), margin-based (Support Vector Machine (SVM)), Decision Tree (DT), ensemble trees (Random Forest (RF), Gradient Boosting Classifier (GBM), HistGradient Boosting Classifier (HistGB), eXtreme Gradient Boosting (XGB)), and neural networks (Multilayer Perceptron (MLP)). This selection balances simple baselines with non-linear models, and overlaps with models used in prior work, enabling direct comparison. We exclude Naive Bayes, SGD and KNN, which perform poorly or scale unfavorably in high-dimensional settings, and omit additional boosting variants (LighGBM, CatBoost) as redundant given XGB, GBM, and HistGB. Our goal is representative, not exhaustive, coverage.

3.4 EXPERIMENTAL SETUP

All datasets were split into train/test subsets using an 80/20 stratified split. For non-tree-based models, the features were standardized priot to feature selection and model training. For each dataset, we compared model performance using

1. all features
2. random feature subsets (baseline) as described in next section
3. embedded (lasso, elastic net) and ensemble-based (random forest importance) feature selection methods
4. reported results from prior published feature selection studies, where available

3.4.1 RANDOM FEATURE SUBSETS

We sampled features subsets of various sizes $(1, 2, \ldots, 50; 60, 70, \ldots, 200; 300, 400, \ldots, 2000)$, resulting in a total of 83 distinct subset sizes. When the total number of features exceeded $20{,}000$,

subsets were drawn without replacement; otherwise, sampling was with replacement. For each subset size, we generated 20 independent subsets by random sampling and repeated the full training–evaluation procedure on each. We then trained and evaluated a classifier on every subset, reporting mean accuracy and AUC across the 20 runs. The plots in figures 1, 2, 3 display these results, with error bars indicating the standard deviation across runs.

3.4.2 Evaluation Metrics

We present our results with Area Under the ROC Curve (AUC-ROC) and Accuracy as the primary metrics. For random subsets, we report the mean performance over 20 runs. For each subset size, standard deviation is shown with an error bar. Each plot has three horizontal reference lines: for AUC (or Accuracy) with all features; for within-2%; and within-5% AUC (or Accuracy).

One interesting metric to explore would be the minimum subset size of randomly-selected features which contain enough information to perform as well at the task as any larger subset - that is, the accuracy becomes asymptotic after that point, indicating a "saturation" of information provided by these random subsets. Interestingly, we notice that this is model-independent, remarkably low in performance variance (across different random selections), and almost always matching full-feature set performance (after all, this is the random subset size with maximum cardinality). We call the random subset feature size at which this happens the *Minimum Sufficient Random Subset Size (MSRSS)*; graphically, this captures the "elbow" at which the information provided by random subsets saturates, resulting in asymptotic, flat performance thereon. We believe that exploring this metric could provide interesting insights into various datasets, especially in computational genomics.

4 Results

We begin by showing Random Forest (RF) results on one representative dataset (or pair, if different datasets exist for train and test) from each type—microarray, bulk RNA-Seq, single-cell RNA-Seq, and imaging—chosen to illustrate breadth. Across 27 out of 30 datasets tested, we find that very small, randomly selected subsets of features ($0.02 - 5\%$ of all features) match or even outperform the predictive performance of full feature set. A full summary appears in table 1, with complete plots for all datasets using Random Forest in appendix B (and results using other models like Decision trees, Support Vector Machine, Logistic Regression etc. are in appendix C).

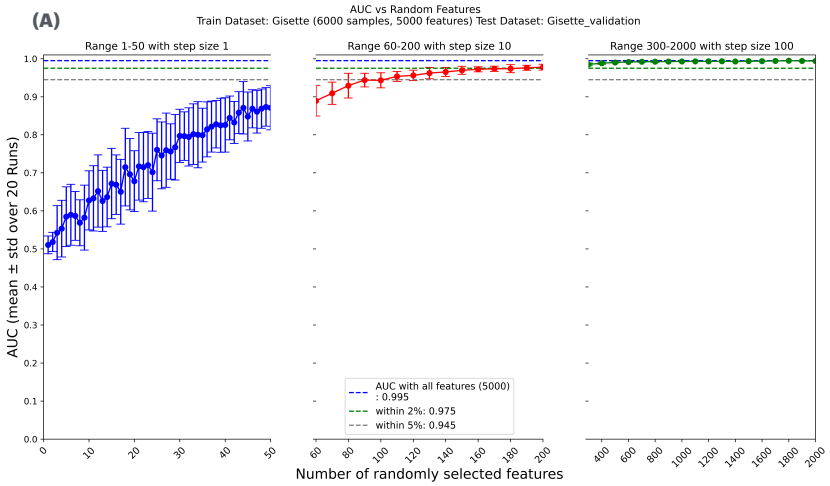


Figure 2: RF results on Gisette **image** dataset. (A) Randomly selected subsets of only 90 features (1.8% of all features) achieve AUC comparable to using the full feature set when training on the Gisette training data and evaluating on the validation set.

Table 2: Comparison with published studies. Accuracy (%) with number of features in parentheses.

Publication	Dataset (samples)	Published FS	Random FS baseline	All features	Standard FS baselines			Remarks
					Lasso	Elastic Net	RF-Imp.	
Chen & Dhahbi (2021)	TCGA (LUAD/LUSC) (1016)	94.2 (500)	93.6 (500)	94.8 (20253)	95.48	95.28	95.18	See Fig. 6 (Appendix B)
Lall & Bandyopadhyay (2020)	ALL/AML (72)	86.0 (35)	80.0 (35)	94.3 (7129)	95.81	94.38	98.57	See Fig. 8 (Appendix B)
Golub (1999); Cilia (2019)	ALL/AML (72)	98.6–99.4 (51)	86.0 (51)	94.3 (7129)	95.71	93.05	97.14	Despite newer FS methods, Golub (1999) performance remains competitive.
Zanella et al. (2022)	GSE4115 (187)	67.9 (5)	57.9 (5)	68.9 (22215)	62.11	66.27	67.31	See Fig. 20 (Appendix B)

Legend. Published FS: results reported in the original study using feature selection. Random FS: random subsets with matched feature counts. All features: random forest trained on the full feature set. Lasso, Elastic Net, RF-Imp.: standard embedded or importance-based feature selection baselines with matched feature counts.

4.1 Random Forest results with cross-dataset evaluation

We start with two independent microarray datasets of lung disease, each measured on an identical set of features. To assess model performance under a strict train-test separation, one dataset was designated as the training set and the other as the external test set. This design avoids overfitting to dataset-specific noise and provides a rigorous test of generalization. Model performance was quantified primarily by the Area Under the ROC Curve (AUC); classification accuracy is reported in appendix B. To ensure robustness and rule out dataset-specific biases, we repeated the experiment after swapping the roles of the two datasets (i.e., training on the second dataset and testing on the first), and we report results for both directions.

To illustrate, we first train a model using GSE18842 (Feltes et al., 2019) (90 samples, 54,765 features) and test it using GSE19804 (Feltes et al., 2019)(114 samples, 54,765 features). As shown in figure 1(A), randomly selected subsets of just 200 features (0.4% of all features), selected without replacement, achieve AUC comparable to using all features. figure 1(B) shows the reverse setting, where the model is trained on GSE19804 and tested on GSE18842. Both plots show similar results. The 2D projections of these two datasets using PCA and t-SNE show that GSE18842 exhibits better class separability than GSE19804. We hypothesize that such low-dimensional separability enables models trained on small random subsets of features to perform well. Accuracy plots for this dataset pair show similar trends in the appendix B: even very small random subsets can match the performance of models using all features.

We also show results for non-disease dataset, Gisette, with cross-dataset evaluation shown in Figure 2. The Gisette image dataset is from the NIPS 2003 FS challenge (Guyon et al., 2004). The task is to discriminate between two confusable handwritten digits: the four and the nine. The challenge organisers provided a separate training and validation set. We find that with 90 (1.8%) randomly selected features, the models could match the AUC with all features. As there are 30% spurious features in this dataset, the results are even more significant – effectively, any random subset of size 60 is enough to discriminate between the two classes in this dataset.

4.2 Random Forest results with intra-dataset splits

In addition to cross-dataset evaluation, we assessed model performance using standard intra-dataset train–test splits of 80:20. For each dataset, we randomly divided the samples into 80% for training and 20% for testing, while ensuring class balance was preserved. For example, figures 1(E) and 1(F) show that for GSE19804 and GSE18842, with a random subset we can match the performance of all features.

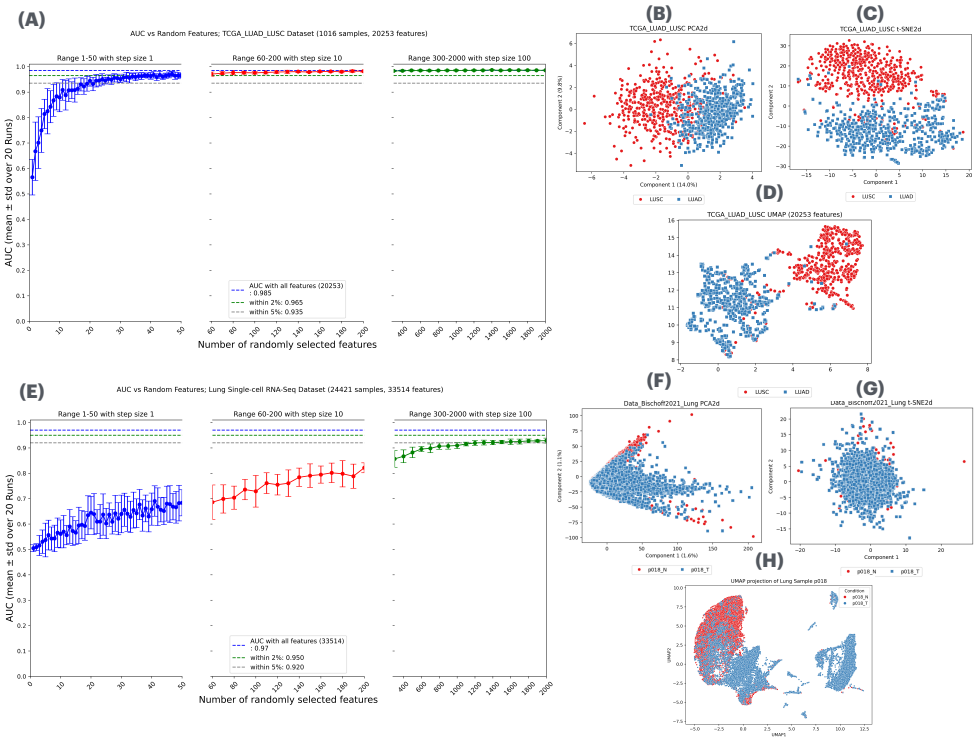


Figure 3: RF results on **bulk and single-cell RNA-Seq** datasets. (A) On the TCGA-LUAD-LUSC bulk RNA-Seq dataset (80:20 split), a random subset of only 50 features($\sim$0.3% of the total) achieves AUC comparable to the full feature set. (B–D) PCA, t-SNE and UMAP visualizations illustrating class separation in the bulk dataset. (E) On the lung cancer single-cell RNA-Seq dataset (80:20 split), random subsets of 1200 features ($\sim$3.6%) achieve AUC within 5% of the full-feature set. (F–H) PCA, t-SNE, and UMAP visualizations showing class separation in the single-cell dataset.

Figure 3 summarizes results for bulk and Single-cell RNA-Seq datasets. For the bulk RNA-Seq data, (Figure 3(A)), Random Forest models trained with just 50 randomly selected features ($\sim$0.22% of all features) achieves performance comparable to the full feature set. For the Single-cell RNA-Seq dataset(Figure 3(E)), subsets of about 1200 ($\sim$3.6% of all features) results in AUC within 5% of the full-feature set. We hypothesize that better low-dimensional separability for TCGA_LUAD_LUSC (bulk RNA-Seq dataset) enables random subsets to perform better. Even for a bulk RNA-Seq dataset with 33 classes (TCGA), random subsets can match the performance of all features as shown in figure 26 in appendix B. **Note:** We also report sparsity via three metrics: (i) global fraction of zeros, (ii) median zero fraction per feature, and (iii) median zero fraction per sample. The scRNA-Seq datasets are extremely sparse ($> 96\%$ global zeros), while bulk rna-seq shows near-zero sparsity. Crucially, random feature subsets still match full-model performance even in scRNA-Seq with $> 96\%$ zeros, indicating the effect is NOT driven by trivially selecting nonzero columns. See Table 4.

We observe similar patterns in three of the five benchmark datasets from the NIPS 2003 Feature Selection Challenge (Guyon et al., 2004). Consistent with the cross-dataset results, we found that models trained on extremely small subsets of randomly selected features often outperformed those trained on the full feature set. In several cases, using just 0.02% - 0.1% of the available features led to improved accuracy and AUC, with performance saturating well before reaching the full dimensionality. Table 1 summarises the results across all 30 datasets, showing that for 28 datasets this pattern holds consistently across diverse data types, including microarray, single-cell RNA-seq, bulk RNA-Seq and benchmarks datasets from NIPS FS challenge.

The only three cases where full feature set performs better than chance are in the categorization of the journalistic text, the prediction of drug activity, and synthetic data. The results for these datasets (Dexter, Dorothea, and Madelon) are shown in Figures 28, 29, and 30.

4.3 Comparison with published studies on FS and standard FS methods

Table 2 provides a comparison of RF results for selected datasets and related published studies. The first column lists the published study followed by the name of the dataset and sample count. The columns list accuracy with (A) selected features from the published study; (B) randomly selected features of the same size as the published study; (C) all features (D-F) features selected using standard FS methods. All results are obtained using random forests trained on the corresponding feature subset. We note that published FS methods do not consistently outperform standard embedded FS baselines, and are often matched or exceeded by them.

5 Conclusion and Future Work

Our findings challenge the conventional, intuitive assumption that more features, and cleverly selected features lead to better classification performance in high-dimensional settings. Across 27 out of 30 diverse datasets, we observe that small, randomly selected feature subsets—sometimes comprising as little as 0.02% of all features—can match or even outperform models trained on the full feature set or on FS sets. This result holds across both cross-dataset and intra-dataset validation and spans multiple data types.

These results reinforce earlier insights on feature redundancy in high-dimensional data and resonate with prior work showing that Random Forests are robust to noise and overfitting due to their ensemble nature and internal feature sampling mechanisms (Breiman, 2001; Díaz-Uriarte & Alvarez de Andrés, 2006). Interestingly, we observe similar results with other non-tree, non-ensemble models. Our findings further demonstrate that even explicit external subsampling of features—performed entirely at random—can yield stable and often superior classification performance. This robustness is reinforced by our observation that the standard deviation in performance decreases consistently with increasing random subset size, suggesting the presence of many equally informative, often non-overlapping, feature combinations.

This phenomenon may have conceptual ties to random subspace methods (Ho, 1998) and the theory of random projections, particularly the Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984), which shows that high-dimensional data can be embedded in lower dimensions while preserving pairwise distances. While our method does not perform explicit projections, the empirical success of randomly selected subspaces suggests that useful structure can be retained without sophisticated transformations. Unlike methods such as Principal Component Analysis (PCA), which apply global, often opaque transformations, random feature selection is both conceptually and computationally simple: the only surprising thing is that it works so well!

Our results also raise critical questions regarding the interpretation of feature importance in gene expression datasets – especially because such results sometimes guide downstream medical research. Our findings suggest that computationally derived feature importance may reflect statistical signals more than underlying biological causality. We do not argue against the search for biologically meaningful genes; on the contrary, we emphasize that identifying causal or mechanistically relevant genes requires biological validation. Features identified as important by computational models should ideally be corroborated through independent experimental methods, such as perturbation assays or wet-lab validation.

For future work, investigating the diversity and overlap among high-performing random subsets may be interesting and reveal deeper insights into the intrinsic dimensionality, redundancy, and structure of high-dimensional biological data. Incorporating random subspace strategies into ensemble learning or active learning pipelines could enhance both performance and generalization, especially in domains like genomics where data is high-dimensional and sample sizes are often limited. Additionally, our MSRSS metric appears to have some interesting properties (e.g., it remains remarkably stable across totally different models - the reasons for this are unclear). It will be interesting to investigate this further and possibly quantify its effect on feature selection.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

REFERENCES

Uri Alon, Naama Barkai, Daniel A. Notterman, Kinneret Gish, Steven Ybarra, David Mack, and Arnold J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, June 1999. doi: 10.1073/pnas.96.12.6745.

Philipp Bischoff, Andreas Trinks, Benedikt Obermayer, Jonathan P. Pett, Jule Wiederspahn, Florian Uhlitz, Xiaoqing Liang, Alexander Lehmann, Philipp Jurmeister, Anna Elsner, Tomasz Dziodzio, Jens-Christian Rückert, Jens Neudecker, Catharina Falk, Dirk Beule, Christine Sers, Markus Morkel, Dieter Horst, Nils Blüthgen, and Frederick Klauschen. Single-cell rna sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. *Oncogene*, 40(50): 6748–6758, December 2021. doi: 10.1038/s41388-021-02054-3.

