A□2025
SUMMER SCHOOL
ai.uni-jena.de

Machine Learning-Based Feature Selection and Classification of NSCLC Using Gene Expression Data

Oleksii Vekha

Friedrich Schiller University Jena

PURPOSE AND RELEVANCE

Non-small cell lung cancer (NSCLC) remains a leading cause of cancer-related mortality causing >1.5 million deaths annually, making accurate early classification crucial. This review examines two computational strategies applied to NSCLC gene expression data: hybrid feature selection with ensemble learning and regularized regression. Their potential for identifying informative gene subsets, improving classification, and supporting biomarker discovery is discussed, highlighting complementary contributions to robust, data-driven diagnostic tools.

DATA SCOURSES

To build and validate machine learning models for NSCLC classification, publicly available gene expression datasets from the GEO database were utilized. These datasets provide diverse and comprehensive molecular profiles necessary for robust biomarker discovery and model training.
Specifically, the following datasets were used:
- **GSE10072, GSE19804, and GSE19188** [1-3],[5]

All main datasets were split into training (70%) and testing (30%) sets for experiments.

Dataset	No. of genes	Classes	No. of samples (normal/NSCLC)	Training/test ratio
GSE10072	250	Normal/Tumor	104(49/55)	74/33
GSE19804	54675	Normal/Tumor	120(60/60)	84/36
GSE19188	54675	Normal/Tumor	156(65/91)	109/47

Table 1. Overview of GEO datasets used for NSCLC classification, including number of genes, sample classes, sample sizes, and training/testing splits.

FEATURE SELECTION METHODS

Before building classifiers, it's crucial to pick out the most important genes from the high-dimensional data. Two main approaches were compared:

1. **Ensemble-based filtering**[1]**:** Statistical filters such as ANOVA (testing differences between groups) and Mutual Information (measuring dependence between features and classes) were applied separately and in combination (**ANOVA+MI**) to select the most relevant genes. This multi-filter strategy was then integrated with ensemble algorithms like Random Forest and others, improving robustness through overlapping feature selection and reducing noise in the final model.

$$MI(C;F) = E(C) - E(C\,|\,F)$$

Eq 1. Mutual Information between class C and feature F [1].

$$f(X_a) = \frac{\sum_{i=1}^{C} n_i \left(\bar{x}_l \cdot - \bar{x}_l \cdot\cdot\right)^2 / (c-1)}{\sum_{i=1}^{C} \sum_{j=1}^{n_j} \left(x_{ij} - \bar{x}_l\right)^2 / (n-c)}$$

Eq 2. ANOVA F-statistic for feature X_a across c classes with $\bar{x}_l$ - mean value of X_a in class I, x_{ij} is the observed values of feature X_a for samples of class I, and $\bar{x}_{l..}$is the mean value of X_a of all samples in the data set [1].

2. **Regularized Logistic Regression:** Logistic regression models were enhanced with advanced regularization terms to address the high dimensionality of gene expression data and the risk of overfitting. The **LogSum + L_2**[3] approach combines a non-convex LogSum penalty—promoting strong sparsity by shrinking small coefficients to zero—with an L_2 term that stabilizes model estimates. **Hybrid $L_{1/2}$ + L_2**[4] regularization balances the sparsity induced by the $L_{1/2}$ norm with the smoothness of the L_2 norm, enabling selection of a minimal yet informative gene subset while maintaining robust decision boundaries.

Gene reduction results:
Across datasets, ensemble-based filtering reduced the feature space to a compact set of candidate genes, while regularized logistic regression narrowed it further to a smaller, highly informative subset. These gene sets were then used for model training and evaluation.

CLASSIFICATION AND EVALUATION

Once the optimal gene subsets were identified, they were used to train models for **NSCLC vs. normal tissue classification**:

1. **Ensemble Methods**[1]**:** Tested 18 model combinations by pairing three feature selection strategies (ANOVA, MI, ANOVA+MI) with six ensemble classifiers (Random Forest, Extra Trees, Gradient Boosting, AdaBoost, XGBoost, CatBoost). Performance was evaluated using 10-fold cross-validation on training data and independent test sets, with metrics including accuracy, F1-score, perecision, and sensitivity.
2. **Regularized Logistic Regression Models**[3-4]**:** Applied selected genes from LogSum + L_2 and Hybrid $L_{1/2}$ + L_2**(HLR)** methods directly in logistic regression. These were compared against standard L_1 regularized logistic regression and LEN models.

