ML-Based Pan-Cancer Tumor Classification using RNA-Seq Data

Beck Schemenauer
bschemen@calpoly.edu

Gavin Lynch
galynch@calpoly.edu

December 2025

Abstract

This study investigates the application of machine learning (ML) models for highly accurate pan-cancer tumor classification using high-dimensional RNA-sequencing (RNA-Seq) gene expression data. A systematic methodology, including Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), and Random Forest-based Feature Selection, was employed to identify the most informative genes across five major cancer types. We demonstrate that various ML models (Logistic Regression, SVM, XGBoost, kNN, and Neural Networks) achieve near-perfect (99-100%) classification accuracy with a minimal set of features (as few as 3-10 top-ranked genes), significantly outperforming random feature baselines. The robust, model-independent signal validates the utility of targeted gene expression panels for accurate cancer diagnostics, with biological interpretation confirming the selected genes' known cancer-related functions.

1 Introduction

There are expected to be over 2 million new cancer diagnoses within the USA alone in 2025. Of these, lung, colorectal, breast, and prostate cancer are expected to account for over 50% of cases in men and women [1]. The ability to identify these tumors and differentiate between them is a huge field of research. Tumor classification is a highly pertinent topic within the bioinformatic and computational sciences communities, with many applications in the medical field. The ability to determine cancer type from RNA-seq data allows for quicker diagnosis and identification of tumors within the body. This can be life-changing for people with these cancers, as improvements to early diagnosis can lead to huge benefits in the effectiveness of treatment. Furthermore, these molecular classification techniques are critical for diagnosing Cancers of Unknown Primary (CUP), where metastatic tumors are identified without a clear origin; in such cases, transcriptomic signatures provide a vital tool for tracing the tissue of origin when standard analysis is inconclusive. Machine diagnosis can also take strain off medical professionals, who may have to analyze test results manually. Even as just a backup tumor type verification model, machine learning has the potential to benefit many parts of the tumor diagnosis pipeline. This research, and previous work on this dataset, aim to predict five tumor types: breast, colorectal, prostate, lung, and kidney. These five cancers are all within the top ten most frequently diagnosed cancer types, making their analysis even more pertinent. By 2030, the number of cancer cases worldwide is expected to rise by over 50% [1]. The ability of machine learning to assist already strained doctors and medical professionals presents a promising path forward in the future of medical diagnoses.

The goal of this paper is to propose a minimum set of genes, identified through feature selection, that can accurately predict cancer tumor type. Identifying tumor type through this Pan-cancer dataset has already been explored in previous publications, but we aim to analyze the minimum number of features required to achieve near-optimal performance. In this dataset's case, near optimal is over 99%, as previous papers have achieved nearly 100% accuracy [2]. This paper will also explore the biological significance behind the top-selected genes and why high expression of those genes is such a good predictor of tumor type. We also propose a feature selection and lightweight model pipeline that achieves a leading performance of 100% accuracy using only 39 selected genes from this dataset.

2 Problem Definition and Algorithms

2.1 Task Definition

The RNA-Seq PANCAN dataset is an open-source gene expression dataset hosted on the UCI Machine Learning repository [3]. This dataset contains 801 samples of gene expression data measured using the Illumina HiSeq platform. Each sample contains the expression data for 20,531 genes and is already cleaned to have no missing features. The target variable consists of five possible classes: BRCA, KIRC, COAD, LUAD, and PRAD. These five categories stand for the five cancer types being analyzed: breast, kidney, colon, lung, and prostate, respectively. This is inherently a multiclass classification problem.

2.2 Algorithm Definitions

2.2.1 Feature Selection

The method of analysis explored in this study begins with a feature selection pipeline, built on top of a Random Forest (RF) estimator. Random Forest is an inherently interpretable algorithm, meaning feature importances can be directly extracted from the model. RF is built on an ensemble of decision trees, which individually split features by those that minimize the Gini impurity. This allows the model to sequentially select the highest predictive features from the dataset. In order to achieve more stable features, the RF model is trained, and features extracted 25 total times. These feature scores are summed and sorted to determine the overall most informative features. These features are then used for the downstream analysis. The implementation details of each cross validation split will be explored in the methodology section, along with a more detail exploration into the results.

