TRANSACTIONS ON EMERGING SCIENCES
Ideas at the Edge of Discovery
ISSN 3136-2710

ORIGINAL RESEARCH ARTICLE

DATASET-DEPENDENT PERFORMANCE OF FILTER-BASED FEATURE SELECTION METHODS: A COMPARATIVE STUDY IN HIGH-DIMENSIONAL CLASSIFICATION

Mehak Fatima
Department of Computer Science,
University of Agriculture, Faisalabad, Pakistan
Corresponding Author Email: mehakfatimaa20@gmail.com

ABSTRACT

This study presents a comprehensive evaluation of filter-based feature selection methods in high-dimensional classification, focusing on the effectiveness of these methods across multiple datasets and classifiers. Three filter-based techniques — Chi-square, Mutual Information, and ANOVA — are evaluated across three benchmark datasets with different characteristics: synthetic noisy data, real-world data, and structured multiclass data. The analysis examines the interactions between feature selection methods, classifiers, and feature subset sizes using multiple performance metrics, including accuracy, precision, recall, and F1-score. The results reveal that the highest-performing experimental configuration was achieved on the ISOLET dataset using Mutual Information with an SVM classifier at the top 40% feature subset, with an accuracy of 95.4%, precision of 95.6%, and F1-score and recall of 95.4%. For the Bioresponse dataset, the best performance was achieved using Mutual Information with Random Forest, achieving an accuracy of 81.4%; for the Madelon dataset, Chi-square performed best with Random Forest, achieving an accuracy of 81.3%. Findings demonstrate that no single method performs best across all datasets, and effectiveness depends on dataset characteristics and classifier selection.

Keywords: *Feature Selection; High-Dimensional Classification; Filter Methods; Chi-square; Mutual Information; ANOVA; Machine Learning*

1. INTRODUCTION

With the rapid growth of data-driven technologies across domains including bioinformatics, cheminformatics, finance, and speech processing, an unprecedented increase in the availability and complexity of data has surfaced. Datasets are usually categorized by high dimensionality (Li et al., 2017; Guyon and Elisseeff, 2003), where the number of features is larger than the number of samples. These high feature spaces can provide valuable information, but they also carry challenges for machine learning models such as increased computational cost, noise, overfitting, redundancy, and the well-known curse of dimensionality (Bellman, 1961), all of which can degrade model performance. As dimensionality increases, the distance between data points (Guyon and Elisseeff, 2003) becomes less meaningful, making it harder for algorithms to identify robust decision boundaries.

ISSN 3136-2710

One of the main issues with high-dimensional data is the presence of noisy and redundant features (Liu and Yu, 2005; Chandrashekar and Sahin, 2014) that are not meaningful for the predictive task. These features can conceal underlying patterns while increasing the risk of overfitting and negatively affecting model interpretability.

Feature selection is one of the fundamental preprocessing steps for addressing these challenges (Guyon and Elisseeff, 2003; Kuhn and Johnson, 2019) by identifying the most relevant subset of features for a given dataset. By removing irrelevant features (Blum and Langley, 1997), feature selection can improve model performance, enhance interpretability, and reduce training time. Feature selection methods are generally categorized into three groups (Jovic et al., 2015; Dash and Liu, 1997): filter methods, wrapper methods, and embedded methods. Filter-based methods are commonly adopted because of their computational efficiency and independence from learning models, making them suitable for high-dimensional datasets (Saeys et al., 2007). These methods evaluate each feature individually based on statistical or information criteria (Vergara and Estévez, 2014), allowing them to be applied prior to model training.

Despite these advantages, filter-based feature selection methods exhibit varying levels of effectiveness depending on the characteristics of the dataset and the choice of classification algorithm. Methods such as Chi-square, ANOVA F-test, and Mutual Information capture different types of relevance among features, including statistical dependence, variance discrimination, and information gain. Because filter-based methods evaluate each feature independently, they may fail where feature interactions (Forman, 2003; Hall and Holmes, 2003) are necessary, leading to differing performance across datasets.

Existing studies mostly evaluate feature selection techniques in isolation or under limited experimental conditions (Bolón-Canedo et al., 2015; Bommert et al., 2020), often focusing on a single dataset or classifier, which restricts the generalizability of their findings and makes it difficult to draw robust conclusions about the effectiveness of different methods. The interactions between feature selection techniques, feature subset sizes, classifier choice, and validation strategies also remain insufficiently explored within a unified experimental framework. This creates a need for a systematic and comprehensive evaluation that examines filter-based feature selection methods across multiple datasets with varying characteristics, using diverse classifiers and evaluation protocols.

In this context, the aim of this study was to perform a comparative analysis of filter-based feature selection methods for high-dimensional classification, specifically investigating the performance of three widely used methods — Chi-square, ANOVA F-test, and Mutual Information — across multiple benchmark datasets representing different domains, including synthetic, chemical, and speech data. Including various classifiers (Tang et al., 2014), evaluation metrics, and feature subset sizes enables a thorough evaluation of how impactful feature selection can be on classification performance under such conditions.

The main purpose of this research is to determine whether filter-based feature selection methods can consistently improve classification performance or whether they are dependent on dataset characteristics and model choices. Evaluating these methods in a unified and reproducible pipeline helps bridge the gap between theoretical understanding and the practical application of feature selection in high-dimensional classification,

ISSN 3136-2710

and provides insight into the strengths and limitations of filter-based methods for selecting appropriate techniques for real-world machine learning problems.

The main contributions of this study are:

- To conduct a comparative evaluation of three filter-based feature selection methods — Chi-square, ANOVA, and Mutual Information — within a unified research methodology.
- To investigate the impact of different feature subset sizes, including both fixed-size and percentage-based selections, on classification performance in high-dimensional datasets.
- To analyze interactions between feature selection methods and multiple machine learning classifiers, understanding how different models respond to reduced feature spaces.
- To evaluate the performance of feature selection techniques across different datasets with diverse characteristics, including noisy synthetic data, real-world chemical data, and structured speech data.
- To determine whether filter-based feature selection methods consistently improve classification performance or exhibit dataset- and model-dependent behaviours.

These findings are expected to provide valuable insight into the strengths and limitations of filter-based methods, thereby guiding practitioners in selecting appropriate techniques for real-world machine learning problems.

2. METHODOLOGY

2.1 Research Framework

To evaluate the impact of filter-based feature selection methods in high-dimensional classification problems, a unified pipeline was created to ensure methodological rigor, reproducibility, and fair comparison (Bommert et al., 2020) across multiple experimental configurations. The pipeline includes data preprocessing, feature selection, model training, performance evaluation, and result aggregation combined into a consistent framework, ensuring that all datasets are processed under the same conditions and allowing a fair and reproducible comparison of different feature selection techniques and machine learning models. The framework includes multiple dimensions of variation, including different feature selection methods, various feature subset sizes, multiple classification algorithms, and diverse evaluation strategies. Each component is controlled so that the effect of feature selection can be analyzed without confounding factors. In addition, a baseline classification setup using the complete set of original features is included, allowing direct comparison between models trained on the full dataset and on reduced feature spaces.

