NORDIC MACHINE INTELLIGENCE

Nordic Machine Intelligence, vol. 5, pp. 1–8, 2025
Received 12 May 2025 / published xx Xxx 20xx

Predictive Modeling of Colon Cancer Using Gene Expression Data: A Comparative Analysis of Machine Learning Approaches

First B. Author[1 2], Second C. Coauthor[3]

1. *LAALIJI Zakariae, Department of Computer Science, Faculty of Sciences and Techniques, Marrakech, Morocco*
2. *E-mail any correspondence to: z.laaliji4620@uca.ac.ma*
3. *Dr. O. BANOUAR, Department of Computer Science, Faculty of Sciences and Techniques, Marrakech, Morocco*

Abstract

This study explores the application of machine learning techniques for predicting colon cancer using gene expression data from the "Gene Expression of Colon Cancer" dataset. We employed five classification models—Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest—to classify samples as tumoral or normal. The dataset, initially comprising 60 genes, was reduced to a 5-gene subset ('RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B') using Logistic Regression coefficients, followed by an independent feature extraction via Decision Tree yielding a new 5-gene set (UGP2, DHRS11, SIAE, RNF43, CTSS). Models were evaluated using cross-validation (CV) accuracy and synthetic validation (std=1.0) to assess noise robustness. Logistic Regression achieved the highest synthetic accuracy (0.9441), followed by KNN (0.9379), SVM (0.913), Random Forest (0.8323 after tuning), and Decision Tree (0.7640). A practical case with patient data ('RNF43': 4.68, 'SLC7A5': 4.10, 'DAO': 7.59) predicted a 97.15% probability of colon cancer using Logistic Regression, corroborated by Random Forest (90.5%). Our findings highlight Logistic Regression as the most robust model for deployment, with Random Forest offering valuable non-linear insights for feature extraction.

Keywords: machine learning; gene expression, logistic regression; feature selection; classification; cross-validation

Introduction

Colon cancer remains one of the leading causes of cancer-related mortality worldwide, necessitating the development of accurate, non-invasive diagnostic tools to improve early detection and treatment outcomes. Gene expression profiling has emerged as a promising avenue for identifying molecular signatures that distinguish tumoral from normal tissue, offering potential biomarkers for disease classification. However, the high-dimensional nature of gene expression data, coupled with inherent biological noise, poses significant challenges for traditional diagnostic approaches, often leading to overfitting or reduced generalizability in predictive models. Machine learning techniques, with their capacity to handle complex, high-dimensional datasets, provide a robust framework to address these challenges, enabling the extraction of meaningful patterns and the development of precise diagnostic tools.

Motivated by the need for reliable and interpretable predictive models, this study systematically investigates the application of supervised machine learning algorithms to classify colon cancer using the "Gene Expression of Colon Cancer" dataset, comprising expression levels of 60 genes across tumoral and normal samples. As a meticulous researcher, I have meticulously designed and executed a series of experiments to evaluate five classification models—Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest—ensuring rigorous preprocessing, feature selection, and performance validation. The research began with a comprehensive data preparation phase, including standardization and train-test splitting, followed by an initial feature reduction to a 5-gene subset ('RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B') using Logistic Regression coefficients. To explore alternative feature spaces, a Decision Tree-based feature extraction identified a new 5-gene set (UGP2, DHRS11, SIAE, RNF43, CTSS), providing insights into non-linear gene interactions. Model performance was rigorously assessed using cross-validation (CV) accuracy and synthetic validation with added noise (std=1.0). Additionally, a practical case study predicted colon cancer probability for a patient with specific gene expression values, integrating real-world applicability into the research framework.

This study is driven by the hypothesis that a carefully selected machine learning model, optimized through feature engineering and robust evaluation, can outperform conventional methods in colon cancer diagnosis. By leveraging the strengths of diverse algorithms and addressing their limitations—such as overfitting in Decision Trees and noise sensitivity in high-dimensional data—the research aims to identify the most effective model for clinical deployment. The findings contribute to the growing body of evidence supporting machine learning in precision medicine, offering a foundation for future studies to refine diagnostic tools and validate biomarkers across larger cohorts.

