

A hybrid feature selection framework: balancing information preservation and multicollinearity in biological datasets

S Mohajer^a, D Kulasiri^a and S Samarasinghe^a

^a Centre for Advanced Computational Solutions (C-fACS), Lincoln University, Christchurch, New Zealand
Email: mohajer.ae@gmail.com

Abstract: This study introduces a hybrid feature selection framework that combines mutual information (MI) analysis with iterative variance inflation factor (VIF) filtering to address the competing demands of predictive performance and multicollinearity management in a biological dataset. Testing on *Solanum* callus induction data (1,081 observations, 16 features), the hybrid approach achieved superior performance with $R^2=0.86$, outperforming VIF filtering alone by 16% and mutual information methods alone by 8%. The framework's key advantage is preserving biologically relevant features with strong predictive value (like relative humidity with $MI=0.66$) that VIF filtering would eliminate due to multicollinearity, while still maintaining statistical independence. This may address an important matter in biological machine learning, where both predictive accuracy and interpretability are essential.

Keywords: Variance inflation factor; mutual information, callus induction

1. INTRODUCTION

Feature selection must balance preserving predictive information with reducing redundancy, particularly multicollinearity. This balance is particularly crucial in agricultural modeling, where models must generalize across genotypes, environments, and management practices. Otherwise, suboptimal feature selection can lead to overfitting and reduced interpretability, producing unreliable predictions that risk agricultural productivity and resource efficiency.

Conventional feature selection methods exhibit significant limitations when applied to agricultural and biological datasets characterized by inherent multicollinearity. Reliance on Variance Inflation Factor (VIF) diagnostics alone can remove features with genuine predictive value solely due to correlation with other variables, whereas mutual information (MI) selection may retain redundant predictors that ultimately degrade model performance (Peng et al. 2005). Moreover, wrapper approaches can be computationally costly and may risk excluding meaningful predictors by prioritizing statistical performance over biological relevance. Model-agnostic tools such as SHAP provide importance estimates but do not address the multicollinearity common in agricultural data, and attributions can be unstable or shared among correlated features (Basu and Maji 2020).

These methodological gaps are particularly problematic in plant tissue culture applications, where optimization directly impacts crop improvement programs, germplasm conservation efforts, and commercial propagation success. Callus induction, a critical tissue culture technique, depends on complex interactions between environmental and hormonal variables, where suboptimal factor selection or parameter tuning directly translate to increased costs, elevated failure rates, and delayed breeding progress. This study presents a novel hybrid feature selection framework that addresses these collective challenges by employing a bifurcated algorithm: (1) protection of features with high information identified through MI analysis and (2) systematic multicollinearity reduction using iterative VIF filtering that preserves protected features. This approach combines the information-theoretic advantages of mutual information with the statistical rigor of VIF analysis.

2. METHODOLOGY AND RESULTS

This methodology was developed and validated using a comprehensive dataset of callus induction experiments across *Solanum* genus, comprising 1,081 observations with 16 experimental variables, including cytokinin and auxin concentrations, environmental conditions, and genetic features. The hybrid method implemented a sophisticated selection process in two steps. Initially, mutual information scores quantify the statistical relationship between each feature and the target variable. Features with extremely high MI scores (>10) are removed to prevent data leakage (features contain target-derived information) because such information create artificially inflated model performance during training that fails to generalize to new data. However, features within the top 10% of MI scores (percentile=10) form a "protected" subset. The second stage employs iterative VIF analysis to eliminate multicollinearity. Unlike traditional methods that remove all features with $VIF > 10$ (Montgomery et al. 2021), our approach only eliminates non-protected features exceeding this threshold. This process continues iteratively

until all remaining non-protected features achieve acceptable VIF values. A baseline Random Forest model was also employed to evaluate the effectiveness of different feature selection approaches and demonstrate the superior performance of our hybrid method. Random Forest was selected as the evaluation framework due to its tolerance of outliers, its capacity to capture non-linear relationships among variables without requiring extensive hyperparameter tuning, and its ensemble nature which provides stable performance estimates that are less sensitive to individual feature perturbations. All categorical variables were appropriately encoded and missing values imputed prior to model training.