Giorgos Borboudakis and Ioannis Tsamardinos. Extending greedy feature selection algorithms to multiple solutions. *Data Mining and Knowledge Discovery*, 35(4):1393–1434, 2021. doi: 10.1007/s10618-020-00731-7. URL https://doi.org/10.1007/s10618-020-00731-7.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Michael E. Burczynski, Ronald L. Peterson, Nancy C. Twine, Karen A. Zuberek, Brigitte J. Brodeur, Laura Casciotti, Venkata Maganti, Pallavi S. Reddy, Arlen Strahs, Frederick Immermann, Walter Spinelli, Udo Schwertschlag, Anne M. Slager, Matthew R. Barnes, Peter Goldschmidt, and Art Dorner. Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *Journal of Molecular Diagnostics*, 8(1): 51–61, February 2006. doi: 10.2353/jmoldx.2006.050079.

Urmila R. Chandran, Chien Ma, Rajiv Dhir, Michele Bisceglia, James Lyons-Weiler, Wei Liang, George Michalopoulos, Michael Becich, and Federico A. Monzon. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer*, 7:64, April 2007. doi: 10.1186/1471-2407-7-64.

J.W. Chen and J. Dhahbi. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Scientific Reports*, 11:13323, 2021. doi: 10.1038/s41598-021-92725-8.

N.D. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, and A. Scotto di Freca. An experimental comparison of feature-selection and classification methods for microarray datasets. *Information*, 10(3):109, 2019. doi: 10.3390/info10030109.

Elaine Coustan-Smith, Guolian Song, Carolyn Clark, Laura Key, Peng Liu, Mahin Mehrpooya, Paul Stow, Xiangning Su, Sheila Shurtleff, Ching-Hon Pui, James R. Downing, Susana C. Raimondi, Frank G. Behm, and Dario Campana. New markers for minimal residual disease detection in acute lymphoblastic leukemia. *Blood*, 117(23):6267–6276, June 2011. doi: 10.1182/blood-2010-12-324004.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006. doi: 10.1186/1471-2105-7-3.

B.C. Feltes, E.B. Chandelier, B.I. Grisci, and M. Dorn. Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, 26(4):376–386, 2019. doi: 10.1089/cmb.2018.0238.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439): 531–537, 1999. doi: 10.1126/science.286.5439.531.

Anne M. Griesinger, David K. Birks, Andrew M. Donson, Vanessa Amani, Laura M. Hoffman, Allen Waziri, Mei Wang, Michael H. Handler, and Nicholas K. Foreman. Characterization of distinct immunophenotypes across pediatric brain tumor types. *Journal of Immunology*, 191(9): 4880–4888, November 2013. doi: 10.4049/jimmunol.1301800.

Oleg V. Grinchuk, Sainath P. Yenamandra, Ramesh Iyer, Manoj Singh, Hwee Hoon Lee, Kiat Hon Lim, Balram Chowbay, and Vladimir A. Kuznetsov. Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. *Molecular Oncology*, 12(1):89–113, January 2018. doi: 10.1002/1878-0261.12148.

Tiffany Gruosso, Vincent Mieulet, Michel Cardon, Beatrice Bourachot, Yann Kieffer, Frederique Devun, Thibaut Dubois, Marie Dutreix, Anne Vincent-Salomon, William H. Miller, and Fatima Mechta-Grigoriou. Chronic oxidative stress promotes h2ax protein degradation and enhances chemosensitivity in breast cancer patients. *EMBO Molecular Medicine*, 8(5):527–549, May 2016. doi: 10.15252/emmm.201505891.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002. doi: 10.1023/A:1012487302797.

I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS)*, pp. 545–552, 2004. URL https://dl.acm.org/doi/10.5555/2976040.2976109.

Natalie D. Hendrix, Rui Wu, Robert Kuick, David R. Schwartz, Eric R. Fearon, and Kathleen R. Cho. Fibroblast growth factor 9 has oncogenic activity and is a downstream target of wnt signaling in ovarian endometrioid adenocarcinomas. *Cancer Research*, 66(3):1354–1362, February 2006. doi: 10.1158/0008-5472.CAN-05-3694.

T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. doi: 10.1109/34.709601.

W.B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. doi: 10.1007/BF02764938.

S. Lall and S. Bandyopadhyay. An l1-norm regularized copula-based feature selection. In *Proceedings of the 2019 International Symposium on Computer Science and Intelligent Control (ISCSIC)*, pp. 30, 2019. doi: 10.1145/3386164.3386177.

Y. Li, K. Kang, J.M. Krahn, N. Croutwater, K. Lee, D.M. Umbach, and L. Li. A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC Genomics*, 18:508, 2017. doi: 10.1186/s12864-017-3906-0.

Shaoheng Liang, Vakul Mohanty, Jinzhuang Dou, Qi Miao, Yuefan Huang, Muharrem Müftüoğlu, Li Ding, Weiyi Peng, and Ken Chen. Single-cell manifold-preserving feature selection for detecting rare cell populations. *Nature Computational Science*, 1(5):374–384, 2021. doi: 10.1038/s43588-021-00070-7.

Rui Z. Liu, Krista Graham, Diedre D. Glubrecht, David R. Germain, John R. Mackey, and Roseline Godbout. Association of fabp5 expression with poor survival in triple-negative breast cancer: implication for retinoic acid therapy. *American Journal of Pathology*, 178(3):999–1008, March 2011. doi: 10.1016/j.ajpath.2010.11.067.

P. Lohavanichbutr, E. Méndez, F. C. Holsinger, T. C. Rue, Y. Zhang, P. F. Nguygen-Tan, P. Wang, C. Chen, D. H. Rice, E. L. Rosenthal, S. M. Schwartz, L. P. Zhao, and C. Chen. A 13-gene signature prognostic of hpv-negative oscc: discovery and external validation. *Clinical Cancer Research*, 19(5):1197–1203, March 2013. doi: 10.1158/1078-0432.CCR-12-2647.

Tzu-Pin Lu, Ming-Hsien Tsai, Jung-Mao Lee, Chih-Pin Hsu, Peng-Chan Chen, Chia-Wen Lin, Jui-Yen Shih, Pan-Chyr Yang, Ching-Kang Hsiao, Liang-Chuan Lai, and Eric Y. Chuang. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiology, Biomarkers & Prevention*, 19(10):2590–2597, October 2010. doi: 10.1158/1055-9965.EPI-10-0365.

Christopher L. Noble, Ali R. Abbas, Julie Cornelius, Charlie W. Lees, Gordon T. Ho, Katherine Toy, Zora Modrusan, Harry F. Clark, Ian D. Arnott, Ian D. Penman, Jack Satsangi, Lutz Diehl, Anthony Fong, David A. van Heel, Kevin G. Smith, Daniel Gaffney, Charles A. Anderson, Daniel Massey, Claire M. Lewis, and et al. Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut*, 57(10):1398–1405, October 2008. doi: 10.1136/gut.2007.143198.

Sungsoo Park, Bonggun Shin, Won Sang Shim, Yoonjung Choi, Kilsoo Kang, et al. Wx: a neural network-based feature selection algorithm for transcriptomic data. *Scientific Reports*, 9:10500, 2019. doi: 10.1038/s41598-019-47016-8.

David A. Quigley and Vessela Kristensen. Gene expression profiling of breast cancer. NCBI Gene Expression Omnibus, GSE70947, 2016. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70947`.

Cheryl A. Roemeling, Derek C. Radisky, Laura A. Marlow, Sarah J. Cooper, Douglas L. Gustafson, Evan P. Stoffregen, Xin Wang, Ke Wu, Amber M. Button, Stacie B. Terra, Jing Yang, Tadashi Kobayashi, Hongkai Ji, Lifeng Wang, Yinhua Wang, Kimberly D. Brown, Roger B. Jenkins, and John A. Copland. Neuronal pentraxin 2 supports clear cell renal cell carcinoma by activating the AMPA-selective glutamate receptor-4. *Cancer Research*, 74(17):4796–4810, September 2014. doi: 10.1158/0008-5472.CAN-14-0107.

Sebastian Roessler, Hai-Ling Jia, Anuradha Budhu, Marc Forgues, Qidong Ye, Jeong-Ho Lee, Snorri S. Thorgeirsson, Zheng Sun, Zhi-Yuan Tang, Li-Xu Qin, and Xinhui W. Wang. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Research*, 70(24):10202–10212, December 2010. doi: 10.1158/0008-5472.CAN-10-2607.

Sophie Rousseaux, Anne Debernardi, Benjamin Jacquiau, Anne-Laure Vitte, Aurélie Vesin, Hélène Nagy-Mignotte, Denis Moro-Sibilot, Pierre-Yves Brichon, Sylvie Lantuejoul, Pierre Hainaut, Elisabeth Brambilla, Elizabeth Myers, Robert J. Hung, James D. McKay, Wan L. Lam, John D. Minna, Veronique Bichsel, Daniel Chubb, Laura Baglietto, Jean Cadet, and et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Science Translational Medicine*, 5(186):186ra66, May 2013. doi: 10.1126/scitranslmed.3005723.

A. Sánchez-Palencia, M. Gómez-Morales, J. A. Gómez-Capilla, V. Pedraza, L. Boyero, R. Rosell, I. Fábregat, A. Carnero, M. Nistal, J. F. García, B. Massuti, A. Blasco, E. Conde, F. Lopez-Rios, L. Paz-Ares, and M. Hidalgo. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer*, 129(2):355–364, July 2011. doi: 10.1002/ijc.25664.

Xavier Solé, Marta Crous-Bou, David Cordero, David Olivares, Elisabet Guinó, Fátima Sánchez-Cabo, Montserrat Pérez-Salvia, Florenci Mateo, Ramon Salazar, Anna Villanueva, Gabriel Capellá, and Víctor Moreno. Discovery and validation of new potential biomarkers for early detection of colon cancer. *PLoS ONE*, 9(9):e106748, 2014. doi: 10.1371/journal.pone.0106748.

Avrum Spira, Jennifer E. Beane, Vikas Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sara Gilman, Yohan-Maura Dumas, Paul Calner, Paola Sebastiani, Sanjay Sridhar, John Beamis, Christopher Lamb, Tracey Anderson, Niall Gerry, John Keane, and Marc E. Lenburg. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13(3):361–366, March 2007. doi: 10.1038/nm1556.

Alexander Statnikov and Constantin F. Aliferis. Analysis and computational dissection of molecular signature multiplicity. *PLOS Computational Biology*, 6(5):1–9, 05 2010. doi: 10.1371/journal.pcbi.1000790. URL https://doi.org/10.1371/journal.pcbi.1000790.

Alexander Statnikov, Nikita I. Lytkin, Jan Lemeire, and Constantin F. Aliferis. Algorithms for discovery of multiple markov boundaries. *Journal of Machine Learning Research*, 14(15):499–566, 2013. URL http://jmlr.org/papers/v14/statnikov13a.html.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012. doi: 10.1038/nature11404.

The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014. doi: 10.1038/nature13385.

The Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Cell*, 173(2):291–304.e10, April 2018. doi: 10.1016/j.cell.2018.03.035.

Shingo Tsukamoto, Takuya Ishikawa, Shigeyuki Iida, Masato Ishiguro, Keishi Mogushi, Hidetoshi Mizushima, Hideki Tanaka, Hiroyuki Uetake, Kenichi Sugihara, Akira Mizutani, Yoshihiko Tanaka, Hitoshi Nakagama, and Takehiko Watanabe. Clinical significance of osteoprotegerin expression in human colorectal cancer. *Clinical Cancer Research*, 17(8):2444–2450, April 2011. doi: 10.1158/1078-0432.CCR-10-2602.

M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, J.R. Marks, and J.R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467, 2001. doi: 10.1073/pnas.201162998.

L. Zanella, P. Facco, F. Bezzo, and E. Cimetta. Feature selection and molecular classification of cancer phenotypes: a comparative study. *International Journal of Molecular Sciences*, 23(16): 9087, 2022. doi: 10.3390/ijms23169087.

A DATASETS SOURCE

Table 3: Summary of datasets used in this study. Citations correspond to the original dataset sources.

No.	Name	(Samples, Features, Classes)	Datatype	Domain (Citation)
1	Colon (Alon)	(62, 2000, 2)	microarray	colon cancer (Alon et al., 1999)
2	ALL/AML	(72, 7129, 2)	microarray	leukemia (Golub et al., 1999)
3	GSE6008	(98, 22283, 4)	microarray	ovarian cancer (Hendrix et al., 2006)
4	GSE18842	(90, 54675, 2)	microarray	lung cancer (Sánchez-Palencia et al., 2011)
5	GSE42743	(103, 54675, 2)	microarray	oral cavity cancer (Lohavanichbutr et al., 2013)
6	GSE19804	(114, 54675, 2)	microarray	lung cancer (Lu et al., 2010)
7	GSE6919_U95B	(124, 12620, 2)	microarray	prostate cancer (Chandran et al., 2007)
8	GSE3365	(127, 22814, 3)	microarray	bowel disease (Burczynski et al., 2006)
9	GSE50161	(130, 54575, 5)	microarray	brain tumours (Griesinger et al., 2013)
10	GSE22820	(139, 33579, 2)	microarray	breast cancer (Liu et al., 2011)
11	GSE53757	(143, 54675, 2)	microarray	kidney cancer (Roemeling et al., 2014)
12	GSE30219	(146, 54675, 2)	microarray	lung cancer (Rousseaux et al., 2013)
13	GSE21510	(147, 54675, 3)	microarray	colorectal cancer (Tsukamoto et al., 2011)
14	GSE45827	(151, 54675, 6)	microarray	breast cancer (Gruosso et al., 2016)
15	GSE76427	(165, 47322, 2)	microarray	liver cancer (Grinchuk et al., 2018)
16	GSE4115	(187, 22215, 2)	microarray	lung epithelial cancer (Spira et al., 2007)
17	GSE44076	(194, 49386, 2)	microarray	colorectal cancer (Solé et al., 2014)
18	GSE11223	(202, 40991, 3)	microarray	colon inflammation (Noble et al., 2008)
19	GSE28497	(281, 22284, 7)	microarray	pediatric leukemia (Coustan-Smith et al., 2011)
20	GSE70947	(289, 35981, 2)	microarray	breast cancer (Quigley & Kristensen, 2016)
21	GSE14250	(357, 22277, 2)	microarray	hepatocellular carcinoma (Roessler et al., 2010)
22	TCGA (LUAD/LUSC)	(1016, 20253, 2)	bulk RNA-Seq	lung cancer (The Cancer Genome Atlas Research Network, 2014; 2012)
23	TCGA pan-cancer	(10223, 20253, 33)	bulk RNA-Seq	pan-cancer (The Cancer Genome Atlas Research Network, 2018)
24	Lung (scRNA-Seq)	(20966, 33514, 9)	scRNA-Seq	lung adenocarcinoma (Bischoff et al., 2021)
25	Lung (scRNA-Seq)	(24421, 33514, 2)	scRNA-Seq	lung adenocarcinoma (Bischoff et al., 2021)
26	Arcene†	(200, 10000, 2)	mass-spectrometry	ovarian vs prostate cancer (Guyon et al., 2004)
27	Dexter†	(600, 20000, 2)	text	corporate acquisition (Guyon et al., 2004)
28	Dorothea†	(1150, 100000, 2)	fingerprint	drug activity prediction (Guyon et al., 2004)
29	Madelon†	(2600, 500, 2)	synthetic	cluster classification (Guyon et al., 2004)
30	Gisette†	(7000, 5000, 2)	image	digit classification (4 vs 9) (Guyon et al., 2004)

Table 4: Dataset statistics and sparsity metrics

dataset	samples	features	global zero frac	median feat zero frac	median sample zero frac
Bischoff 2021 (Lung cancer sc-RNA-Seq)	21052	33514	96.4%	99.9%	96.5%
TCGA Pan-cancer (bulk RNA-Seq)	10223	20253	0.0%	0.0%	0.0%

B RESULTS WITH RANDOM FOREST FOR ALL DATASETS

Additional figures supporting the main text are provided here (figs. 4–34). They include RF results across datasets.

