Classification and Identification of Biomarkers from 3 Cancer Diseases

Erin Xu

2026-03-20

Problem Statement

We address the challenge of balancing a high-dimensional feature space with relatively limited sample sizes in microarray-based gene expression data, while identifying biologically interpretable genomic markers, using traditional machine learning models. This problem is particularly important in the context of cancer diagnosis and clinical outcome prediction, where accurate and biologically relevant classification models are essential for reliable patient care.

Gene expression technologies, such as DNA microarrays and RNA sequencing, enable the quantification of thousands of gene expression levels simultaneously, resulting in a high-dimensional setting where the number of features far exceeds the number of samples ($p \gg n$). The dataset used in this study reflects this imbalance, with 12,042 features and 886 samples, and additionally exhibits class imbalance (Class 1: 348; Class 2: 130; Class 3: 408) (Kaggle, 2025). These characteristics increase the risk of overfitting, reduce generalizability, and may bias models toward majority classes.

While gene expression profiles capture biologically meaningful differences across cancer types, they also introduce potential confounding. Variation in cell-type composition across tissues may drive apparent separation in expression space that is not directly related to cancer-specific mechanisms (Guyon et al., 2002). This is especially relevant for this dataset, which includes glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC), and ovarian cancer (OV), originating from distinct tissues. In addition, microarrays are sensitive to experimental conditions such as hybridization temperature and sample quality (Eschrich & Yeatman, 2004). Because this dataset joins multiple studies, it may also be affected by batch effects, where technical variation rather than biological signal drives model performance (UCSC Xena, 2026).

Statistical Analysis

Let $X \in \mathbb{R}^{886\times 12042}$ denote the data matrix, where rows correspond to samples and columns to log-transformed gene expression measurements. Due to variation in feature magnitudes and potential batch effects, z-score normalization was applied. Standardization is also important for scale-invariant models such as SVM and L2-regularized logistic regression.

Prior work suggests that linear models remain highly effective for multiclass classification of microarray data, often outperforming more complex approaches (Alharbi & Vakanski, 2023). This was supported by exploratory data analysis using principal component analysis (PCA), kernel PCA (KPCA), t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).

PCA was used to obtain a low-dimensional representation by projecting onto orthogonal directions of maximum variance. After centering $X_c = XH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, SVD yields $X_c = U\Sigma V^\top$. The first three components explain 43.5%, 7.49%, and 4.37% of the variance (55.6% total). Notably, 372

components explain 95% of the variance (approximately 3.1% of the original feature space), indicating substantial redundancy despite limited interpretability.

Kernel PCA with a radial basis function (RBF) kernel was used to explore nonlinear global structure. The kernel parameter γ controls the locality of the similarity function; we used the heuristic $\gamma = 1/p$ as a baseline as suggested by the sci-kit learn documentation (Pedregosa et al., 2011). The first three kernel principal components accounted for 21.56%, 5.72%, and 2.5% of variance explained (29.78% cumulative), and 657 components were required to reach 95%, which is roughly twice of the results of linear PCA. This suggests that nonlinear structure is more diffuse and not concentrated in a small number of dominant directions, and linear structure captures the primary variance.

t-SNE and UMAP are commonly used for visualization of microarray data to identify potential nonlinear structure (Kobak & Berens, 2019). t-SNE preserves local neighborhood structure by modeling pairwise similarities as conditional probabilities (Maaten & Hinton, 2008). UMAP additionally preserves aspects of global topology (fuzzy simplicial complex), and had very clear boundary divides, but if not seriously hyperparameterized, produces downstream artifacts (McInnes et al., 2018). So, given the simplicity of linear PCA, along with its strong cumulative variance explained across the first three components and comparable levels of class mixing, we find that nonlinear methods such as t-SNE and UMAP do not provide meaningfully improved cluster separation (Figure 1). This suggests nonlinear structure is not the primary driver of separation. As such, we choose not to sacrifice interpretability for more complex representations that may introduce overfitting without clear performance gains due to the importance of biological context. While are various methods to interpret UMAP, such as centered Log-Ratio (CLR) followed by Euclidean distance transformed data or cosine-space embeddings to produce compositionally valid distance metrics, specialization indicates that should be better explored in future work (Quinn et al., 2018). Finally, a sample-wise mean expression plot by class (Figure 2) shows mostly unimodal and globally distinct expression levels across cancer types.

