WORLD FORUM ON ACADEMIC INNOVATION AND DIGITAL LITERACY

STABILITY ANALYSIS OF FEATURE SELECTION ALGORITHMS UNDER DIFFERENT DATA SPLITS AND RANDOM SEEDS

G.Kh. Tashmatov

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

E-mail: ganijon.tashmatov@gmail.com

Research Article | Artificial Intelligence / Machine Learning

Abstract

Feature selection is widely used to improve interpretability, reduce dimensionality and support robust deployment of machine-learning models, especially in domains such as medical artificial intelligence where explanations are expected to be reproducible. However, predictive performance alone does not reveal whether an algorithm selects a stable and interpretable subset of features. This study evaluates the stability of seven feature-selection algorithms under changes in random seed, train/test split and cross-validation folds: ANOVA F-test, mutual information, chi-square, recursive feature elimination, L1-based SelectFromModel, Random Forest importance and Extra Trees importance. Experiments were conducted on four real open classification data sets from scikit-learn (Breast Cancer Wisconsin, Wine, Iris and Digits) and one explicitly synthetic data set containing informative, redundant and noisy features. For each data set and algorithm, ten random seeds were used. Accuracy, precision, recall, F1-score, ROC-AUC, pairwise Jaccard similarity between selected feature subsets and feature-selection frequencies were computed. The results show that high predictive performance does not necessarily imply stable feature selection. L1-based selection achieved the highest mean ROC-AUC (0.9819) and F1-score (0.9380), while ANOVA F-test achieved the highest seed-level Jaccard stability (0.9126). RFE with a linear classifier produced high mean ROC-AUC (0.9816) but the lowest average

stability (0.6722). The Spearman correlation between mean ROC-AUC and seed-level Jaccard similarity was close to zero (-0.0252), indicating that performance and stability capture different dimensions of model quality.

Keywords: feature selection stability; Jaccard similarity; robust machine learning; medical AI; random seeds; scikit-learn pipelines

1. Introduction

Feature selection is a central preprocessing and model-building step in artificial intelligence and machine learning. It reduces the number of variables, can improve generalization, and often makes a model easier to interpret. In high-stakes applications, such as medical AI, the selected features may be interpreted as candidate biomarkers, clinical indicators or domain-relevant explanations. Therefore, a feature selector should not only support strong predictive performance but should also be stable: small changes in the training sample, data split or random seed should not completely change the selected subset of variables. Recent reviews emphasize that feature-selection stability is an important quality criterion, not a secondary implementation detail [1,2].

The central problem addressed in this paper is that classification metrics such as accuracy and ROC-AUC may remain high even when the selected features vary substantially across runs. This creates a risk for interpretability: two models may perform almost identically while pointing to different explanatory variables. In practical AI systems, such instability can reduce trust, complicate model auditing and weaken reproducibility. The objective of this study is to measure this phenomenon using a controlled, reproducible Python experiment based on real open data sets and one synthetic noisy benchmark.

The contribution of the paper is threefold. First, it compares seven commonly used feature-selection algorithms under identical scikit-learn pipelines. Second, it evaluates both predictive performance and feature-selection stability using pairwise Jaccard

similarity across random seeds and cross-validation folds. Third, it reports feature-selection frequency charts to show which variables are repeatedly selected and which appear sensitive to sampling variation.

2. Related Work

Modern feature-selection research increasingly treats stability as a core requirement. Büyükkeçeci and Okur [1] reviewed feature-selection stability measures and emphasized that stability can be defined as the robustness of selected features under small data perturbations. Khaire and Dhanalakshmi [2] reviewed stability issues across algorithms and domains, highlighting the need to compare feature selectors not only by prediction quality but also by consistency of selected subsets. Łukaszuk et al. [3] investigated stability in multi-omics analysis and explicitly evaluated accuracy, stability and reproducibility together, which is particularly relevant for biomedical machine learning.

Stability is also linked to interpretability and robustness. Lazebnik and Rosenfeld [4] argued that stable selected features can improve explanations and reduce sensitivity to adversarial or sampling effects. Recent comparative work on bio-inspired and machine-learning-based feature selection has also focused on the trade-off between accuracy and subset consistency [5]. In parallel, reproducibility studies in machine learning have shown that random seeds and implementation choices can materially influence reported results [6,7]. In clinical classification, seed sensitivity is especially important because a small change in random initialization or split can affect fairness or model conclusions [7]. More recent work has proposed procedures for controlling seed instability and improving reproducibility [8].

