Feature Selection for High-Dimensional Noisy Data

A Comparative Study of Filter, Wrapper, Embedded, and Hybrid Approaches on the NIPS 2003 Madelon Dataset

K. Janith Chathuranga Mahanama
Dept. of CS & Engineering
University of Moratuwa
Sri Lanka
janith.23@cse.mrt.ac.lk

***Abstract*—Feature selection is a critical preprocessing step in machine learning pipelines, particularly for high-dimensional datasets where many features are irrelevant or redundant. This study presents a systematic investigation of feature selection methods applied to the NIPS 2003 Feature Selection Challenge Madelon dataset: an artificial benchmark designed with 500 features, of which only 20 are informative (5 truly relevant, 15 redundant combinations) and the remaining 480 are pure noise. This study implements and compares four families of feature selection: Filter methods (ANOVA F-score, Mutual Information, Pearson Correlation), Wrapper methods (Recursive Feature Elimination, Sequential Forward Selection), Embedded methods (Random Forest and Extra Trees importance, L1-regularized models), and Hybrid multi-stage pipelines. All experiments are based solely on classical machine learning methods (no deep learning). Evaluation is performed on a held-out validation set using Balanced Accuracy and Binary Error Rate (BER) as primary metrics, with AUC-ROC as a secondary measure. Results demonstrate that all feature selection approaches substantially outperform the no-selection baseline (Balanced Accuracy 0.5817), with the best Hybrid approach achieving Balanced Accuracy 0.8900 (BER 0.1100, AUC 0.9657), representing a gain of +0.3083 over the baseline. A clear performance hierarchy emerges: Hybrid $>$ Embedded $>$ Wrapper $>$ Filter $>$ No Selection, with tree based feature ranking consistently emerging as the most effective selection signal.**

***Index Terms*—feature selection, dimensionality reduction, MADELON, NIPS 2003, filter methods, recursive feature elimination, random forest, hybrid pipelines, high dimensional data**

I. Introduction

High dimensional supervised learning suffers acutely from the curse of dimensionality [1]. When a large fraction of input features are irrelevant to the target, models overfit noise, distance based algorithms lose discriminative power, and even regularized linear models may fail to converge to a well generalized solution. Feature selection, the process of identifying and retaining only the most predictive input variables, directly addresses these problems without requiring architecture changes to the learning algorithm itself.

The NIPS 2003 Feature Selection Challenge [2] provides a carefully controlled benchmark for studying this problem. Its Madelon dataset is an artificial binary classification task with 500 continuous input features. By construction, only 5 features carry true discriminative information, 15 further features are linear combinations of the true features (introducing redundancy), and the remaining 480 dimensions are Gaussian noise. The task is therefore not merely to classify accurately, but to recover a small informative subset from a severely contaminated feature space.

This study investigates four established strategies for feature selection in this setting. *Filter methods* score features using univariate statistics independent of any classifier. *Wrapper methods* evaluate feature subsets by wrapping a specific learner. *Embedded methods* extract selection information as a byproduct of model fitting. *Hybrid methods* compose elements from the above families into multi-stage pipelines that trade computational cost for selection quality.

The primary contributions of this paper are as follows:

- A systematic sweep of three filter statistics (ANOVA F-score, Mutual Information, Pearson Correlation) across nine feature count values ($k = 5$ to 500) to empirically identify the optimal cutoff and characterise performance vs dimensionality trade offs.
- A cost-performance analysis comparing wrapper and embedded configurations, revealing that computational investment alone does not predict selection quality.
- A novel model aligned hybrid pipeline (Random Forest importance pre-ranking followed by direct Random Forest classification) that achieves the best performance among classical ML methods: Bal. Acc. 0.8900, BER 0.1100, AUC 0.9657.
- Empirical evidence for a consistent performance hierarchy (Hybrid $>$ Embedded $>$ Wrapper $>$ Filter $>$ No Selection) across all family types, with cross-validation confirming the robustness of the leading pipeline.

The rest of this paper is structured as follows. Section II reviews related work. Section III describes the dataset and experimental setup. Section IV describes all implemented methods and justifies the final recommendation. Section V presents all experimental results. Section VI interprets the findings and proposes improvements. Section VII concludes.

II. Related Work

A. The NIPS 2003 Feature Selection Challenge

Guyon et al. [2] introduced five benchmark datasets to probe feature selection algorithms under controlled conditions: Arcene (mass spectrometry), Dexter (text classification),

Dorothea (drug discovery), Gisette (digit recognition), and Madelon (artificial). Madelon is the most widely studied because its ground truth feature structure is fully known and reproducible. The challenge used BER as the primary metric and attracted many participating teams using a wide range of selection methods, from simple correlation filters to complex SVM based recursive elimination and Bayesian feature selection approaches. Top results achieved BER in the range 0.06 - 0.11 using Bayesian neural networks. This study targets this performance band using only classical machine learning methods [10] without deep learning and expensive computational resources.