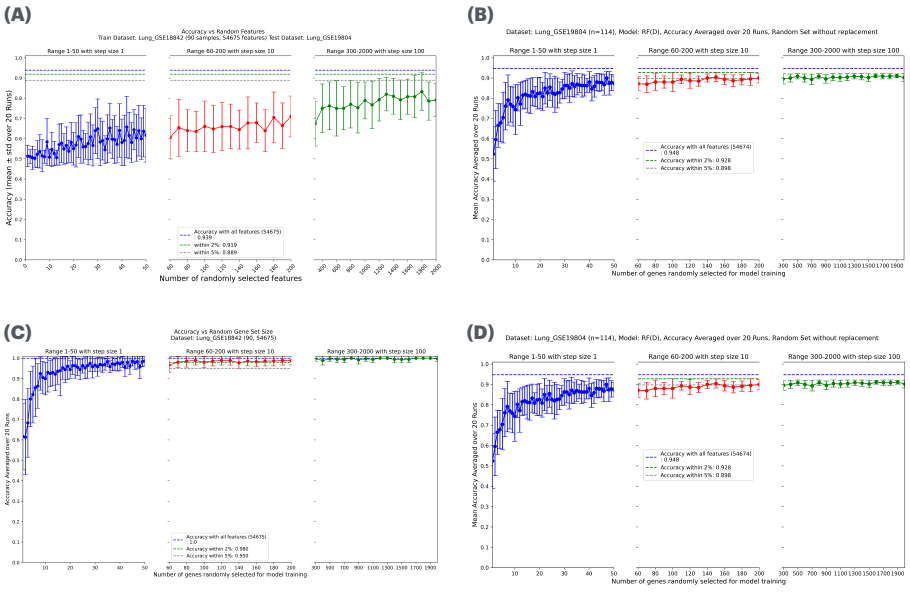


Figure 4: Random Forest performance with lung microarray dataset pairs (mean and standard deviation are reported over 20 runs) (A) RF models trained on GSE18842 and tested on GSE19804 show that randomly selected subsets never achieve accuracy comparable to using all features. (B) RF models trained on 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. (D) Similarly, for GSE19804, 200 features suffice to match accuracy with all features.

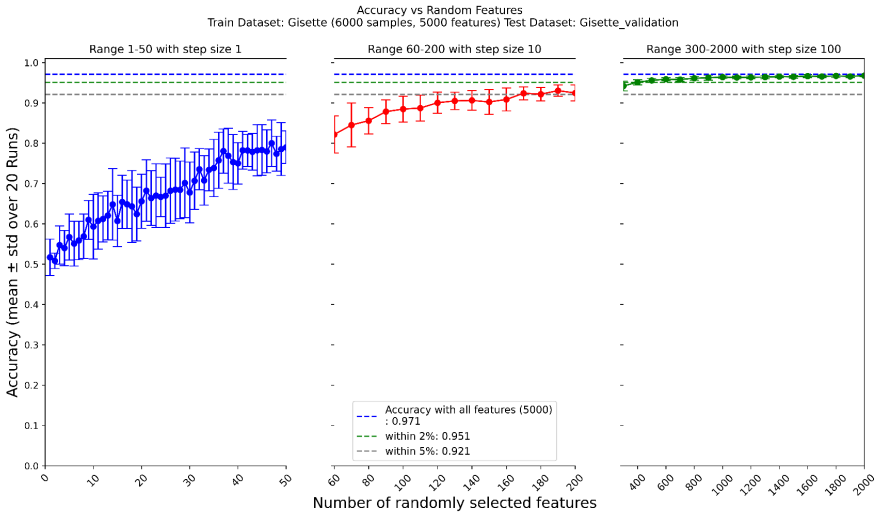


Figure 5: Random Forest performance with Gisette image dataset (mean and standard deviation are reported over 20 runs). The task of GISETTE is to discriminate between two confusable handwritten digits: the four and the nine. The model is trained on Gisette train dataset and tested on Gisette validation dataset shows that randomly selected subsets of just 200 achieve accuracy comparable to using all features.

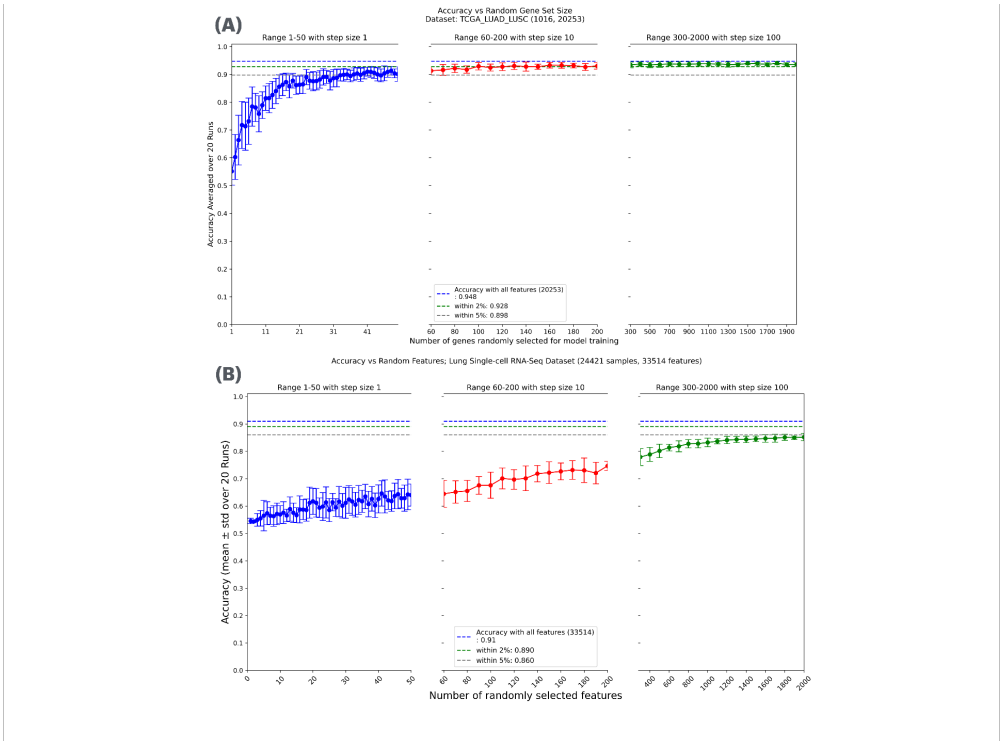


Figure 6: Random Forest performance with bulk RNA-Seq and Single-cell RNA-Seq datasets (mean and standard deviation are reported over 20 runs). (A) Models trained and tested on TCGA-LUAD-LUSC bulk RNA-Seq dataset (80:20 split) shows that a random subset of size 50 (<0.3%) is able to match within-5% accuracy 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.

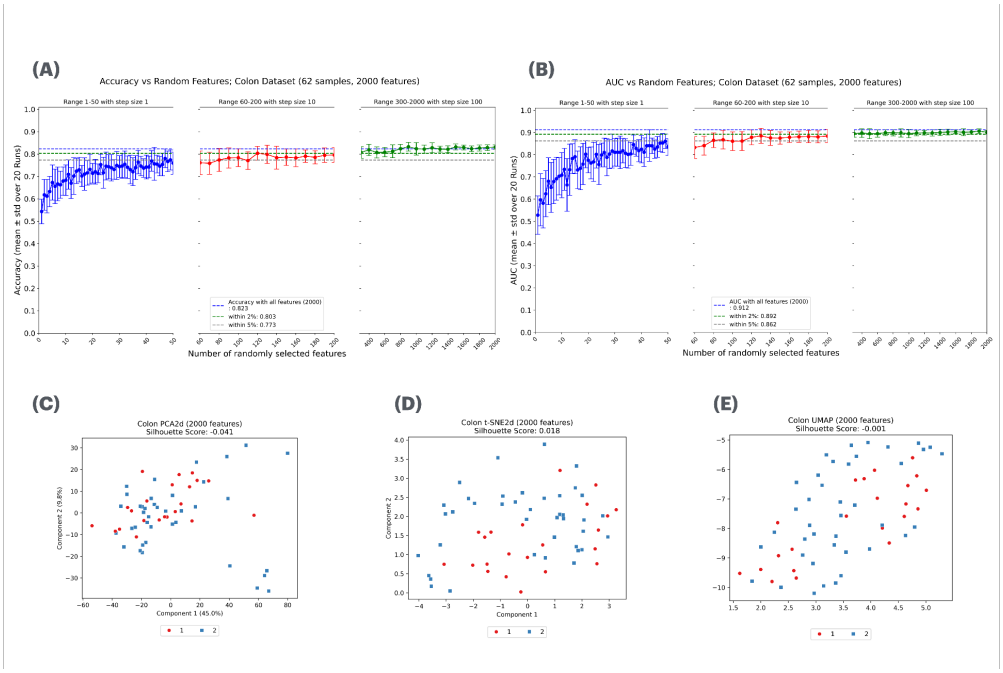


Figure 7: Random Forest performance with Colon microarray dataset (mean and standard deviation are reported over 20 runs). (A) & (B) models trained and tested on 80:20 split shows that a random subset of size $\sim$100 is able to match accuracy and AUC with all features, respectively. (C) & (D) & (E) PCA, t-SNE and UMAP plots showing class separation.

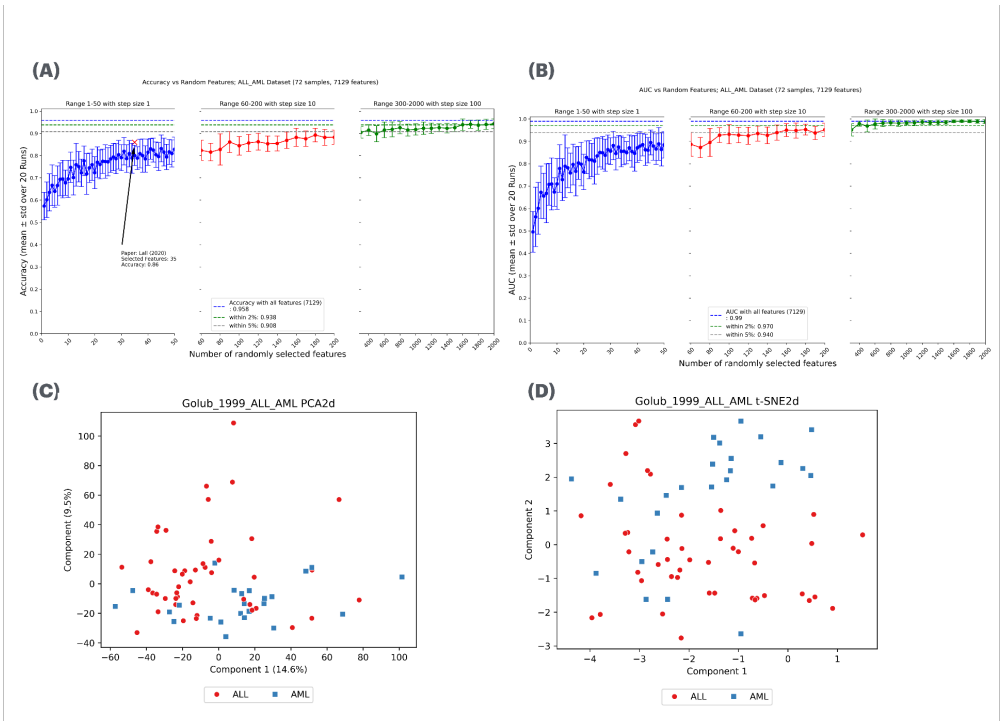


Figure 8: Random Forest performance with ALL/AML Leukemia microarray dataset (mean and standard deviation are reported over 20 runs). (A) & (B) models trained and tested on 80:20 split shows that a random subset of size $\sim$200 is able to match accuracy and AUC with all features, respectively. (C) & (D) PCA, t-SNE plots showing class separation.

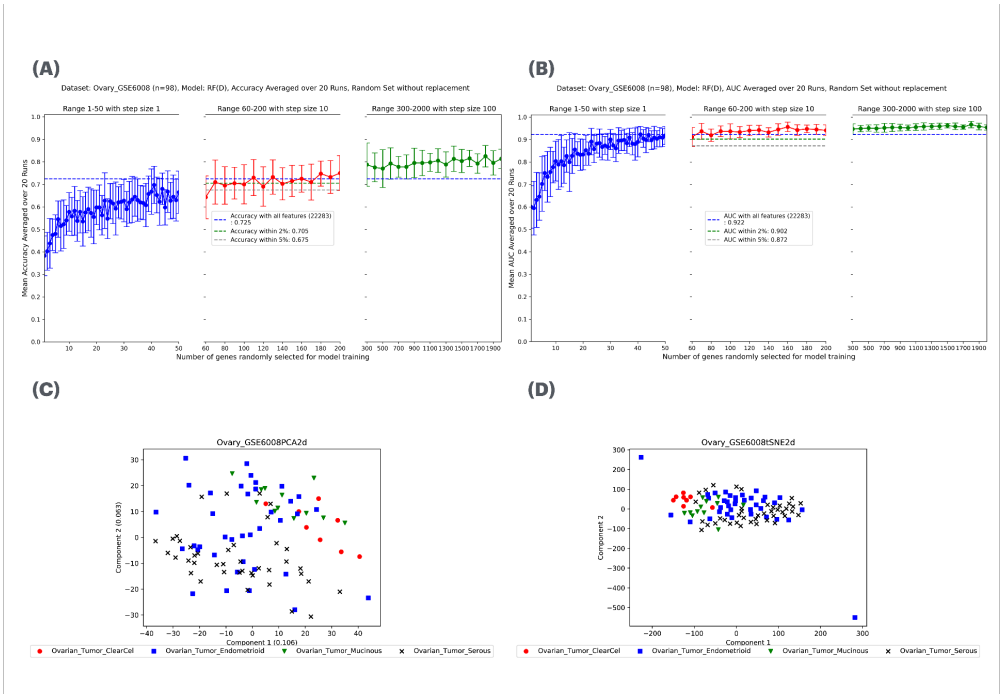


Figure 9: Random Forest performance with Ovary (GSE6008) microarray dataset (mean and standard deviation are reported over 20 runs). (A) & (B) models trained and tested on 80:20 split shows that a random subset is able to match accuracy and AUC with all features, respectively. (C) & (D) PCA, t-SNE plots showing class separation.

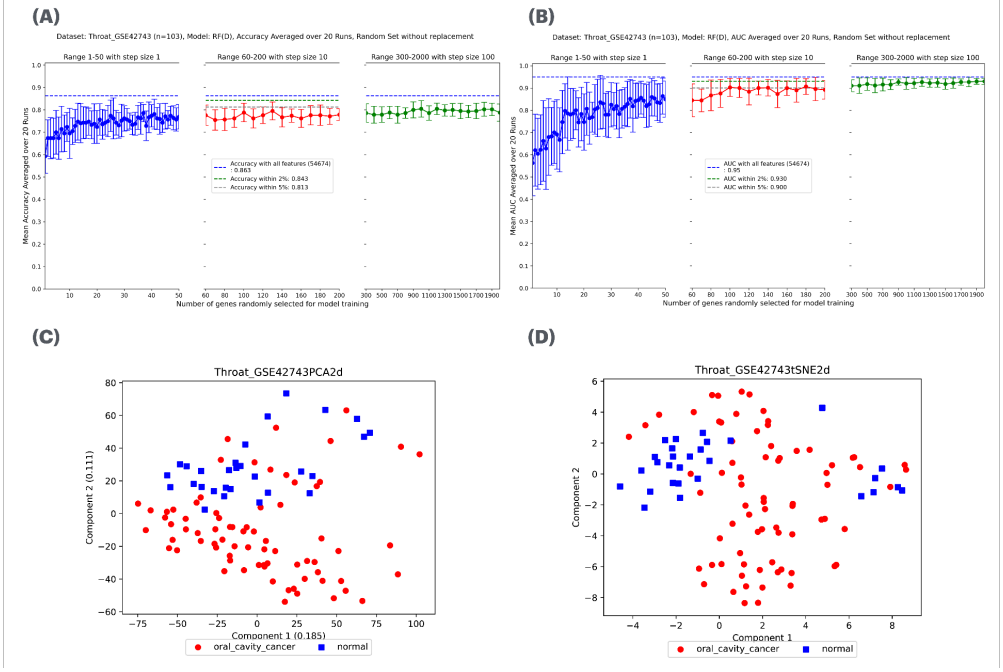
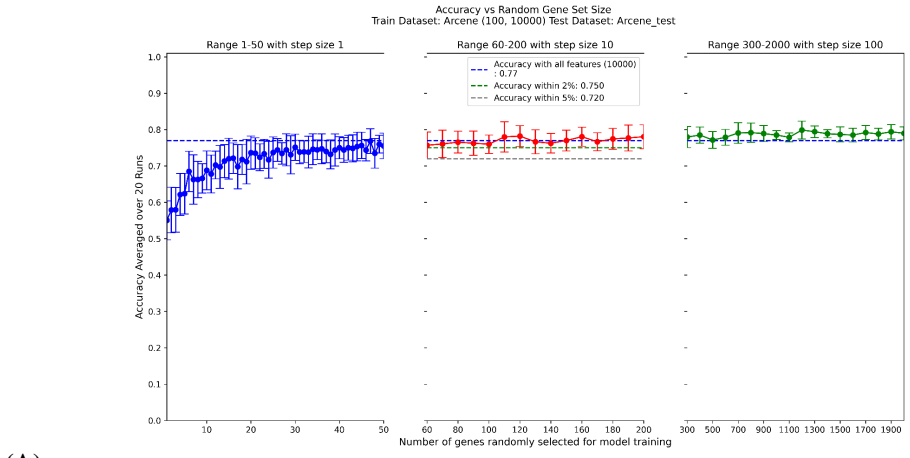
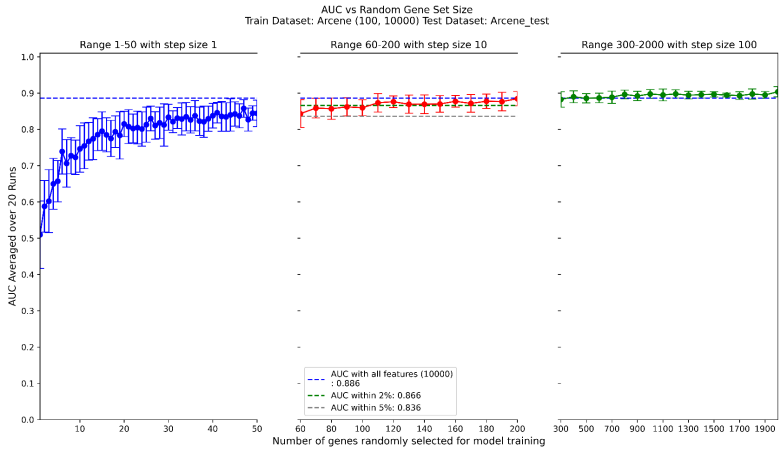


Figure 10: Random Forest performance with Oral/Throat (GSE42743) microarray dataset (mean and standard deviation are reported over 20 runs). (A) & (B) models trained and tested on 80:20 split shows that a random subset is able to match within-5% accuracy and AUC with all features, respectively. (C) & (D) PCA, t-SNE plots showing class separation.