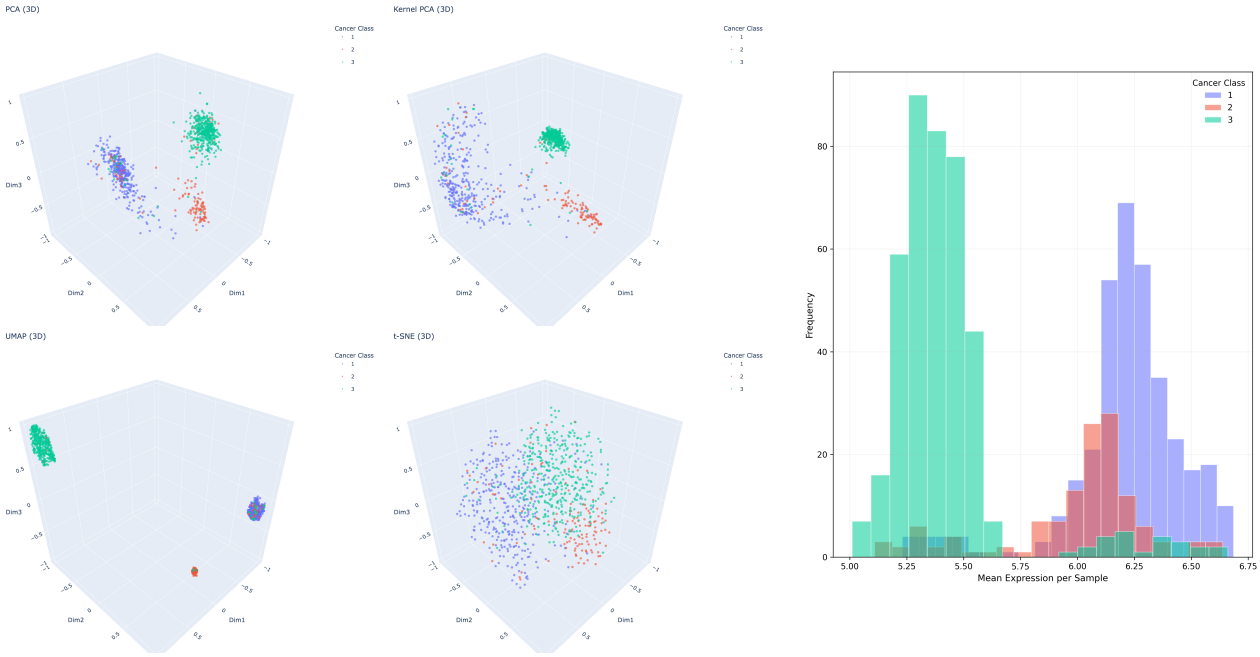

Figure 1 (left): Visualizations for PCA, Kernel PCA, UMAP, and t-SNE (left to right, top to bottom) **Figure 2 (right):** Distribution of sample-wise mean expression by cancer type. *Note: high-res code visualizations in Appendix B.*

As a result, we focus on comparing a set of traditional machine learning models, including support vector machines (SVMs), logistic regression, k-nearest Neighbors (kNNs) and random forests (RFs) when no gene selection is performed and when several popular gene selection methods are used.

Linear SVM and L2-regularized logistic regression were used as baseline linear models, offering interpretability and strong performance in high-dimensional settings. RFs were included as a nonlinear ensemble method to capture complex feature interactions, while KNN and RBF SVM modeled nonlinear decision boundaries through local similarity and kernel transformations. Together, these models span linear and nonlinear approaches for comprehensive comparison.

Gene selection reduces dimensionality and computational cost while identifying informative features. Two approaches were used: filter and wrapper methods. Filter methods rank genes independently using statistical heuristics, while wrapper methods evaluate subsets using predictive models (Xing et al., 2001).

Filter methods included signal-to-noise ratio (S2N) and Kruskal-Wallis (KW). S2N ranks genes by mean differences normalized by within-class variability, while KW is a non-parametric test robust to non-Gaussianity and suitable for multiclass settings (Golub et al., 1999). However, both operate univariately and do not capture gene interactions. They are widely used and effective (Statnikov et al., 2008). Wrapper methods included SVM-based recursive feature elimination (SVM-RFE) and a RF-based backward elimination procedure (RFVS); both are greedy backward selection methods that iteratively remove features based on model-derived importance and do not guarantee globally optimal subsets. SVM-RFE ranks features using coefficients from a Linear SVM, enabling multivariate selection of compact and discriminative gene sets (Guyon et al., 2002). RFVS leverages ensemble learning to capture nonlinear relationships while remaining robust to noise (Díaz-Uriarte & Alvarez de Andrés, 2006).