Medical AI guidelines and reviews further motivate stability analysis. Healthcare prediction models are expected to be validated, reproducible and resistant to overfitting [9]. A model that repeatedly changes its selected clinical or biological features across data splits may be difficult to justify, even if it reports a high ROC-

AUC. The present study therefore evaluates feature selection as a two-dimensional problem: predictive performance and subset stability.

The implementation follows scikit-learn documentation for feature-selection methods and pipelines. The scikit-learn user guide distinguishes univariate F-tests, mutual information and chi-square statistics, and notes the different assumptions behind these functions [10]. The Pipeline API is used to combine preprocessing, feature selection and classification into reproducible workflows [11].

3. Materials and Methods

3.1 Study Design

The experiment was designed as a repeated evaluation over data sets, feature-selection methods and random seeds. For each data set, the same list of ten random seeds was used: 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100. For each seed, a stratified 70/30 train/test split was created. A feature selector was fitted on the training data only, selected features were recorded and a linear classifier was evaluated on the held-out test set. Separately, three-fold stratified cross-validation on the training set was used to estimate fold-level feature-selection stability. This design measures two kinds of instability: sensitivity to train/test splits and sensitivity to cross-validation folds.

3.2 Data Sets

The study used four real open benchmark data sets available in scikit-learn and one explicitly synthetic data set. The synthetic data set was included to create a controlled noisy environment with informative, redundant and random noise features. It is not presented as a real-world data source.

Table 1. Data sets used in the experiment.

Dataset	Type	Samples	Features	Classes	k	Source
Breast Cancer Wisconsin	Real/open	569	30	2	10	scikit-learn/UCI WDBC

Digits	Real/open	1797	64	10	20	scikit-learn/UCI Digits
Iris	Real/open	150	4	3	2	scikit-learn Iris
Synthetic noisy data	Synthetic	1000	80	2	20	Synthetic make_classification
Wine	Real/open	178	13	3	5	scikit-learn/UCI Wine

3.3 Feature-Selection Algorithms

Seven feature-selection algorithms were compared. The number of selected features k was fixed per data set to make the selected subsets comparable across algorithms. For Breast Cancer Wisconsin, Wine, Iris, Digits and SyntheticNoisy80, k was set to 10, 5, 2, 20 and 20, respectively.

Table 2. Feature-selection methods and implementation choices.

Method	scikit-learn implementation	Rationale
ANOVA F-test	SelectKBest(f_classif)	Univariate filter; measures linear class-feature dependency.
Mutual information	SelectKBest(mutual_info_classif)	Nonparametric filter; can detect nonlinear dependency.
Chi-square	MinMaxScaler + SelectKBest(chi2)	Univariate filter for non-negative feature values.
RFE (linear classifier)	RFE with a linear estimator	Wrapper method; recursively removes less

		important features.
L1-based SelectFromModel	LinearSVC with L1 penalty	Embedded sparse linear model; selects features with non-zero or largest coefficients.
Random Forest importance	SelectFromModel(RandomForestClassifier)	Embedded tree-ensemble importance method.
Extra Trees importance	SelectFromModel(ExtraTreesClassifier)	Embedded randomized tree-ensemble importance method.

3.4 Evaluation Metrics

Predictive performance was evaluated by accuracy, macro/weighted precision, recall, F1-score and ROC-AUC. For binary data sets, ROC-AUC was computed using the decision score of the positive class. For multiclass data sets, weighted one-vs-rest ROC-AUC was computed from normalized decision scores.

Feature-selection stability was measured by pairwise Jaccard similarity. For two selected feature subsets A and B, the Jaccard similarity is defined as $J(A, B) = |A \cap B| / |A \cup B|$. A value of 1 means that two runs selected exactly the same feature subset, while a value close to 0 means that the selected features are almost disjoint. For each method and data set, the mean pairwise Jaccard similarity was computed across all

random-seed runs. A second stability value was computed across cross-validation folds within each seed and then averaged.

Feature-selection frequency was also calculated for each feature as the number of times it was selected divided by the total number of runs. This frequency is useful for interpretation because it distinguishes consistently selected variables from occasional or unstable variables.