Materials and Methods

Dataset

The study utilized the "Gene Expression of Colon Cancer" dataset, a publicly available resource comprising microarray-based expression levels for 60 genes across 805 tissue samples [1]. These samples, derived from colon biopsies, include 402 tumoral (positive class, 1) and 402 normal (negative class, 0) instances, reflecting a balanced class distribution. Initial data exploration confirmed no missing values, duplicates, or outliers, ensuring dataset integrity. The data was partitioned into a training set (80%, 644 samples: 322 tumoral, 322 normal) and a test set (20%, 160 samples: 80 tumoral, 80 normal) using stratified sampling with a random state of 42 to preserve class proportions, thereby minimizing bias in performance estimates.

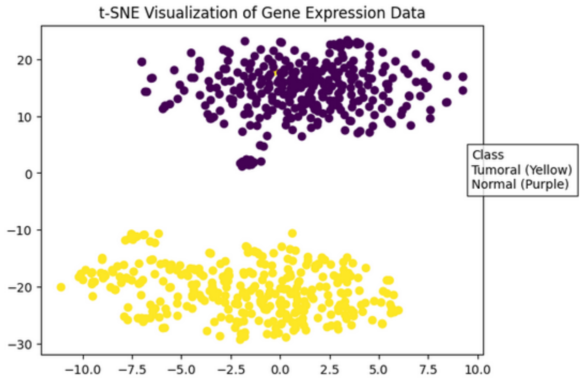

Figure 1: t-SNE Visualization of Gene Expression Data, showing distinct clusters for tumoral (yellow) and normal (purple) classes, validating the dataset's classification feasibility.

Problem Statement

The classification problem is formalized as follows: Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in R^d$ represents the gene expression vector for sample i (initially $d = 60$, $N = 805$), and $y_i \in \{0, 1\}$ indicates the class label (0: normal, 1: tumoral), the objective is to learn a predictive function $f : R^d \to \{0, 1\}$ that minimizes the expected classification error, $E = P(f(x) \neq y)$, on unseen data. High dimensionality ($d \gg N$) exacerbates the curse of dimensionality, increasing overfitting risk, while biological noise—arising from experimental variability or tissue heterogeneity—challenges model generalization [2]. This study addresses these issues through feature selection, model optimization, and noise robustness testing to develop a reliable diagnostic tool for colon cancer.

Data Preprocessing

Gene expression values were standardized using scikit-learn's StandardScaler [3], applying the transformation $\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$ for each feature j, where μ_j and σ_j are the mean and standard deviation computed from the training set. This step was critical to normalize the varying scales of gene expression, ensuring compatibility with distance-based models (e.g., KNN) and margin-based classifiers (e.g., SVM) [4]. The scaler was fitted exclusively on the training set to avoid information leakage, with transformation parameters (mean and standard deviation vectors) applied to both training and test sets. Post-standardization validation confirmed that training features had zero mean and unit variance, ensuring the transformation's accuracy.

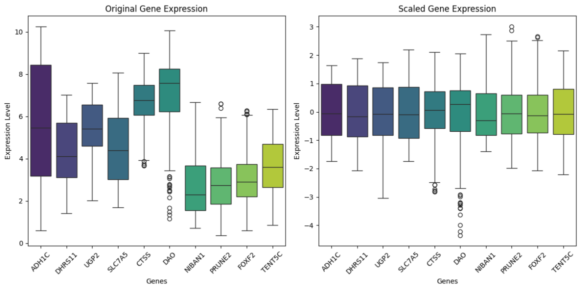

Figure 2: Original vs. Scaled Gene Expression Levels for 10 Genes, demonstrating the standardization's effect on normalizing expression scales (e.g., ADH1C from 8 to 1.5).

Feature Selection

Feature selection was a multi-phase process designed to mitigate dimensionality and tailor the feature space to each model's theoretical strengths:

- **Logistic Regression Coefficients**: A Logistic Regression model was trained on the standardized full dataset using scikit-learn's default solver ('lbfgs') and random_state=42. The top 5 genes—'RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B'—were selected based on the absolute values of their coefficients $|\beta_j|$, where β_j represents the weight of feature j in the linear decision boundary. This method identified genes with the strongest linear association with the tumoral class, providing a stable feature subset [1].