B. Feature Selection Taxonomy

Guyon and Elisseeff [3] provide the canonical taxonomy dividing feature selection into filter, wrapper, and embedded families, later extended to include hybrid approaches by Liu and Motoda [11]. Filter methods such as ANOVA F-score and mutual information [12] are fast and model agnostic but ignore feature interactions and assume each feature can be evaluated independently. Wrapper methods [4], including RFE [5] and Sequential Feature Selection (SFS), use a learner as a black box to evaluate candidate subsets, capturing feature interactions at the cost of repeated model fitting. Embedded methods such as L1 regularization [6] and tree based feature importance [7] integrate selection into the learning objective, offering a middle ground between the speed of filters and the interaction awareness of wrappers.

C. Tree Based Feature Importance

Random Forest [7] computes mean decrease in impurity (MDI) as a by-product of training, producing a feature importance ranking at near zero additional cost. Subsequent work has demonstrated MDI is biased toward high cardinality continuous features in mixed type datasets [14], but this bias is minimal in the controlled Madelon setting where all features are continuous and drawn from the same marginal distribution. Extra Trees [8] use random splits rather than optimal splits, further reducing variance in importance estimates at the cost of slight bias increase. XGBoost [9] extends gradient boosted trees with histogram based binning, providing an additional importance signal evaluated in hybrid pipelines and demonstrating strong performance on structured tabular data.

D. Hybrid and Multi Stage Selection

Hybrid approaches have been shown to outperform single family methods in several high dimensional domains including genomics and text classification [11]. The typical strategy is a two-stage pipeline: a fast filter or embedded method pre selects a candidate set of M features ($M \ll p$, the original dimensionality), after which an expensive wrapper or evaluator refines to a final set of k features ($k < M$). This strategy is motivated by the observation that wrappers scale poorly with input dimensionality (runtime is $O(p \cdot t_{\text{fit}})$ for RFE) but excel at fine grained discrimination within a smaller candidate pool. Prior work [3], [4] has explored filter to wrapper pipelines. This study extends this to include embedded to embedded and embedded to wrapper combinations with a model alignment analysis.

E. Comparison with Original Benchmark Baselines

The reference paper [2] reports that naive application of SVM on all features yields BER approximately 0.40 - 0.42 on Madelon, matching this study's observed Linear SVM baseline of BER 0.4200. Challenge winning methods achieved BER of $\sim$0.07 by utilizing Bayesian neural networks and Dirichlet diffusion trees. While the best classical pipeline result (BER 0.1100) does not quite reach the absolute limit set by these neural architectures, it is highly competitive with the top tier ensemble submissions from the challenge while using selected hyperparameters and standard library implementations, confirming that a well designed feature selection pipeline can reach top tier performance without deep learning and bespoke kernel engineering.

III. Dataset and Experimental Setup

A. Madelon Dataset

The Madelon dataset was obtained directly from the NIPS 2003 Feature Selection Challenge. Dataset statistics are summarized in Table I. The training split was used for all model fitting and cross-validation; the validation split was used exclusively for final performance reporting.

TABLE I
Madelon Dataset Summary

Property	Value
Training samples	2,000
Validation samples	600
Test samples	1,800 (unlabeled)
Total features	500
Truly informative features	5
Redundant features	15
Pure noise features	480
Class balance (train & valid.)	Balanced (50/50)
Feature type	Continuous

B. Preprocessing

All 500 features were standardized to zero mean and unit variance using `the StandardScaler` fitted to the training set and applied to both splits. No imputation was necessary as the dataset contains no missing values.

C. Evaluation Protocol

Because the dataset is perfectly class balanced, accuracy and balanced accuracy are numerically identical throughout this study. The primary metrics are:

- **Balanced Accuracy** $= \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$
- **Balanced Error Rate** (BER) $= 1 - \text{Bal. Acc.}$
- **AUC-ROC**: area under the receiver operating characteristic curve

BER matches the primary metric used in the original challenge [2]. An RBF-kernel SVM (C = 1.0, `gamma='scale'`)

serves as the primary comparative evaluator across filter and wrapper families to maintain consistency with the benchmark paper. For embedded and most hybrid pipelines, native models are used primarily for feature ranking, while RBF SVM remains the main evaluator for unified cross family comparison. Only specific additional model specific hybrid rows are evaluated directly with RF/XGBoost. All results are reported on the held-out validation set, five-fold cross-validation on training data is additionally reported for multiple reference pipelines.