Cross-validation (10-fold) was used for training evaluation, followed by testing on separate validation datasets. Reported metrics included accuracy, sensitivity and specificity.

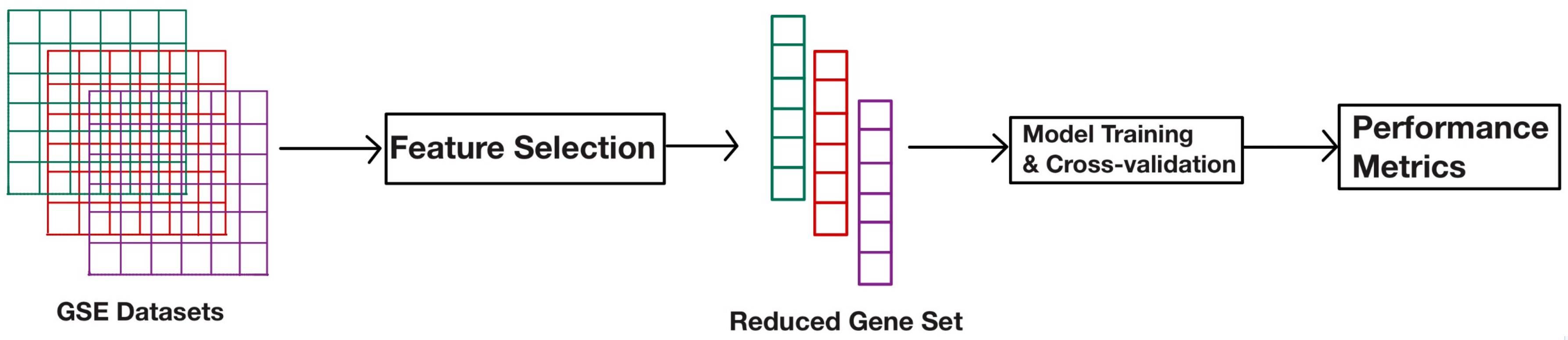


Figure 1. Feature Selection & Classification Pipeline.