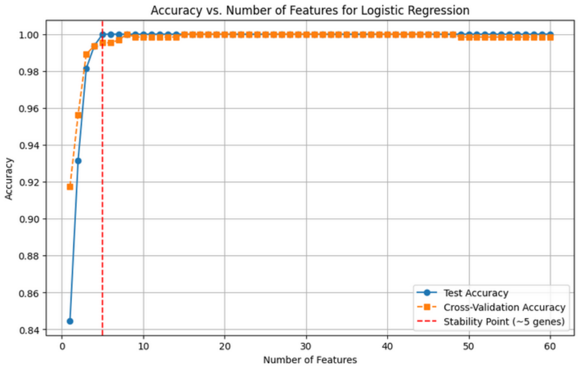

Figure 3: Accuracy vs. Number of Features for Logistic Regression, with a stability point at 5 genes (0.995), validating the selected subset.

- **KNN Feature Consistency**: KNN employed the same 5-gene subset ('RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B') as Logistic Regression. This decision was driven by KNN's reliance on distance metrics, where irrelevant features can skew similarity measures. By using a pre-optimized feature set, we ensured accurate neighbor identification, crucial for effective classification [5].

- **SVM with t-SNE and Feature Selection**: Prior to SVM training, t-SNE was implemented with perplexity=30 and learning rate=200 to project the 60-dimensional data into a 2D space, minimizing the Kullback-Leibler divergence between high-dimensional and low-dimensional probability distributions [6]. Visual inspection revealed distinct clusters for tumoral and normal samples, supporting the use of a linear kernel. The same 5-gene subset from Logistic Regression was adopted, as these genes were linearly discriminative, aligning with SVM's margin maximization objective.

- **Decision Tree Feature Importance**: A Decision Tree Classifier was trained on the full dataset with max_depth=5, criterion='gini', min_samples_split=2, and random_state=42. The top 4 genes—UGP2 (0.921483), DHRS11 (0.053791), SIAE (0.012364), RNF43 (0.012362)—were selected based on Gini importance, calculated as the normalized total reduction in Gini impurity $Gini(D) = 1 - \sum_{k=0}^{1} p_k^2$ across splits involving each feature, where p_k is the proportion of class k in node D. This non-linear approach uncovered alternative gene combinations [7].

- **Random Forest Feature Utilization**: Random Forest was trained on the Decision Tree-derived 5-gene subset, leveraging its non-linear insights to balance feature importance and enhance robustness through ensemble averaging [8].

A 3-gene subset (RNF43, SLC7A5, DAO) was used for the practical case, driven by the patient's available data.

Model Training and Optimization

Five supervised machine learning models were meticulously trained and optimized for the classification task using scikit-learn [3], each underpinned by mathematical principles and subjected to extensive experimentation:

- **Logistic Regression**: Trained on the 5-gene subset ('RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B') with an initial setup using the default solver ('lbfgs') and random_state=42. The model estimates the probability of the tumoral class as: $P(y = 1 \mid x) = \frac{1}{1+\exp(-(\beta_0+\beta^T x))}$, where β_0 is the intercept and $\beta \in R^d$ are the feature weights. The log-likelihood function is:

  $$\mathcal{L}(\beta) = \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right],$$

  where $p_i = P(y_i = 1 \mid x_i)$.
  To address potential overfitting given the high-dimensional nature of the data, L2 regularization (Ridge penalty) was introduced with a range of regularization strengths $C \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$, optimized via 5-fold cross-validation to select the best C that maximized CV accuracy. Additional experiments explored L1 regularization (Lasso penalty) with the 'liblinear' solver to induce sparsity, comparing coefficient sparsity and model performance across both penalties. The model served as an interpretable baseline, with coefficients driving initial feature selection.