The workflow shown in Figure 1 highlights the execution of each step sequentially, from the raw datasets to result aggregation.

ISSN 3136-2710

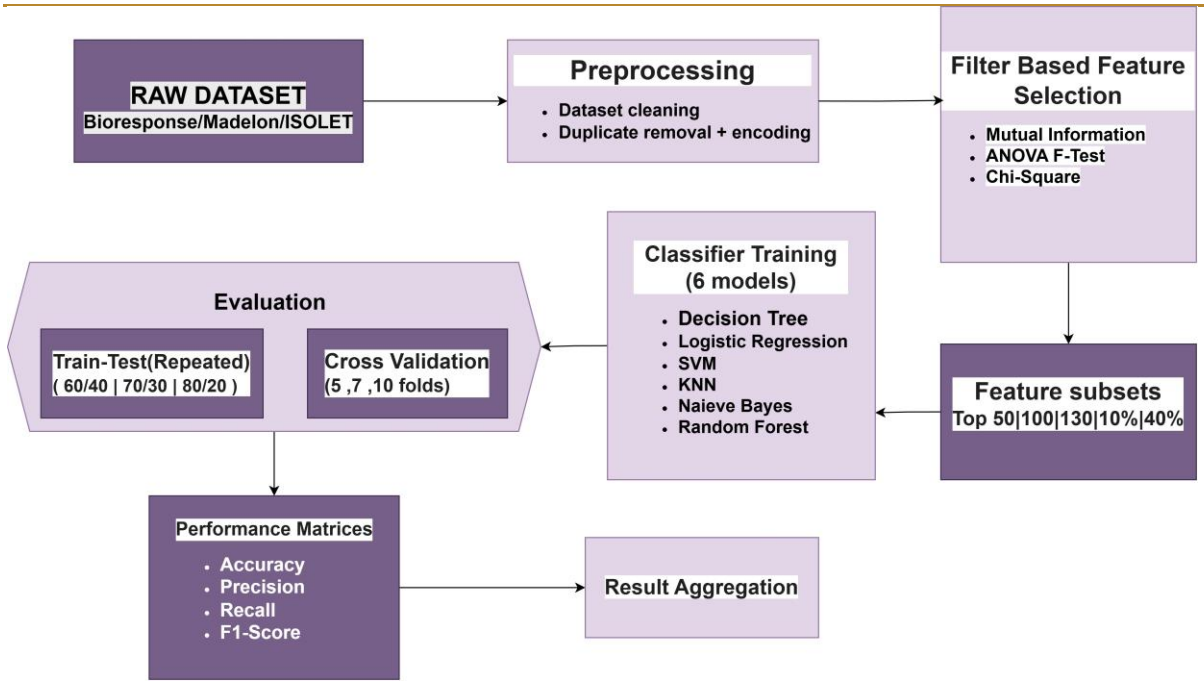

Figure 1. Proposed methodology pipeline.

2.2 Dataset Collection

For the comparative analysis of feature selection techniques, as shown in Table 1, three datasets with diverse features, instances, and class structures were selected.

2.2.1 Bioresponse Dataset

Bioresponse, which consists of 1,773 features and 3,751 instances, originates from chemical and drug response studies, representing a high-dimensional problem with noisy and redundant features and a high feature-to-instance ratio.

2.2.2 Madelon Dataset

The Madelon dataset, with 500 features and 2,000 instances, is a synthetic dataset specifically designed for feature selection, consisting of numerous irrelevant and redundant features, making it ideal for assessing how well feature selection methods can isolate informative attributes (Guyon, 2003).

2.2.3 ISOLET Dataset

The ISOLET dataset consists of 617 features and 7,797 instances across 26 classes, and is a spoken letter recognition dataset. The multiclass nature of ISOLET presents a more complex challenge for feature selection, as features must be informative across all classes to support accurate predictions.

Together, the three datasets provide a diverse testbed covering synthetic, real-world, and multiclass high-dimensional scenarios (Fanty and Cole, 1991).

ISSN 3136-2710

Dataset	Features	Instances	Classes	Domain
Madelon	500	2000	2	Synthetic / ML challenge
Bioresponse	1773	3751	2	Chemical / drug response
ISOLET	617	7797	26	Speech recognition

Table 1. Summary of datasets used in the study.

2.3 Data Preprocessing

The datasets are initially available in Attribute-Relation File Format (.arff), which is commonly used in machine learning dataset repositories such as OpenML (Vanschoren et al., 2014). Data in these files is first converted into a tabular format using data-handling libraries to enable processing; any byte-encoded attributes are then decoded into standard string format to ensure compatibility with subsequent operations. All feature values are then transformed into numeric format to meet the input requirements of machine learning algorithms.

A further preprocessing step involved the identification and removal of duplicate instances, as these can introduce bias by over-representing patterns in the data, affecting model training and evaluation. Removing duplicates ensures a more accurate representation of the data distribution and improves the consistency and reliability of the datasets used for experimental evaluation.

2.4 Baseline Classification Setup

A baseline classification framework was created to act as a reference for the effectiveness of feature selection methods, using the full set of features without any dimensionality reduction. This behaves as a control experiment, representing classification model performance when all available information is utilized. The setup employs the exact same classifiers and evaluation strategies used in the feature selection experiments to ensure a fair and unbiased comparison between baseline and feature-selected models. Including a baseline is important, as it allows evaluation of whether feature selection leads to improvements in performance, has a minimal effect, or degrades model accuracy due to the removal of important features.

2.5 Feature Selection Methods

Three filter-based feature selection methods — Mutual Information (Brown et al., 2012), ANOVA F-test, and Chi-square — were used due to their computational efficiency, scalability to high-dimensional data, and independence from specific classifiers, making them suitable for comparative evaluation.

- Mutual Information measures non-linear dependencies between features and the target while capturing relationships that linear methods might miss. It can capture both linear and non-linear relationships, making it useful in complex datasets where feature interactions may not be strictly linear (Vergara and Estévez, 2014).

ISSN 3136-2710

- ANOVA F-test compares the variance between class means relative to within-class variance, identifying features that best separate classes. Features that show significant differences are assigned higher importance scores (James et al., 2021).
- Chi-square evaluates the statistical dependence of feature values and target labels, selecting features with a higher association. Features that show stronger dependence on the target are considered most relevant. The Chi-square test requires non-negative feature values, for which appropriate normalization is applied prior to feature evaluation (James et al., 2021).

All feature selection methods were applied using a single ranking approach to evaluate each feature independently.

2.6 Feature Subset Configuration

Multiple feature subset configurations were defined to analyze the impact of feature dimensionality on classification performance; configurations included both fixed-size and percentage-based subsets, enabling a comprehensive exploration of feature reduction strategies. The fixed subsets include the top 50, 100, and 130 features. The percentage-based subsets included the top 10% and top 40% of features, allowing the number of selected features to scale proportionally with the number of features in each dataset, making the approach suitable for datasets of varying dimensionality (Bolón-Canedo et al., 2015).