IV. Methods

This section describes all feature selection families implemented, the specific configurations evaluated, and the justification for the final recommended approach.

A. Filter Methods

Filter methods assign a score s_j to each feature x_j independently of any classifier. Three statistics were evaluated:

ANOVA F-score measures the ratio of between-class to within-class variance:

$$F_j = \frac{\text{SS}_{\text{between}}/(K-1)}{\text{SS}_{\text{within}}/(N-K)} \tag{1}$$

where $K = 2$ classes and N is the number of training samples.

Pearson Correlation computes the point-biserial correlation between x_j and binary label y:

$$r_j = \frac{\bar{x}_{j,+} - \bar{x}_{j,-}}{s_j}\sqrt{\frac{n_+ n_-}{N^2}} \tag{2}$$

For binary targets, ANOVA with two classes is monotonic in the squared point biserial correlation, so ranking by F_j is equivalent to ranking by $|r_j|$, which in this study's experiments yielded the same selected subsets at evaluated k values [3].

Mutual Information (MI) estimates $I(X_j;Y) = H(Y) - H(Y|X_j)$ using a k-nearest-neighbour estimator [13]. MI is theoretically capable of capturing nonlinear dependencies, but the kNN MI estimator can have higher variance under limited sample size and heavy noise feature spaces, which can reduce ranking stability.

A sweep over $k \in \{5, 10, 15, 20, 30, 50, 100, 200, 500\}$ selected features was conducted for each statistic, with the RBF SVM as the downstream evaluator.

B. Wrapper Methods

Wrapper methods use a learner as an oracle to score feature subsets.

Recursive Feature Elimination (RFE) [5] fits a `LinearSVC` and removes the lowest ranked features by absolute coefficient magnitude at each iteration (with step size > 1 in this implementation) until k features remain. In the final unified comparison, evaluation is done for $k \in \{10, 15, 20, 30\}$.

RFECV extends RFE with 5 fold cross-validation to automatically select k, optimizing cross-validated balanced accuracy (equivalently minimizing BER).

Sequential Forward Selection (SFS) greedily adds the feature that most improves cross-validated balanced accuracy at each step. It was run from a pre-filtered candidate set of 120 features (MI top 120) with an RBF SVM oracle and target $k = 15$.

C. Embedded Methods

Random Forest (RF) Importance. A Random Forest model is first fitted on the training set, and Mean Decrease in Impurity (MDI) is used to rank features. In the embedded sweep, top k subsets with $k \in \{5, 10, 15, 20, 30, 50, 100\}$ were evaluated (with $k = 15$ retained in the final unified comparison). For cross-family comparability, selected subsets were evaluated using the common downstream RBF SVM classifier.

Extra Trees (ET) Importance. The same ranking and selection protocol was applied using Extremely Randomized Trees [8]. Compared with RF, ET increases split randomization, which can improve robustness in noisy settings. Top k subsets were evaluated over the same k grid as RF.

L1-Logistic Regression and L1-LinearSVC. Sparse linear models were trained and features were ranked by absolute coefficient magnitude. In the final unified configuration, L1-Logistic Regression used $C = 0.1$ and L1-LinearSVC used $C = 0.05$, retaining the top 15 features. In the broader embedded sweep, multiple C values were tested and the resulting sparse subsets were compared.

D. Hybrid Multi-Stage Pipelines

Hybrid pipelines combine fast coarse screening with targeted refinement to exploit complementary strengths of different selectors. The implemented structures are:

1) **Stage 1 (coarse pre-filter):** reduce from 500 features to $M \in \{20, 80, 100, 120\}$ using ANOVA, MI, or RF importance.
2) **Stage 2 (fine selector- most hybrid rows):** apply a higher cost refiner to obtain $k = 15/20$ features. Refiners used in this study include LinearSVC RFE, Logistic L1 coefficient ranking, Sequential Forward Selection (RBF SVM or KNN), and tuned XGBoost (histogram) importance ranking.
3) **Stage 3 (evaluation):** for unified cross family comparison, most hybrid rows are evaluated with the common RBF SVM downstream classifier. In addition, model specific rows evaluate the RF top 20 subset directly with RF or XGBoost ($k = 20$), i.e., without a separate Stage 2 refinement.