3.5 Statistical Analysis and Reproducibility

To summarize the relationship between prediction quality and stability, Spearman rank correlations were computed between mean ROC-AUC and seed-level Jaccard similarity and between mean F1-score and seed-level Jaccard similarity across method-level summaries. The observed correlations were -0.0252 for ROC-AUC versus stability and -0.0723 for F1-score versus stability. The full experiment is reproducible from the supplied Python scripts, README and requirements file. All random seeds are fixed and all selected feature subsets are saved as CSV tables.

4. Results

Across all data sets, methods and random seeds, predictive performance was high and relatively similar. However, stability differed substantially between methods. Table 3 summarizes the overall results averaged across five data sets and ten seeds per data set.

Table 3. Overall mean performance and stability by feature-selection method.

Method	Accuracy	F1	ROC-AUC	ROC-AUC SD	Seed Jaccard	CV Jaccard	AUC rank	Stability rank
ANOVA F	0.9263	0.9285	0.9802	0.0067	0.9126	0.9111	6	1
Chi-square	0.9246	0.9268	0.9799	0.0067	0.8597	0.8537	7	2



Extra Trees	0.9306	0.9327	0.9807	0.0067	0.8367	0.8639	5	3
L1 SelectFromModel	0.9368	0.9380	0.9819	0.0059	0.7478	0.7279	1	6
Mutual information	0.9272	0.9294	0.9807	0.0065	0.8213	0.8344	4	5
RFE (linear)	0.9312	0.9326	0.9816	0.0062	0.6722	0.6035	2	7
Random Forest	0.9302	0.9324	0.9808	0.0063	0.8263	0.8499	3	4

The L1-based selector achieved the highest mean ROC-AUC (0.9819) and F1-score (0.9380), while ANOVA F-test achieved the highest seed-level Jaccard similarity (0.9126). The RFE method with a linear classifier had the second-best mean ROC-AUC (0.9816) but the lowest seed-level Jaccard stability (0.6722) and the lowest CV-fold stability (0.6035). These results show that strong classifier performance can coexist with unstable selected feature subsets.

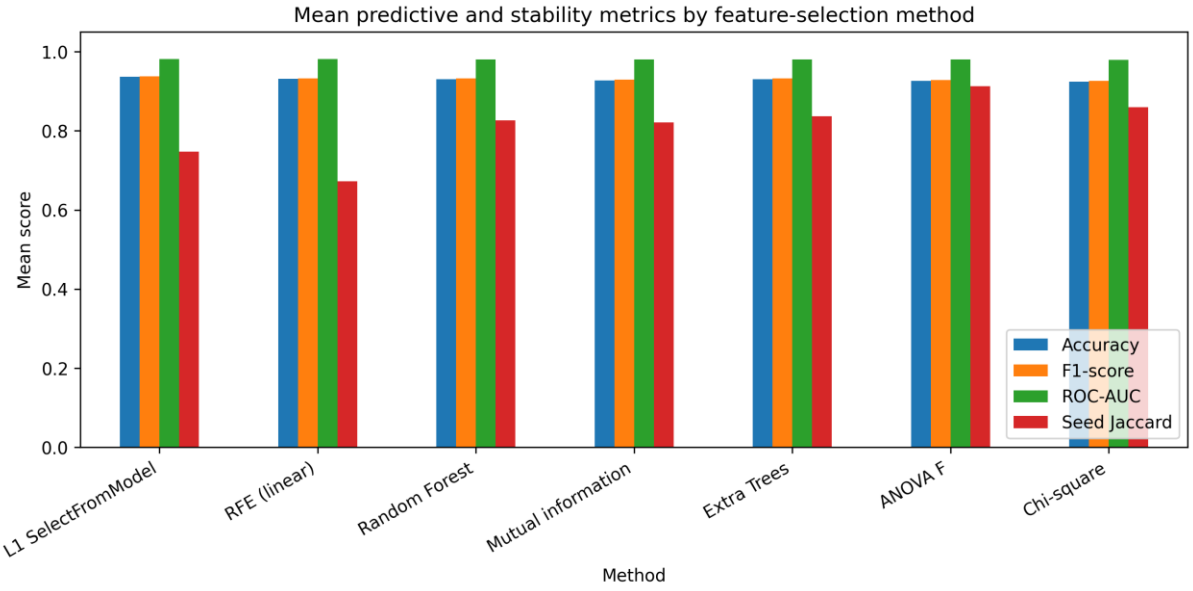


Figure 1. Mean accuracy, F1-score, ROC-AUC and seed-level Jaccard similarity by method.