Algorithm 1 Stratified Nested Feature Selection & Ranking

Require: Expression Matrix $X \in \mathbb{R}^{N \times M}$, Labels $y \in \mathbb{R}^N$
Require: Outer Folds $K_{out} = 5$, Inner Folds $K_{in} = 5$
Require: Random Forest Hyperparameters θ
Ensure: Global Ranked Feature List L_{ranked}
Initialize global importance vector $\mathbf{I}_{global} \leftarrow \mathbf{0}^M$
for $k = 1$ to K_{out} **do** ▷ Outer Loop
  Split X, y into X_{train}, X_{test} via Stratified K-Fold
  Initialize fold importance vector $\mathbf{I}_{fold} \leftarrow \mathbf{0}^M$
  for $j = 1$ to K_{in} **do** ▷ Inner Loop on X_{train}
    Split X_{train} into $X_{inner\_train}, X_{inner\_val}$
    $Model \leftarrow \text{TrainRF}(X_{inner\_train}, y_{inner\_train}, \theta)$
    Get feature importances $\mathbf{i} \leftarrow Model.\text{feature\_importances\_}$
    $\mathbf{I}_{fold} \leftarrow \mathbf{I}_{fold} + \mathbf{i}$ ▷ Sum across inner folds
  end for
  $\mathbf{I}_{global} \leftarrow \mathbf{I}_{global} + \mathbf{I}_{fold}$ ▷ Aggregate for stability
end for
$L_{ranked} \leftarrow \text{ArgsortDescending}(\mathbf{I}_{global})$
return L_{ranked}

2.2.2 Logistic Regression

The first model evaluated was Multinomial Logistic Regression, selected as a robust linear baseline. As a generalized linear model, it separates classes via a linear decision boundary in the high-dimensional feature space. While less complex than ensemble or neural methods, its performance serves as a crucial reference point; if this linear baseline achieves high accuracy, it indicates that the classes are linearly separable and more computationally expensive non-linear models may be unnecessary. For this linear baseline, we employ multinomial logistic regression with ℓ_2 regularization. For a sample $\mathbf{x} \in \mathbb{R}^d$ and K classes, the model

estimates the probability that $\mathbf{x}$ belongs to class k using the softmax function:

$$P(y = k \mid \mathbf{x}) = \frac{e^{\mathbf{w}_k^\top \mathbf{x} + b_k}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^\top \mathbf{x} + b_j}} \tag{1}$$

where $\mathbf{w}_k$ and b_k are the weight vector and bias for class k. The model parameters $\mathbf{W}$ are learned by minimizing the regularized cross-entropy loss:

$$\mathcal{J}(\mathbf{W}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}(y^{(i)} = k) \log \Big(P(y^{(i)} = k \mid \mathbf{x}^{(i)}) \Big) + \frac{\lambda}{2} \|\mathbf{W}\|_F^2 \tag{2}$$

where N is the number of samples, $\mathbb{1}(\cdot)$ is the indicator function, and λ controls the strength of the regularization to prevent overfitting on the high-dimensional gene feature space.

2.2.3 Support Vector Machines

The next model, Support Vector Machines (SVM), is another linear classifier that creates hyperplanes through the high dimensional feature space to split the data. Logistic Regression tries to optimize a probabilistic objective, while SVM creates a geometric distinction between classes. It constructs a hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$ that separates classes while maximizing the margin, the distance between the hyperplane and the nearest data points (support vectors). SVM only considers datapoints close to this hyperplane, meaning it is more stable to outliers that are still classified correctly. Instead of optimizing the regularized Cross entropy loss, SVM optimizes hinge loss.

$$\min_{\mathbf{w},b} \quad \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{\text{Regularization}} \quad + \quad C \sum_{i=1}^{N} \underbrace{\max\Big(0, 1 - y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b)\Big)}_{\text{Hinge Loss}} \tag{3}$$

The first term maximizes the geometric margin (by minimizing the norm of $\mathbf{w}$), helping to prevent overfitting on the high-dimensional gene expression features. The second term, the Hinge Loss, ensures that the model pays a linear penalty only when a sample is misclassified or violates the margin. The hyperparameter C controls the trade-off: a high C enforces a strict margin (low bias, high variance), while a low C allows more margin violations for a simpler decision boundary (higher bias, low variance).

