

DOI: <https://doi.org/10.38035/gcir.v1i2><https://creativecommons.org/licenses/by/4.0/>

Received: 01 September 2025 | Revised: 10 September 2025 | Accepted: 20 September 2025

The Impact of Feature Selection on Machine Learning Model Performance

Jaja Miharja¹¹Institut Teknologi dan Bisnis Dewantara, Bogor, Indonesia, jaja.miharja@dewantara.ac.idCorresponding Author: jaja.miharja@dewantara.ac.id¹

Abstract: Feature selection can reduce input dimensionality and improve interpretability, but its effect on predictive performance depends on the data and learning algorithm. This study experimentally compares three classifiers Logistic Regression, Support Vector Machine (SVM) with a radial basis function kernel, and Random Forest under three feature configurations using the Wisconsin Breast Cancer Diagnostic dataset. A stratified five-fold cross-validation design was used, with scaling and feature selection fitted only within each training fold to reduce data leakage. The conditions included all 30 predictors, the 10 highest-ranked predictors by ANOVA F-score, and the 10 highest-ranked predictors by mutual information. Mean accuracy, precision, recall, F1-score, and ROC-AUC were recorded, treating malignant cases as the positive class. SVM with all features achieved the highest mean accuracy (97.72%) and F1-score (0.9688), while Logistic Regression with all features achieved the highest ROC-AUC (0.9953). Reducing the input to 10 features lowered mean accuracy for all three classifiers, although the magnitude of the change varied by model. The findings indicate that feature selection should be evaluated jointly with the classifier and the intended performance objective rather than assumed to improve prediction automatically. Because the experiment uses one benchmark dataset, the results are methodological evidence rather than clinical validation.

Keywords: Feature Selection, Machine Learning, Data Science, Classification, Model Evaluation

INTRODUCTION

Machine learning has become a standard component of contemporary data science, supporting classification, prediction, pattern discovery, and decision support across scientific and industrial domains. In supervised learning, a model learns a mapping between input variables and a target outcome from labeled examples. Model performance, however, depends not only on the algorithm but also on how the data are represented, measured, and prepared. Irrelevant or redundant predictors can increase computational cost, complicate interpretation, and sometimes impair generalization. Feature selection addresses this problem by retaining a subset of the original variables while discarding others (Guyon & Elisseeff, 2003; Chandrashekar & Sahin, 2014).

Feature selection is related to, but distinct from, feature extraction. Feature extraction transforms existing measurements into a new representation, whereas feature selection preserves selected original variables. This distinction matters in applied data science because original variables often have direct meaning for domain specialists. A smaller feature set may also reduce measurement or storage requirements when variables are costly to obtain. Nevertheless, fewer variables do not automatically produce better predictions: a discarded feature may contain complementary information, and different classifiers respond differently to redundancy, scaling, nonlinear relationships, and feature interactions (Saeys et al., 2007; Pudjihartono et al., 2022).

Feature selection methods are often grouped into filter, wrapper, and embedded families. Filter methods score features using data-driven criteria before fitting a predictive model. ANOVA F-score evaluates univariate differences in target-associated means and is commonly used for numeric predictors in classification. Mutual information estimates statistical dependence between a feature and the target and can capture nonlinear dependencies, although its estimation can be sensitive to sample size and implementation choices. Wrapper methods evaluate subsets through model performance, while embedded methods perform selection as part of model training (Kohavi & John, 1997; Yu & Liu, 2004; Büyükkeçeci & Okur, 2023).

An important methodological risk is information leakage. If features are ranked using the entire dataset before cross-validation, information from a held-out fold influences the feature set used to evaluate that fold. The resulting performance estimate can be optimistic because the test fold is no longer independent of the model-selection process. Feature selection should therefore be placed inside the validation pipeline and fitted only on the training partition of each fold. The broader literature on model selection and cross-validation emphasizes that evaluation procedures must reproduce the separation between model development and assessment (Cawley & Talbot, 2010; Varma & Simon, 2006).