E. Selection of the Best Approach

The recommended final pipeline is **RF20-EvalRF**: Random Forest importance pre-ranking to the top 20 features, followed by a freshly trained Random Forest classifier on those 20 features (with an additional 15 feature variant). This choice is justified on four grounds:

1) *Performance*: It achieves the highest Balanced Accuracy (0.8900) and lowest BER (0.1100) among all evaluated configurations.
2) *Model alignment*: Using the same model family for both ranking and evaluation aligns feature selection with the downstream learner's inductive bias.
3) *Computational efficiency*: In the current execution environment, Stage 1 (RF ranking) plus Stage 3 (RF evaluation) runs in approximately 7.45 s, versus about 472 s for SFS RBF at lower performance.
4) *Noise robustness*: Bootstrap aggregation and random feature subsampling in Random Forest reduce sensitivity to noisy dimensions, yielding stable and high utility rankings in this high noise setting.

V. Results

A. Baseline: No Feature Selection

Table II reports validation performance for six classifiers trained on all 500 features. The RBF SVM, the reference model for this benchmark achieves only Bal. Acc. 0.5817 (BER 0.4183), near random chance. The Decision Tree's comparatively higher baseline (0.7517) reflects its implicit feature gating at each split node.

TABLE II
BASELINE PERFORMANCE (ALL 500 FEATURES, VALIDATION SET)

Classifier	Bal. Acc.	BER	AUC
Decision Tree	0.7517	0.2483	0.7517
Gaussian NB	0.5917	0.4083	0.6457
RBF SVM	0.5817	0.4183	0.6288
Linear SVM	0.5800	0.4200	0.6021
Logistic Regression	0.5800	0.4200	0.5994
KNN (k=5)	0.5067	0.4933	0.5239

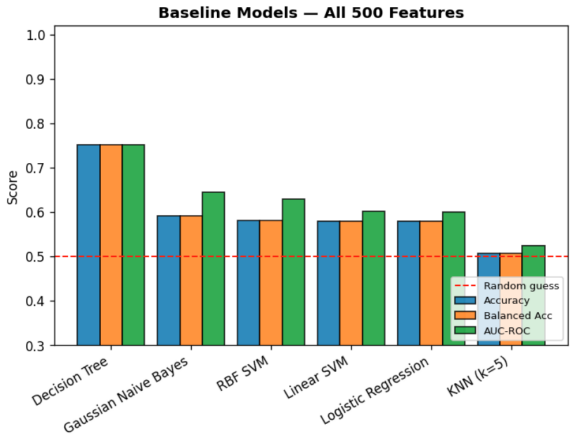

Fig. 1. Baseline Summary

B. Filter Methods

Table III presents the RBF SVM validation accuracy across the full k-sweep. ANOVA and Pearson produce identical rankings and peak sharply at $k = 15$ (Bal. Acc. 0.8250, BER 0.1750). For both ANOVA and Pearson, performance then degrades monotonically for larger tested feature budgets ($k \geq 20$), consistent with progressive inclusion of noise dimensions. MI consistently underperforms ANOVA/Pearson across all k values and reaches its best score at $k = 20$ (0.6667).

TABLE III
FILTER SWEEP: RBF SVM BALANCED ACCURACY VS. FEATURE COUNT

k	ANOVA	Mut. Info	Pearson
5	0.7050	0.5933	0.7050
10	0.8133	0.6400	0.8133
15	**0.8250**	0.6600	**0.8250**
20	0.7583	**0.6667**	0.7583
30	0.7183	0.6417	0.7183
50	0.6633	0.6100	0.6633
100	0.6300	0.6050	0.6300
200	0.5950	0.5983	0.5950
500	0.5817	0.5817	0.5817

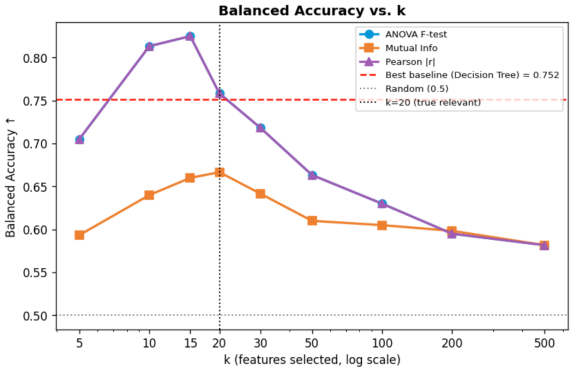

Fig. 2. Filter sweep performance

Table IV demonstrates that the ANOVA-15 feature subset improves all six classifiers, but with very different magnitudes. KNN's improvement is the most dramatic (+0.3433 Bal. Acc.), confirming that high dimensional noise disproportionately harms distance based methods. Logistic Regression and Linear SVM improve only modestly, indicating that substantial nonlinearity remains in the reduced subspace.

