Evaluating Hyperparameter Optimization in Machine Learning Algorithms for Cancer Driver Gene Classification

Ana Laura Schardosim[1], **Kamille Konarzewski**[1],
Renan Soares de Andrades[1,2], **Mariana Recamonde-Mendoza**[1,2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA)
Porto Alegre – RS – Brazil

{ana.schardosim,kamille.kpimentel,rsandrades,mrmendoza}@inf.ufrgs.br

Abstract. *Cancer driver genes (CDGs) play a central role in tumorigenesis and represent important targets for diagnosis and therapy. In this study, we evaluate the impact of hyperparameter optimization on the predictive performance of traditional machine learning algorithms using multi-omics data. We perform systematic searches across different configurations to identify the most effective settings for CDG classification. Our results show that optimized models consistently outperform their default counterparts, particularly in recall and precision-recall metrics, with ensemble methods showing the most pronounced gains. These findings indicate that traditional algorithms, when carefully tuned, can represent promising approaches for identifying CDGs from genomics data.*

1. Introduction

Cancer is a major global health challenge, responsible for nearly 10 million deaths in 2022 and projected to affect more than 28 million people by 2040 [Bray et al. 2018]. Its development is strongly linked to somatic mutations in cancer driver genes (CDGs), which provide growth advantages to tumor cells and promote carcinogenesis [Vogelstein et al. 2013, Stratton et al. 2009]. Identifying these genes is a critical goal in cancer research, but distinguishing drivers from the vast majority of passenger mutations remains difficult due to mutation heterogeneity across tumor types.

Machine learning (ML) offers powerful tools to address this challenge, enabling the integration of multi-omics data for CDG prediction. A critical but sometimes underexplored aspect of applying ML to biomedical data is the choice of hyperparameters. Default configurations often lead to suboptimal results, especially when dealing with imbalanced datasets such as CDG prediction. Hyperparameter optimization allows models to better adapt to the characteristics of the data, improving generalization and robustness across evaluation metrics.

In this work, we investigate the impact of hyperparameter optimization on the predictive performance of six traditional ML algorithms applied to CDG classification. By comparing default and optimized settings across accuracy, recall, precision, F1, AUC-ROC, and AUC-PR, we highlight the improvements achieved through tuning and discuss their implications for more reliable cancer gene discovery.

2. Materials and Methods

This section presents the experimental design for CDG prediction, detailing the omics data employed, the labeling of positive and negative CDGs, and the procedures for model training and evaluation. All datasets were obtained from publicly available sources.

2.1. Omics data

To evaluate the impact of omics data on predictive performance, we used multi-omics datasets from TCGA, comprising over 8,000 samples across 16 cancer types and including Single Nucleotide Variants (SNVs), Copy Number Alterations (CNAs), DNA methylation, and gene expression. Following established preprocessing procedures [Schulte-Sasse et al. 2021], mutation rates, copy number changes, promoter methylation levels, and differential expression levels (comparing cancer and control samples) were computed for each gene, normalized, and structured into feature matrices by cancer type. The four matrices were then integrated into a single dataset, representing the molecular profiles in terms of mutations, CNA, DNA methylation and gene expression of more than 13,000 genes and serving as the basis for CDG prediction experiments.

2.2. Positive and negative examples of CDGs

In this study, genes were classified into two categories: driver genes (positive labels) and non-cancer driver, also known as passenger genes (negative labels). Driver genes were compiled from curated sources, including the Network of Cancer Genes (NCG), the Tier 1 list from the Cancer Gene Census (CGC), and the OncoKB database, resulting in 907 unique entries after removing duplicates. Non-cancer genes were defined through stringent filtering, excluding any gene with known associations to cancer or other diseases, as recorded in OMIM. This labeling strategy yielded a substantially imbalanced dataset (6.6% of driver genes and 93.4% of passengers), with driver genes underrepresented, posing a challenge for predictive modeling.

2.3. Model training and evaluation

The dataset was partitioned via stratified sampling into 80% training-validation and 20% test data, ensuring class balance across splits. A nested cross-validation scheme was applied: the inner loop performed hyperparameter tuning with Optuna using 5-fold stratified cross-validation and 20 trials per model, while the outer split reserved data for evaluating the best hyperparameter configuration. Six algorithms were optimized and evaluated, including decision tree, random forest, gradient boosting, histogram gradient boosting, k-nearest neighbors, and multi-layer perceptron. Considering the class imbalance, AUC-PR was prioritized as the optimization metric, but recall, precision, F1, AUC-ROC, and accuracy were also measured for better assessment.

3. Experiments and Results

Table 1 presents the performance metrics obtained on the independent test set for all evaluated algorithms. Cross-validation results are not included due to space constraints but were analyzed to investigate potential overfitting.