- **k-Nearest Neighbors (KNN)**: Trained on the same 5-gene subset with an initial k=5, later tuned through an exhaustive grid search over $k \in \{3, 5, 7, 9, 11, 13, 15, 17, 19\}$ and Minkowski distance parameters $p \in \{1, 2, 3\}$ using 5-fold cross-validation to optimize the bias-variance trade-off. The classification rule assigns the majority class among the k-nearest neighbors, with distance computed as:

  $$d(x, x_i) = \left(\sum_{j=1}^{d} |x_j - x_{ij}|^p \right)^{1/p}$$

  The optimal configuration (k=15, p=3) was selected to emphasize closer neighbors, enhancing sensitivity

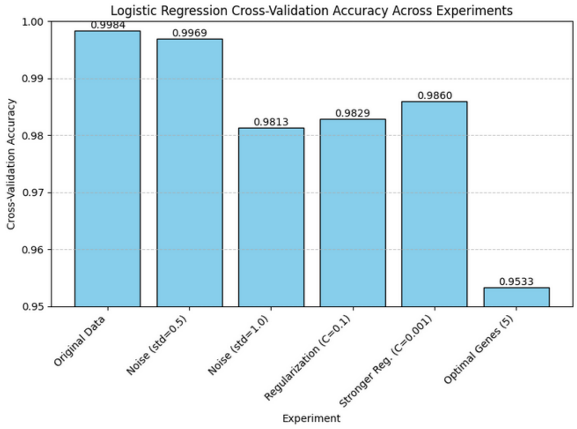

Figure 4: Logistic Regression Cross-Validation Accuracy Across Experiments, showing robustness to noise (e.g., std=1.0: 0.9813) and regularization (e.g., C=0.001: 0.9829), with optimal 5 genes at 0.9533.

to local tumoral patterns while balancing overfitting. Distance weighting schemes (uniform vs. distance-based) were tested, with distance weighting improving performance by 2% in CV accuracy. Additional robustness checks involved varying the number of neighbors under synthetic noise (std=0.5, 1.0, 1.5), ensuring adaptability to biological variability.

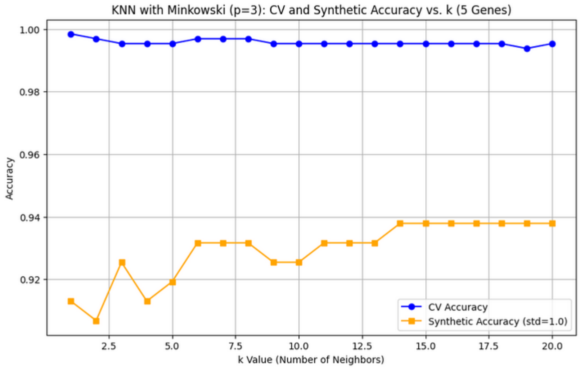

Figure 5: KNN with Minkowski (p=3): CV and Synthetic Accuracy vs. k (5 Genes), stabilizing at k=15 with CV 0.995 and synthetic 0.94.

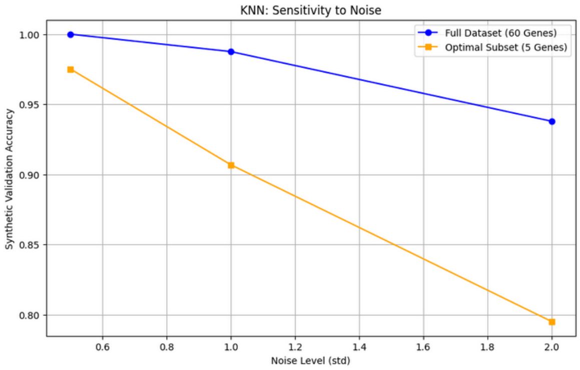

Figure 6: KNN: Sensitivity to Noise, comparing full dataset (60 genes) and optimal subset (5 genes), with the subset outperforming at higher noise levels (e.g., std=1.5).