The hybrid feature selection method successfully identified 11 optimal features while effectively managing the balance between information preservation and multicollinearity reduction (Table 1). The unchanged pH VIF (4249.40→4249.40) indicates immediate elimination due to extreme multicollinearity. Features with moderate initial VIF values (LP: 6.33→5.38, BAP: 1.97→1.86) showed minor improvements, while protected features experienced dramatic reductions (RH: 794.44→9.73), demonstrating the cascade effects from targeted feature removal. It validates our hypothesis that strategic elimination enables preservation of valuable features. MI selection alone retained all 16 features, achieving moderate performance ($R^2=0.80$) but failing to address multicollinearity issues that degrade precision (RMSE=0.65) (Table 2). VIF selection alone eliminated the critical RH variable (VIF=490.92, MI=0.66) due to multicollinearity, while paradoxically retaining DCP (MI=0.02) simply due to lower correlation, resulting in inferior performance ($R^2=0.74$). The hybrid approach strategically preserved RH through the protection mechanism while eliminating 5 genuinely redundant features (pH, Sucrose, Temperature, Agar, DCP), achieving optimal performance ($R^2=0.86$) with fewer features than either comparison method.

Table 1 Hybrid Feature Selection Performance Summary

Feature	Initial VIF	MI Score	Protected	Final VIF	Final Selection
2,4-D	1.54	0.07	FALSE	1.51	Yes
Agar	433.04	0.18	FALSE	267.18	No
BAP	1.97	0.16	FALSE	1.86	Yes
DCP	23.66	0.02	FALSE	17.31	No
GDG	4.42	0.52	TRUE	3.60	Yes
IAA	1.20	0.08	FALSE	1.19	Yes
IBA	1.10	0.02	FALSE	1.09	Yes
KIN	1.09	0.02	FALSE	1.07	Yes
LP	6.33	0.17	FALSE	5.38	Yes
NAA	1.65	0.14	FALSE	1.55	Yes
pH	4249.4	0.16	FALSE	4249.4	No
RH	794.44	0.66	TRUE	9.73	Yes
Sucrose	3434.2	0.05	FALSE	1517.5	No
TDZ	1.12	0.01	FALSE	1.05	Yes
Temp.	625.39	0.11	FALSE	364.10	No
ZT	1.06	0.04	FALSE	1.02	Yes

Auxin hormones (2,4-D, IAA, NAA, IBA), Cytokinin hormones (BAP, KIN, ZT, TDZ), DCP (donor culture plant), GDG (geographically diverse germplasm), LP (light photoperiod), RH (relative humidity), and Temp. (temperature)

Table 2 Comparative Analysis of Feature Selection Methods

Method	Features Selected	R^2	MAE	RMSE	Key Eliminated Features
MI-only	16	0.80	0.32	0.65	None (retains all features)
VIF-only	12	0.74	0.39	0.52	pH, Sucrose, RH, Temp., Agar
Hybrid	11	0.86	0.25	0.35	pH, Sucrose, Temp., Agar, DCP

3. CONCLUSIONS

This study represents the importance of preserving valuable predictive features even when they exhibit correlation complications through strategic information protection and iterative filtering, thereby avoiding the loss of critical data. The hybrid feature selection can improve callus induction protocols, thereby raising genetic transformation and plant regeneration rates in target genotypes. Extending the approach to precision agriculture models for yield prediction, irrigation scheduling, and pest-risk assessment promises broader gains in sustainability and food security.

ACKNOWLEDGEMENTS

This study was conducted via the New Zealand eScience Infrastructure (NeSI) high-performance computing facilities, which are supported by the Ministry for Business Innovation and Employment (MBIE).

REFERENCES

- Basu, I and Maji S (2022) February. Multicollinearity correction and combined feature effect in shapley values. In *Australasian Joint Conference on Artificial Intelligence* (pp. 79-90). Cham: Springer International Publishing.
- Montgomery DC, Peck EA and Vining GG (2021) *Introduction to Linear Regression Analysis*. John Wiley & Sons.
- Peng H, Fuhui L and Chris D (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005): 1226-1238.