2.7 Classification Models

To evaluate feature selection effectiveness across different learning paradigms, six classification algorithms were employed: Decision Tree, Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbors, and Random Forest (Hastie et al., 2009). These classifiers were selected to represent diverse modelling approaches: Decision Trees and Random Forests capture non-linear interactions between features, Logistic Regression and Support Vector Machines are sensitive to feature scaling and dimensionality, Naive Bayes is a probabilistic classifier that assumes feature independence, and K-Nearest Neighbors is an instance-based method that relies mainly on distance metrics.

2.8 Pipeline Design

A unified machine learning pipeline was developed to integrate preprocessing, feature selection, and classification into a single, consistent workflow. The pipeline begins with feature scaling; for most feature selection methods, standardization is performed using StandardScaler, while for the Chi-square method, MinMaxScaler (Géron, 2022) is used to transform feature values into a non-negative range, satisfying the requirements of the Chi-square test.

Feature selection is then applied using the three filter-based methods, each of which evaluates features independently. A subset selection step follows, using the SelectKBest approach. The selected features are then passed to the classification model for training and prediction.

2.9 Evaluation Strategy

ISSN 3136-2710

For a reliable and comprehensive assessment of model performance, two evaluation strategies were employed: repeated train-test splits and stratified cross-validation. In the repeated train-test approach, the dataset is divided into training and testing sets using three different ratios — 60:40, 70:30, and 80:20 — each repeated multiple times to account for variability, with results averaged across runs to obtain stable performance estimates.

Stratified K-fold cross-validation (Kohavi, 1995) is also used to provide a more robust and generalizable assessment. In this evaluation, the dataset is divided into k folds, each preserving the class distribution of the original dataset. All models were trained and evaluated across all folds, and their results were averaged to produce stable performance estimates.

2.10 Performance Metrics

Model performance was evaluated using multiple metrics — accuracy, precision, recall, and F1-score (Sokolova and Lapalme, 2009) — to capture different aspects of classification effectiveness.

- Accuracy measures the overall proportion of correctly classified instances, providing a general overview of performance after dimensionality reduction.
- Precision evaluates the proportion of correctly predicted instances among those predicted as positive for each class.
- Recall measures the proportion of positive instances correctly identified by the classifier, retaining class-related information after feature reduction.
- F1-score combines precision and recall into a single balanced measure of performance and is used to analyze the correctness of predictions.

3. RESULTS

3.1 Overall Performance Trends Across Datasets

The overall performance trends across the three datasets — ISOLET, Bioresponse, and Madelon — indicate clear variability in the effectiveness of filter-based feature selection methods, showing that performance is strongly dependent on dataset characteristics. The aggregated results in Figure 2 show that ISOLET achieved the highest average accuracy, around 0.771, followed by Bioresponse at 0.741, while Madelon showed the lowest average performance at 0.633, suggesting that the intrinsic structure and signal-to-noise ratio of a dataset play a critical role in determining classification success.

ISSN 3136-2710

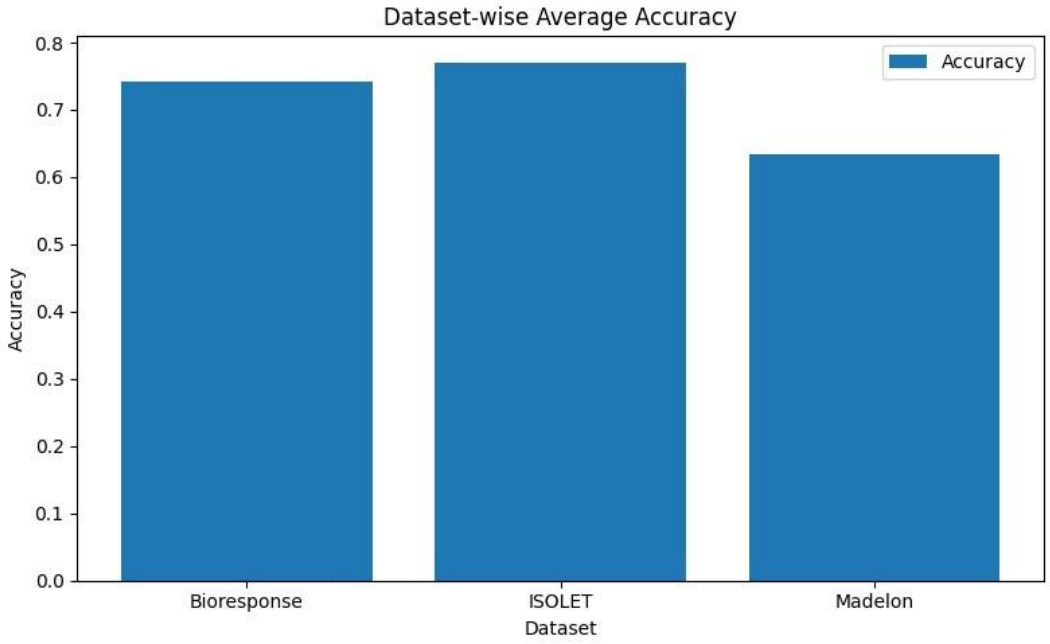

Figure 2. Datasets comparison.

Performance variability is higher in the ISOLET dataset compared to the others, exhibiting a maximum accuracy of 0.954 and a minimum of 0.268, indicating a very wide performance range. This suggests that while the dataset contains highly informative features capable of producing excellent classification results, it is also highly sensitive to inappropriate feature selection. The Bioresponse dataset shows stable behaviour across methods and classifiers, with a range of about 0.598 to 0.814. The Madelon dataset shows a maximum accuracy of about 0.813 but still has the lowest average performance because of its synthetic nature and high level of noise, making it difficult for models to identify meaningful patterns.

These interactions between methods, classifiers, and dataset characteristics show notable variability: in ISOLET, SVM combined with Mutual Information achieves performance as high as 95.4%, while other combinations result in degraded performance, reinforcing that feature selection can enhance or diminish performance depending on how well it aligns with the dataset structure.

Dataset	Average Accuracy	Best Accuracy	Worst Accuracy
ISOLET	0.7710	0.9540	0.2680
Bioresponse	0.7410	0.8140	0.5980
Madelon	0.6330	0.8130	0.5310

Table 2. Dataset summary.

3.2 Comparative Analysis of Feature Selection Methods

ISSN 3136-2710

A comparative evaluation of Chi-square, Mutual Information, and ANOVA F-test shows differing behaviours across all datasets, as shown in Figure 3, highlighting each method's ability to capture feature relevance under different data conditions (Chandrashekar and Sahin, 2014; Urbanowicz et al., 2018). As shown in Table 3, in the ISOLET dataset ANOVA achieves the highest average accuracy at about 0.810 and Mutual Information at 0.783, while Chi-square performs comparatively worse, at around 0.720, indicating that Chi-square is less effective at preserving distinctive features in this dataset, leading to performance degradation.