2.2.4 K-Nearest Neighbors

The next type of model tested was a K-Nearest Neighbors (KNN) classifier. This model is a clustering model, meaning it identifies classes by creating distinct "clusters" of points that are most related. Unlike SVM and Logistic Regression, KNN does not train weights; it simply compares test data points to similar data points in its training classes and assigns them to the same class. To achieve this, the KNN must create high-dimensional groupings of data points, which is often computationally expensive as the number of features increases [4].

We utilize the standard Euclidean distance to quantify similarity between the test sample $\mathbf{x}$ and a training sample $\mathbf{x}^{(i)}$:

$$d(\mathbf{x}, \mathbf{x}^{(i)}) = \|\mathbf{x} - \mathbf{x}^{(i)}\|_2 = \sqrt{\sum_{j=1}^{d} (x_j - x_j^{(i)})^2} \tag{4}$$

2.2.5 Feed Forward Neural Network

To test a more modern approach, this study utilizes a simple feed-forward neural network (NN) as a proof of concept for deep learning techniques in this domain. Neural Networks create layers of "neurons" that learn weights and biases for each connection, allowing complex patterns to be understood. This network has n input neurons, corresponding to the number of input genes, and five output neurons corresponding to the number of possible cancer types in this dataset.

2.2.6 Extreme Gradient Boosting

Extreme Gradient Boosting, or XGBoost, is a powerful tree-based model built on top of decision trees, similar to Random Forest. Because of this, XGBoost maintains the explainability that makes RF so useful in biological domains. XGBoost trains decision tree "stumps" (low-depth decision trees) to classify the gradient of the loss function. This sequential approach allows the model to incrementally correct the errors (residuals) of previous trees, effectively performing gradient descent in the function space. Unlike standard gradient boosting, XGBoost includes a unique regularized objective function that penalizes model complexity (both tree depth and leaf weights), which is critical for preventing overfitting on high-dimensional data like RNA-Seq [5].

2.3 Expectations

While this dataset and problem have been explored extensively in previously published research, this study hopes to apply feature selection to improve the state-of-the-art accuracy on this dataset, using as few features as possible. Based on these previous studies, we expect to be able to achieve relatively high accuracy on a few genes after applying rigorous feature selection. We hope that we will be able to push the boundary forward and find genes that are able to explain the variance in cancer type well. We also expect most of our model types to easily be able to classify the different cancer types, especially once the dimensionality is reduced from 20k genes to something around 100. Because of this, there is likely to be little difference between model performance. Although gene counts can be low, around 10, the differences in model architectures may present advantages or disadvantages over others.

3 Experimental Evaluation

3.1 Methodology

The input data was structured into two matrices: an expression matrix $X \in \mathbb{R}^{801\times 20531}$ containing gene expression values (samples as rows, genes as columns), and a label vector $y \in \{1,\ldots,5\}^{801}$ representing the five tumor types (BRCA, COAD, KIRC, LUAD, PRAD). Labels were encoded as categorical integers prior to model training.

Feature selection was performed using a Random Forest classifier configured with `n_estimators = 500` to ensure stable aggregation of feature importances, `max_features = "sqrt"` to maintain computational scalability in a high-dimensional setting, `n_jobs = -1` for full CPU utilization, and `class_weight = "balanced_subsample"` to correct for class imbalance during tree construction.

We implemented a stratified nested cross-validation to obtain unbiased performance estimates and stable feature rankings:

Outer loop: 5-fold stratified CV to generate train/test partitions.

Inner loop: For each outer-train split, a 5-fold inner CV was performed.

Feature importance: A Random Forest model was trained on each inner-fold training set using distinct random seeds. Feature importances (one importance value per gene) were extracted from each of the 5 inner models.

Importance aggregation: Importances were summed across the 5 inner folds to obtain a stable importance score per gene for that outer fold. Genes were ranked by these summed importance scores.