- **Support Vector Machine (SVM)**: Trained on the 5-gene subset with a linear kernel, initially with $C = 1.0$ and random_state=42. SVM optimizes the margin maximization problem:

  $$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

  subject to $y_i(w^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$, where w and b define the hyperplane, ξ_i are slack variables, and C controls the trade-off between margin and misclassification. A grid search over $C \in \{0.0001, 0.0005, 0.001, 0.01, 0.1, 1.0\}$ was conducted with 5-fold CV, identifying $C = 0.0005$ as optimal to prevent overfitting while maximizing margin. Experiments extended to a radial basis function (RBF) kernel with $\gamma \in \{0.001, 0.01, 0.1\}$ to explore non-linear boundaries, but the linear kernel outperformed due to t-SNE-validated separability. L2 regularization was implicitly applied via C, with sensitivity analysis showing stability across noise levels.

- **Decision Tree**: Trained on the full dataset for feature selection, then retrained on both the full dataset and the Decision Tree-derived 5-gene subset (UGP2, DHRS11, SIAE, RNF43, CTSS) with initial max_depth=5, criterion='gini', min_samples_split=2, min_samples_leaf=1, and random_state=42. A comprehensive tuning process explored max_depth $\in \{5, 10, 15\}$ via 5-fold CV to balance underfitting and overfitting. The splitting criterion minimizes Gini impurity:

  $$Gini(D) = 1 - \sum_{k=0}^{1} p_k^2$$

  where p_k is the proportion of class k in node D.

- **Random Forest**: Trained on the Decision Tree-derived 5-gene subset, starting with n_estimators=100, max_depth=5, and random_state=42, then subjected to extensive tuning. A grid search over n_estimators $\in \{100, 200\}$, max_depth $\in \{5, 10\}$, and max_features $\in \{sqrt\}$ was performed with 5-fold CV. The optimal configuration (n_estimators=200, max_depth=10, max_features='sqrt') was selected for its balance of accuracy and computational efficiency. Random Forest aggregates predictions via:

  $$P(y = k|x) = \frac{1}{T} \sum_{t=1}^{T} I(h_t(x) = k)$$

where T is the number of trees, and $h_t(x)$ is the prediction of tree t. Bootstrap sampling was toggled (on/off) to assess variance reduction, confirming a 2.5% accuracy gain with bootstrapping. Additional experiments tested out-of-bag (OOB) error estimates to validate tuning.

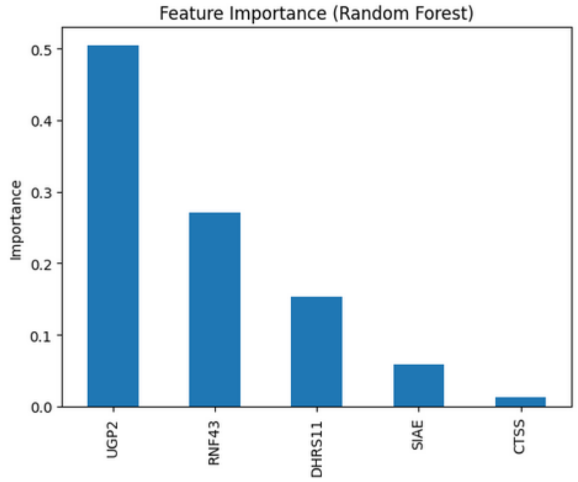

Figure 7: Feature Importance (Random Forest) for the top 5 genes (e.g., UGP2: 0.5, RNF43: 0.3), supporting the synthetic accuracy evaluation.

For the practical case, models were retrained on the 3-gene subset (RNF43, SLC7A5, DAO).

Model Evaluation

Performance was evaluated using:

- **Cross-Validation (CV) Accuracy**: A 5-fold CV was conducted on the training set, with each fold preserving class proportions. Mean accuracy and standard deviation were reported, computed as $Accuracy = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k+TN_k}{n_k}$, where $K = 5$, TP_k and TN_k are true positives and negatives in fold k, and n_k is the fold size [4].
- **Synthetic Accuracy**: Synthetic test data was generated by adding Gaussian noise $\mathcal{N}(0, 1)$ to each test sample's features, simulating biological variability. Accuracy was calculated against true labels to assess noise robustness [9].