(A)



(B)

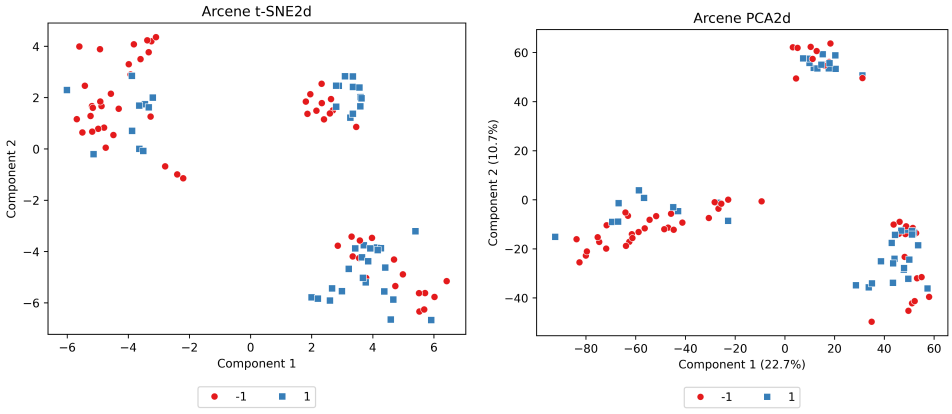


Figure 11: Random Forest performance with Arcene mass-spectrometry dataset (mean and standard deviation are reported over 20 runs). The task of ARCENE is to distinguish cancer versus normal patterns from mass-spectrometric data. (A) Models trained and tested on 80:20 split shows that a random subset of size ∼50 (0.5% of all features) is able to match within-5% Accuracy and AUC of all features. (B) PCA, t-SNE plots showing class separation.

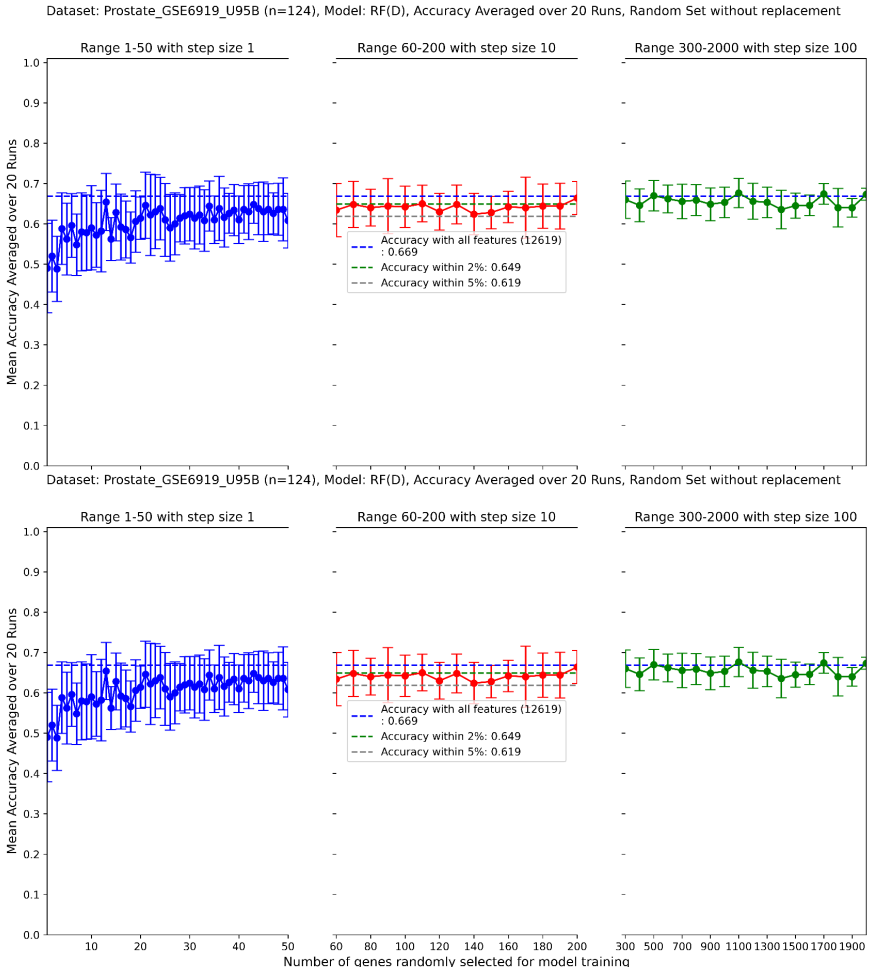


Figure 12: RF results with Prostate (GSE6919_U95B) dataset. Models trained and tested on 80:20 split shows that a random subset of size ∼50 (0.4% of all features) is able to match within-5% Accuracy and AUC of all features.

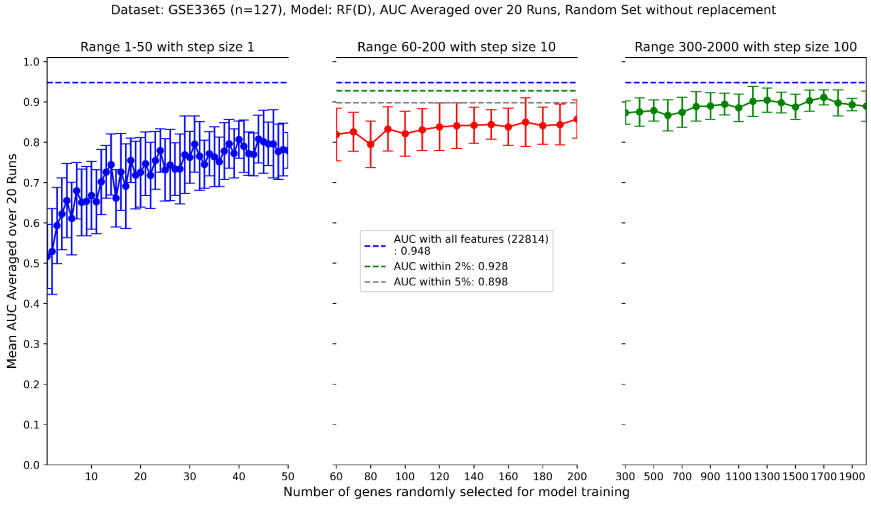


Figure 13: RF results with Bowel (GSE3365) dataset. Models trained and tested on 80:20 split shows that a random subset of size ∼500 (∼2.2% of all features) is able to match within-5% Accuracy and AUC of all features.

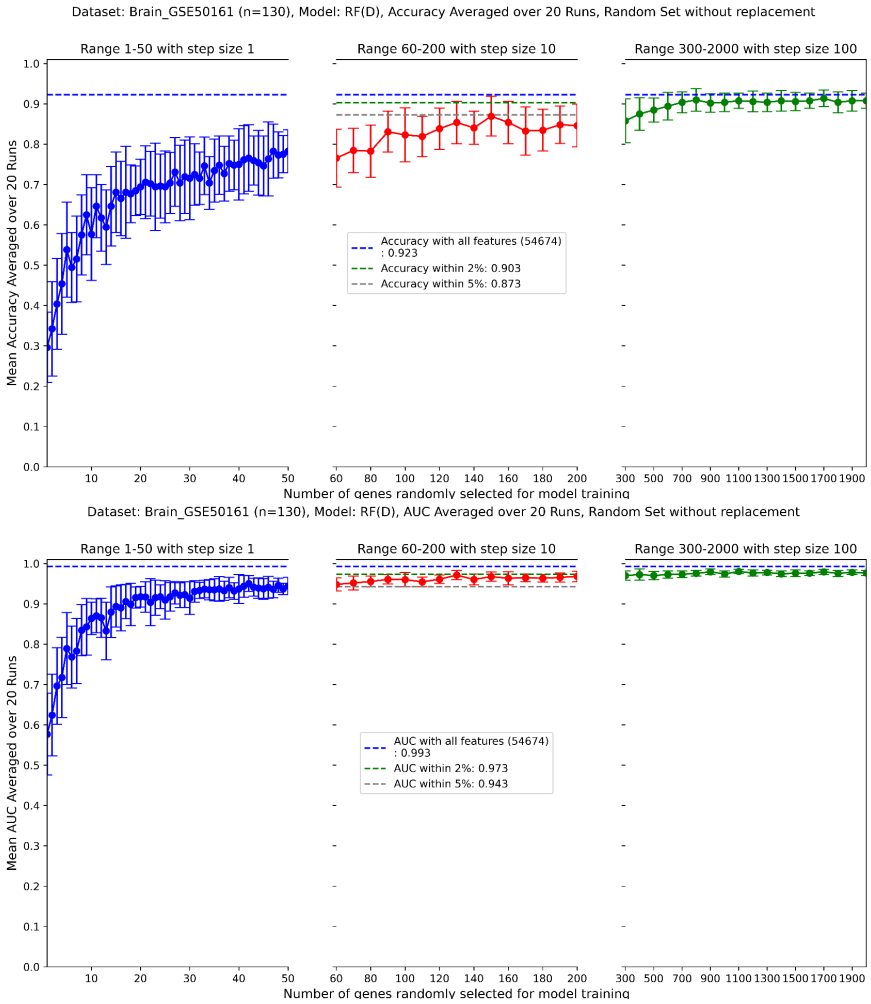


Figure 14: Random Forest performance with Brain (GSE50161) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset of size ∼50 (∼0.09% of all features) is able to match within-5% Accuracy and AUC of all features.

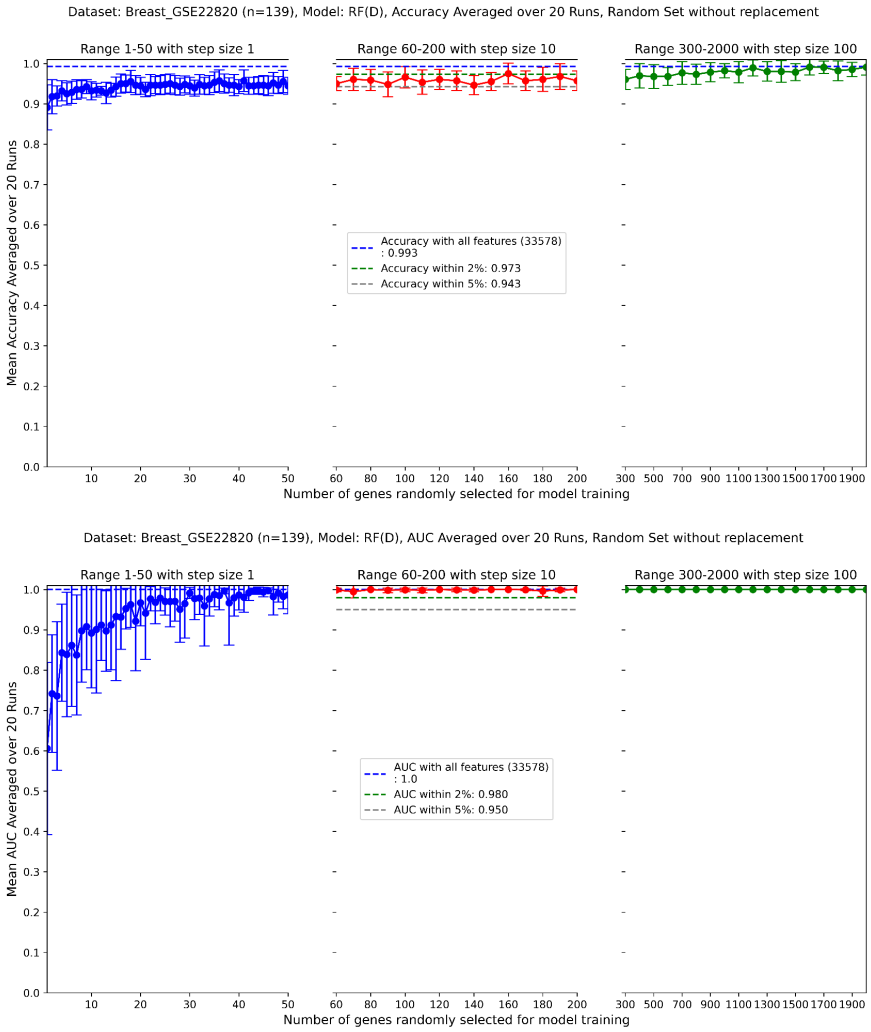


Figure 15: Random Forest performance with Breast (GSE22820) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset of size $\sim$50 ($\sim$0.14% of all features) is able to match within-5% Accuracy and full AUC of all features. (The unusually high accuracy with just one feature is because there is a severe class imbalance in this dataset).

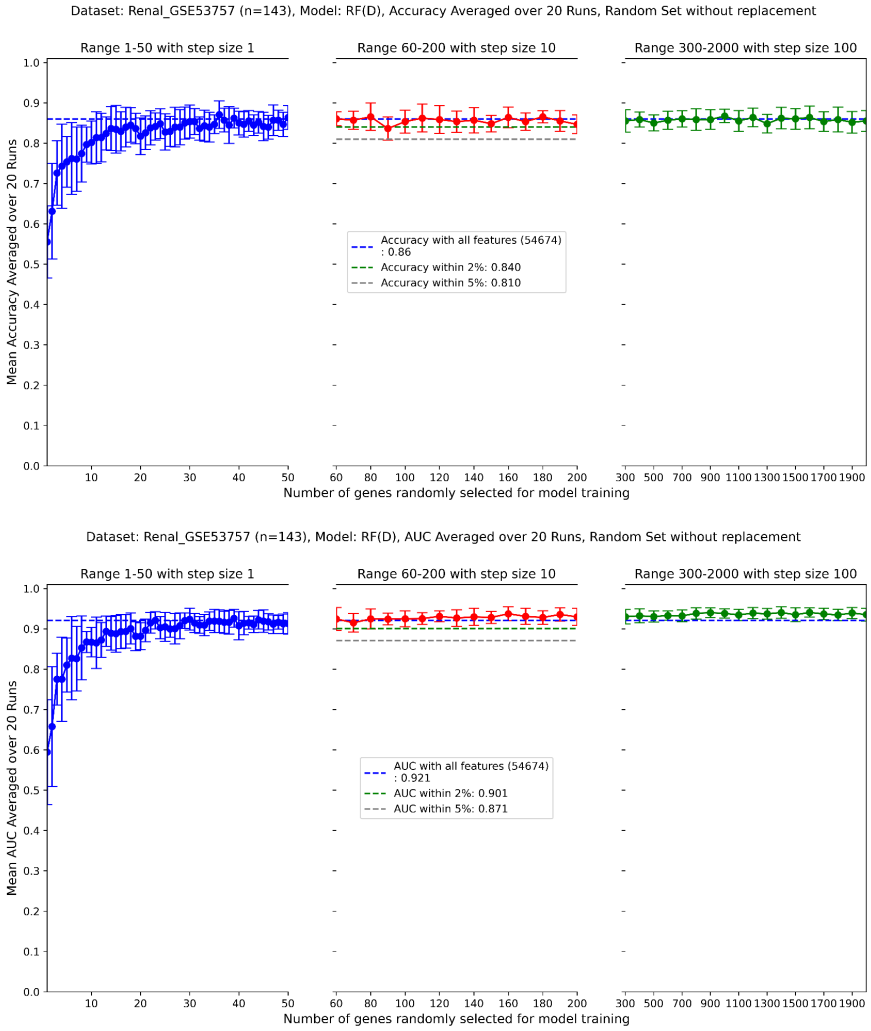


Figure 16: Random Forest performance with Renal (GSE53757) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset of size $\sim$30 ($\sim$0.06% of all features) is able to match full Accuracy and full AUC of all features.