Top-N evaluation: For each $N \in [1, 50]$, the top-N ranked genes were selected. A classifier was trained on the outer-train split (restricted to those N genes), then evaluated on the outer-test split. Mean accuracy and standard deviation were recorded across outer folds.

Global importance ranking: Summed importances from all outer folds were aggregated (5 outer $\times$ 5 inner = 25 RF models total), yielding a final ranked gene list for downstream biological interpretation.

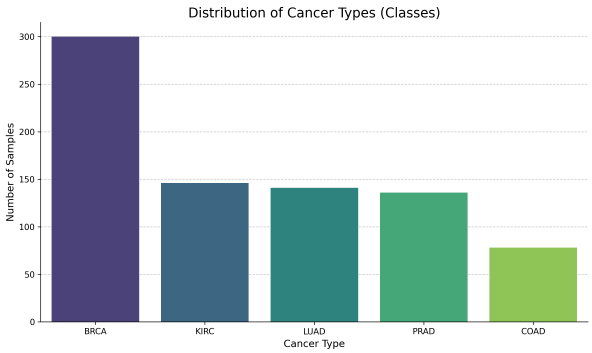

Figure 1: Class distribution across the five selected cancer types.

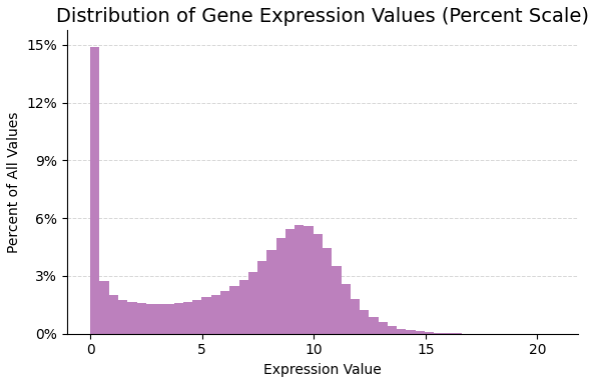

Figure 2: Distribution of gene expression values.

Classifiers evaluated in this framework included multinomial logistic regression (LR), support vector machines (SVM), k-nearest neighbors (KNN), a simple feed-forward neural network (NN), and XGBoost (XGB). All linear and distance-based models were trained using standardized features via a `StandardScaler` inside a scikit-learn pipeline.

3.2 Results

Prior to applying the full modeling pipeline, we conducted exploratory data analysis (EDA) and principal component analysis to characterize the dataset. Class distribution was moderately imbalanced, with BRCA (n=300), KIRC (146), LUAD (141), PRAD (136), and COAD (78) samples (Figure 1). Gene expression values were approximately Gaussian for many genes, with substantial mass near zero due to low expression in subsets of samples (Figure 2). While these observations suggest that transformations or resampling strategies could be considered, none were strictly required for the models used.

PCA (projected to three dimensions) revealed clear separation among the five tumor classes (Figure 3), indicating strong underlying structure in the data. This led us to expect very high classification performance, especially after feature selection. Although multiclass imbalance often motivates metrics such as balanced accuracy or macro-F1, the strong PCA separation suggested that simple accuracy would be highly informative. Accuracy proved sufficient, with most models achieving near-perfect performance using only small numbers of top-ranked genes (Table 1). Logistic regression performed best overall, reaching 100% accuracy at $N = 39$ selected genes. We also tracked the standard deviation across outer folds to quantify variability due to both feature-subset instability and train/test partitioning. Results are summarized in the grouped bar chart (Figure 4).

Across models, performance was consistently high. Even the lowest performing model (NN) approached the same accuracy as other classifiers, though with higher variance at low N. Variability decreased as more genes were added, reflecting increased model stability. Accuracy curves flattened early, showing that adding additional genes beyond the top $\sim$10 yielded minimal improvements. Across all models, 99–100% accuracy was achieved using fewer than 50 genes, and 3–10 genes were typically sufficient to reach near-maximum performance. To evaluate the effectiveness of our Random Forest-based feature selection, we compared against models trained on randomly selected genes (Figure 5).