Figure 3. Method comparison across datasets.

In the Bioresponse dataset, all three methods perform similarly — ANOVA at around 0.746, Mutual Information at 0.742, and Chi-square at 0.737 — showing minimal difference. This suggests that the dataset does not strongly differentiate between feature selection methods, as informative features appear to be distributed moderately across the dataset, allowing even a statistical method like Chi-square to retain enough information for classification.

For the Madelon dataset, the results show a slightly different pattern: Chi-square achieves an average accuracy of 0.643, outperforming both ANOVA (0.634) and Mutual Information (0.623). Performance remains low overall, indicating that none of the feature selection techniques are particularly effective in handling the high noise and synthetic nature of this dataset.

Dataset	ANOVA	Chi-square	Mutual Information
ISOLET	0.8096	0.7202	0.7827
Bioresponse	0.7457	0.7367	0.7415
Madelon	0.6342	0.6427	0.6225

Table 3. Method-wise performance across datasets.

ISSN 3136-2710

Method performance is inconsistent across the three datasets: ANOVA performs best in ISOLET but does not outperform the other two methods in Bioresponse, nor is it best in Madelon; Chi-square performs worst in ISOLET but comparatively better in Madelon. This suggests that effectiveness depends on each method's underlying assumptions together with the dataset's statistical properties.

3.3 Dataset-Specific Analysis

3.3.1 Madelon Dataset

The Madelon dataset shows the lowest performance of the three datasets, with an average accuracy of around 0.633, despite achieving a maximum performance of around 0.813 under some configurations. This behaviour is tied to the dataset's synthetic design, where a large proportion of features are intentionally irrelevant and noisy. The comparative results across feature selection methods indicate that Chi-square achieves the highest average accuracy (0.643), outperforming ANOVA (0.634) and Mutual Information (0.623), although the differences between methods are relatively small.

From a classifier perspective, Random Forest achieves the highest average performance (0.743), followed by Decision Tree (0.723), while classifiers such as Logistic Regression (0.566) and SVM (0.567) perform significantly worse. The best observed performance on the Madelon dataset across methods is shown in Table 4.

Method	Best Classifier	Feature Set	Accuracy
ANOVA	Random Forest	Top 10%	0.8130
Chi-square	Random Forest	Top 10%	0.8130
Mutual Information	Random Forest	Top 50	0.7810

Table 4. Best observed performance on the Madelon dataset.

The effect of feature subset size, shown in Figure 4, further clarifies this pattern: smaller subsets such as the top 10% and top 50 perform better than larger subsets such as the top 100 and top 130, indicating that aggressive feature reduction helps remove noise. However, these improvements are marginal, suggesting that even reduced subsets still contain a significant amount of irrelevant information.

ISSN 3136-2710

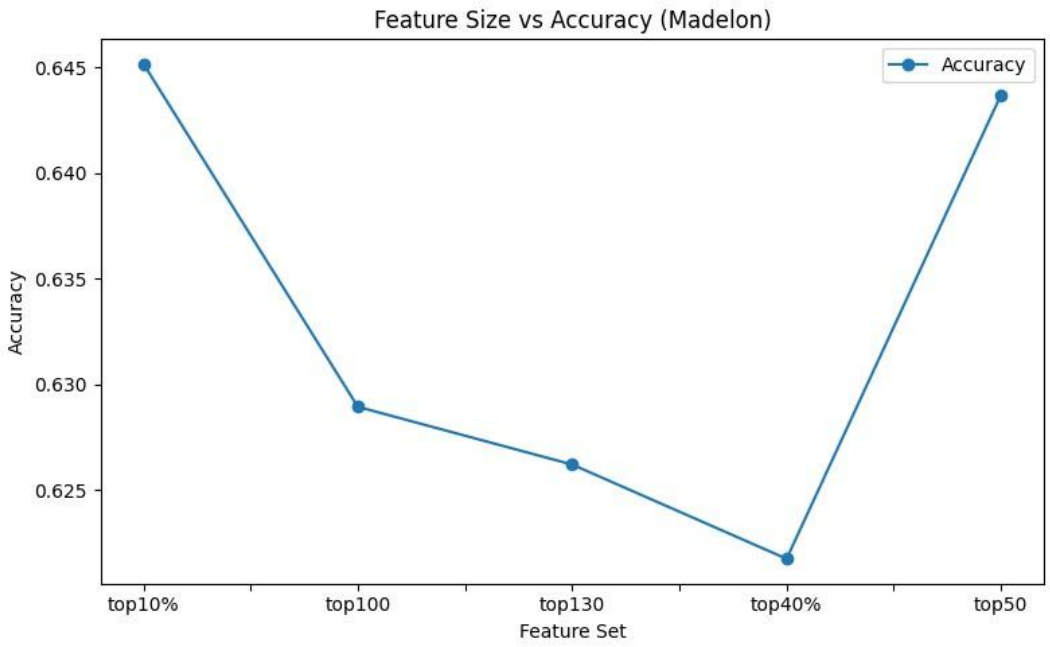

Figure 4. Feature size vs. accuracy (Madelon).

This confirms that in noise-dominated environments, improving model robustness is more impactful than refining the feature selection strategy.

3.3.2 Bioresponse Dataset

The Bioresponse dataset shows more moderate and stable performance across all experimental configurations, with an average accuracy of approximately 0.741 and a relatively narrow performance range between 0.598 and 0.814. This stability suggests that the dataset contains a balanced mixture of informative and redundant features, where feature selection has a limited but noticeable impact. Unlike ISOLET and Madelon, differences between feature selection methods are minimal in this dataset: ANOVA (0.746), Mutual Information (0.742), and Chi-square (0.737) produce almost identical results, indicating that the dataset does not strongly differentiate between methods.

From a classifier perspective, Random Forest achieves the highest performance (0.796), followed by KNN (0.760) and Logistic Regression (0.756). The best observed performance on the Bioresponse dataset across methods is shown in Table 5.

Method	Best Classifier	Feature Set	Accuracy
ANOVA	Random Forest	Top 40%	0.8140
Chi-square	Random Forest	Top 40%	0.8040
Mutual Information	Random Forest	Top 40%	0.8140

Table 5. Best observed performance on the Bioresponse dataset.

ISSN 3136-2710

The analysis of feature subset size, shown in Figure 5, shows very limited variation in performance, with all subsets producing accuracies within a narrow range of approximately 0.734 to 0.747.