TABLE IV
ANOVA TOP-15 FEATURE SUBSET: ALL CLASSIFIERS (VALIDATION)

Classifier	Bal. Acc.	BER	AUC
KNN (k=5)	**0.8500**	0.1500	0.9258
RBF SVM	0.8250	0.1750	0.9146
Decision Tree	0.7917	0.2083	0.7917
Logistic Regression	0.6100	0.3900	0.6444
Linear SVM	0.6067	0.3933	0.6453
Gaussian NB	0.6017	0.3983	0.6481

Five-fold cross-validation on the training set further evaluates generalization of representative pipelines. Table V reports CV results spanning the observed performance range.

TABLE V
5-FOLD CROSS-VALIDATION ON TRAINING SET (BALANCED ACCURACY)

Pipeline	Mean	Std	Mean BER
RBF SVM + ANOVA (k=15)	0.7885	0.0220	0.2115
RBF SVM + MI (k=15)	0.6170	0.0267	0.3830
LinSVM + ANOVA (k=15)	0.5785	0.0193	0.4215
LinSVM (all 500 feat.)	0.5400	0.0158	0.4600

The ANOVA+RBF CV mean (0.7885, std 0.0220) is 0.0365 points below its validation score (0.8250), corresponding to approximately 1.66 standard deviations. This indicates mild optimism in the single validation split, but does not alter the method ranking. The all features Linear SVM (CV mean 0.5400) remains the clearest quantification of noise induced degradation for linear classifiers. Low standard deviations across pipelines (≤ 0.0267) indicate stable relative ordering across folds.

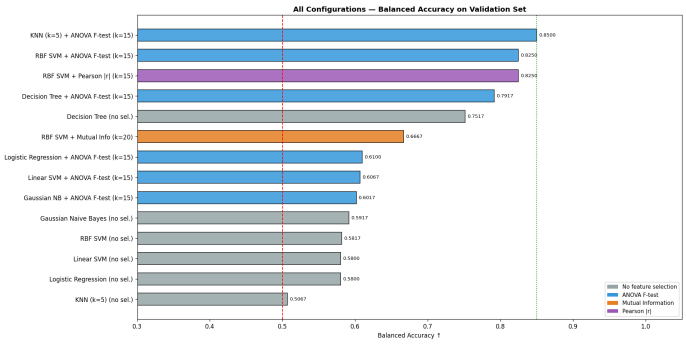

Fig. 3. Filter and Baseline Summary

TABLE VI
WRAPPER METHODS: VALIDATION PERFORMANCE AND RUNTIME

Method	k	Bal. Acc.	BER	Time
RFE (k=10)	10	**0.8333**	0.1667	∼5s
RFE (k=15)	15	0.8217	0.1783	∼5s
RFE (k=20)	20	0.7833	0.2167	∼5s
RFE (k=30)	30	0.7350	0.2650	∼6s
RFECV (auto)	5	0.6800	0.3200	∼17s
SFS-RBF (120→15)	15	0.8050	0.1950	∼472s

C. Wrapper Methods

Table VI reports wrapper results. For cross-family comparability, all wrapper selected subsets were evaluated with the common downstream RBF SVM classifier. RFE at $k = 10$ achieves the best wrapper performance (Bal. Acc. 0.8333, BER 0.1667), slightly exceeding the best filter result. Performance degrades as k increases beyond 10, mirroring the filter sweep pattern. RFECV's automated selection of only 5 features results in underfitting (0.6800). The SFS-RBF approach, despite being approximately two orders of magnitude slower than RFE, achieves only 0.8050, lower than both RFE at $k = 10$ and RFE at $k = 15$.

The top 10 features selected by RFE overlap substantially with the ANOVA top 15 set (5 of 10 features in common), indicating that both methods recover part of the same informative core. The remaining RFE specific features may explain the modest performance gain of RFE over ANOVA+RBF, consistent with backward elimination exploiting multivariate structure that univariate ANOVA does not model directly.

D. Embedded Methods

Table VII shows that tree based embedded methods outperform the best single family filter and wrapper configurations. Random Forest (top 15) achieves the strongest embedded result (Bal. Acc. 0.8600, BER 0.1400, AUC 0.9448). In contrast, sparse linear embedded models remain weak in this setting (Bal. Acc. around 0.60), consistent with their difficulty in retaining correlated informative structure under strong L1 sparsity.

Extra Trees and Random Forest are close overall but not uniformly ordered across all k: they are tied at $k = 20$ (0.8533), Random Forest is better at $k = 15$. The strongest tree based embedded performance occurs in the $k = 15$–20 range, aligning with the empirically observed informative dimensionality in this dataset.