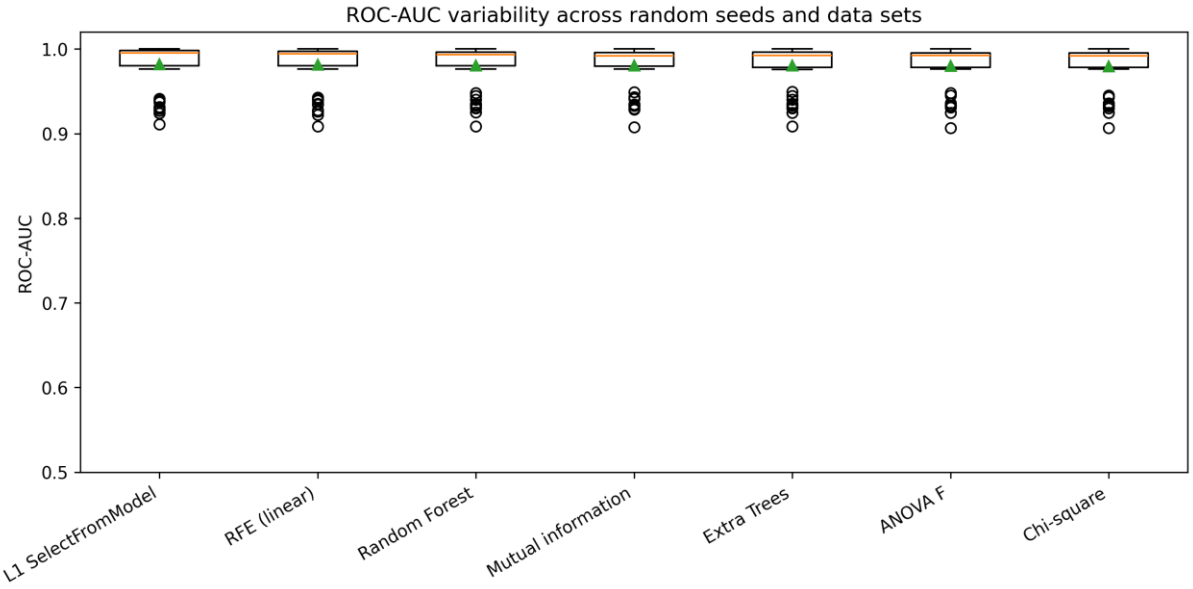


Figure 2. Distribution of ROC-AUC values across random seeds and data sets.

The boxplot in Figure 2 shows that ROC-AUC variation across random seeds is small for most methods. This makes the instability of selected features more important: if performance values appear almost identical, stability can become the deciding criterion for interpretability and reproducibility.

Table 4. Examples of low feature-selection stability despite high predictive performance.

Dataset	Method	ROC-AUC	F1	Seed Jaccard	CV Jaccard
Synthetic noisy data	L1-based SelectFromModel	0.9310	0.8774	0.5074	0.4490
Synthetic noisy data	RFE (linear classifier)	0.9319	0.8776	0.5442	0.4690
Breast Cancer Wisconsin	RFE (linear classifier)	0.9926	0.9704	0.5467	0.3794



Digits	RFE (linear classifier)	0.9951	0.9242	0.6215	0.5195
Breast Cancer Wisconsin	L1-based SelectFromModel	0.9926	0.9727	0.6254	0.5867
Wine	RFE (linear classifier)	0.9957	0.9537	0.6487	0.6496

Table 4 illustrates the main risk considered in this paper. For example, on the synthetic noisy data, L1-based selection and RFE both produced ROC-AUC values above 0.93 but selected inconsistent feature subsets, with seed-level Jaccard similarities of 0.5074 and 0.5442, respectively. On Breast Cancer Wisconsin, RFE also achieved ROC-AUC above 0.99 while having a seed-level Jaccard similarity of only 0.5467.