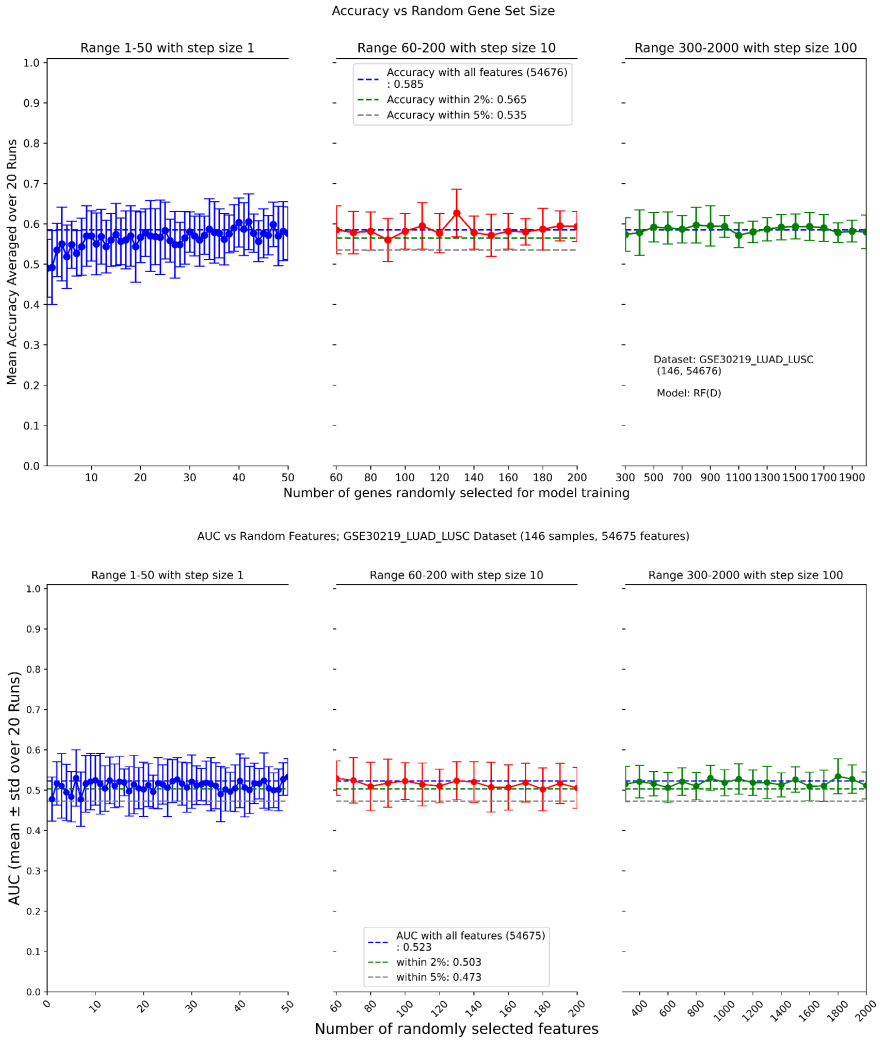


Figure 17: Random Forest performance with Lung Cancer (GSE30219) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match full Accuracy and full AUC of all features.

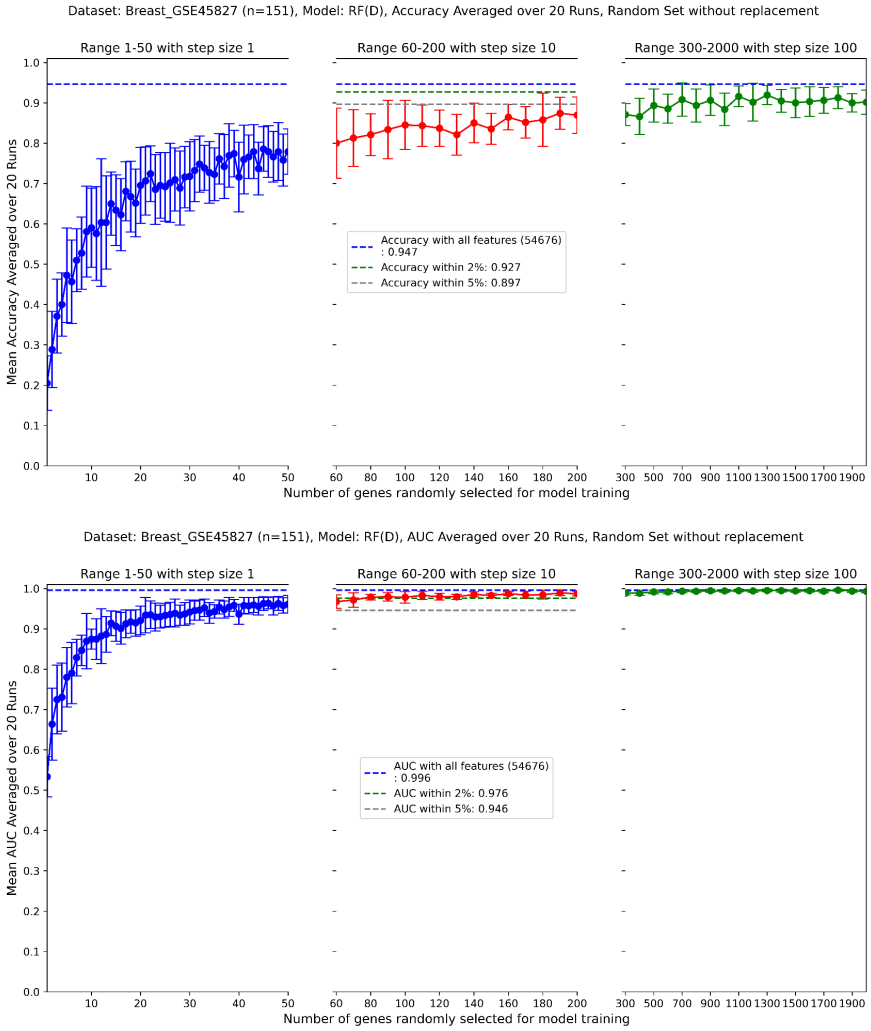


Figure 18: Random Forest performance with Breast Cancer (GSE45827) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match within-5% Accuracy and full AUC of all features.

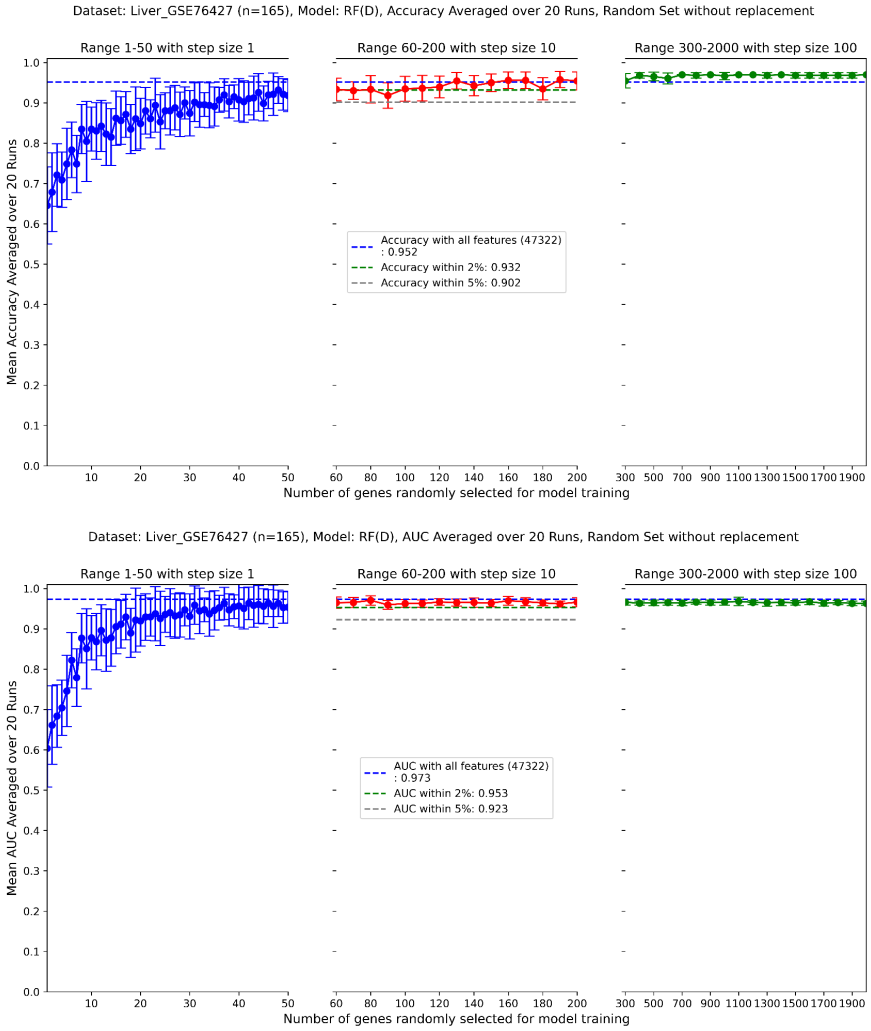


Figure 19: Random Forest performance with Liver Cancer (GSE76427) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match full Accuracy and full AUC of all features.

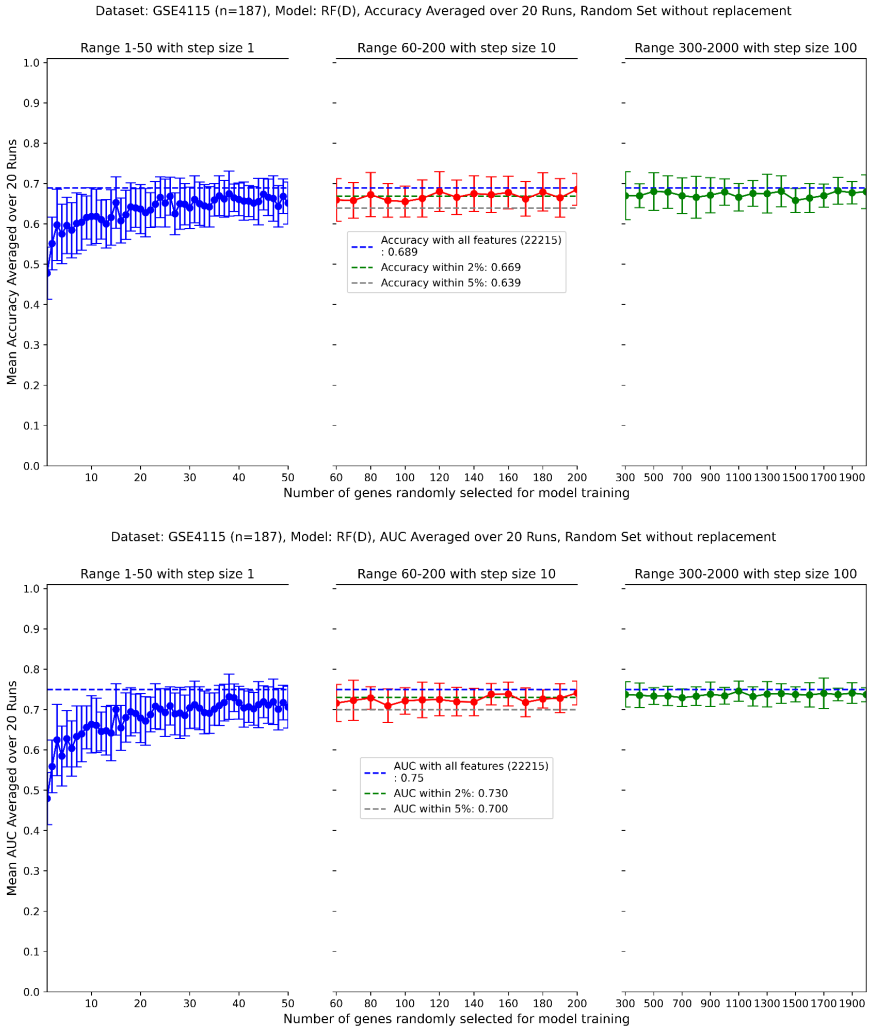


Figure 20: Random Forest performance with Lung Cancer (GSE4115) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match within-2% Accuracy and AUC of all features.

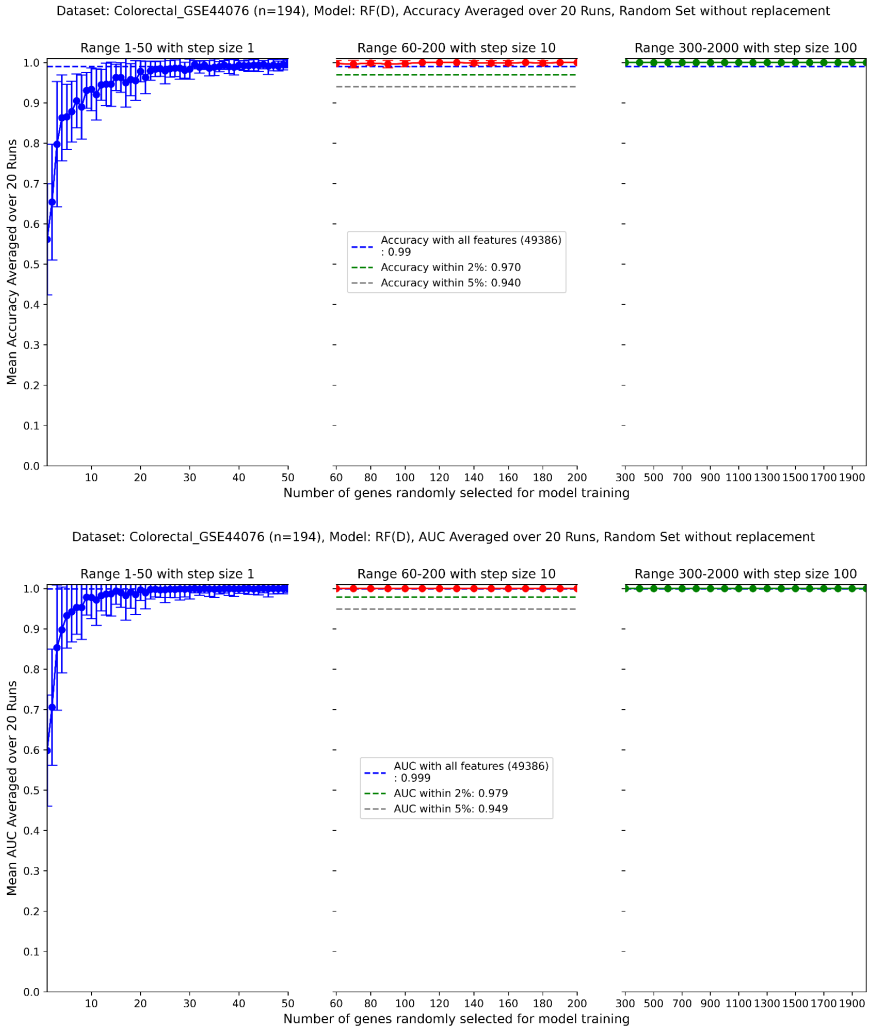


Figure 21: Random Forest performance with Colorectal Cancer (GSE44076) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match full Accuracy and full AUC of all features.

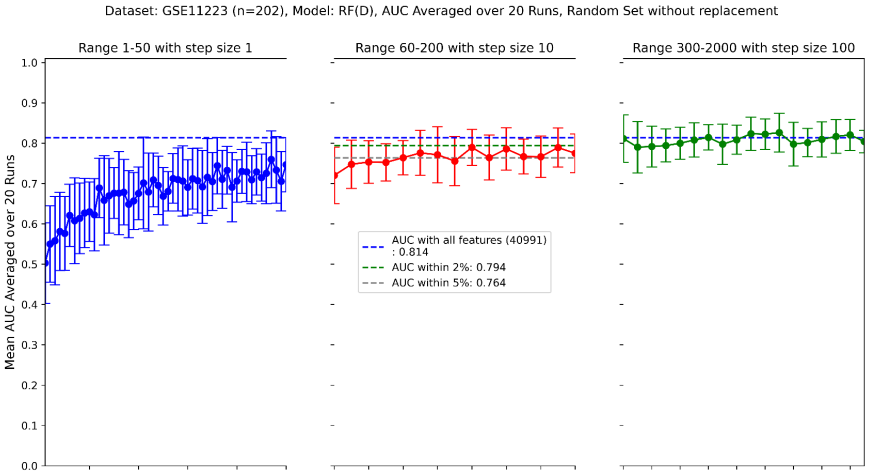


Figure 22: Random Forest performance with Colon Cancer (GSE11223) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match full AUC of all features.