TABLE VII
EMBEDDED METHODS: VALIDATION PERFORMANCE

Method	Bal. Acc.	BER	AUC
RF Importance (top 15)	**0.8600**	0.1400	0.9448
RF Importance (top 20)	0.8533	0.1467	0.9446
Extra Trees (top 20)	0.8533	0.1467	0.9446
Extra Trees (top 15)	0.8500	0.1500	0.9373
Extra Trees (top 30)	0.8067	0.1933	0.8986
RF Importance (top 30)	0.7933	0.2067	0.8726
L1-LinearSVC ($C = 0.05$, 50 features)	0.6033	0.3967	0.6168
L1-Log. Reg. ($C = 0.1$, 50 features)	0.5950	0.4050	0.6091

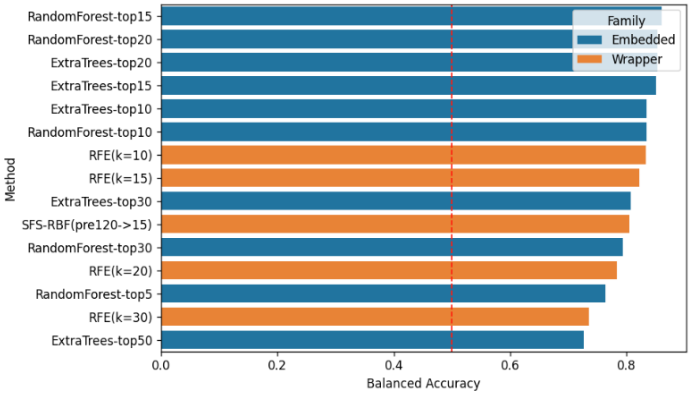

Fig. 4. Wrapper and Embedded top 15 Methods by Balanced Accuracy

E. Hybrid Methods

Table VIII presents the hybrid pipeline results. The RF20-EvalRF configuration achieves the highest overall performance: Bal. Acc. 0.8900 (BER 0.1100), with the highest AUC at 0.9657 for the $k = 20$ variant. Relative to the best single-family embedded method (RF top-15 at 0.8600), this is a +0.0300 Bal. Acc. gain and relative to the best filter baseline (ANOVA-15 at 0.8250), the gain is +0.0650.

The MI→L1 hybrid (0.6633) remains the weakest hybrid result, indicating that upstream dimensionality reduction alone does not resolve the sparse linear selector's limitation in this dataset. Model aligned RF→RF variants are the strongest overall, while several mixed family hybrids remain competitive in a narrow second tier.

TABLE VIII
HYBRID PIPELINE RESULTS (VALIDATION SET)

Pipeline	k	**Bal. Acc.**	**BER**	**AUC**
RF20-EvalRF	15	**0.8900**	**0.1100**	0.9634
RF20-EvalRF	20	**0.8900**	**0.1100**	**0.9657**
RF20-EvalXGB	20	0.8617	0.1383	0.9415
RF80-SFSRBF15	15	0.8567	0.1433	0.9417
ANOVA100-SFSKNN15	15	0.8567	0.1433	0.9404
RF80-RFE15	15	0.8550	0.1450	0.9422
ANOVA120-XGBhist15	15	0.8550	0.1450	0.9371
RF20-EvalXGB	15	0.8533	0.1467	0.9396
ANOVA100-RFE15	15	0.7900	0.2100	0.8872
MI120-L1-15	15	0.6633	0.3367	0.7681

A compact second tier is observed around 0.8550–0.8567 Bal. Acc. (RF80-SFSRBF15, ANOVA100-SFS KNN15, RF80-RFE15, ANOVA120-XGBhist15), suggesting that once Stage 1 screening reduces the space to a mostly informative candidate pool, multiple Stage 2 refiners can reach similar performance. For compute constrained settings, ANOVA100-RFE15 (0.7900) is substantially cheaper than SFS based hybrids, but incurs a clear accuracy gap versus the best hybrid result.

F. Unified Performance Summary

Table IX presents the best result per family and improvement magnitudes over the baseline.