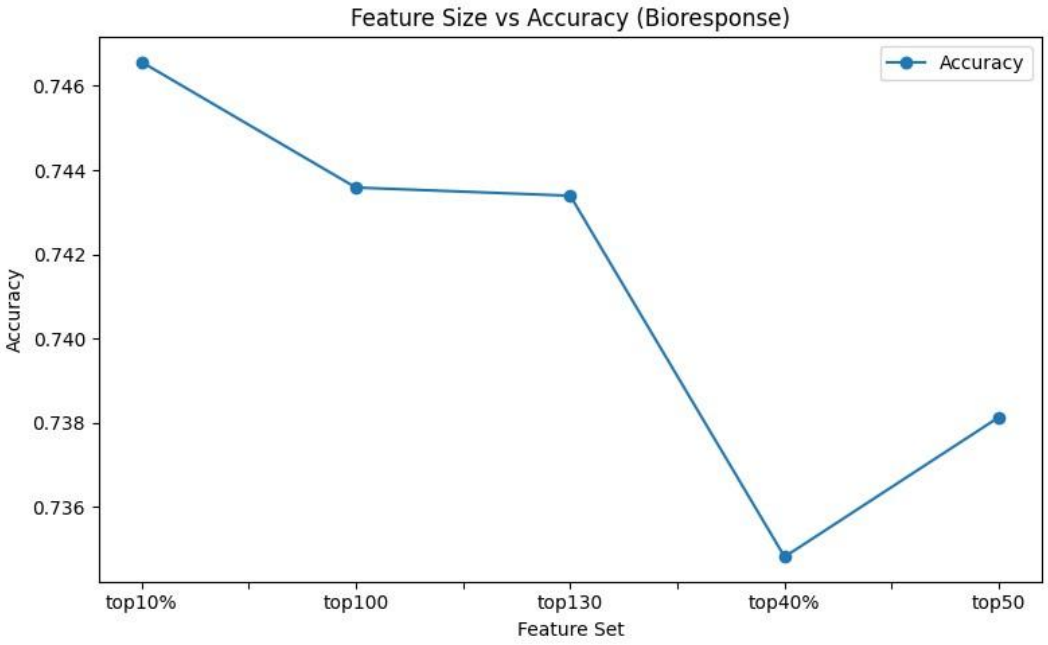

Figure 5. Feature size vs. accuracy (Bioresponse).

3.3.3 ISOLET Dataset

The ISOLET dataset shows the highest overall classification performance among all datasets, achieving a maximum accuracy of 0.954 as shown in Table 6, while also demonstrating the largest performance variability, with results dropping as low as 0.268. This wide performance range highlights the dataset's extreme sensitivity to feature selection. This variability is further reflected in the average performance values, where ANOVA achieves 0.810, Mutual Information 0.783, and Chi-square a notably lower 0.720 — a substantial performance gap of about 9 percentage points between the best- and worst-performing methods — indicating that the choice of feature selection method plays a critical role in determining classification outcomes for this dataset. This behaviour may be attributable to the structured and correlated nature of the ISOLET feature space.

This effect is also evident in the feature subset analysis shown in Figure 6, where smaller subsets such as the top 50 result in significantly lower performance, whereas larger subsets such as the top 130 and top 40% achieve substantially higher accuracy.

ISSN 3136-2710

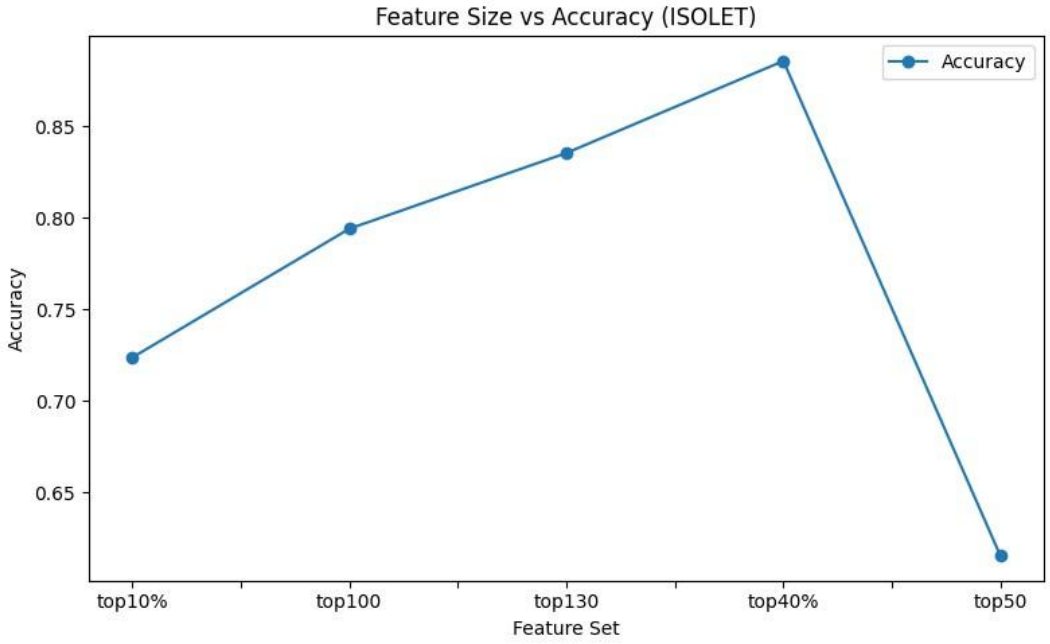

Figure 6. Feature size vs. accuracy (ISOLET).

Classifier-wise performance further reinforces this observation, as models such as Random Forest (0.835), SVM (0.833), and Logistic Regression (0.826) all achieve strong average performance, suggesting that the dataset is inherently learnable.

Method	Best Classifier	Feature Set	Accuracy
ANOVA	SVM	Top 40%	0.9480
Chi-square	KNN	Top 40%	0.9120
Mutual Information	SVM	Top 40%	0.9540

Table 6. Best observed performance on the ISOLET dataset.

3.4 Effect of Feature Subset Size

The analysis of feature subset sizes reveals a clear, dataset-dependent relationship between the number of selected features and classification performance. As shown in Table 7, the choice of subset size plays a critical role in determining model accuracy across all datasets, even though the nature of this relationship varies significantly depending on the underlying data characteristics. As shown in Figure 7, ISOLET performance improves as the number of selected features increases: smaller subsets such as the top 50 result in low accuracy, while larger subsets such as the top 130 and top 40% yield significantly better results. Larger subsets are therefore necessary to preserve the underlying structure of this dataset.

The Bioresponse dataset shows lesser sensitivity to feature subset size, with performance remaining relatively stable across configurations. This suggests that the dataset does not rely heavily on a specific subset of

ISSN 3136-2710

features and that most features contribute moderately to the predictive task. For the Madelon dataset, results show a slight advantage for smaller subsets, with the top 10% and top 50 performing better than larger subsets. This is likely due to the difficulty of identifying informative features in a noisy dataset.