Breast cancer diagnostic classification provides a useful benchmark for demonstrating this methodological issue because the Wisconsin Breast Cancer Diagnostic dataset contains a moderate number of samples, numeric predictors, and a binary outcome. The features are computed from digitized images of fine-needle aspirates, and the task is to distinguish malignant from benign cases (Street et al., 1993). Published studies have compared machine learning algorithms and feature-selection strategies on breast cancer data, showing that reported performance can depend on the algorithm, selected variables, and experimental protocol (Aamir et al., 2022; Sharma & Mishra, 2022; Strelcenia & Prakoonwit, 2023). Those reports motivate careful experimental design, but benchmark accuracy alone should not be interpreted as evidence of clinical readiness.

Despite extensive work on feature selection, the practical question remains conditional: how much predictive performance is retained when a subset of features is used, and does the answer vary by classifier and selection criterion? A useful comparison should hold the dataset and validation design constant while varying the feature configuration. It should also report more than accuracy because a single aggregate score may hide changes in sensitivity, precision, or ranking performance. ROC-AUC measures ranking discrimination across thresholds, whereas precision and recall describe the consequences of a chosen classification threshold (Fawcett, 2006; Saito & Rehmsmeier, 2015).

This study addresses three research questions. RQ1: How does selecting 10 features affect the predictive performance of Logistic Regression, SVM, and Random Forest relative to using all 30 features? RQ2: Do ANOVA F-score and mutual information produce different performance patterns? RQ3: What trade-off emerges between dimensionality reduction and classification performance on this benchmark dataset? The study compares nine model-configuration combinations using stratified five-fold cross-validation, with feature selection and scaling contained within each pipeline. The contribution is a transparent and reproducible

benchmark of a defined experimental setting, including results across several complementary metrics and an explicit discussion of limitations.

METHODS

Research Design

This study uses a quantitative experimental design. The independent experimental factor is the feature configuration: all 30 original predictors, the 10 highest-ranked predictors by ANOVA F-score, or the 10 highest-ranked predictors by mutual information. The classifier is a second factor with three levels: Logistic Regression, SVM with an RBF kernel, and Random Forest. The outcomes are cross-validated accuracy, precision, recall, F1-score, and ROC-AUC. The experiment is intended to compare configurations under a controlled protocol rather than to develop a clinical decision system.

Dataset and Target Definition

The Wisconsin Breast Cancer Diagnostic dataset was loaded using scikit-learn's `load_breast_cancer` interface. It contains 569 observations and 30 numeric features derived from image measurements; 212 observations are malignant and 357 are benign. To make the evaluation clinically interpretable at the metric level, the target was recoded so that malignant cases are positive (1) and benign cases are negative (0). This explicit recoding means that precision, recall, and F1-score quantify performance for malignant cases. The dataset is a public benchmark, and no new patient data were collected. Its use here does not imply external validation or suitability for clinical decisions.

Data Preparation and Feature-Selection Methods

Logistic Regression and SVM were preceded by `StandardScaler` because these models can be sensitive to predictor scale. Random Forest was not scaled because tree-based splitting is not dependent on linear scaling in the same way. For the two feature-selection conditions, `SelectKBest` retained 10 variables. ANOVA F-score ranked each numeric feature using a univariate F-test, while mutual information estimated statistical dependence between each feature and the target. Mutual information estimation used a fixed `random_state` of 42. The selector was fitted separately inside each training fold, avoiding feature-ranking leakage from held-out observations.

Learning Algorithms

Logistic Regression provides a regularized linear decision boundary and a probabilistic output. The implementation used the scikit-learn default L2 regularization with `max_iter=5000` and `random_state=42`. SVM used the radial basis function kernel, which can represent nonlinear decision boundaries; `probability=True` was enabled to obtain probabilities for ROC-AUC calculation. Random Forest used 300 trees and `random_state=42`, with parallel tree fitting enabled. Hyperparameters were held fixed across feature configurations to isolate the effect of feature selection rather than to optimize each configuration independently. This choice limits the scope of the conclusions: different hyperparameter tuning strategies could alter relative performance.