As expected, random gene sets achieved high accuracy only when many genes were included, and performance was unstable across folds. In contrast, feature-selected genes achieved comparable or higher accuracy with far fewer features, and their performance variance was substantially lower. Importantly, the selected genes provided a consistent, interpretable signal reflective of true biological differentiation. These results confirm that RF-based feature selection captured meaningful biological information rather than noise.

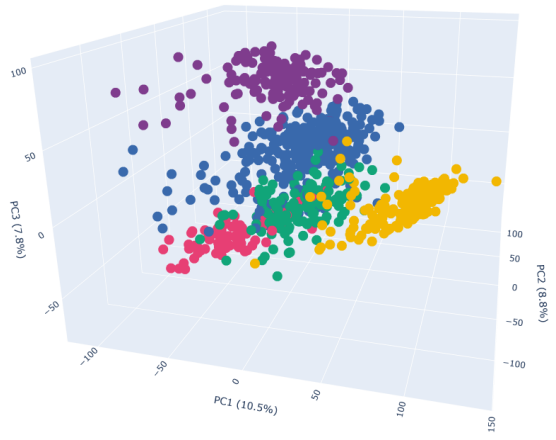

Figure 3: Three-dimensional PCA projection showing clear separation among the five tumor classes.

Table 1: Nested CV mean accuracy at selected numbers of features (N).

N	Logistic Reg.	SVM	XGBoost	KNN	Neural Net
1	63.54%	63.17%	58.42%	54.17%	55.05%
2	83.27%	82.40%	78.52%	80.65%	76.91%
3	93.01%	92.51%	90.76%	91.01%	81.90%
4	95.75%	96.63%	95.88%	96.25%	92.88%
5	97.50%	97.00%	97.13%	97.25%	95.75%
10	99.00%	99.00%	98.75%	99.13%	97.75%
39	100.00%	99.88%	99.62%	99.50%	98.88%
40	100.00%	100.00%	99.62%	99.50%	99.13%

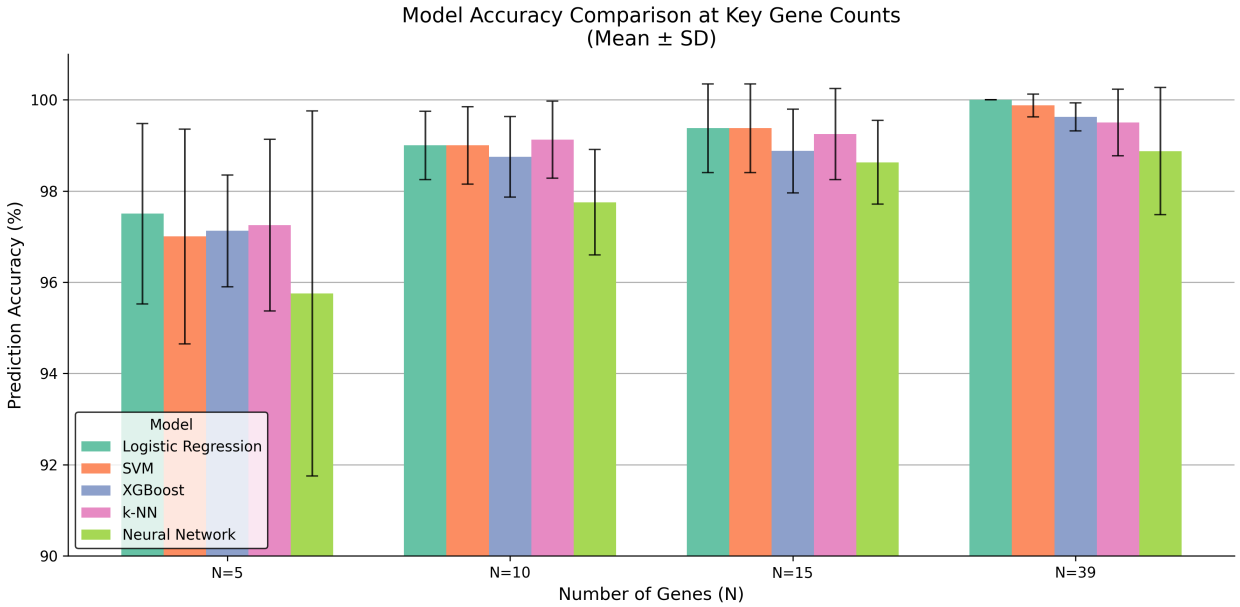

Figure 4: Mean accuracy and standard deviation for each model and feature-set size.