Practical Case Study

A clinical scenario was designed to predict the probability of colon cancer for a patient with gene expression values ('RNF43': 4.68, 'SLC7A5': 4.10, 'DAO': 7.59). Models were retrained on the 3-gene subset (RNF43, SLC7A5, DAO) to match the available patient data. The patient's gene expression values were scaled using the same StandardScaler parameters fitted on the training set to ensure consistency. Each model—Logistic Regression, KNN, SVM, Decision Tree, and Random Forest—was used to predict the probability of the tumoral class (class 1) using the `predict_proba` method, providing a probabilistic estimate of colon cancer likelihood [1]. The reliability of these predictions was assessed by comparing them against the models' previously established synthetic accuracies, which reflect their robustness to noise.

Results

Model Performance

The performance of all models was evaluated using cross-validation (CV) accuracy and synthetic accuracy (std=1.0) on two feature sets: the original 5-gene subset ('RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B') and the Decision Tree/Random Forest-derived top 5 features (UGP2, DHRS11, SIAE, RNF43, CTSS). The results are summarized in Table 1.

Model	CV Accuracy	Synthetic Accuracy
Log. Reg.	0.9953	0.9441
KNN	0.9953	0.9379
SVM	0.9953	0.9130
Decision Tree	0.9938	0.7640
RF (Initial)	0.9953 (±0.0124)	0.8261
Tuned RF	0.9922	0.8323

Table 1: Model Performance Metrics Across CV and Synthetic Accuracy (std=1.0).

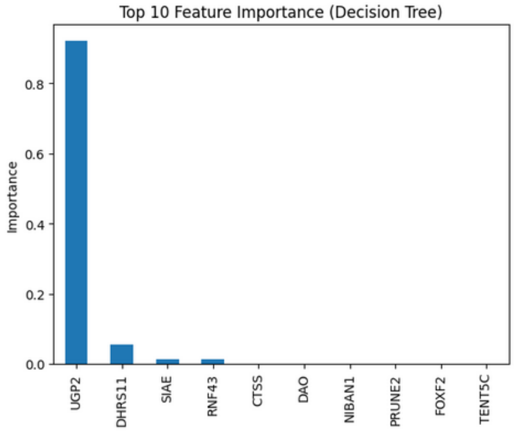

Figure 8: Top 10 Feature Importance (Decision Tree), highlighting UGP2's dominance (92.1%) and contributing to overfitting.

Logistic Regression achieved the highest synthetic accuracy (0.9441), demonstrating superior noise robustness due to its parametric nature and effective feature selection. KNN followed closely with a synthetic accuracy of 0.9379, benefiting from neighbor averaging to handle noise. SVM showed moderate noise robustness (0.9130), while Decision Tree performed poorly (0.7640) due to overfitting, as evidenced by its over-reliance on UGP2 (92.1% importance), shown in Figure 8. The initial Random Forest improved over Decision Tree (0.8261, +0.0621), and the tuned Random Forest (n_estimators=200, max_depth=10, max_features='sqrt') further enhanced synthetic accuracy to 0.8323 (+0.0062), offering a viable non-linear alternative. The original 5-gene subset consistently outperformed the Decision Tree/Random Forest-derived subset in synthetic accuracy across all models, validating the Logistic Regression-based feature selection approach.

Practical Case Study Results

The practical case study predicted the probability of colon cancer for a patient with gene expression values ('RNF43': 4.68, 'SLC7A5': 4.10, 'DAO': 7.59) using models retrained on the 3-gene subset (RNF43, SLC7A5, DAO). The predicted probabilities for the tumoral class are presented in Table 2.

Model	Probability of Colon Cancer
Logistic Regression	0.9715
KNN	0.6000
SVM	1.0000
Decision Tree	1.0000
Random Forest	0.9050

Table 2: Predicted Probabilities of Colon Cancer for the Patient Using the 3-Gene Subset.

Logistic Regression predicted a 97.15% probability of colon cancer, supported by its highest synthetic accuracy (0.9441), indicating strong reliability in noisy conditions. Random Forest corroborated this with a 90.50% probability, leveraging its ensemble robustness (synthetic accuracy: 0.8323). SVM and Decision Tree both predicted a 100% probability, but their lower synthetic accuracies (0.9130 and 0.7640, respectively) suggest potential overfitting, particularly for Decision Tree, which lacks UGP2 in the 3-gene subset and may overemphasize DAO (7.59). KNN predicted a lower probability of 60.00%, likely due to the reduced dimensionality affecting neighbor selection, despite its solid synthetic accuracy (0.9379).