All feature selection was performed within training folds to prevent information leakage. Model evaluation used nested 5-fold stratified cross-validation, where the outer loop estimates generalization and the inner loop performs hyperparameter tuning. This ensures unbiased performance estimates in high-dimensional settings.

To reduce computational cost, wrapper methods were first applied after prefiltering to the top 500 genes using S2N within each training fold. SVM-RFE selected 50 or 100 genes by iteratively removing 50% of features, while RFVS removed the bottom 20% based on importance at each step, retaining the subset with the best out-of-bag performance (see Appendix B).

To comprehensively compare the performances of the machine learning techniques, Macro F1-score, Balanced Accuracy, and confusion matrices were used, which each account for class imbalance. Macro F1 is the harmonic mean of precision and recall averaged across all classes and assigns equal weight to each class, while Balanced Accuracy measures average recall across classes. Accuracy alone would be biased towards larger classes. Let TP_k, FP_k, FN_k, and TN_k denote the true positives, false positives, false negatives, and true negatives for class $k \in \{1, 2, 3\}$.

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \quad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k}, \quad F1_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$$

$$\text{Macro F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k, \quad \text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}$$

Results and Conclusion

Many configurations across baseline, filter, and wrapper approaches achieved identical performance (Macro F1 of 0.866), indicating that model performance is largely insensitive to the choice of algorithm or feature selection strategy. There is a 10-way tie (Table 1) for first place amongst highest Macro F1 and Balanced Accuracy, with filter feature selection and linear boundary models dominating. Out of all the wrapper methods tested, only RFVS is part of the first place tie. SVM-RFE methods are some of the lowest ranked. This suggests that the classification task is relatively well-separated in the feature space as hypothesized. This indicates that both linear and nonlinear models are able to capture the underlying structure effectively, and that additional model complexity does not yield meaningful

improvements. Linear models (logistic regression, Linear SVM) performed comparably to nonlinear models (RBF SVM, RF), indicating that nonlinear decision boundaries are not necessary to achieve strong performance. As multiple models achieved identical performance up to three decimal places, differences among top-ranked models should be interpreted as negligible, but ultimately, when it comes to model simplicity and computational efficiency, linear models are preferable to more complex feature selection and modeling strategies.

The difference between the best filter method and best wrapper method outside of first place has a margin of 0.016 Macro F1 units (1-2 misclassified samples). The additional complexity of wrapper methods may not be justified for this dataset, and the original high-dimensional feature space already contains sufficient discriminative information. The slightly lower performance of SVM-RFE may be due to the aggressive feature elimination process discarding informative features early, which would be more helpful on a more complicated dataset.

The presence of multiple models achieving identical performance suggests that the classification task is relatively well-separated (in which the initial PCA EDA strongly supports), and that different modeling approaches converge to similar decision boundaries. The confusion matrices are homogenous across the tied models, so the top 4 ranked models, RF as a baseline, KW Linear SVM, S2N RBF SVM and KW KNN confusion matrices are shown (Figure 3). The lower rates seem to be due to misclassification in class 2, as evidenced by its substantially lower class-wise accuracy and higher confusion with both class 1 and class 3 across the 3 lower performing models. It is important to note that class 2 is the smallest class (130 samples). Although models incorporated `class_weight='balanced'` in their hyperparameter tuning grid to partially compensate for class imbalance, misclassification of class 2 remained substantial. This suggests that some samples assigned to class 2 may share molecular characteristics with other classes, or may be mislabeled due to dataset integration artifacts. Upon further inspection, there are only 3 points in which these models disagree on (Table 2). Most notably sample 442: true class 3, RF, RBF SVM, and KNN predict 1, and Linear SVM predicts 2, so no model predicted correctly here. Upon visual inspection (Figure 4), we deduct that this sample may be systematically mislabeled, but it also lies on a very ambiguous decision boundary with GBM, which are microscopically nuanced across different models. On the other hand, sample 5 is well within the green cluster.

To identify potential biomarkers, we examined the frequency of gene selection across the top-performing models. We focused on genes that appeared consistently across multiple models and considered more robust candidate biomarkers. Table 2 groups genes by how consistently they were selected across the top 10 models used in nested cross-validation. A gene appearing in the 90% row was chosen by 9 out of 10 models, and 70% means 7 out of 10.