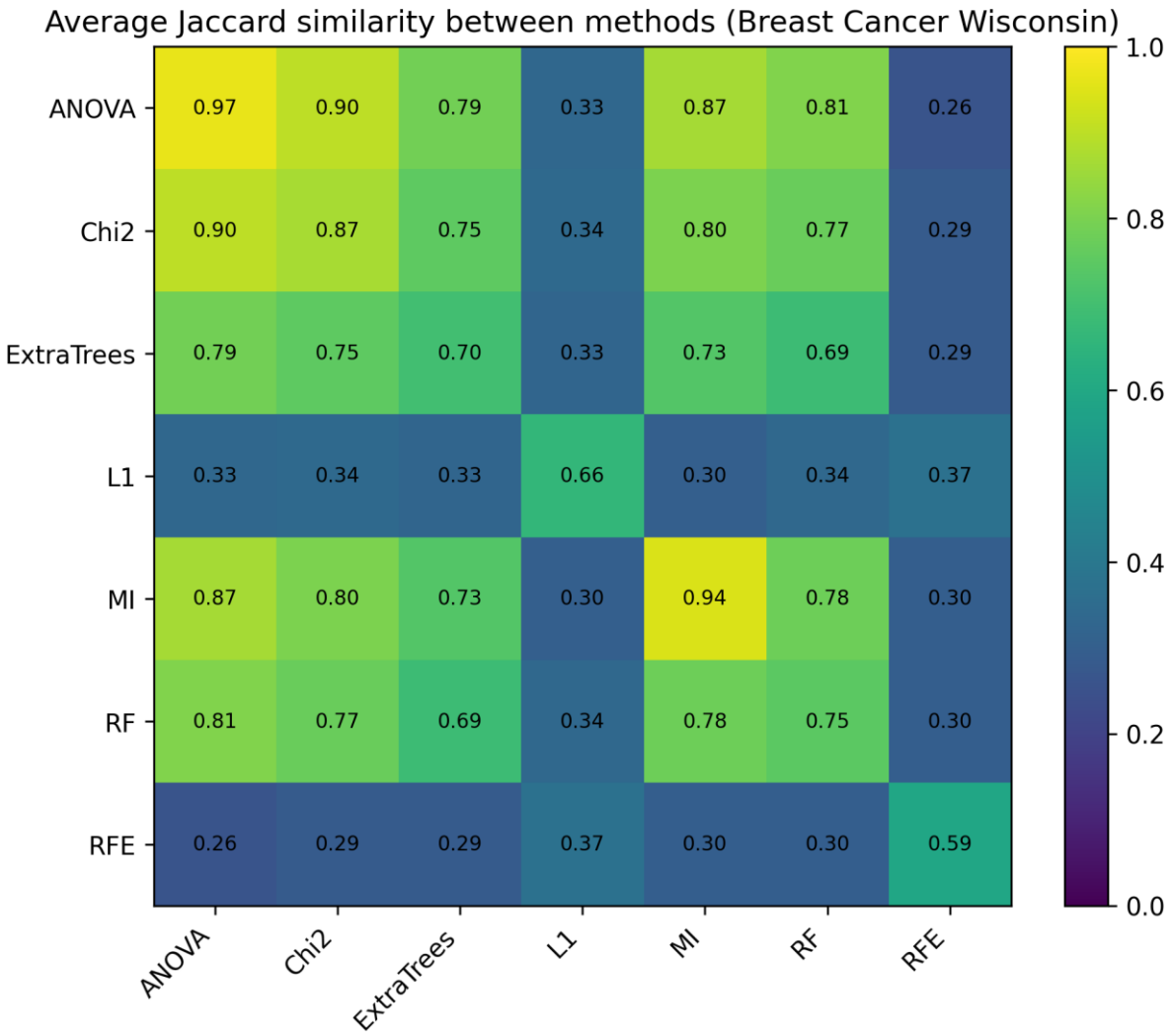


Figure 3. Heatmap of pairwise Jaccard similarity for Breast Cancer Wisconsin selected feature subsets.