Validation Protocol

Performance was estimated using stratified five-fold cross-validation with shuffling and `random_state=42`. In each iteration, four folds were used to fit the preprocessing, selector (when present), and classifier, while the remaining fold was held out for evaluation. Stratification preserved approximately the same class proportions across folds. Mean scores and sample standard deviations across the five held-out folds were recorded. The folds are not

independent replications in the strict statistical sense, so the standard deviations are descriptive measures of fold-to-fold variation rather than formal confidence intervals.

Evaluation Metrics

Accuracy is the fraction of correct predictions. Precision is the fraction of predicted malignant cases that are malignant, while recall is the fraction of malignant cases correctly identified. F1-score is the harmonic mean of precision and recall. ROC-AUC evaluates the ability to rank malignant cases above benign cases across classification thresholds. These metrics were selected because a useful comparison should consider both threshold-dependent classification performance and threshold-independent ranking performance. Precision and recall can move in different directions, so the preferred model depends on the costs of false positives and false negatives in the intended application.

Implementation and Reproducibility

The experiment was run in Python with scikit-learn version 1.8.0. The fixed random seed was 42 for the cross validation splitter, mutual-information estimation, and stochastic estimators. The main data-processing steps were implemented as scikit-learn Pipeline objects, so feature ranking and scaling were fitted only within each training fold. The generated result table contains mean metrics and standard deviations across folds. Hardware and operating system details were not controlled as a benchmark and are therefore not used to make claims about runtime or computational efficiency improvements.

Ethical Considerations

The analysis uses an openly available, de-identified benchmark dataset and does not involve new human-participant recruitment or access to identifiable clinical records. The dataset is used solely for methodological evaluation. No recommendation for diagnosis, screening, or treatment is made, and any clinical deployment would require separate validation, clinical governance, fairness assessment, and regulatory review.

RESULTS AND DISCUSSION

This section reports the observed cross-validation results and interprets them in relation to the three research questions. The analysis is restricted to the dataset, classifiers, feature-selection methods, and fixed validation protocol described above.

Dataset profile

The dataset included 569 cases, with 357 benign cases (62.7%) and 212 malignant cases (37.3%). It contains 30 numeric predictors. Selecting 10 features therefore reduces the input dimensionality by 66.7%. This reduction is substantial in feature-count terms, but it should not be treated as a direct measure of training-time savings because runtime was not systematically benchmarked.

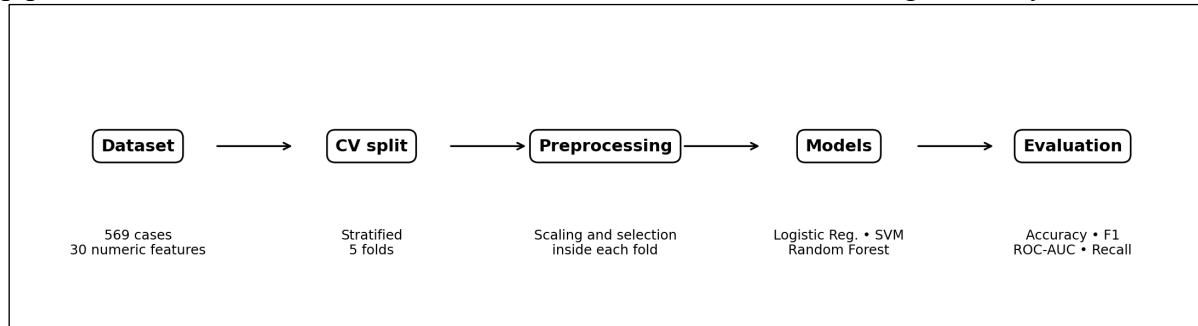
Table 1. Dataset characteristics

Characteristic	Value
Observations	569
Original predictors	30 numeric features
Benign cases	357 (62.7%)
Malignant cases	212 (37.3%)
Selected subset size	10 features (66.7% reduction)
Validation	Stratified 5-fold cross-validation