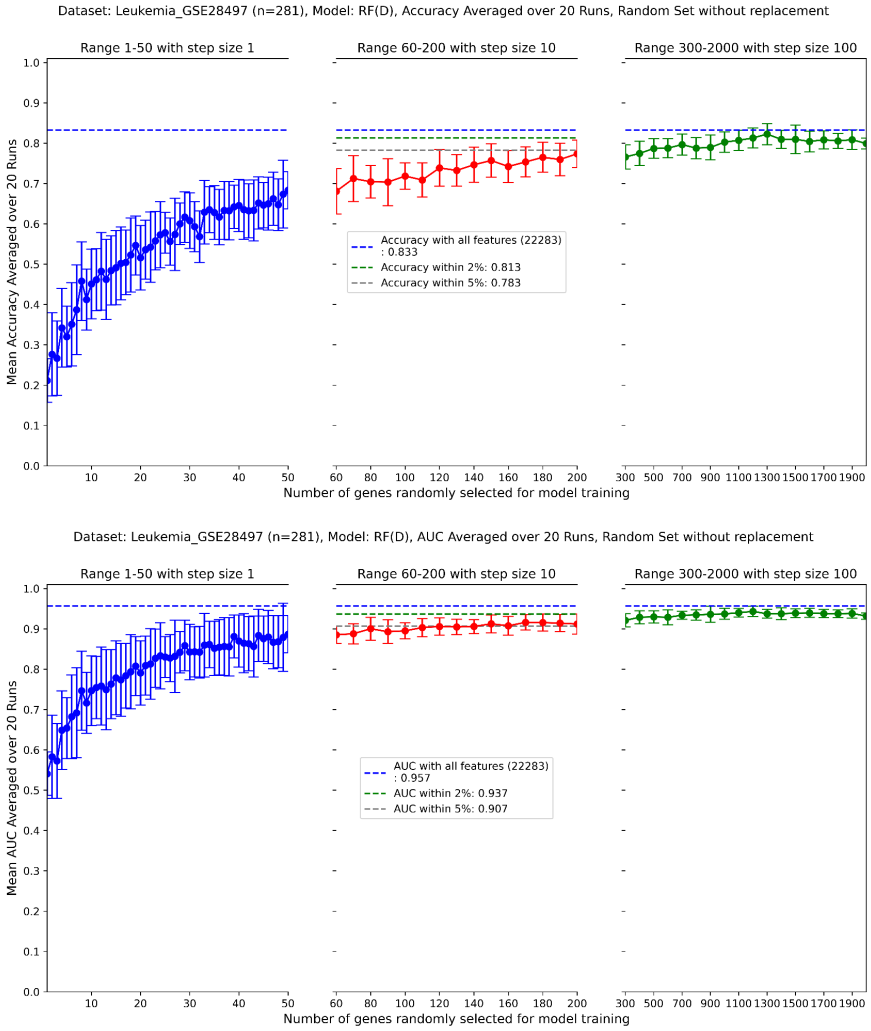


Figure 23: Random Forest performance with Leukemia (GSE28497) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match within-2% Accuracy and AUC of all features.

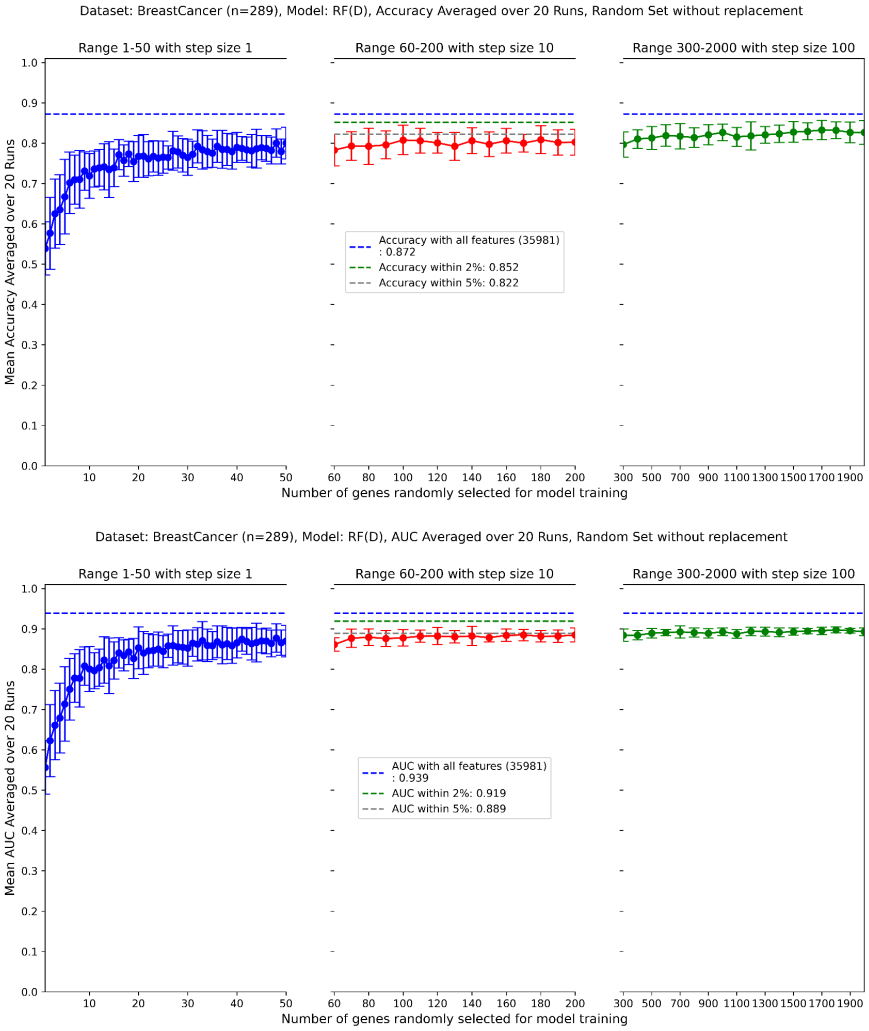


Figure 24: Random Forest performance with Breast Cancer (GSE70947) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a random subset is able to match within-5% Accuracy and AUC of all features.

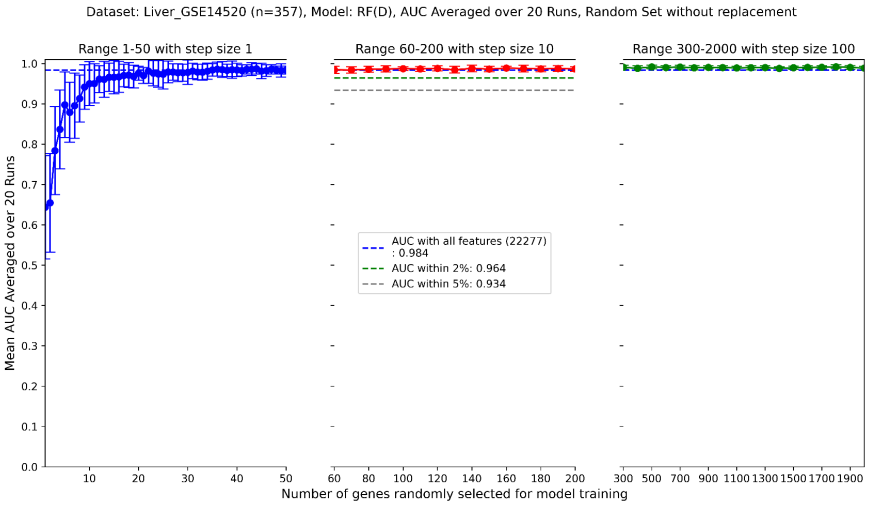


Figure 25: Random Forest performance with Liver Cancer (GSE14520) dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a small random subset is able to match full Accuracy and AUC of all features.

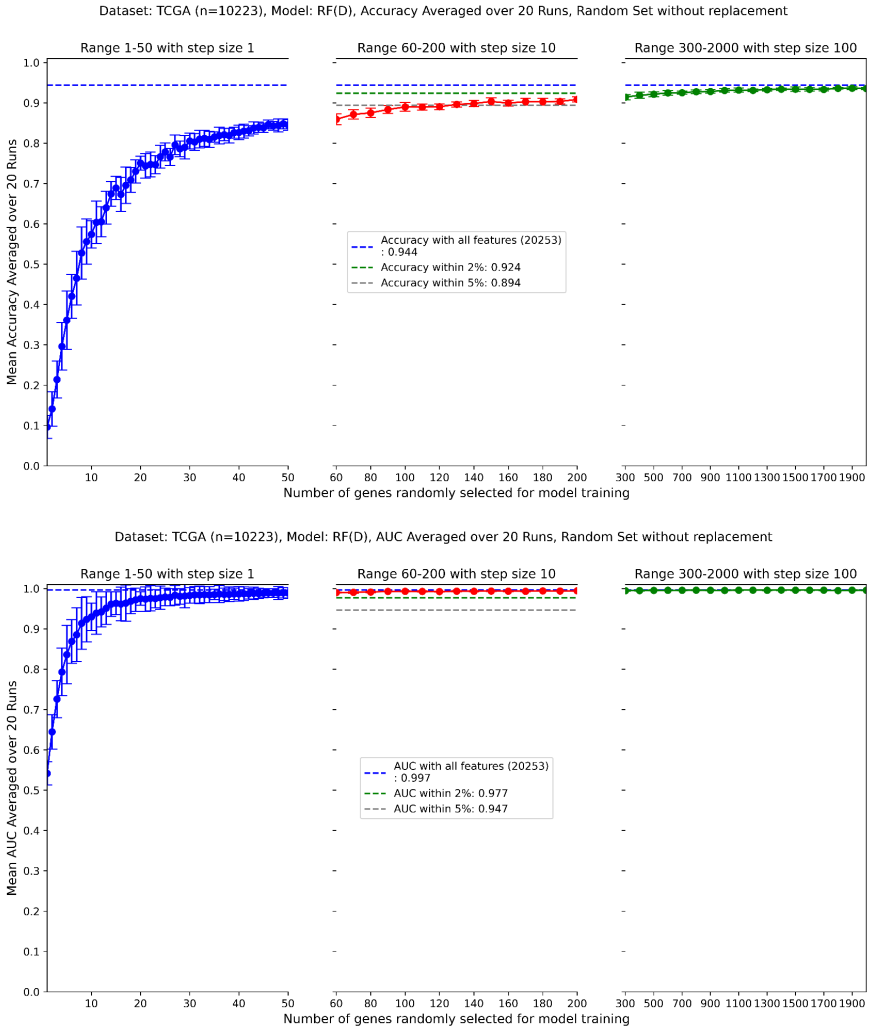


Figure 26: Random Forest performance with bulk RNA-Seq TCGA dataset with 33 classes (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a small random subset is able to match full Accuracy and AUC of all features.

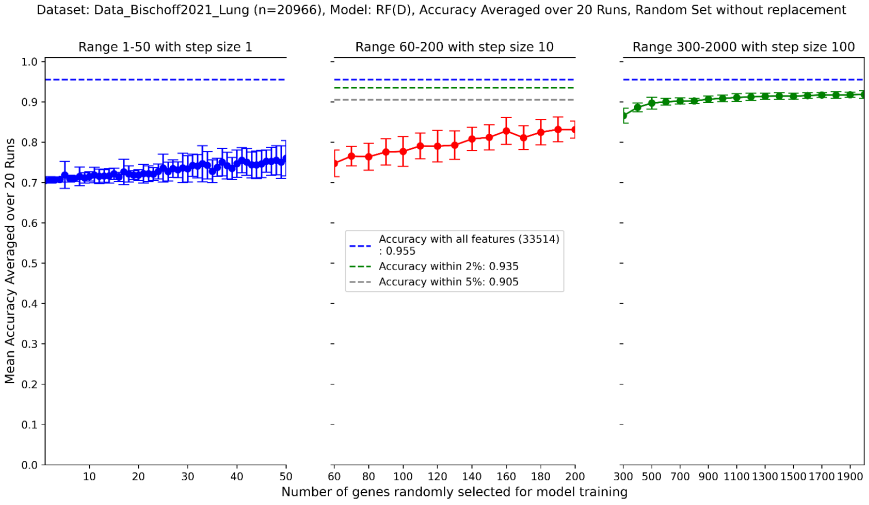


Figure 27: Random Forest performance with single-cell RNA-Seq Lung dataset with 9 classes (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that a small random subset is able to match within-5% Accuracy and AUC of all features.

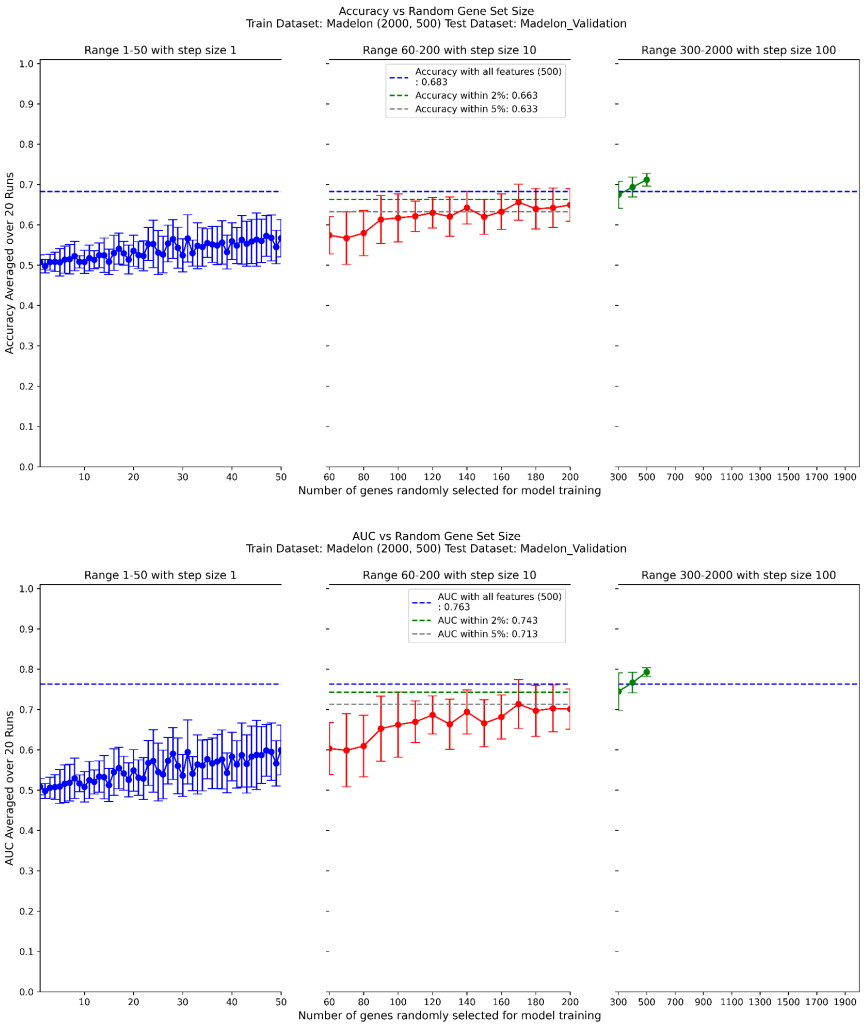


Figure 28: Random Forest performance with Madelon dataset (mean and standard deviation are reported over 20 runs). The task of MADELON is to classify random data. Models trained and tested on 80:20 split shows that a random subset is able to match within-5% Accuracy and AUC of all features. (As there are only 500 features in this dataset, there is no result beyond 500).

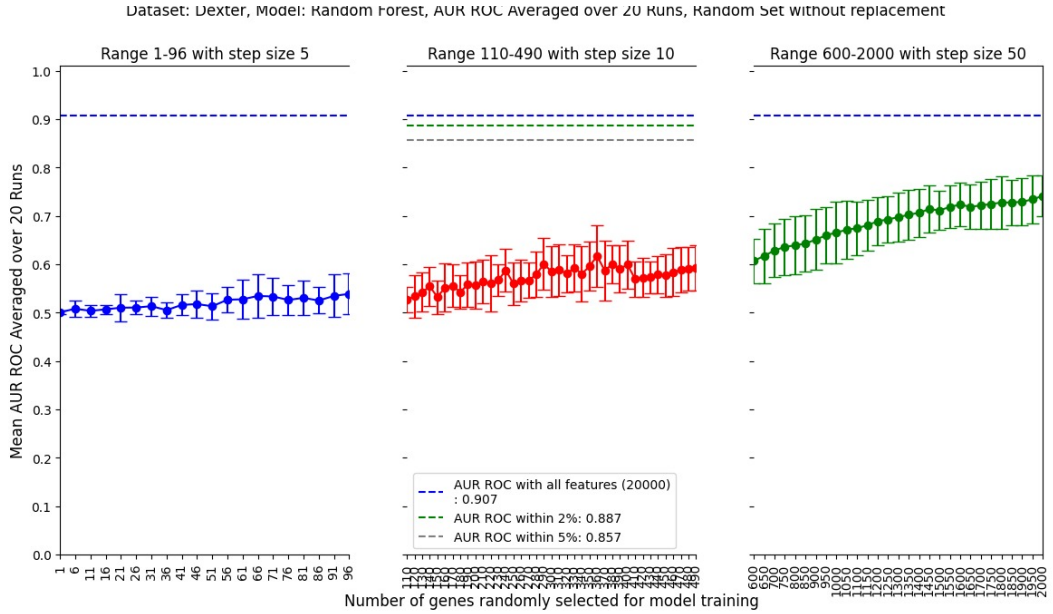


Figure 29: Random Forest performance with Dexter dataset (mean and standard deviation are reported over 20 runs). The task of DEXTER is to filter texts about "corporate acquisitions". Models trained and tested on 80:20 split shows that a random subset is NOT able to match AUC of all features.