3.3 Discussion

Our pipeline effectively breaks down the structure of tumor gene expression data and demonstrates that a small, stable subset of genes can differentiate cancer types with extremely high accuracy. Because these tumors originate in distinct tissues, their transcriptomic profiles reflect underlying biological differences, making high classification performance plausible. CDX1 is primarily involved in intestinal development and function and is therefore strongly associated with colorectal tumors. SFTPB encodes a key lung surfactant protein, making it highly specific to lung tissue and related disease processes. TMEFF2 acts as a tumor suppressor and is particularly linked to prostate cancer biology. Finally, TRPS1 encodes a transcription factor involved in connective tissue regulation and has known roles in breast cancer proliferation. Together, these genes provide biologically meaningful signals that align with the tissue origins of the tumor classes, enabling robust classification performance even with very small feature sets. We further visualized gene-to-sample relationships using a hierarchically clustered heatmap (Figure 6). Rows correspond to the selected top-ranked

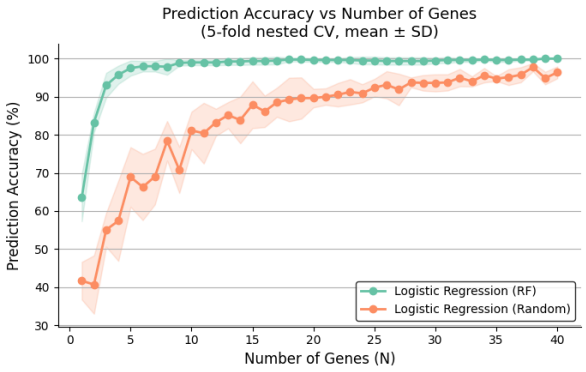

Figure 5: Performance comparison between Random Forest-based feature selection and baseline models trained on randomly selected gene subsets.

genes and columns to patient samples. The standardized expression gradient (red = up-regulation, blue = down-regulation) revealed distinct, tumor-specific expression signatures. Samples clustered cleanly into their cancer-type groups, and genes exhibited well-defined expression patterns that aligned with known biological roles.

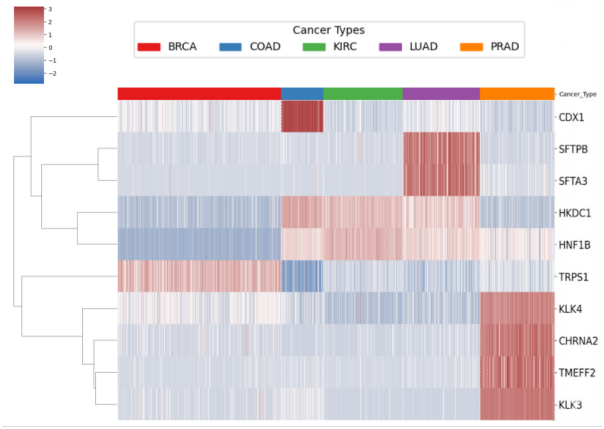

Figure 6: Gene-to-sample relationships across tumor classes.

Together, these results highlight the strength of combining Random Forest-based feature selection with nested cross-validation in high-dimensional transcriptomic analysis. The method yields both stable predictive performance and biologically interpretable gene subsets suitable for downstream investigation.

4 Related Work

Several studies have applied machine learning to the same TCGA Pan-Cancer RNA-Seq dataset, consistently reporting high accuracy but differing in feature selection and interpretability. He et al. [2] and Alanazi et al. [6] both reached near-perfect accuracy (99.8% and 99.995%, respectively), yet their models relied on thousands of genes (after light ANOVA filtering). Hybrid approaches such as Akter et al. [7] identified $\tilde{1}$00-gene signatures achieving $\tilde{9}$9.9% accuracy, but still did not explore truly minimal gene sets or provide detailed biological interpretation due to missing metadata.