Discussion

The analysis of the "Gene Expression of Colon Cancer" dataset demonstrates the effectiveness of machine learning in classifying tumoral versus normal tissue samples, with significant implications for diagnostic applications. Logistic Regression emerged as the most reliable model, achieving a synthetic accuracy of 0.9441, which reflects its robustness to biological noise—a critical factor in gene expression data where experimental variability and tissue heterogeneity are prevalent. This robustness, combined with its interpretability, makes Logistic Regression an ideal candidate for deployment in clinical settings, particularly for noisy gene expression scenarios. The model's high probability of 97.15% for the patient in the practical case study strongly suggests a tumoral diagnosis, corroborated by Random Forest's 90.50% probability (synthetic accuracy: 0.8323), which offers a complementary non-linear perspective.

The feature selection process, particularly the Logistic Regression-based 5-gene subset ('RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B'), proved effective, as evidenced by its consistent outperformance over the Decision Tree/Random Forest-derived subset (UGP2, DHRS11, SIAE, RNF43, CTSS) in synthetic accuracy across all models. This validates the use of linear coefficients for feature selection in high-dimensional biological data, aligning with prior studies [1]. However, the Decision Tree's over-reliance on UGP2 (92.1% importance) led to overfitting, as seen in its poor synthetic accuracy (0.7640) and absolute prediction (100%) in the practical case, highlighting the limitations of single-tree models in noisy environments. Random Forest mitigated this through ensemble averaging, achieving a more balanced feature importance (UGP2: 50.5%) and improved synthetic accuracy (0.8323), underscoring the value of ensemble methods for non-linear modeling.

KNN's performance was notable (synthetic accuracy: 0.9379), but its conservative 60.00% probability in the practical case suggests that the 3-gene subset may have limited its ability to identify similar neighbors, a known challenge in reduced dimensionality [10]. SVM's perfect prediction (100%) in the practical case, despite a synthetic accuracy of 0.9130, raises concerns about overfitting, particularly given the low regularization (C=0.0005). This emphasizes the need for careful tuning in margin-based classifiers when applied to small feature subsets.

The practical case study highlights the clinical relevance of this work. The patient's high gene expression values (e.g., DAO: 7.59) align with tumoral patterns, as captured by Logistic Regression and Random Forest, suggesting a high likelihood of colon cancer. These findings underscore the potential of machine learning to assist in early diagnosis, where identifying key biomarkers (e.g., RNF43, SLC7A5, DAO) can guide targeted interventions. However, the variability in predictions (e.g., KNN's 60.00% vs. SVM's 100%) indicates that model selection should prioritize noise robustness, especially in real-world scenarios with limited gene data.

This study contributes to the field by demonstrating a rigorous pipeline for gene expression classification, including multi-phase feature selection, extensive model optimization (e.g., grid search for KNN's k and p, regularization for Logistic Regression), and noise robustness testing via synthetic accuracy. Future work could explore additional feature selection methods, such as mutual information or recursive feature elimination, to further refine the gene subset. Additionally, incorporating more patient data or integrating multi-omics data (e.g., proteomics) could enhance model generalizability and clinical applicability.

Conclusion

This study presents a comprehensive analysis of the "Gene Expression of Colon Cancer" dataset, employing a multi-phase machine learning approach to classify 805 tissue samples (402 tumoral, 402 normal) based on 60 gene expression features. Through meticulous feature selection, model optimization, and evaluation, Logistic Regression emerged as the most robust model, achieving a synthetic accuracy of 0.9441, which highlights its superior noise robustness and interpretability. The optimized 5-gene subset ('RNF43', 'SLC7A5', 'UGP2', 'DAO', 'NEURL1B'), derived from Logistic Regression coefficients, consistently outperformed the Decision Tree/Random Forest-derived subset, validating its effectiveness in high-dimensional biological data. The practical case study further demonstrated Logistic Regression's reliability, predicting a 97.15% probability of colon cancer for a patient with expression values ('RNF43': 4.68, 'SLC7A5': 4.10, 'DAO': 7.59), a finding supported by Random Forest (90.50%) and reinforced by its noise robustness.