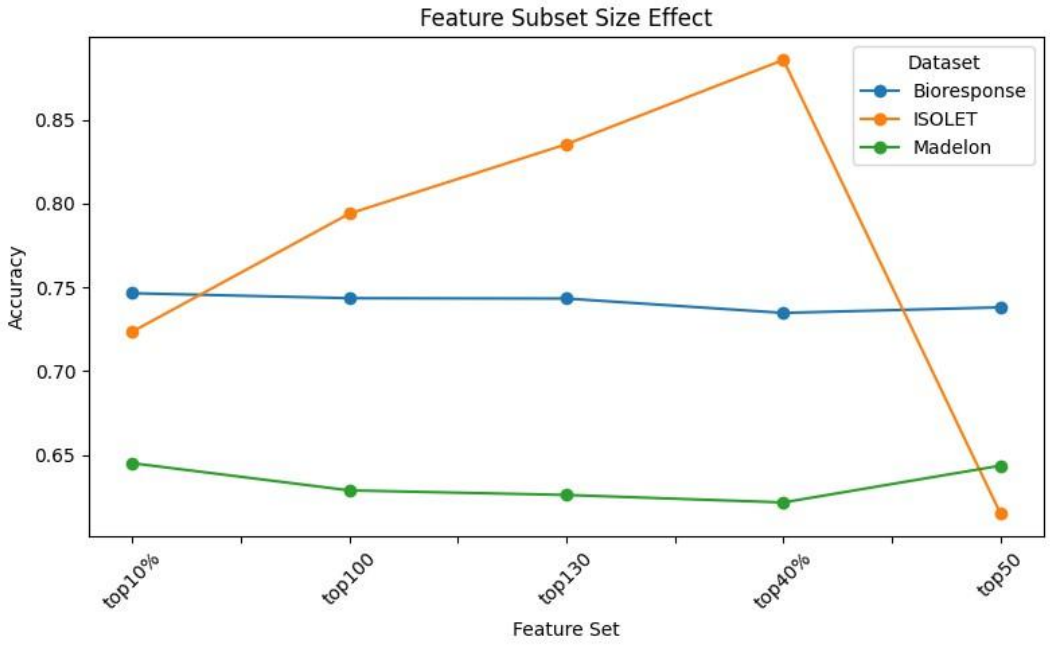

Figure 7. Feature size vs. accuracy (all datasets).

This behaviour indicates that feature subset size acts as a dataset-sensitive hyperparameter rather than a fixed design choice: where information is concentrated in a small number of features, removing many features can be beneficial, whereas where information is distributed across many features, selecting a larger subset is necessary for classification performance.

Dataset	Top 10%	Top 50	Top 100	Top 130	Top 40%
ISOLET	0.7236	0.6152	0.7940	0.8354	0.8857
Bioresponse	0.7466	0.7381	0.7436	0.7434	0.7348
Madelon	0.6452	0.6437	0.6289	0.6262	0.6217

Table 7. Effect of feature subset size on classification performance.

3.5 Classifier-Wise Performance Analysis

The classifier-wise analysis highlights significant variation in how learning algorithms behave when feature selection is applied. As shown in Figure 8, the evaluated models — Random Forest, SVM, and Logistic Regression — consistently achieve strong performance, although their behaviour varies depending on the dataset.

ISSN 3136-2710

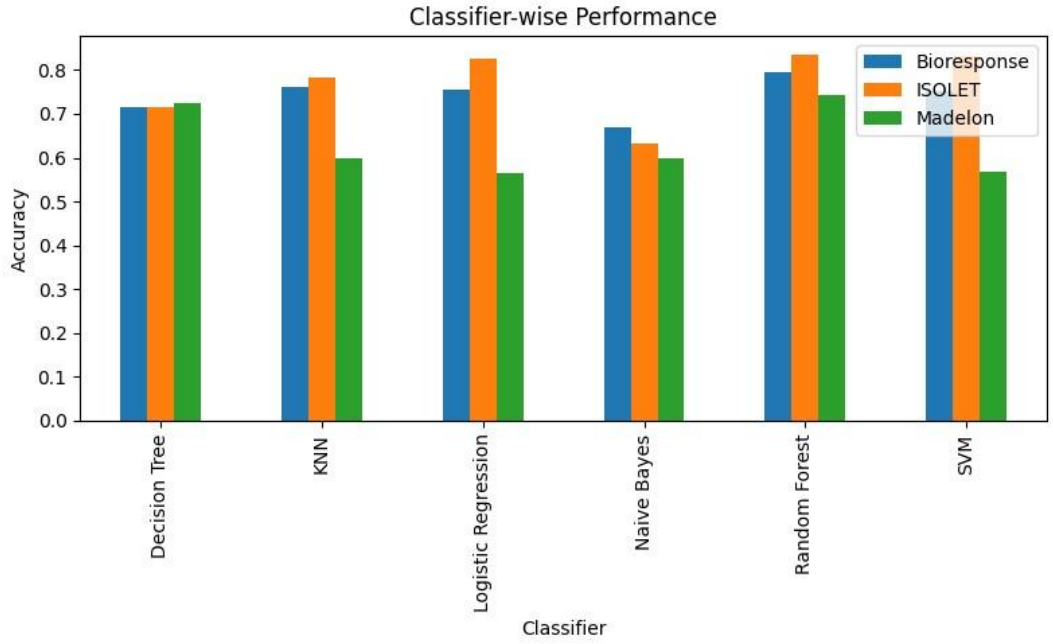

Figure 8. Classifier-wise performance across datasets.

Random Forest demonstrates high accuracy but shows some sensitivity to feature selection. SVM and Logistic Regression maintain relatively stable performance, as they are less affected by moderate feature reduction when the underlying data structure supports clear decision boundaries. Distance-based methods such as KNN show improvement (Forman, 2003) with feature selection, since reducing irrelevant features enhances distance calculations. Naive Bayes consistently underperforms across datasets, reflecting the limitations of its independence assumption in high-dimensional feature spaces.

The results in Table 8 indicate that classifier performance is closely linked to how models utilize feature information, and that the impact of feature selection varies depending on the learning mechanism of each classifier.

Dataset	DT	KNN	LR	NB	RF	SVM
ISOLET	0.7158	0.7838	0.8260	0.6309	0.8355	0.8329
Bioresponse	0.7143	0.7601	0.7557	0.6685	0.7962	0.7531
Madelon	0.7235	0.5992	0.5659	0.6001	0.7433	0.5669

Table 8. Classifier-wise performance across datasets.

3.6 Cross-Dataset Comparative Insights

The cross-dataset analysis reveals that the effectiveness of filter-based feature selection methods is not consistent and is strongly influenced by dataset characteristics and classifier behaviour (Li et al., 2017; Bolón-

ISSN 3136-2710

Canedo et al., 2015). Rather than providing uniform improvements across all datasets, the results indicate that feature selection performance depends on how predictive information is distributed within the feature space.

Across the evaluated datasets, three distinct behavioural patterns emerge. In structured datasets (ISOLET), where features exhibit interdependencies, performance becomes highly sensitive to feature selection, as independent evaluation fails to preserve important relationships. In moderately complex datasets (Bioresponse) with balanced feature contributions, performance remains relatively stable, indicating that feature selection has limited impact. In noise-dominated datasets (Madelon), the presence of a large number of irrelevant features reduces the effectiveness of filter methods, leading to inconsistent improvement.