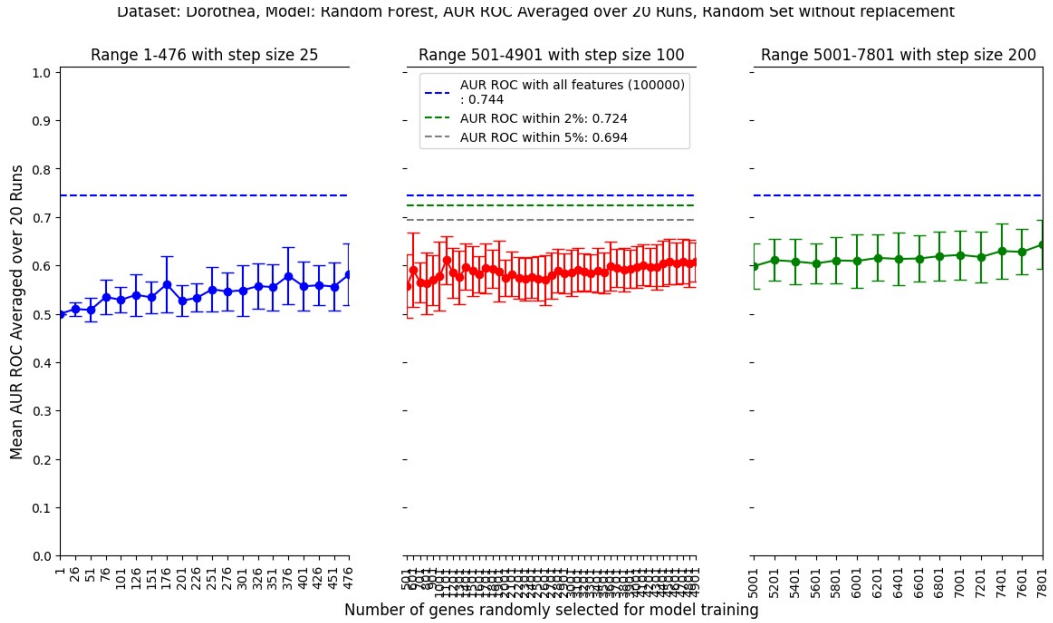


Figure 30: Random Forest performance with Dorothea dataset (mean and standard deviation are reported over 20 runs). The task of DOROTHEA is to predict which compounds bind to Thrombin. Models trained and tested on 80:20 split shows that a random subset is NOT able to match AUC of all features.

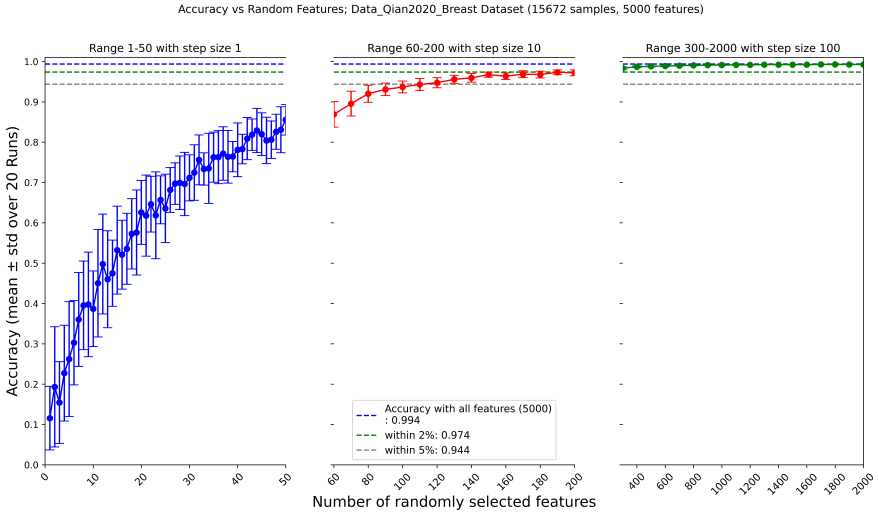


Figure 31: Random Forest performance with lung cancer sc-RNA-Seq Qian 2020 dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that any random subset of 100 features (2% of 5000) achieves within 5% accuracy of 5000 HVGs.

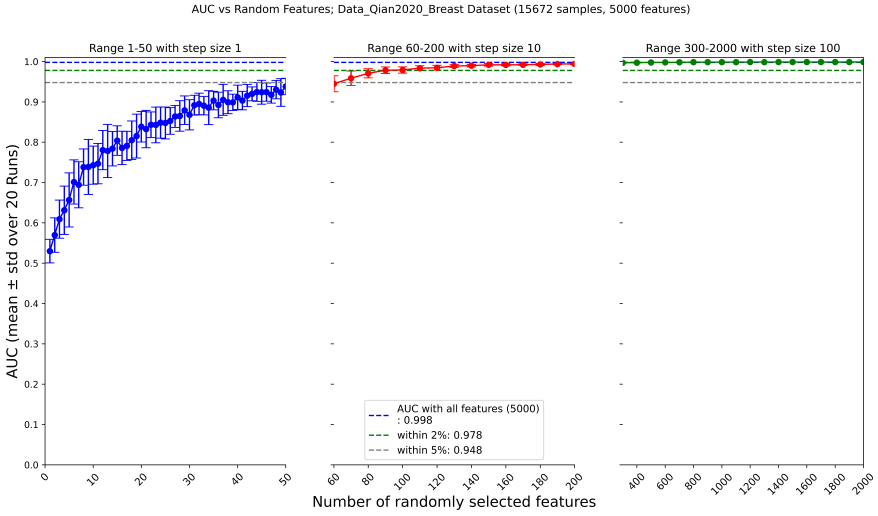


Figure 32: Random Forest performance with lung cancer sc-RNA-Seq Qian 2020 dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that any random subset of 100 features (2% of 5000) achieves within 5% AUC of 5000 HVGs.

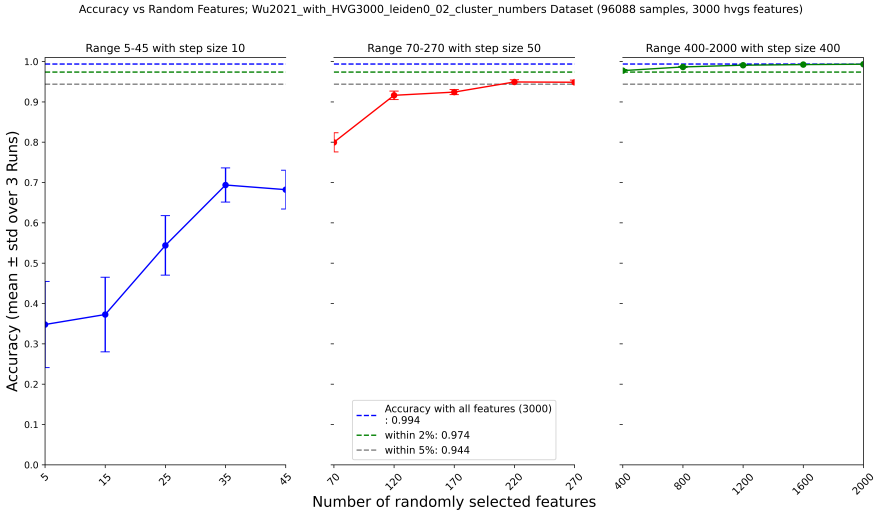


Figure 33: Random Forest performance with lung cancer sc-RNA-Seq Wu 2021 dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that any random subset of 200 features achieves within 5% accuracy of 3000 HVGs.

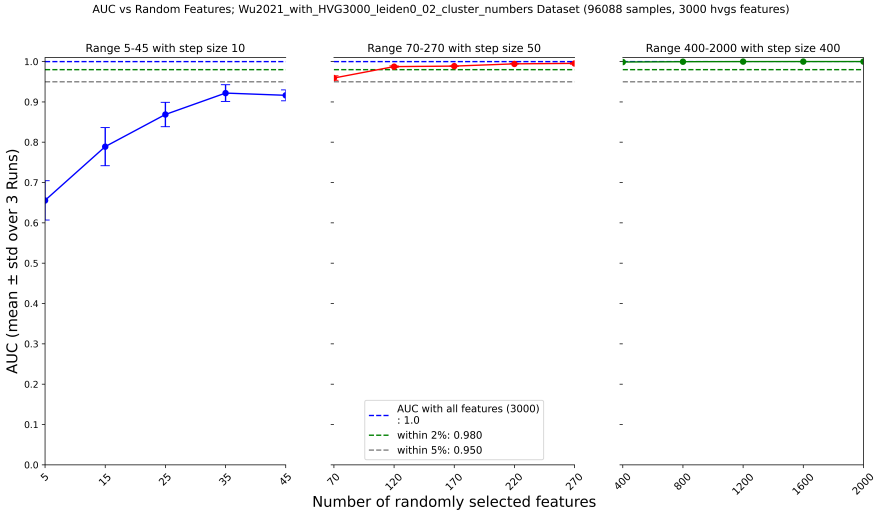


Figure 34: Random Forest performance with lung cancer sc-RNA-Seq Wu 2021 dataset (mean and standard deviation are reported over 20 runs). Models trained and tested on 80:20 split shows that any random subset of 200 features achieves within 5% AUC of 3000 HVGs.

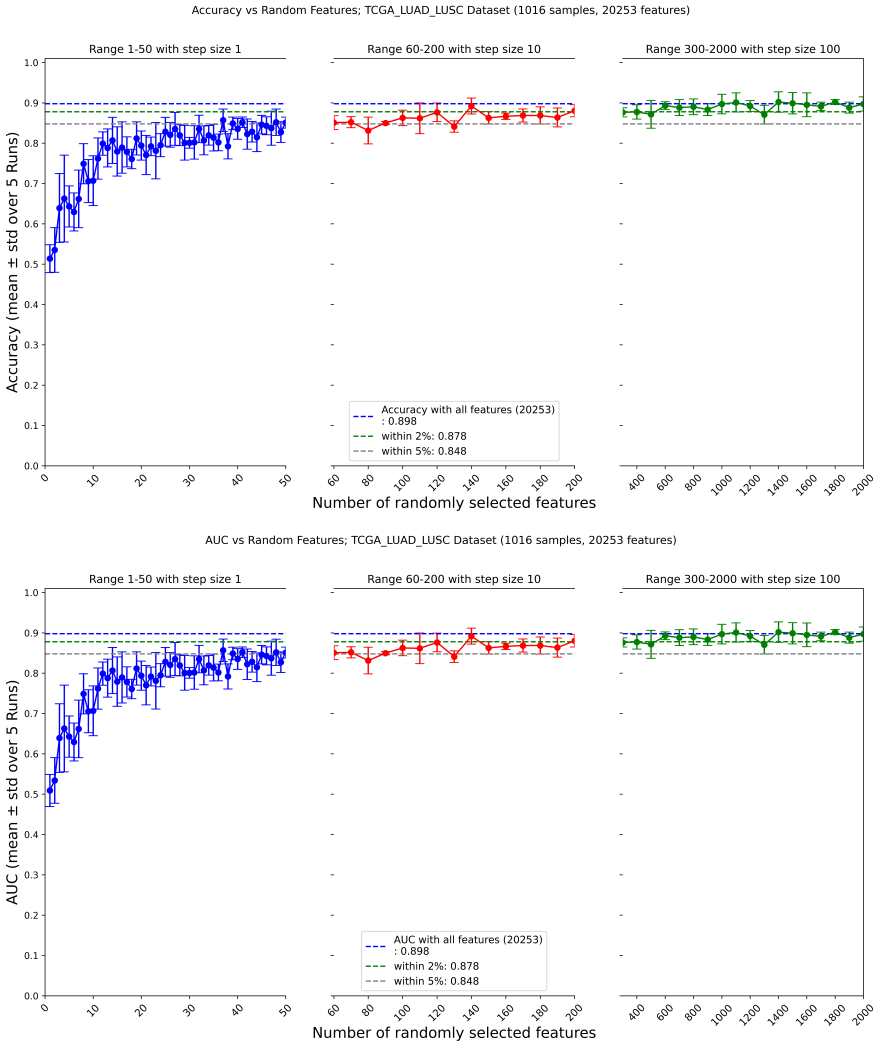


Figure 35: Decision Tree (DT) results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.

C RESULTS WITH OTHER MODELS

Additional figures supporting the main text are provided here (figs. 35–42).

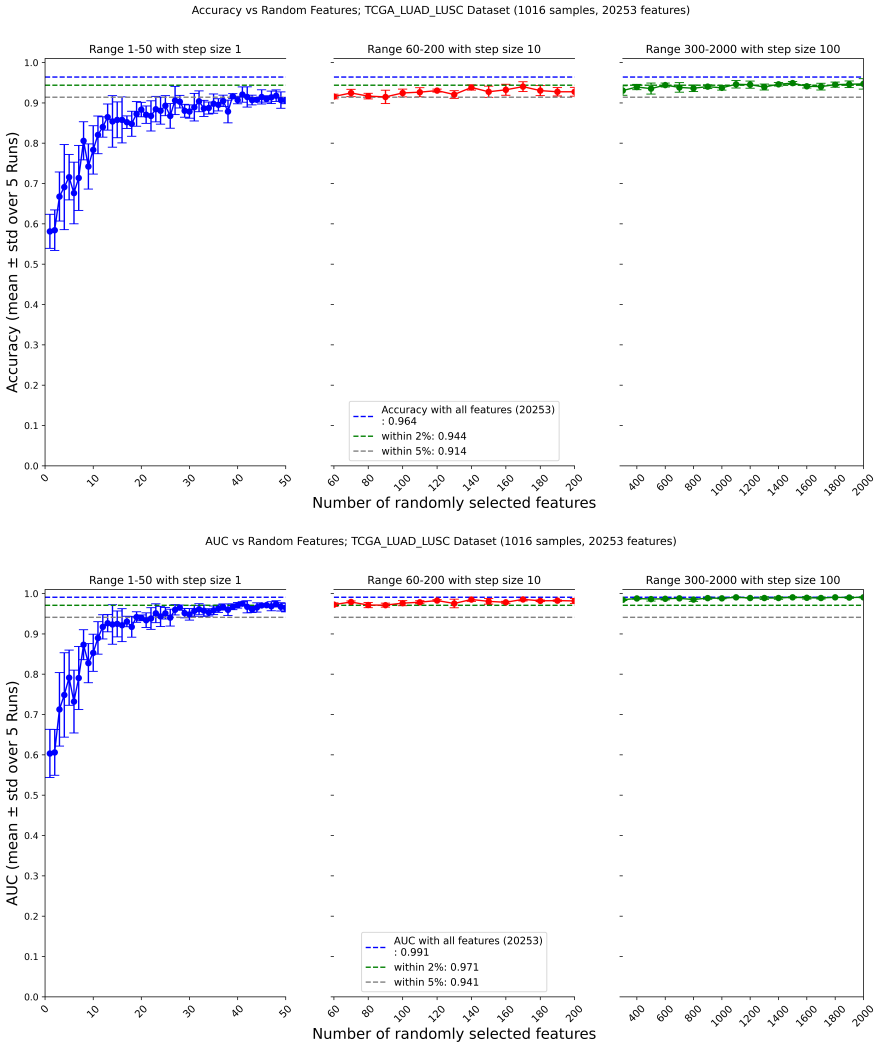


Figure 36: HistGradient Boosting (HistGB) classifier results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.