Source: Research Result

Experimental Workflow

The experimental workflow separates data partitioning, fold-specific preprocessing, feature selection, model fitting, and held-out evaluation. The workflow was implemented as a pipeline to ensure that the selected variables were derived from training data only.



Source: Research Result

Figure 1. Experimental workflow for the feature-selection comparison

Overall Predictive Performance

Table 2 summarizes mean performance across the five held-out folds. The SVM with all 30 features achieved the highest mean accuracy (0.9772, or 97.72%) and the highest mean F1-score (0.9688). Logistic Regression with all features achieved the highest mean ROC-AUC (0.9953). These are close results, and differences should not be overstated: the experiment did not perform a formal hypothesis test, and the standard deviations show that fold-level estimates vary.

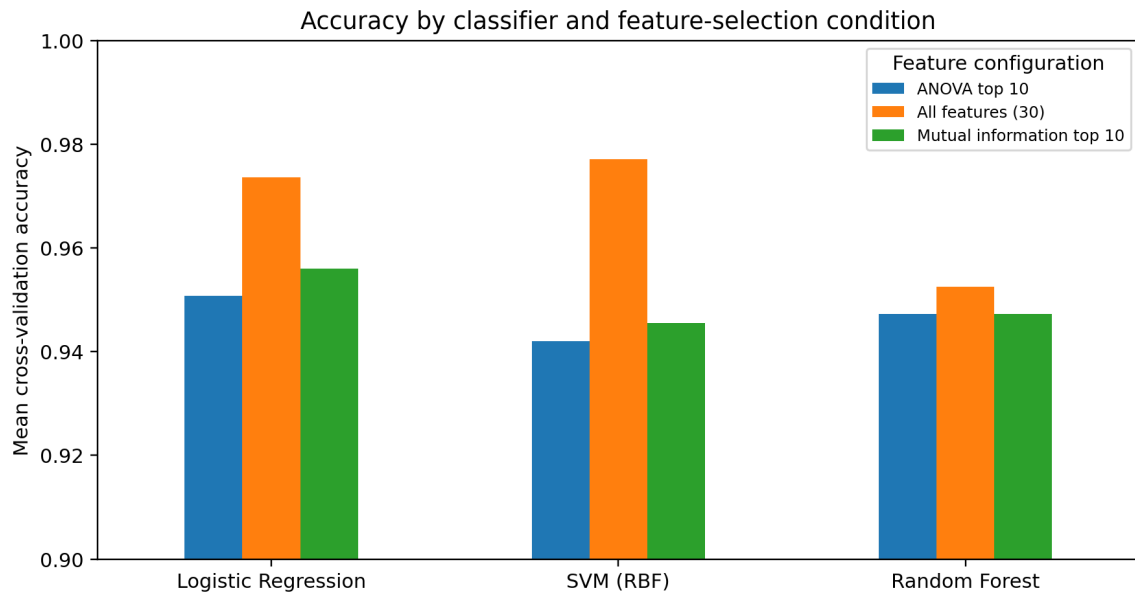
Table 2. Cross-Validated Performance by Classifier and Feature Configuration