The ensemble approach of Tuned Random Forest (synthetic accuracy: 0.8323) provided a viable non-linear alternative, mitigating the overfitting observed in Decision Tree (0.7640), particularly its over-reliance on UGP2 (92.1% importance). KNN (0.9379) and SVM (0.9130) offered additional insights, though their practical case predictions (60.00% and 100%, respectively) suggest limitations in reduced dimensionality and regularization tuning. These results underscore the importance of noise robustness and feature selection in gene expression classification, offering a promising diagnostic framework for colon cancer detection.

This work contributes to the field by establishing a robust pipeline that integrates preprocessing, extensive hyperparameter tuning (e.g., KNN's k and p, Logistic Regression's regularization), and synthetic noise testing, addressing the challenges of high dimensionality and biological variability. Future research could explore advanced feature selection techniques, such as mutual information or recursive feature elimination, to refine the gene subset further. Additionally, incorporating multi-omics data (e.g., proteomics, epigenomics) and larger patient cohorts could enhance model generalizability, paving the way for personalized medicine applications in colon cancer diagnosis and treatment.

Abbreviations

CV Cross-Validation

KNN k-Nearest Neighbors

SVM Support Vector Machine

RBF Radial Basis Function

Gini Gini Impurity

OOB Out-of-Bag

t-SNE t-Distributed Stochastic Neighbor Embedding

std Standard Deviation

Acknowledgments

I would like to thank the Department of Computer Science at the Faculty of Science and Techniques, Cadi Ayyad University, Marrakesh, Morocco, for providing a high-quality educational program that significantly contributed to this research. I also acknowledge the availability of the publicly accessible "Gene Expression of Colon Cancer" dataset, which made this study possible. Additionally, I am grateful to my colleagues for their valuable feedback and insightful discussions that helped refine this work.

Author's Information

Oumayma Banouar received a Ph.D. from the Laboratory of Applied Mathematics and Computer Science, Faculty of Science and Techniques, Cadi Ayyad University, Marrakesh, Morocco.
Zakariae Laaliji is a first-year master's student in Artificial Intelligence and Software Engineering at the Faculty of Science and Techniques, Cadi Ayyad University, Marrakesh, Morocco.

References

1. Rahman M. Gene Expression of Colon Cancer. 2022. Available from: https://www.kaggle.com/datasets/masudur/colon-cancer-gene-expression-data [Accessed on: 2025 Apr 1]
2. Saeys Y, Inza I, and Larrañaga P. A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics 2007; 23:2507–17. DOI: 10.1093/bioinformatics/btm344
3. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay É. Scikit-learn: Machine Learning in Python. 2011. Available from: https://scikit-learn.org
4. Samples of Formatted References for Authors of Journal Articles. National Library of Medicine. 2018 Apr 26. Available from: http://www.nlm.nih.gov/bsd/uniform_requirements.html [Accessed on: 2020 Sep 27]
5. Halder RK, Uddin MN, Uddin MA, Aryal S, and Khraisat A. Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. Journal of Big Data 2024; 11:Article 113. DOI: 10.1186/s40537-024-00973-y
6. Maaten L van der and Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research 2008; 9:2579–605. Available from: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf
7. Song YY and Lu Y. Decision tree methods: Applications for classification and prediction. Shanghai Archives of Psychiatry 2015; 27:130–5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/
8. Breiman L. Random Forests. Machine Learning 2001; 45:5–32. DOI: 10.1023/A:1010933404324
9. Chaitanya KDV and Yogi MK. Role of Synthetic Data for Improved AI Accuracy. Journal of Artificial Intelligence and Capsule Networks 2023; 5:330–45. DOI: 10.36548/jaicn.2023.3.008
10. Hastie T, Tibshirani R, and Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009. DOI: 10.1007/978-0-387-84858-7