Diseases are driven by pathways, not single genes, so sets of genes can act as proxies for the same process. A lot of the genes in the 90% do not immediately fit GBM, LUSC or OV biology but reflect tumor microenvironment, inflammation, metabolism or broad cancer biology, rather than tumor-specific identity. For example, DRD3, ADRA1A, MC2R, GNRHR, OPRM1, TAS2R13 from the 90% selected are canonical G-protein-coupled receptors (GPCRs), a large family of transmembrane receptors that regulate many cell functions, including cell proliferation, survival and motility, are key players in tumour growth, angiogenesis and metastasis (Dorsam & Gutkind, 2007). KSR1, PPARA, NOX1 are signaling related pathways with GPCRs or functional oncogenic contributors.

We confirm with prior studies using microarray data that OLIG1/2, GFAP, and PTPRZ1 are key regulatory factors in GBM (Tian et al., 2025). We also propose PDGFB as a predictive feature. Since platelet-derived growth factors (PDGFs) signalling are a family of mitogens that become co-expressed in GBM, the subunit B gene may also be related. As further evidence, mutations in PDGFB are already associated with meningioma (Black et al., 1994). Of course, further statistical testing on PDGFs should be completed with biologists. For OV, we confirm that WT1, CLDN3, and SOX17 are markers (Goto et al., 2024). Because the LUSC signal was weak from the confusion matrices due to class imbalance, established markers like TP63, SOX2, and KRT5/5-6-related squamous programs are noticeably absent. Only CLDN3 and PDGFB have context in lung cancer. Fortunately, there is new

evidence that AVPR1A is expressed in small lung cancer cell lines, so we can support the hypothesis and propose this as a predictive gene (Zhao et al., 2019).

The appearance of genes like MLANA is suspicious, since MLANA is primarily known as a potential prognostic melanocytic/melanoma marker, not a canonical GBM/LUSC/OV discriminator (Tang et al., 2026). Because there remain context biomarkers or noise or dataset artifacts, these are signs of batch effects and tissue contamination and reflect feature selection instability. This highlights the need for additional validation to distinguish robust biomarkers from features driven by noise or sampling variability. Again, although several selected genes are biologically plausible, stability analysis and permutation testing of significance of differences are necessary to determine whether these genes are consistently selected beyond random variation, so, these findings are limited to this dataset and may not generalize to external cohorts or RNA-seq data without further validation.

Rank	Branch	Model	Macro F1	Balanced Accuracy
1	Filter	S2N_k100_KNN	0.866 ± 0.026	0.844 ± 0.029
1	Filter	S2N_k100_LinearSVM	0.866 ± 0.026	0.844 ± 0.029
1	Filter	S2N_k100_RBF_SVM	0.866 ± 0.026	0.844 ± 0.029
1	Filter	S2N_k50_LinearSVM	0.866 ± 0.026	0.844 ± 0.029
1	Filter	S2N_k50_LogisticEN	0.866 ± 0.026	0.844 ± 0.029
1	Wrapper	RFVS_RF_k100	0.866 ± 0.026	0.844 ± 0.029
1	Wrapper	RFVS_RF_k50	0.866 ± 0.026	0.844 ± 0.029
1	Filter	KW_k100_LogisticEN	0.866 ± 0.026	0.844 ± 0.029
1	Filter	KW_k100_RBF_SVM	0.866 ± 0.026	0.844 ± 0.029
1	Baseline	RandomForest	0.866 ± 0.026	0.844 ± 0.029
2	Filter	KW_k100_LinearSVM	0.865 ± 0.026	0.844 ± 0.029
3	Filter	S2N_k50_RBF_SVM	0.864 ± 0.026	0.843 ± 0.028
4	Filter	KW_k100_KNN	0.864 ± 0.026	0.843 ± 0.029
5	Filter	S2N_k50_KNN	0.864 ± 0.024	0.843 ± 0.028
5	Filter	KW_k50_LogisticEN	0.862 ± 0.024	0.842 ± 0.028
6	Filter	KW_k50_KNN	0.862 ± 0.025	0.842 ± 0.027
7	Filter	S2N_k100_LogisticEN	0.861 ± 0.035	0.841 ± 0.034
8	Filter	KW_k50_RBF_SVM	0.861 ± 0.027	0.840 ± 0.030
9	Baseline	LogisticL2	0.857 ± 0.031	0.838 ± 0.031
10	Filter	KW_k50_LinearSVM	0.851 ± 0.032	0.832 ± 0.033
11	Wrapper	SVM_RFE_k50	0.850 ± 0.025	0.832 ± 0.025
12	Baseline	LinearSVM	0.843 ± 0.029	0.830 ± 0.031
13	Wrapper	SVM_RFE_k100	0.838 ± 0.029	0.826 ± 0.029