Figure 9 shows a ranked distribution highlighting another important observation — the interaction between feature selection method and classifier behaviour: distance-based models tend to benefit from reduced dimensionality, while models that rely on feature interactions are more sensitive to feature removal. These findings demonstrate that filter-based feature selection methods do not provide universally consistent (Li et al., 2017) improvements, and their effectiveness is inherently constrained by dataset-specific properties and their inability to capture feature interactions.

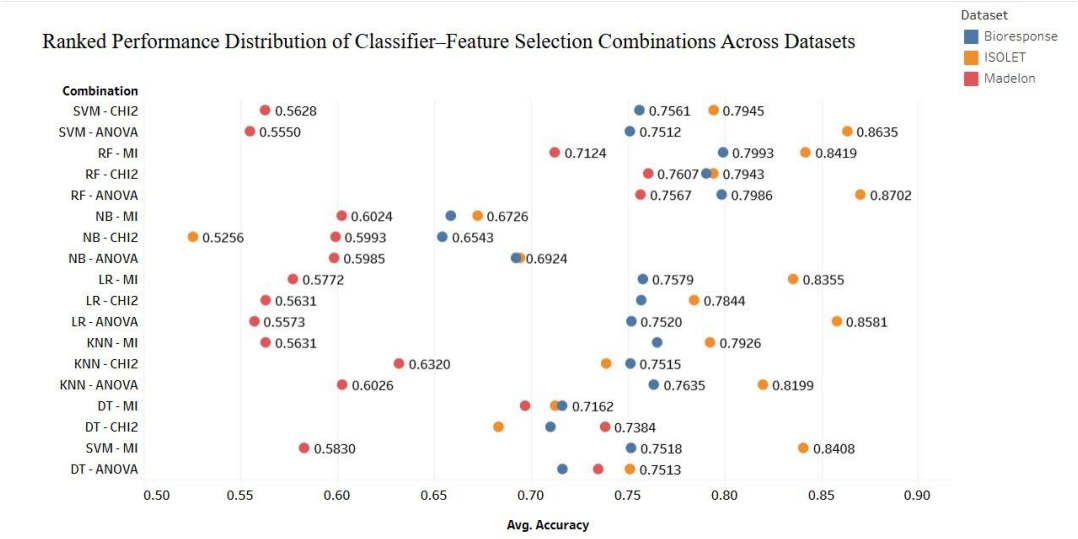

Figure 9. Ranked performance distribution of classifier–feature selection combinations across datasets.

3.7 Evaluation Metrics Consistency Analysis

In addition to accuracy, further evaluation metrics — precision, recall, and F1-score — were used to provide a detailed analysis of classification performance across all datasets, feature selection methods, and classifiers. These metrics are important for high-dimensional classification problems because they capture different aspects of model behaviour. The review of findings shows that all evaluation metrics follow consistent trends across configurations, as shown in Figure 10: models and feature selection methods that achieve higher accuracy also show correspondingly higher precision, recall, and F1-score. This consistency indicates that model performance degradation and improvement are systematic rather than dependent on a single measure,

ISSN 3136-2710

meaning accuracy can be considered a representative metric for comparative analysis without affecting the interpretation of the conclusions.

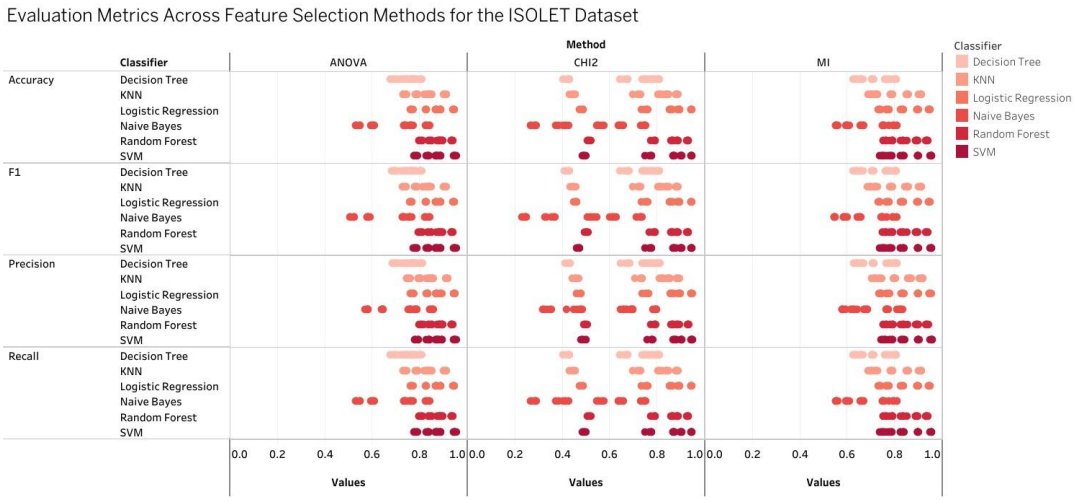

Figure 10. Evaluation metrics consistency — ISOLET dataset.

The discussion in this study therefore relies primarily on accuracy to provide a clear and concise comparison of feature selection methods and classifiers across datasets, improving stringency and readability while maintaining consistency with the overall findings.

3.8 Baseline vs. Feature Selection Comparison

A comparison between baseline performance and feature-selected models provides critical insight into the impact of filter-based feature selection methods. The results indicate that feature selection does not consistently improve classification performance and, in several cases, leads to noticeable degradation, reinforcing the importance of dataset-specific considerations.

As shown in Table 9, ISOLET baseline performance is already very high, with Logistic Regression achieving approximately 0.962 and Random Forest reaching 0.949. After feature selection, the performance of these models decreases — Logistic Regression drops to 0.950 and Random Forest to 0.938 — resulting in negative differences of about 0.011 and 0.010, respectively. This indicates that the full feature set already contains highly informative features, and removing any subset, particularly through univariate methods, leads to a loss of useful information. An exception is observed with KNN, where performance improves from 0.888 to 0.912, suggesting that this classifier benefits from reduced dimensionality due to its sensitivity to irrelevant features in distance calculations.

Classifier	Baseline Accuracy	Best FS Accuracy	Feature Subset
Logistic Regression	0.9620	0.9500	Top 40%
Random Forest	0.9490	0.9380	Top 40%

ISSN 3136-2710

SVM	0.9550	0.9540	Top 40%
KNN	0.8880	0.9120	Top 40%
Naive Bayes	0.7200	0.6310	Top 50
Decision Tree	0.9000	0.8700	Top 100

Table 9. ISOLET dataset — baseline vs. best feature selection performance.

As shown in Figure 11, differences between baseline and feature selection in the Bioresponse dataset are comparatively small, indicating a more stable relationship: feature selection neither meaningfully improves nor degrades performance for most classifiers, which aligns with earlier observations that the dataset contains a balanced distribution of informative features. The Madelon dataset shows mixed behaviour, with feature selection providing slightly improved results in certain configurations but not leading to consistent gains. These results demonstrate that feature selection does not always guarantee performance improvement and may be detrimental when the original feature space is already well structured, indicating that feature selection should not be applied blindly but must be carefully evaluated against baseline performance.