Classifier	Feature configuration	n	Accuracy (mean ± SD)		Precision	Recall	F1-score	ROC-AUC
Logistic Regression	All features (30)	30	0.9737	± 0.0186	0.9863	0.9436	0.9633	0.9953
Logistic Regression	ANOVA top 10	10	0.9508	± 0.0181	0.9483	0.9200	0.9322	0.9893
Logistic Regression	Mutual information top 10	10	0.9561	± 0.0175	0.9627	0.9197	0.9396	0.9914
SVM (RBF)	All features (30)	30	0.9772	± 0.0182	0.9810	0.9576	0.9688	0.9945
SVM (RBF)	ANOVA top 10	10	0.9420	± 0.0281	0.9550	0.8868	0.9183	0.9850
SVM (RBF)	Mutual information top 10	10	0.9455	± 0.0258	0.9623	0.8915	0.9234	0.9887
Random Forest	All features (30)	30	0.9526	± 0.0146	0.9461	0.9295	0.9355	0.9895
Random Forest	ANOVA top 10	10	0.9473	± 0.0255	0.9344	0.9248	0.9283	0.9850
Random Forest	Mutual information top 10	10	0.9473	± 0.0248	0.9352	0.9247	0.9288	0.9848

Source: Research Result

Accuracy Comparison

Figure 2 visualizes mean accuracy across all nine configurations. The full-feature condition was best for each classifier in this experiment. For Logistic Regression, mean accuracy decreased from 0.9737 to 0.9508 with ANOVA selection and 0.9561 with mutual information. For SVM, accuracy decreased from 0.9772 to 0.9420 and 0.9455, respectively. Random Forest showed smaller differences: 0.9526 with all features and 0.9473 with either reduced-feature condition.



Source: Research Result

Figure 2. Mean accuracy for each classifier and feature-selection condition.

RQ1: Effect of selecting 10 features

The observed results do not support the assumption that reducing the feature count necessarily improves predictive performance. The best mean accuracy was achieved using all 30 features for each of the three classifiers. The absolute accuracy reduction for Logistic Regression was 2.29 percentage points with ANOVA and 1.76 percentage points with mutual information. For SVM, the corresponding reductions were 3.52 and 3.17 percentage points. For Random Forest, both reduced-feature conditions were 0.53 percentage points below the all-feature score.

The larger decline for SVM suggests that the removed predictors contained useful information for the nonlinear decision boundary learned by this model under the selected settings. However, this interpretation is tentative: the experiment does not identify causal feature contributions, and a different subset size or tuning strategy could change the result. Random Forest was comparatively stable in accuracy across feature configurations, which may reflect its ability to learn useful splits from a reduced set, but the differences are small enough that they should be interpreted descriptively.

These findings are consistent with the general principle that feature selection is a model-dependent operation. Feature selection can improve performance when it removes noisy or redundant predictors, but it can also reduce the information available to a classifier. The review by Büyükkeçeci and Okur (2023) highlights that stability and selection method are important aspects of feature selection, beyond the single best predictive score. In a methodological comparison, it is therefore useful to report both retained performance and the number of variables removed.

RQ2: Differences between ANOVA and mutual information

Mutual information outperformed ANOVA in mean accuracy for Logistic Regression (0.9561 versus 0.9508) and SVM (0.9455 versus 0.9420). For Random Forest, both selection methods achieved the same rounded accuracy of 0.9473. The results indicate a modest advantage for mutual information in two of the three model comparisons, but not a universal advantage across all classifiers or metrics.