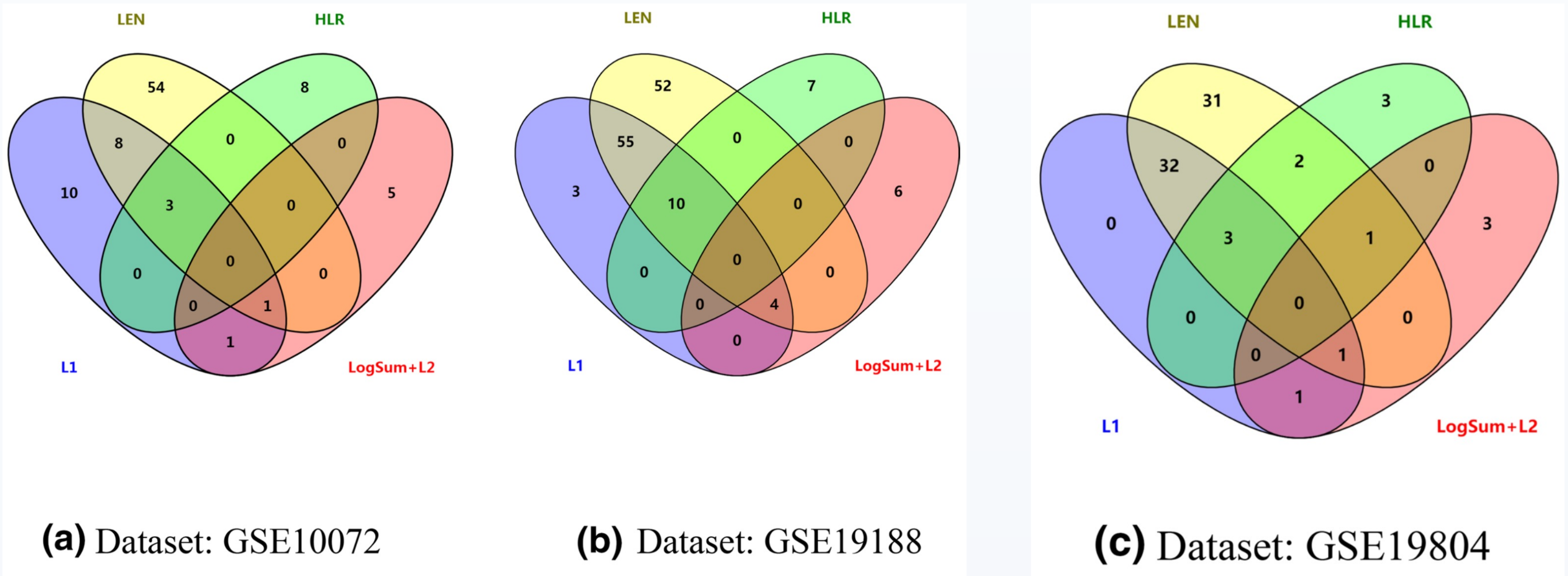
RESULTS

Dataset	FS Method	Classification Alg.	Accurasy	No. of selected genes
GSE10072	ANOVA + MI	Random Forest	100%	unknown
	ANOVA + MI	CatBoost	100%	unknown
	embedded	LogSum + L2	99.15%	7
GSE19804	MI	XGBoost	97.22%	8
	ANOVA + MI	Random Forest	97.22%	unknown
	ANOVA + MI	AdaBoost	97.22%	unknown
GSE19188	ANOVA	Random Forest	100%	10
	MI	Random Forest	100%	2
	MI	Extra Trees	100%	7

Table 2. Top-3 performing models (based on test accuracy) for each dataset [1],[3-4].

Across datasets, top accuracies exceeded 97%, with several methods (ANOVA + MI + Random Forest/CatBoost, MI + XGBoost/Extra Trees) reaching 100%. Embedded LogSum + L_2 also performed strongly, achieving up to 99.15% with only 6–10 genes. In contrast, L_1 and LEN feature selection in logistic regression showed lower test accuracies (≈46–96%). Overall, **ensemble-based classifiers** delivered **more consistent results** (≈94–100%) than regression-based models, confirming their robustness for NSCLC vs. normal tissue classification.

The top-performing models identified key NSCLC markers such as **TUBB1**, **CLDN18**, **EGFR** , and **MUC1**. Ensemble-based classifiers generally selected broader but overlapping gene sets, while regression-based approaches (L_1, LEN, LogSum + L_2) yielded more compact and partially unique subsets.



(a) Dataset: GSE10072 (b) Dataset: GSE19188 (c) Dataset: GSE19804

Figure 2. Venn diagram analysis of the results of L1, LEN, HLR and LogSum + L2 regularization methods [3].

CONCLUSIONS

This study demonstrates that combining hybrid feature selection with ensemble learning provides robust and high-performing models for NSCLC classification. Regression-based models, though less accurate, offered more compact gene signatures, highlighting a balance between performance and interpretability. These results support the proposed pipeline as a reliable tool for biomarker discovery, with future efforts directed toward validation on larger, more diverse patient cohorts and functional characterization of the selected genes to confirm their clinical utility.

REFERENCES

[1] Fhira Nhita and Isman Kurniawan, “Classification of Non-Small Cell Lung Cancer Based on Gene Expression in Cases of Smokers and Non-Smokers using Ensemble Methods with Statistical Based Feature Selection“ , 2022.

[2] Tcharé Adnaane Bawa, Yalçın Özkan and Çiğdem Selçukcan Erol, “Reanalysis of Non-Small-Cell Lung Cancer Microarray Gene Expression Data”, 2020.

[3] Xiao-Ying Liu, Sheng-BingWu, Wen-Quan Zeng, Zhan-JiangYuan & Hong-Bo Xu, “LogSum+L2 penalized logistic regression model for biomarker selection and cancer classification”, 2020.

[4] Hai-Hui Huang, Xiao-Ying Liu, Yong Liang, “Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid L1/2 +2 Regularization”, 2016.

[5] Fei Yuana,* , Lin Lub , Quan Zou, “Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms”, 2020.

CONTACT INFORMATION

E-Mail:alexeiveha@gmail.com LinkedIn:www.linkedin.com/in/vekha-oleksii-387100325