TABLE IX
BEST-PER-FAMILY SUMMARY AND BASELINE IMPROVEMENT

Family	**Method**	**Bal.Acc**	**BER**	**ΔBER**
Baseline	RBF SVM (all 500)	0.5817	0.4183	—
Filter	ANOVA k=15 + SVM	0.8250	0.1750	−0.2433
Wrapper	RFE k=10 + SVM	0.8333	0.1667	−0.2516
Embedded	RF top-15 + RBF SVM	0.8600	0.1400	−0.2783
Hybrid	RF20-EvalRF	**0.8900**	**0.1100**	−0.3083

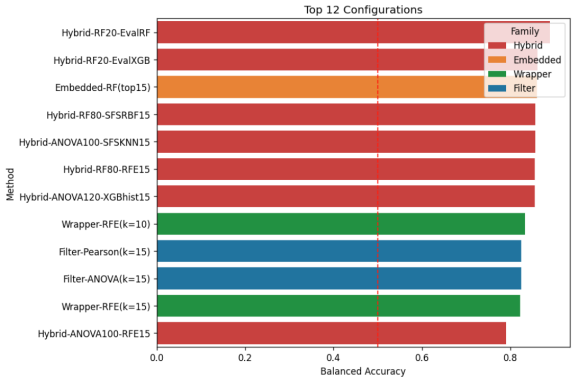

Fig. 5. Top 12 configurations across all feature selection families

VI. DISCUSSION

A. Why Feature Selection is Decisive Here

The weak baseline of RBF SVM on all 500 features (Bal. Acc. 0.5817) illustrates how strongly 480 noise dimensions can degrade kernel based methods. The RBF kernel is driven by pairwise Euclidean distances in the full feature space. When most dimensions are non informative, distance variance from noise can dominate variance from informative dimensions. Under a simplified isotropic view, the informative dimension share in squared distance is approximately $\frac{20}{500} = 0.04$, indicating severe dilution of signal. Restricting to a compact informative subset (about 15 - 20 features) substantially restores discriminative geometry for margin based classification.

B. Alignment of Filter Optima with Dataset Structure

The ANOVA peak at $k = 15$ is highly informative. Madelon is constructed with 20 informative observed features (5 true informative features plus 15 redundant linear combinations) and 480 noise features. The empirical optimum at $k = 15$ indicates that ANOVA recovers a high quality informative core before larger k values introduce progressively more non informative dimensions. In the tested sweep, performance declines for larger budgets ($k = 20, 30, 50, \ldots$), consistent with this interpretation.

C. Mutual Information vs. ANOVA

Although MI is theoretically able to capture nonlinear dependencies, it underperforms ANOVA in this study. ANOVA reaches 0.8250 at $k = 15$, while MI peaks at 0.6667 (at $k = 20$ and near-peak 0.6600 at $k = 15$). A likely cause is estimation instability of nonparametric kNN MI in noisy high-dimensional settings with limited sample size ($N = 2000$). By contrast, ANOVA's univariate F-test is computationally stable and empirically robust in this dataset.

D. Wrapper Trade Offs

As shown in Figure 6, RFE provides a stronger accuracy–runtime trade-off than SFS in this benchmark. RFE at $k = 10$ reaches Bal. Acc. 0.8333 in approximately 5 s, whereas SFS-RBF at $k = 15$ reaches only 0.8050 with a substantially higher runtime ($\sim$ 472s). RFECV selects $k = 5$ and underfits (0.6800), missing the manually observed RFE peak at $k = 10$. This suggests that automated CV-based subset sizing can be conservative for this dataset.

From a practical perspective, these results indicate that computationally intensive SFS wrappers may be difficult to justify when faster alternatives (RFE and tree-based embedded methods) achieve equal or better accuracy.

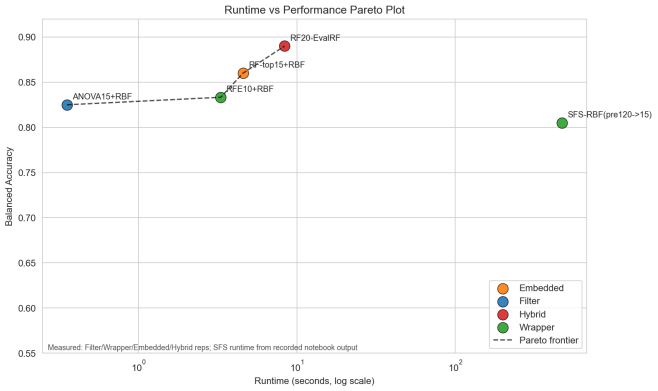

Fig. 6. Runtime vs. performance Pareto plot (log-scale runtime).

E. L1 Sparsity Failures on Correlated Features

L1 Logistic Regression (Bal. Acc. 0.5883) and L1 LinearSVC (0.6067) perform poorly relative to tree based embedded methods. This is consistent with a known limitation of pure L1 regularization under correlated informative groups, optimization may select one representative and zero out others [6]. In Madelon, redundant informative variables are correlated with the core informative structure, making pure L1 sparsity brittle. Elastic net (L1+L2) is a natural alternative.