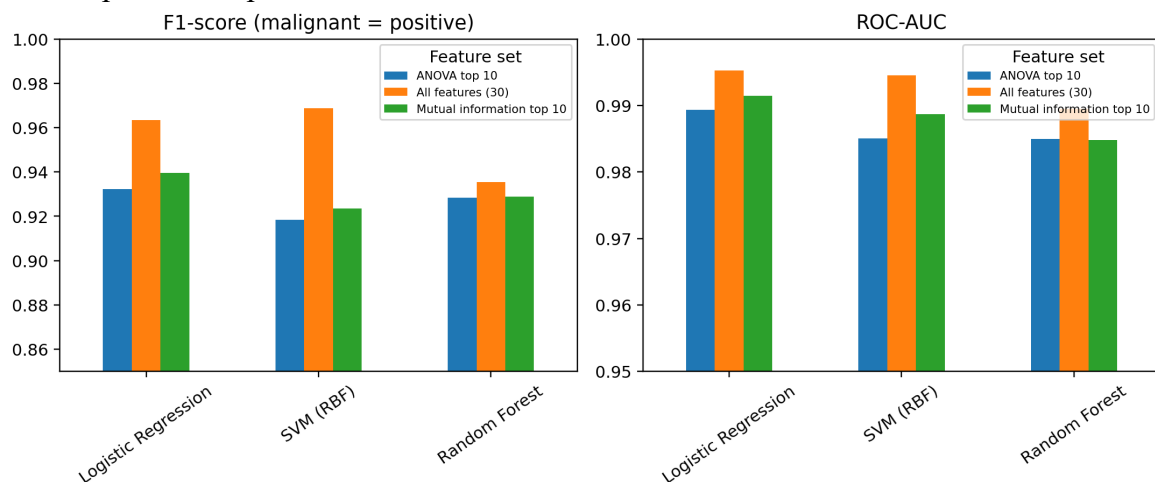
ANOVA F-score evaluates a univariate relationship based on class-associated means and can be effective when a feature has a strong mean difference between classes. Mutual information measures statistical dependence and can capture relationships that are not limited to linear mean differences, although estimation quality depends on sample size and data

characteristics. Neither method directly models all interactions among variables, and both may select correlated features that provide overlapping information. Thus, the best filter method cannot be chosen from its name alone; it should be evaluated inside the full learning pipeline.

Precision, recall, F1-score, and ROC-AUC

The metric comparison adds nuance to the accuracy results. SVM with all features had mean recall of 0.9576 and F1-score of 0.9688 for malignant cases. Logistic Regression with all features had slightly lower recall (0.9436) and F1-score (0.9633), but a higher ROC-AUC (0.9953 compared with 0.9945 for SVM). This difference illustrates that ranking ability across thresholds and performance at a particular threshold are related but distinct. A model may rank cases very well while making a somewhat different set of hard classifications at its default threshold.

Under ANOVA selection, SVM recall declined to 0.8868, which was the lowest recall among the nine configurations. In this dataset, that indicates a larger proportion of malignant cases were missed at the classifier's decision threshold. For a hypothetical diagnostic-support context, missed malignant cases could be more consequential than an equivalent number of false alarms; however, this study does not establish an operating threshold or clinical utility. Any application should choose thresholds according to validated clinical requirements and should report class-specific confusion matrices and calibration where relevant.