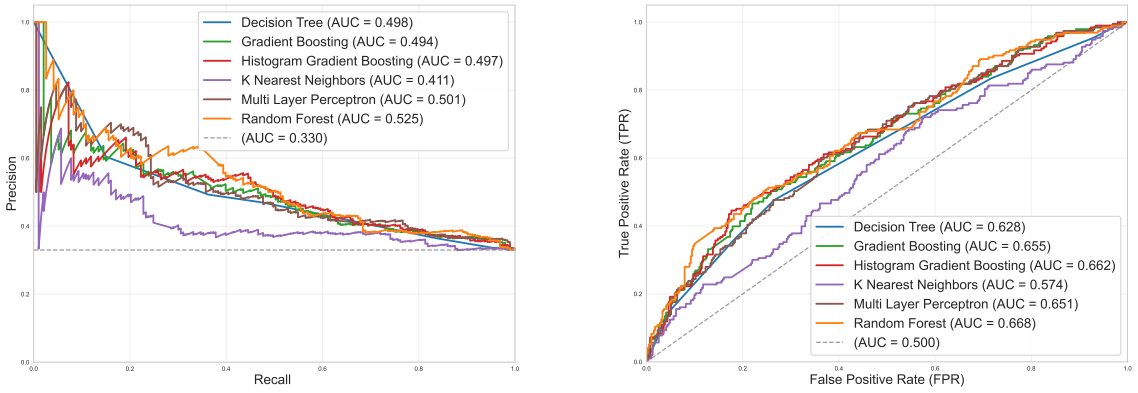
The results indicate varying levels of success across different models, with ensemble methods generally outperforming individual classifiers. After hyperparameter optimization, the Multi-Layer Perceptron achieved the highest F1-score (0.39) and recall

Table 1. Performance comparison between traditional ML algorithms.

Algorithms Default	Test set metrics					
	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUC-PR
Decision Tree	0.6113	0.4096	0.3990	0.4042	0.5598	0.3646
Random Forest	0.6832	0.5952	0.1295	0.2128	0.6357	0.4604
Gradient Boosting	0.6969	0.6600	0.1710	0.2716	0.6629	0.5118
Histogram Gradient Boosting	0.6661	0.4894	0.2383	0.3206	0.6352	0.4612
K-nearest Neighbors	0.6284	0.3929	0.2280	0.2885	0.5559	0.3592
Multi-layer Perceptron	0.6301	0.4320	0.3782	0.4033	0.5900	0.4031

Algorithms Tuned	Test set metrics					
	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUC-PR
Decision Tree	0.6866	0.6042	0.1503	0.2407	0.6276	0.4351
Random Forest	0.6901	0.7000	0.1088	0.1883	0.6679	0.5257
Gradient Boosting	0.6935	0.6207	0.1865	0.2869	0.6551	0.4961
Histogram Gradient Boosting	0.6866	0.6087	0.1451	0.2343	0.6618	0.4992
K-Nearest Neighbors	0.6764	0.5556	0.1036	0.1747	0.5737	0.4132
Multi-Layer Perceptron	0.6815	0.5315	0.3057	0.3882	0.6554	0.5005

(0.31), while Gradient Boosting and Random Forest showed competitive overall performance, reaching the highest accuracy (0.69) and AUC metrics (AUC-ROC = 0.67; AUC-PR = 0.53 for Random Forest). Histogram Gradient Boosting also ranked consistently among the top three models across all metrics, confirming its balanced trade-off between precision, recall, and discriminative capacity. In contrast, Decision Tree and K-Nearest Neighbors presented comparatively weaker results, with lower AUC and F1 values, reflecting limited ability to handle class imbalance



(a) AUC-PR Curve **(b) AUC-ROC Curve**

Figure 1. Curves Comparing Tuned Algorithms

When trained with default configurations, ensemble models achieved moderate AUC-ROC values (0.63–0.66) and AUC-PR values (0.46–0.51), but recall and F1-scores remained low (recall ¡ 0.24; F1 ¡ 0.32). After hyperparameter tuning, however, most algorithms improved precision and discriminative stability, though recall gains were modest. Overall, ensemble-based approaches maintained the best balance across evaluation metrics.

Cross-validation results revealed consistent trends across folds, with all models

showing higher training than validation metrics, particularly for AUC-ROC and AUC-PR, indicating varying degrees of overfitting. The KNN model exhibited the most pronounced gap, with mean training AUC-ROC of 0.9995 versus 0.5658 in validation, and AUC-PR of 0.9980 versus 0.4153, evidencing strong overfitting and poor generalization. In contrast, models such as Decision Tree and MLP showed smaller discrepancies between training and validation metrics, reflecting more balanced but overall lower performance.

Training and optimization times varied significantly across algorithms with Multi-Layer Perceptron and Gradient Boosting being the most time-consuming, requiring approximately 14 minutes. The Random Forest followed with a moderate training time of about 5.5 minutes. In contrast, the Histogram Gradient Boosting (58.00s), K-Nearest Neighbors (38.78s), and Decision Tree (32.64s) were significantly faster, highlighting their computational efficiency compared to more complex models.

4. Conclusion

In this study, we systematically evaluated the impact of hyperparameter optimization on the predictive performance of traditional machine learning algorithms for CDG classification. Our findings showed that models trained with default configurations, although achieving moderate accuracy, exhibited imbalanced predictions, particularly reflected in low recall and F1-scores. Ensemble methods such as Random Forest, Gradient Boosting, and Histogram Gradient Boosting displayed the most consistent improvements after tuning, achieving balanced accuracy, precision, and AUC metrics across evaluations. These results emphasize the crucial role of hyperparameter optimization in enhancing model generalization and mitigating class imbalance across evaluation metrics. Future work will expand the range of model families and optimization strategies explored and investigate interpretability approaches to better understand the biological signals underlying predictions and to derive insights relevant to cancer research.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by grants from FAPERGS [22/2551-0000390-7 (Project CIARS)] and CNPq [308075/2021-8].

References

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Schulte-Sasse, R., Budach, S., Hnisz, D., and Marsico, A. (2021). Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, 3(6):513–526.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127):1546–1558.