F. Model Alignment in Hybrid Pipelines

RF20-EvalRF outperforms RF20-EvalXGB (+0.0283) and RF80-SFSRBF15 (+0.0333), supporting a model alignment principle: when ranking and final evaluation share the same model family, selected features are better matched to the downstream learner's inductive bias. In this benchmark, RF based ranking followed by RF evaluation yields the best hybrid result.

G. Geometry of Madelon: 5-D Hypercube Structure

Madelon is commonly described as originating from clusters on vertices of a 5 dimensional hypercube (32 vertex clusters), then embedded into a much higher dimensional observed space with added redundancy and many probes [3]. This geometry helps explain three empirical observations: (1) univariate filters can recover a strong core but not all interaction structure, (2) linear sparse models struggle to capture the effective nonlinear decision geometry, and (3) nonlinear/tree based selectors and evaluators are advantaged once noise dimensions are removed.

H. Potential Improvements

Several extensions are likely to improve BER beyond 0.1100:

Hyperparameter optimization. No exhaustive automated search (GridSearchCV/RandomizedSearchCV) was performed. Instead, fixed non default settings and manual sweeps were used. Nested CV with Bayesian optimization (Optuna/Hyperopt) over model and subset size hyperparameters is a high value next step.

Ensemble feature selection. Averaging rankings across multiple RF fits (different seeds/bootstraps) can improve stability. Stability selection [15] is a principled option.

Elastic net regularization. Replacing pure L1 with elastic net in embedded/hybrid stages can better preserve correlated informative groups.

Permutation importance. Permutation importance [7] can be used as a bias reduced alternative to MDI in marginal feature regions [14].

Cross validated hybrid selection. Computing Stage 1 rankings per fold and aggregating can provide more stable subsets and uncertainty aware selection.

VII. Conclusion

This paper presented a comparative evaluation of feature selection strategies on the NIPS 2003 Madelon benchmark, a 500 feature binary classification dataset with 480 noise dimensions. Four method families were implemented and analyzed. Filter (ANOVA F-score, Mutual Information, Pearson Correlation), Wrapper (RFE, RFECV, SFS), Embedded (Random Forest, Extra Trees, L1-Logistic, L1-LinearSVC), and Hybrid multi stage pipelines.

Five central empirical findings emerge. First, feature selection is essential in this setting. The no selection RBF-SVM baseline is weak (Bal. Acc. 0.5817, BER 0.4183), consistent with severe signal dilution in high noise dimensions. Second, a stable performance hierarchy is observed at the family best level: Hybrid $>$ Embedded $>$ Wrapper $>$ Filter $>$ Baseline. Third, ANOVA at $k = 15$ provides the strongest filter baseline (BER 0.1750) with very low computational overhead and strong alignment with the dataset's informative subset size range. Fourth, tree based embedded methods are the strongest single family approaches. Random Forest top 15 reaches BER 0.1400 and AUC 0.9448, while pure L1 sparse models remain weak, consistent with correlated feature limitations.

Fifth, the best overall result is achieved by a model aligned hybrid. RF importance pre-ranking to 20 features followed by RF evaluation, yielding Bal. Acc. 0.8900, BER 0.1100, and AUC 0.9657. This represents a substantial improvement over both baseline and single family alternatives. The model alignment principle, using the same model family for ranking and final evaluation is a practical design heuristic for high noise tabular pipelines.

From an operational perspective, higher computational cost does not guarantee better subsets. SFS-RBF ($\sim$472s) underperforms faster alternatives such as RFE based wrappers and tree based embedded/hybrid routes. For high dimensional noisy tabular data, tree based embedded or hybrid methods should be prioritized before expensive sequential wrapper searches.

Future work should examine elastic net regularization for

correlated informative groups, Bayesian hyperparameter optimization under nested cross validation, stability selection for robust ranking aggregation across multiple RF fits, and permutation importance as a bias reduced alternative to MDI. The hybrid template developed here is potentially transferable to real world high dimensional applications such as biomarker discovery, financial feature engineering, and industrial sensor analytics.

REFERENCES

[1] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.

[2] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result analysis of the NIPS 2003 feature selection challenge," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 17, 2004.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[10] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] H. Liu and H. Motoda, *Computational Methods of Feature Selection*. CRC Press, 2007.

[12] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 1991.

[13] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[14] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[15] N. Meinshausen and P. Bühlmann, "Stability selection," *J. Royal Statistical Society: Series B*, vol. 72, no. 4, pp. 417–473, 2010.