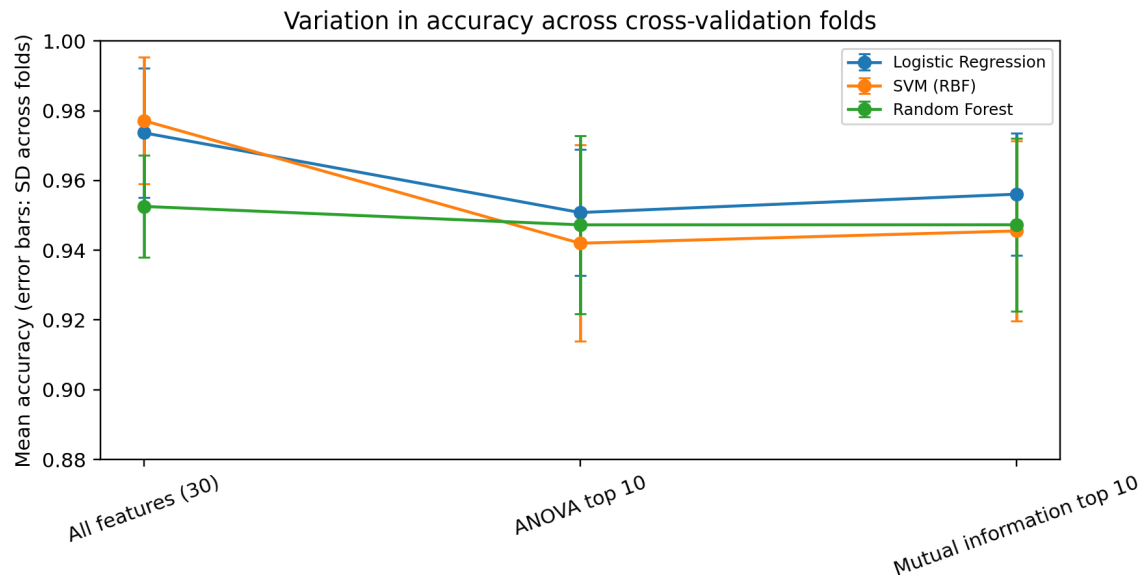


Source: Research Result

Figure 3. Mean F1-score and ROC-AUC across model configurations.

Variation Across Folds

Figure 4 displays mean accuracy with the sample standard deviation across the five validation folds. SVM with all features achieved high mean accuracy but still showed fold-to-fold variation ($SD = 0.0182$). Random Forest with all features had $SD = 0.0146$, the smallest accuracy standard deviation among the all-feature conditions. The reduced-feature SVM configurations showed larger variability ($SD = 0.0281$ for ANOVA and 0.0258 for mutual information), suggesting that their performance was more sensitive to which observations were held out in this particular split design.



Source: Research Result

Figure 4. Mean accuracy and fold-to-fold standard deviation.

Implications for Data Science Practice

For practitioners, the key implication is to treat feature selection as a decision that should be justified empirically. A two-thirds reduction in feature count may be attractive when data collection is costly or interpretation is important, but it should be accepted only if the resulting model meets the task's performance requirements. In this experiment, the reduced models did not preserve the full-feature accuracy, and the recall reduction for SVM with ANOVA was notable. A practitioner might still choose a reduced model for a separate operational reason, but that decision would require an explicit trade-off analysis.

The findings also reinforce the importance of leakage-resistant evaluation. A feature-ranking procedure performed before cross-validation would use information from all observations, including those later treated as held out. Placing the selector inside the pipeline helps ensure that each test fold remains unseen during both feature selection and model fitting. This is especially important when many features, selection methods, or hyperparameter combinations are being explored, because repeated model selection can produce optimistic estimates if evaluation is not nested or otherwise separated from tuning (Cawley & Talbot, 2010; Varma & Simon, 2006).

The study's novelty is not the claim that one feature-selection method is universally superior. Rather, it provides a controlled, reproducible demonstration that the effect of feature selection is conditional on the classifier and metric, and that dimensionality reduction can come with a measurable performance cost. This is a useful caution for applied data science workflows that treat feature selection as an automatic preprocessing step. The final choice should balance predictive quality, feature acquisition cost, interpretability, robustness, and the consequences of different types of errors.

Limitations

Several limitations constrain the conclusions. First, only one public dataset was used, so the results cannot be generalized to unrelated domains such as finance, education, manufacturing, or time-series prediction. Second, only one reduced subset size (10 features) and two filter methods were evaluated. Wrapper methods such as recursive feature elimination and embedded methods such as L1-regularized models or tree-based selection may produce different outcomes. Third, model hyperparameters were fixed rather than optimized separately within a nested validation design. Fourth, five folds provide a compact estimate of generalization but not a formal measure of statistical significance; the fold scores are also not

fully independent. Finally, computational time, feature-selection stability, fairness, calibration, and decision-analytic utility were not measured. The results should therefore be understood as a controlled benchmark comparison, not a complete evaluation of feature selection in practice.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion

This study experimentally compared Logistic Regression, SVM, and Random Forest with all 30 predictors, ANOVA-selected top 10 features, and mutual-information-selected top 10 features on the Wisconsin Breast Cancer Diagnostic dataset. Under stratified five-fold cross-validation, SVM with all features achieved the highest mean accuracy (97.72%) and F1-score (0.9688), while Logistic Regression with all features achieved the highest ROC-AUC (0.9953). Selecting 10 features reduced mean accuracy in all three classifiers, although the magnitude of the reduction differed. Mutual information modestly outperformed ANOVA in accuracy for Logistic Regression and SVM, while the two selectors produced the same rounded Random Forest accuracy. The results answer the research questions by showing that the effect of feature selection is classifier-dependent, metric-dependent, and not guaranteed to be positive.