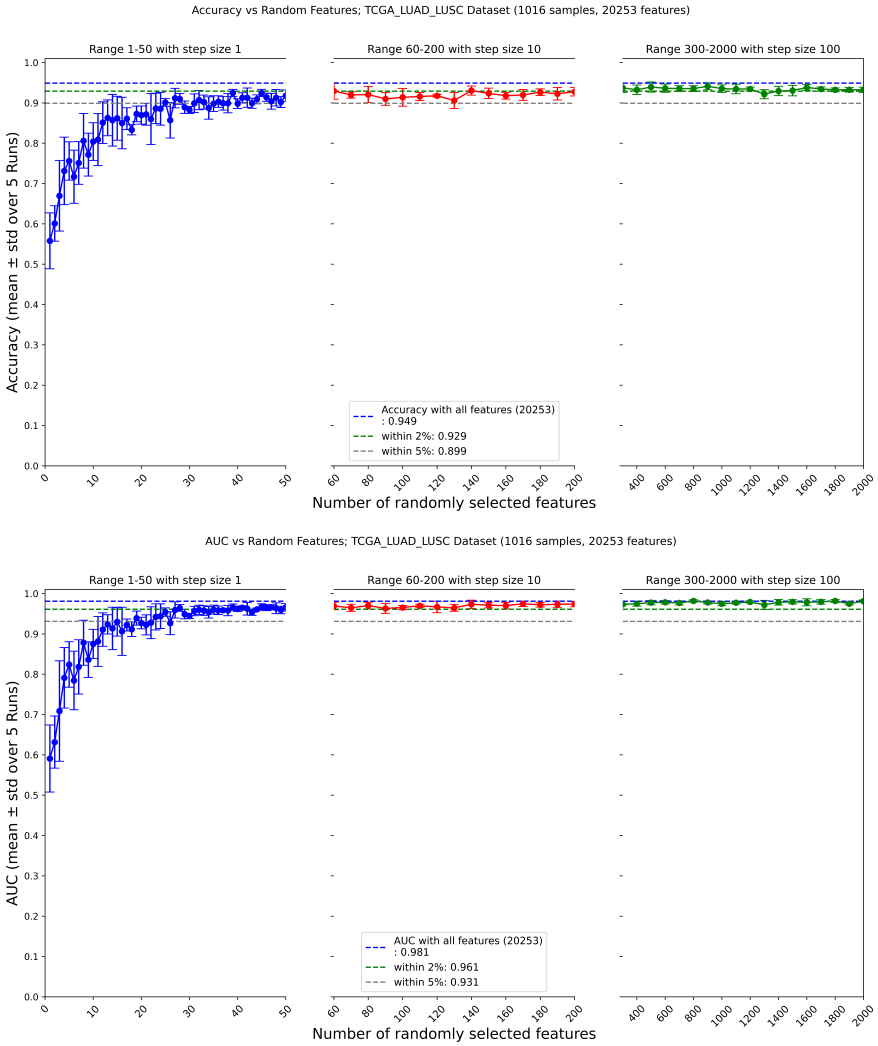


Figure 37: Logistic Regression (LR) results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.

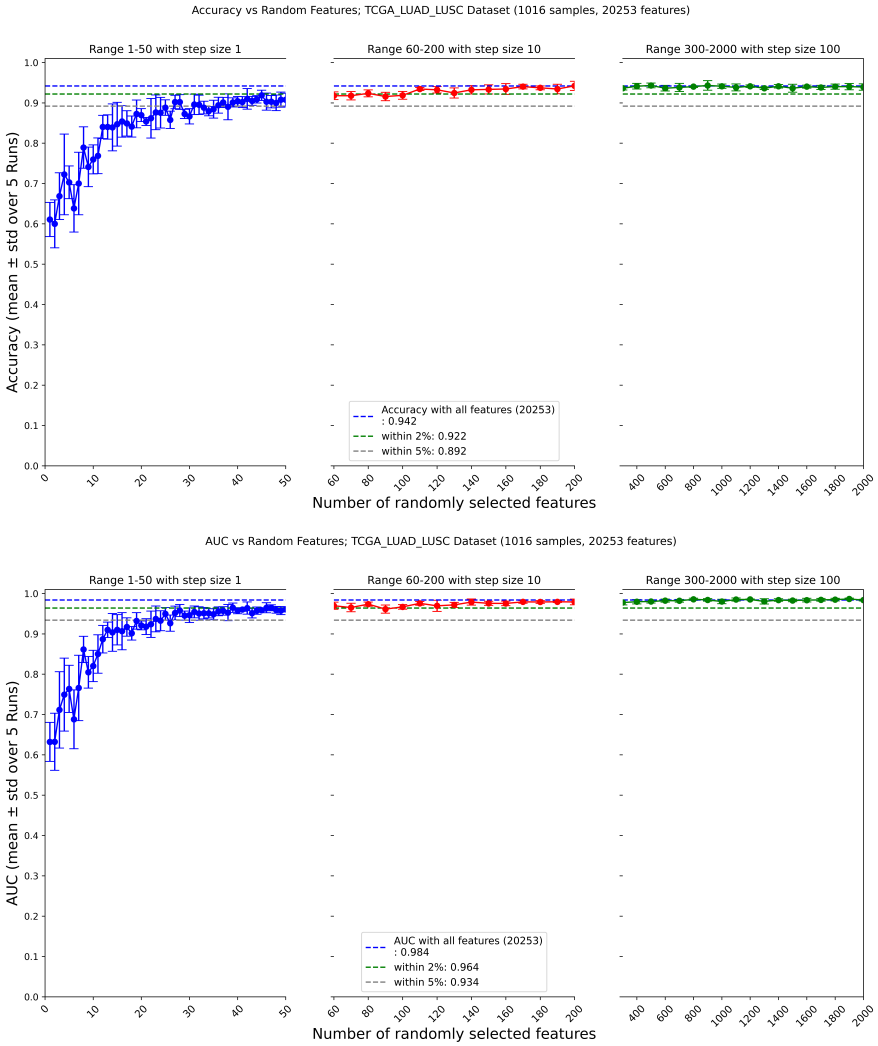


Figure 38: Multilayer Perceptron (MLP) results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.

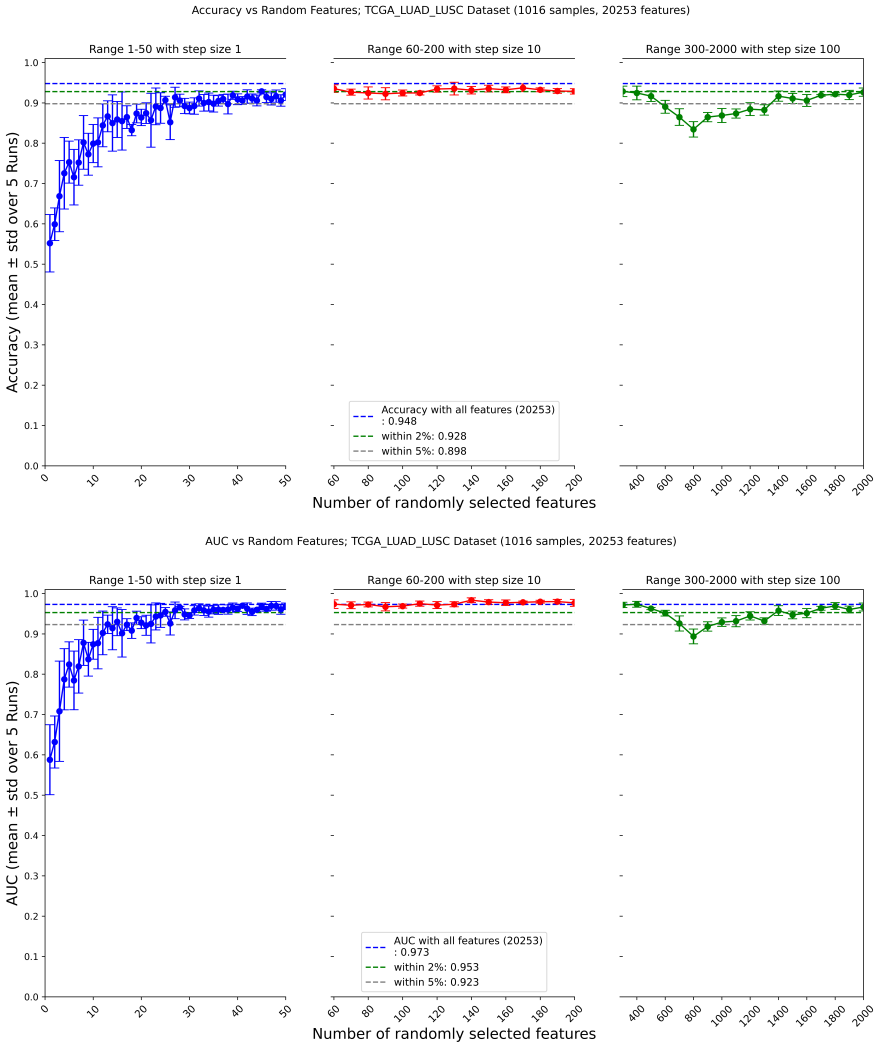


Figure 39: Ridge classifier results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.

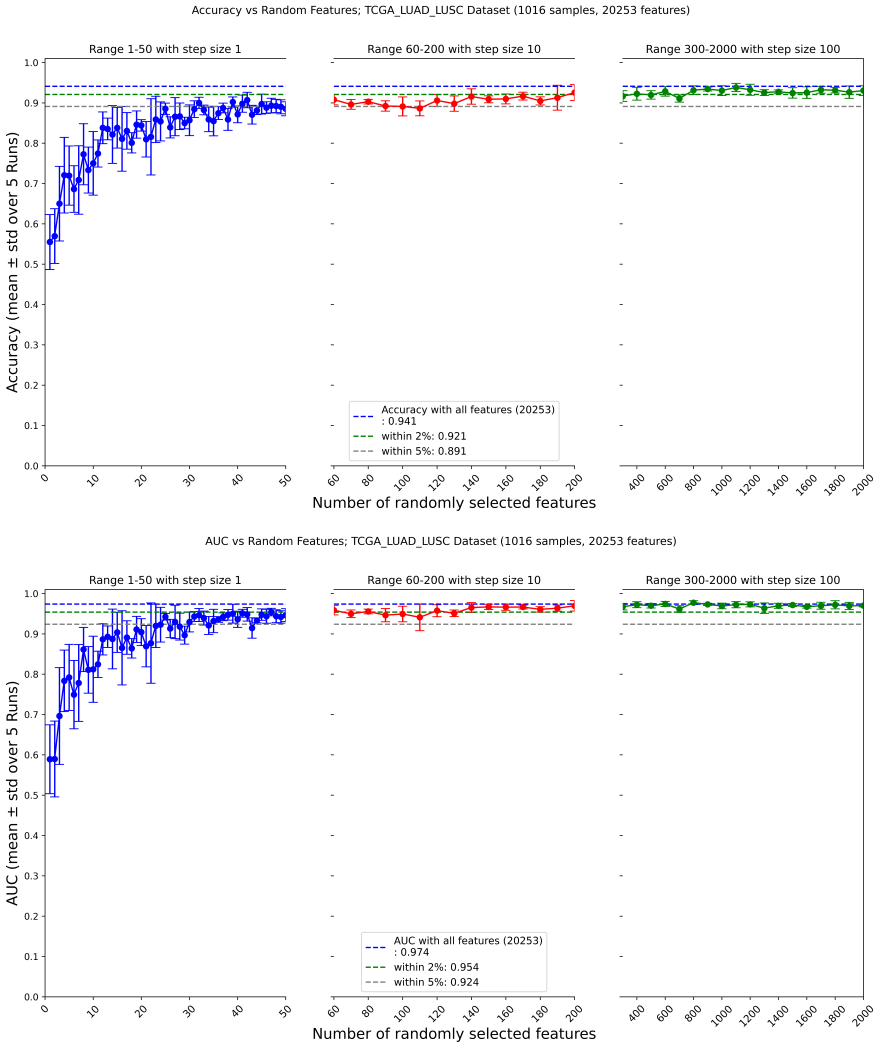


Figure 40: SGD classifier results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.

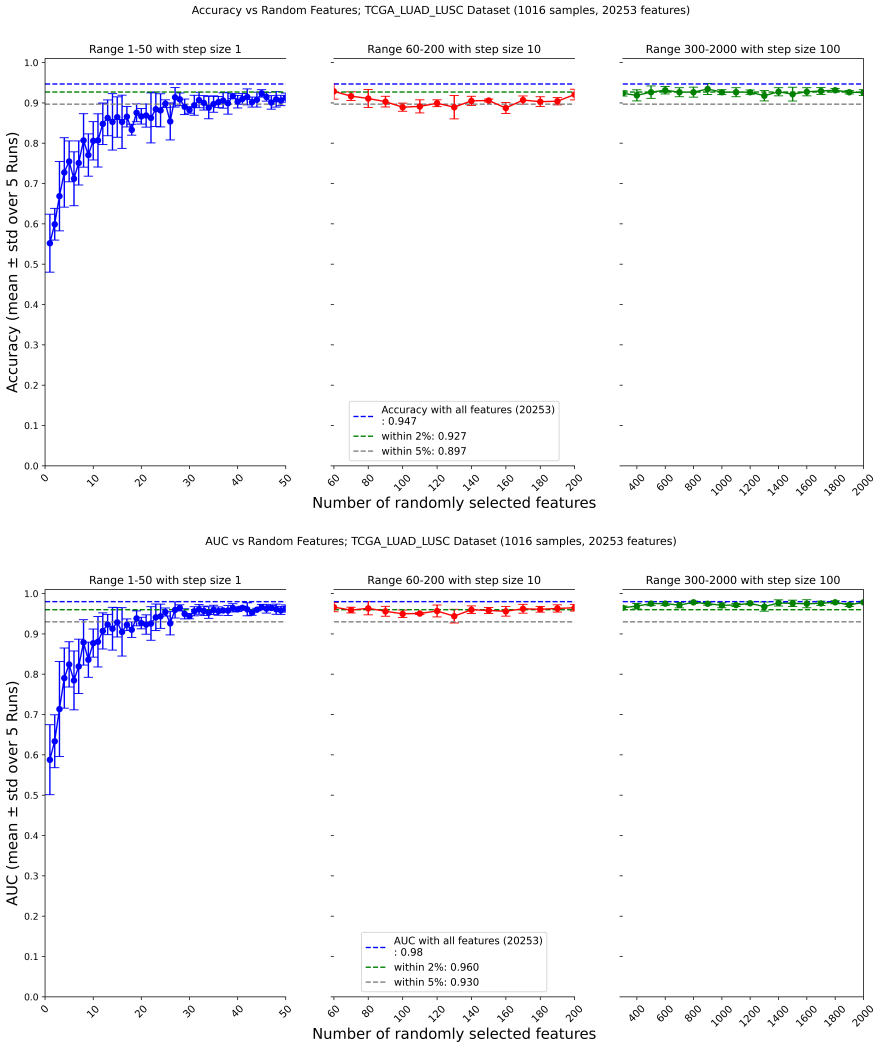


Figure 41: Support Vector Machine (SVM) results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.

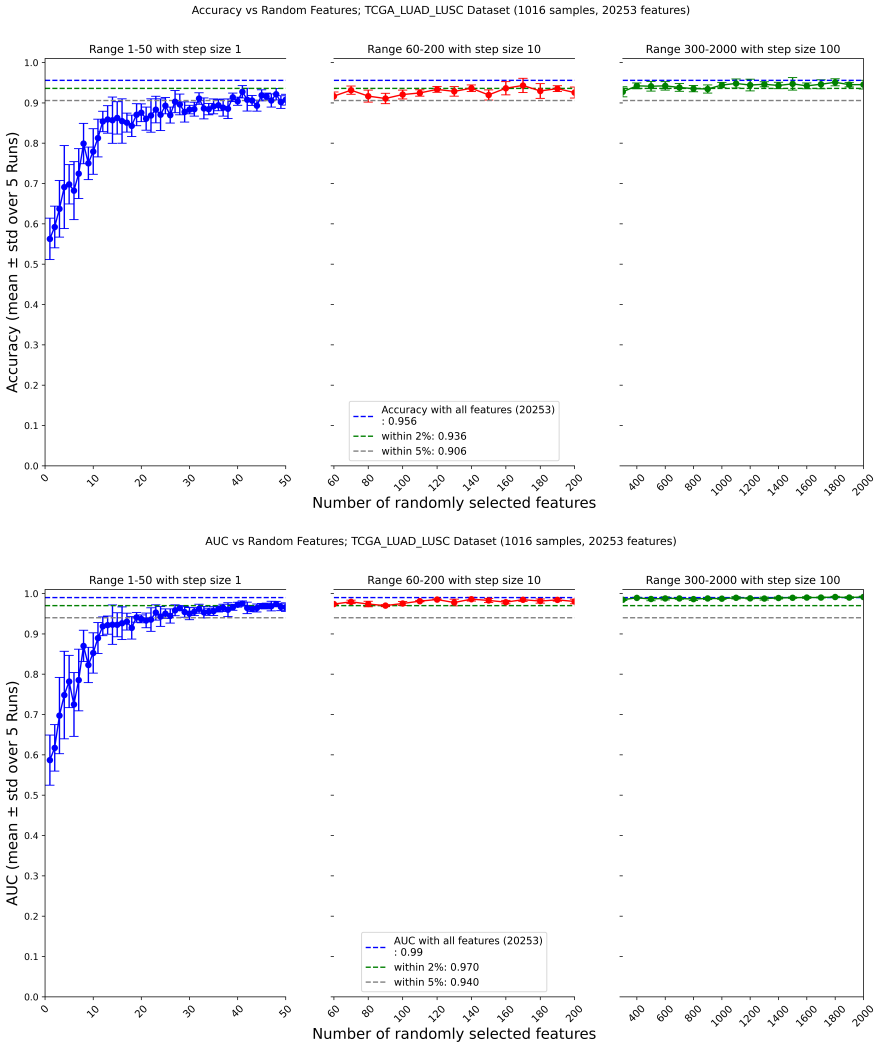


Figure 42: eXtreme Gradient Boosting (XGB) results with TCGA(LUAD/LUSC) bulk RNA-Seq dataset. Trained and tested on 80:20 split, the plot shows that a random subset is able to match accuracy and AUC with all features, respectively.