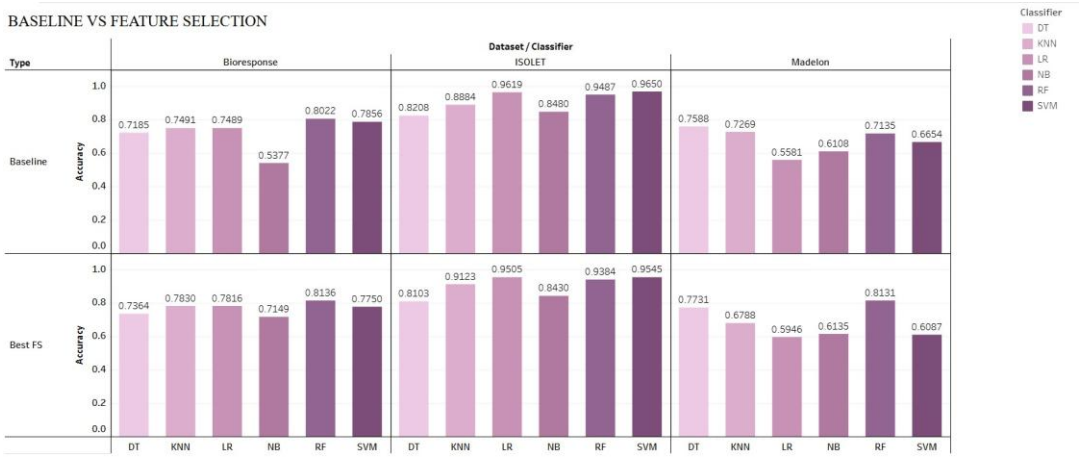

Figure 11. Baseline vs. best feature selection.

4. CONCLUSION

This study presented a comparative evaluation of filter-based feature selection methods across multiple high-dimensional datasets, focusing on their impact on classification performance under different data conditions. The findings reveal that while feature selection can influence model performance, its effectiveness is not consistent and depends strongly on dataset characteristics, feature distribution, and classifier behaviour.

Across the evaluated datasets, structured data with correlated features such as ISOLET achieved the highest accuracy, about 95.4% with ANOVA, and an average accuracy around 77%, indicating high sensitivity to feature reduction; in such cases, retaining a larger subset of features gives better performance, as aggressive reduction leads to the loss of important inter-feature relationships. The Bioresponse dataset exhibited

ISSN 3136-2710

relatively stable performance, with an average accuracy of 74.10% across different configurations, suggesting that feature selection plays a limited role when informative features are already distributed across the dataset. The Madelon dataset, characterized by high noise, showed only minimal improvements, with an average accuracy of 63.3%, highlighting the limitations of filter-based methods in identifying truly relevant features in noisy environments.

The comparison of Chi-square, ANOVA, and Mutual Information further illustrates that no single method consistently outperforms the others. Each method captures different aspects of feature relevance, resulting in performance variation across datasets. The results confirm that filter-based feature selection methods are inherently dataset-dependent and should not be applied as a one-size-fits-all solution. The effectiveness of filter-based feature selection should always be evaluated in light of dataset characteristics and the model selected.

REFERENCES

Bellman, R. (1961) *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

Blum, A. L. and Langley, P. (1997) 'Selection of relevant features and examples in machine learning', *Artificial Intelligence*, 97(1–2), pp. 245–271.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2015) 'Feature selection for high-dimensional data', *Progress in Artificial Intelligence*, 4(2), pp. 65–75.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J. and Lang, M. (2020) 'Benchmark for filter methods for feature selection in high-dimensional classification data', *Computational Statistics & Data Analysis*, 143, p. 106839.

Brown, G., Pocock, A., Zhao, M.-J. and Luján, M. (2012) 'Conditional likelihood maximisation: A unifying framework for information theoretic feature selection', *Journal of Machine Learning Research*, 13, pp. 27–66.

Chandrashekar, G. and Sahin, F. (2014) 'A survey on feature selection methods', *Computers & Electrical Engineering*, 40(1), pp. 16–28.

Dash, M. and Liu, H. (1997) 'Feature selection for classification', *Intelligent Data Analysis*, 1(3), pp. 131–156.

Fanty, M. and Cole, R. (1991) 'Spoken letter recognition', in *Advances in Neural Information Processing Systems*, Vol. 3.

Forman, G. (2003) 'An extensive empirical study of feature selection metrics for text classification', *Journal of Machine Learning Research*, 3, pp. 1289–1305.

Géron, A. (2022) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 3rd edn. Sebastopol, CA: O'Reilly Media.

Guyon, I. (2003) 'Design of experiments for the NIPS 2003 variable selection benchmark', in *NIPS 2003 Workshop on Feature Extraction and Feature Selection*.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, 3, pp. 1157–1182.

Hall, M. and Holmes, G. (2003) 'Benchmarking attribute selection techniques for discrete class data mining', *IEEE Transactions on Knowledge and Data Engineering*, 15(6), pp. 1437–1447.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. 2nd edn. New York: Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) *An Introduction to Statistical Learning*. 2nd edn. New York: Springer.

Jovic, A., Brkić, K. and Bogunović, N. (2015) 'A review of feature selection methods with applications', in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205.

ISSN 3136-2710

Kohavi, R. (1995) 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 2, pp. 1137–1143.

Kuhn, M. and Johnson, K. (2019) *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton, FL: CRC Press.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. and Liu, H. (2017) 'Feature selection: A data perspective', *ACM Computing Surveys*, 50(6).

Liu, H. and Yu, L. (2005) 'Toward integrating feature selection algorithms for classification and clustering', *IEEE Transactions on Knowledge and Data Engineering*, 17(4), pp. 491–502.

Saeys, Y., Inza, I. and Larrañaga, P. (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, 23(19), pp. 2507–2517.

Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', *Information Processing & Management*, 45(4), pp. 427–437.

Tang, J., Alelyani, S. and Liu, H. (2014) 'Feature selection for classification: A review', in C. C. Aggarwal (ed.) *Data Classification: Algorithms and Applications*. Boca Raton, FL: CRC Press, pp. 37–64.

Urbanowicz, R. J., Meeker, M., Cava, W. L., Olson, R. S. and Moore, J. H. (2018) 'Relief-based feature selection: Introduction and review', *Journal of Biomedical Informatics*, 85, pp. 189–203.

Vanschoren, J., van Rijn, J. N., Bischl, B. and Torgo, L. (2014) 'OpenML: Networked science in machine learning', *SIGKDD Explorations*, 15(2), pp. 49–60.

Vergara, J. R. and Estévez, P. A. (2014) 'A review of feature selection methods based on mutual information', *Neural Computing and Applications*, 24(1), pp. 175–186.