Recommendations

Future research should replicate this comparison on multiple datasets from distinct domains and evaluate several subset sizes rather than only top 10 features. Filter methods should be compared with wrapper and embedded approaches, including recursive feature elimination, sequential selection, and model-based selection. Where model tuning is conducted, nested cross-validation or an independent test set should be used to reduce selection bias. Studies should report fold-level distributions, class-specific metrics, confusion matrices, feature-selection stability, and controlled runtime measurements. For high-stakes domains, evaluation should also include calibration, subgroup performance, threshold selection, and domain-expert review before any real-world deployment.

Author Contributions

The author was responsible for conceptualization, methodology, implementation of the experiment, data analysis, interpretation of results, and preparation of the manuscript. The author should review and approve the final submitted version.

AI Disclosure Statement

AI-assisted tools were used to support drafting, language refinement, and code preparation. The author is responsible for independently checking the experimental implementation, verifying all references, reviewing the interpretation of results, and approving the final manuscript. This statement should be updated to match the journal's disclosure policy and the actual use of AI tools.

Acknowledgment

No external funding or institutional support is claimed in this draft. Add any applicable funding, institutional, or technical acknowledgments before submission.

REFERENCES

- Aamir, S., et al. (2022). Predicting breast cancer leveraging supervised machine learning techniques. *Computational and Mathematical Methods in Medicine*, 2022, Article 5869529. <https://doi.org/10.1155/2022/5869529>
- Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143, 106839. <https://doi.org/10.1016/j.csda.2019.106839>

- Büyükkeçeci, M., & Okur, M. C. (2023). A comprehensive review of feature selection and feature selection stability in machine learning. *Gazi University Journal of Science*, 36(4), 1506–1520. <https://doi.org/10.35378/gujs.993763>
- Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107. <https://www.jmlr.org/papers/v11/cawley10a.html>
- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28. <https://doi.org/10.1016/j.compeleceng.2013.11.024>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, Article 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182. <https://www.jmlr.org/papers/v3/guyon03a.html>
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X)
- Kuzudisli, C., Bakir-Gungor, B., Bulut, N., Qaqish, B. F., & others. (2023). Review of feature selection approaches based on grouping of features. *PeerJ*, 11, e15666. <https://doi.org/10.7717/peerj.15666>
- 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. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://www.jmlr.org/papers/v12/pedregosa11a.html>
- Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O’Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2, Article 927312. <https://doi.org/10.3389/fbinf.2022.927312>
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517. <https://doi.org/10.1093/bioinformatics/btm344>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Sharma, A., & Mishra, P. K. (2022). Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *International Journal of Information Technology*, 14, 1603–1611. <https://doi.org/10.1007/s41870-021-00671-5>
- Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology* (Vol. 1905, pp. 861–870). SPIE. <https://doi.org/10.1117/12.148698>
- Strelcenia, E., & Prakoonwit, S. (2023). Effective feature engineering and classification of breast cancer diagnosis: A comparative study. *BioMedInformatics*, 3(3), 42. <https://doi.org/10.3390/biomedinformatics3030042>

- Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85, 189–203. <https://doi.org/10.1016/j.jbi.2018.07.014>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. <https://doi.org/10.1186/1471-2105-7-91>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5, 1205–1224. <https://www.jmlr.org/papers/v5/yl04a.html>
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 532–538). Springer. https://doi.org/10.1007/978-0-387-39940-9_565