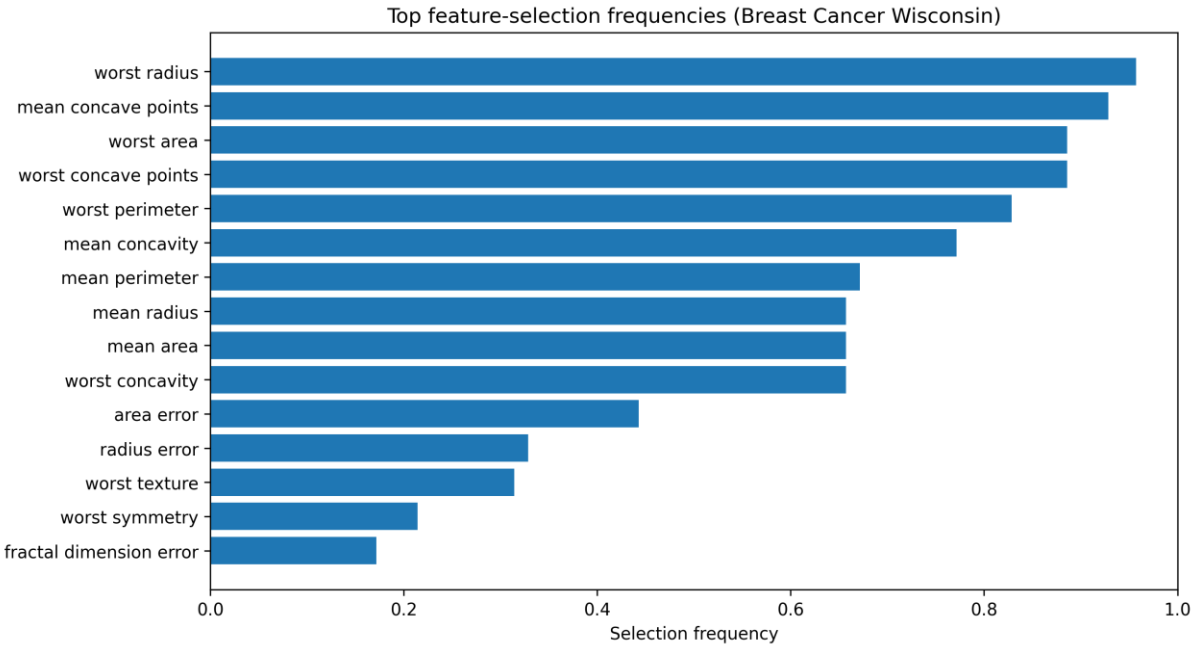


Figure 4. Feature-selection frequency for the Breast Cancer Wisconsin data set.