The strongest prior result comes from Jiménez et al. [8, 9], who achieved 100% accuracy using multivariate feature-ranking methods that model gene–gene interactions. However, these methods scale quadratically and required 6–7 hours on this dataset ($\approx$ 400 minutes per method), making them impractical for rapid

cross-validation or repeated analyses. Moreover, while they statistically validated the robustness of their selected genes, the biological interpretation of the top markers was left as future work.

Together, existing literature highlights two persistent gaps: (1) achieving perfect or near-perfect accuracy with very small gene subsets, and (2) integrating fast, scalable feature selection with direct biological interpretation. Our pipeline addresses both. Using an efficient Random Forest–based ranking within nested cross-validation, we obtain 99–100% accuracy with as few as 3–10 genes, reaching 100% accuracy using only 39 genes, while running in seconds rather than hours. Importantly, we also perform gene-level biological analysis, showing that the top-ranked markers correspond to well-established tissue-specific cancer genes. This combination of minimal gene signatures, rapid execution, and biological interpretability distinguishes our approach from prior work.

5 Future Work

One topic of future work that could build on this study's progress would be to implement hyperparameter tuning within each classifier model. The models were left with their default parameters, but with some tuning, the number of genes needed for near-optimal performance could be reduced. The models might also show more differences between each other if hyperparameters were optimally tuned. Moreover, the whole pipeline could be repeated with varying random seeds, and the stability of prediction and accuracy could be compared. Genes that offer high stability of accuracy could be preferred depending on the application of the model. Lastly, future work could do in silico analysis of certain genes to analyze what specific mutations or variations are linked to cancer, and what proteins they code for.

6 Conclusion

This study demonstrates the efficacy of integrating interpretable machine learning with robust feature selection to differentiate primary tumor types using transcriptomic data. By employing a stratified nested cross-validation framework, we identified a minimal signature of genes that enables near-perfect classification accuracy across five major cancer types, significantly reducing the dimensionality from over 20,000 features to fewer than 50. Our results highlight that complex, computationally expensive models are not strictly necessary for this task; simple linear classifiers utilizing highly informative, tissue-specific markers are sufficient to achieve state-of-the-art performance.

Crucially, the selected features, such as *CDX1* and *SFTPB*, are not merely random features, but biologically relevant drivers of tissue identity. This alignment between statistical importance and biological function validates the potential for developing cost-effective, targeted gene expression panels that offer rapid and accurate diagnostic capabilities in clinical settings. By prioritizing model interpretability and feature stability, this work provides a scalable foundation for future computational oncology tools capable of supporting early and accurate cancer diagnosis.

References

[1] May 2025. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/statistics

[2] Y. He, R. Bockmon, and M. Modey, "Classification of cancer types based on gene expression data," in *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 2175–2182.

[3] S. Fiorini, "gene expression cancer RNA-Seq," UCI Machine Learning Repository, 2016, DOI: https://doi.org/10.24432/C5R88H.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer Science & Business Media, 2009.

[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016, p. 785–794. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[6] S. A. Alanazi, N. Alshammari, M. Alruwaili, K. Junaid, M. R. Abid, and F. Ahmad, "Integrative analysis of rna expression data unveils distinct cancer types through machine learning techniques," *Saudi Journal of Biological Sciences*, vol. 31, 2024, article ID: 103918; Available online Dec. 2023.

[7] S. Akter, R. O. Adesola, and S. Basnet, "Machine learning approach to identify significant genes and classify cancer types from rna-seq data," *Global Medical Genetics*, vol. 12, 2025, article ID: 100079.

[8] F. Jiménez, G. Sánchez, J. Palma, L. Miralles-Pechuán, and J. A. Botía, "Multivariate feature ranking with high-dimensional data for classification tasks," *IEEE Access*, 2022.

[9] F. Jiménez, G. Sánchez, J. Palma, L. Millares, and J. Botía, "Multivariate feature ranking of gene expression data," 2022, precursor manuscript to the IEEE Access version.