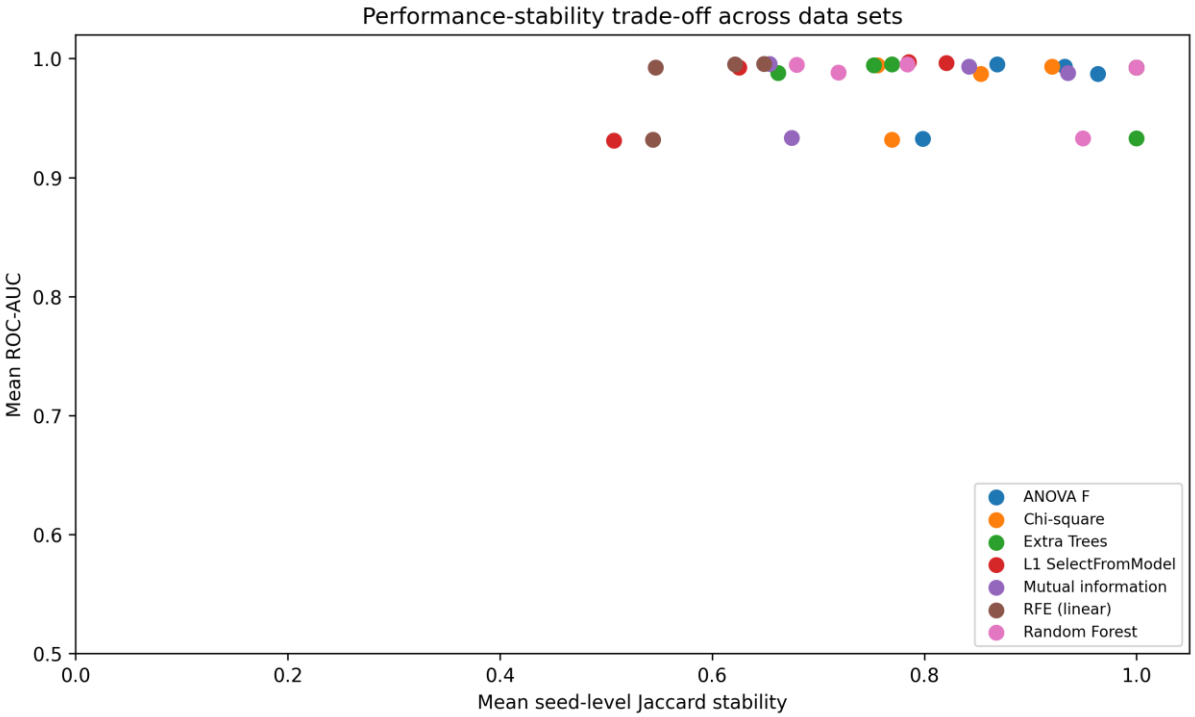


Figure 5. Relationship between ROC-AUC and seed-level Jaccard stability across method-data set combinations.

The feature-frequency chart demonstrates that some features are selected in nearly all

runs, while other features are selected only occasionally. In applied interpretation, high-frequency features can be considered more robust candidates than features that appear in only a small number of runs.

Table 5. Top Breast Cancer Wisconsin features by overall selection frequency.

Feature	Selected runs	Total runs	Frequency
worst radius	67	70	0.957
mean concave points	65	70	0.929
worst area	62	70	0.886
worst concave points	62	70	0.886
worst perimeter	58	70	0.829
mean concavity	54	70	0.771
mean perimeter	47	70	0.671
mean radius	46	70	0.657
mean area	46	70	0.657
worst concavity	46	70	0.657

5. Discussion

The experiment confirms that feature-selection stability should be reported alongside predictive metrics. The strongest example is the contrast between L1-based selection and ANOVA F-test. L1-based selection produced the best mean ROC-AUC and F1-score, but its selected feature subsets were less stable than those of ANOVA F-test and chi-square. Conversely, ANOVA F-test was not the best method by ROC-AUC rank, but it provided the most stable feature subsets across seeds.

This finding is important for interpretability. A model can be accurate because many correlated or redundant features carry similar information. When the data split changes,

a selector may choose a different member of the same correlated group. Prediction may remain strong, but the interpretation changes. In medical AI, this can lead to inconsistent conclusions about which measurements or biomarkers matter. Therefore, reporting only ROC-AUC can overstate the reliability of a selected feature set.

Filter methods such as ANOVA F-test and chi-square were relatively stable in this experiment because they score features individually and deterministically once the training data is fixed. Embedded and wrapper methods can exploit interactions and may improve performance, but they can also be more sensitive to sample variation, model regularization and correlated features. Tree-based methods were intermediate: Random Forest and Extra Trees importance provided good performance and moderate-to-high stability, but not the highest stability overall.

The near-zero Spearman correlations between performance and stability indicate that these criteria measure different aspects of feature-selection quality. For practical ML pipelines, this means that a model-selection process should consider at least two objectives: predictive utility and subset reproducibility. If the aim is interpretation, a slightly lower ROC-AUC may be acceptable when the selected features are substantially more stable.

6. Limitations

This study has several limitations. First, the real data sets are benchmark data sets rather than large-scale clinical or industrial data. Second, only one synthetic noisy data set was used; additional synthetic settings could vary the number of redundant features, class imbalance and noise level. Third, ten random seeds provide a useful stability estimate but do not exhaust all possible splits. Fourth, the downstream classifier was kept linear for consistency; nonlinear classifiers might change the performance-stability trade-off. Fifth, the experiment used Jaccard similarity as the main subset-overlap measure, while alternative stability measures such as Kuncheva or Nogueira indices could be added in future work.

7. Conclusion

This paper evaluated the stability of seven feature-selection algorithms under different random seeds, train/test splits and cross-validation folds. The results show that high accuracy or ROC-AUC does not automatically imply stable and interpretable feature selection. L1-based selection and RFE achieved strong predictive metrics but were less stable than simpler filter methods in several settings. ANOVA F-test was the most stable method by average seed-level Jaccard similarity, while tree-based importance methods achieved a balanced compromise between predictive performance and stability. For AI/ML research and especially for medical AI applications, feature-selection studies should report both predictive metrics and stability metrics. Jaccard similarity and feature-selection frequency charts are simple, reproducible tools for making this evaluation transparent.

Data and Code Availability

All code, generated CSV files, figures, Excel result tables, README and requirements are supplied with this manuscript package. Real data sets are loaded through scikit-learn; the SyntheticNoisy80 data set is generated with scikit-learn make_classification and is explicitly synthetic.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

Author Contributions

G.Kh. Tashmatov: conceptualization, methodology, software implementation, experimental analysis, visualization and manuscript preparation.

References

[1] Büyükkeçeci, M.; Okur, M.C. A Comprehensive Review of Feature Selection and

Feature Selection Stability in Machine Learning. Gazi University Journal of Science 2023, 36(4), 1506-1520. https://doi.org/10.35378/gujs.993763

[2] Khaire, U.M.; Dhanalakshmi, R. Stability of feature selection algorithm: A review. Journal of King Saud University - Computer and Information Sciences 2022, 34(4), 1060-1073. https://doi.org/10.1016/j.jksuci.2019.06.012

[3] Łukaszuk, T.; Krawczuk, J.; Żyła, K.; Kęsik, J. Stability of Feature Selection in Multi-Omics Data Analysis. Applied Sciences 2024, 14(23), 11103. https://doi.org/10.3390/app142311103

[4] Lazebnik, T.; Rosenfeld, A. A new definition for feature selection stability analysis. Annals of Mathematics and Artificial Intelligence 2024, 92, 753-770. https://doi.org/10.1007/s10472-024-09936-8

[5] Barbieri, M.C.; Grisci, B.I.; Dorn, M. Analysis and comparison of feature selection methods towards performance and stability. Expert Systems with Applications 2024, 249, 123667. https://doi.org/10.1016/j.eswa.2024.123667

[6] Gundersen, O.E.; Coakley, K.; Kirkpatrick, C.; Gil, Y. Sources of Irreproducibility in Machine Learning: A Review. arXiv 2022. https://doi.org/10.48550/arXiv.2204.07610

[7] Amir, S.; van de Meent, J.-W.; Wallace, B.C. On the Impact of Random Seeds on the Fairness of Clinical Classifiers. arXiv 2021. https://doi.org/10.48550/arXiv.2104.06338

[8] Williams, N.; Schuler, A. Improving reproducibility by controlling random seed stability in machine learning. arXiv 2026. https://doi.org/10.48550/arXiv.2604.17694

[9] de Hond, A.A.H.; Leeuwenberg, A.M.; Hooft, L.; Kant, I.M.J.; Nijman, S.W.J.; van Os, H.J.A.; Aardoom, J.J.; Debray, T.P.A.; Schuit, E.; van Smeden, M.; Reitsma, J.B.; Steyerberg, E.W.; Chavannes, N.H.; Moons, K.G.M. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a

scoping review. npj Digital Medicine 2022, 5, 2. https://doi.org/10.1038/s41746-021-00549-7

[10] Scikit-learn Developers. Feature selection. scikit-learn User Guide, version 1.9.0 documentation. Accessed 2 July 2026. https://scikit-learn.org/stable/modules/feature_selection.html

[11] Scikit-learn Developers. sklearn.pipeline.Pipeline. scikit-learn API Reference, version 1.9.0 documentation. Accessed 2 July 2026. https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

[12] Scikit-learn Developers. Toy data sets: load_breast_cancer, load_wine, load_iris and load_digits. scikit-learn API Reference, version 1.9.0 documentation. Accessed 2 July 2026. https://scikit-learn.org/stable/datasets/toy_dataset.html

[13] MDPI. Instructions for Authors: Information. Accessed 2 July 2026. https://www.mdpi.com/journal/information/instructions

[14] Jiang, Y.; Wang, X.; Liang, C.; Jiang, B.; Li, H. Utilizing stability criteria in choosing feature selection methods yields reproducible results in microbiome data. arXiv 2020. https://doi.org/10.48550/arXiv.2012.00001

[15] Nogueira, S.; Sechidis, K.; Brown, G. On the stability of feature selection algorithms. Journal of Machine Learning Research 2018, 18(174), 1-54.

[16] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011, 12, 2825-2830.