Table 1: Nested Cross-Validated Performance of Classification Models Across Baseline, Filter-Based, and Wrapper-Based Feature Selection Approaches. *Abbreviations:* S2N: Signal-to-Noise ratio filter; KW: Kruskal–Wallis filter; k50/k100: number of selected genes; LinearSVM: Linear SVM; RBF_SVM: RBF kernel SVM; LogisticL2: L2-regularized logistic regression; LogisticEN: elastic net logistic regression; KNN: k-nearest neighbors; RandomForest: random forest; SVM_RFE: SVM recursive feature elimination; RFVS: random forest variable selection.

Figure 3: Confusion matrices for the top 4 models.

Sample ID	Test Index	True Class	RandomForest	KW_k100 Linear SVM	S2N_k50 RBF SVM	KW_k100 KNN
S5708912	5	3	3	3	3	2
S7143107	386	1	1	1	2	1
S2917620	442	3	1	2	1	1

Table 2: Disagreement points for the top 4 models

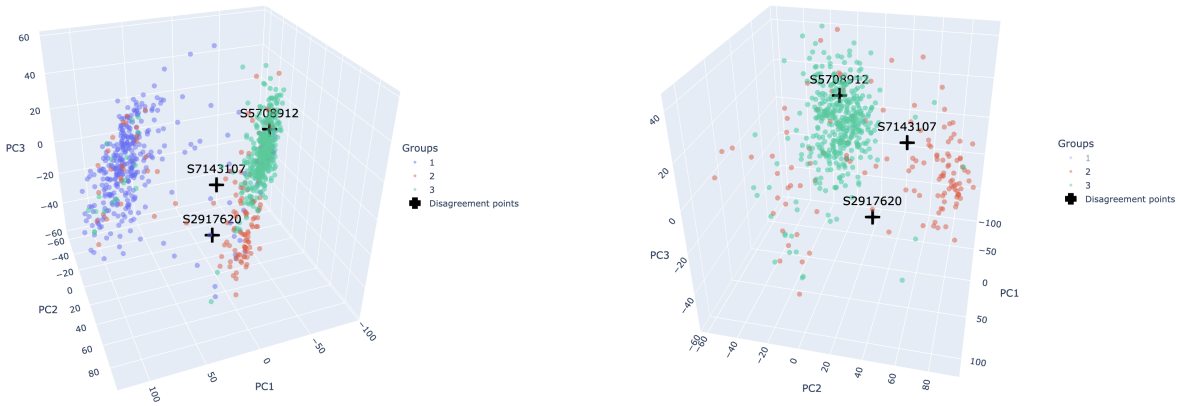

Figure 4: 3D PCA plot of labeled disagreement points (left) and with class 1 masked (right).

Selected (%)	Genes
90% (9/10)	DRD3, SORBS1, ADRA1A, APAF1, SLC12A4, TCL6, SIGLEC8, MAGEB4, FETUB, CYLC2, LZTS1, COL4A3, SLC25A31, FCAR, BMP8A, GNRHR, PPARA, NOX1, KCNJ14, BMP10, GTDC1, IGSF9B, HTR7P, TAS2R13, SEMG1, ZNF674, SLC30A4, MC2R, OPRM1, MUSK, C10orf12, CD28, ADAM22, ADAM7, CCL1, HRH4, MLANA, USH2A, FAM12A, IFNA2, SLC9A7, KSR1, PADI4, NLRP3, PTPRJ, CHRD, GRIN2B, CRHR2
70% (7/10)	CHRNB3, GYPA, SEPT7, ALDOB, PKLR, NOX3, ZNF507, TRGV7, SIGLEC6, GABRQ, CD84, HS3ST3B1, OSBPL7, OR2J2, POU3F1, CNR1, IL25, MUPCDH, IFNA6, ESR2, CHRNA4, ZSCAN12, PTGDR, MPL, LHX5, MASP2, FCGR2C, BPESC1, CACNA1I, KIAA0913, CACNA1C, LIMK1, KLRA1, AFM, P2RY4, VENTXP1, CDX4, GFAP, TSPAN5, TRPM6, TRIM10, HTR5A, PDE10A, DST, NLRP1, PAX3, OLIG2, SCRT1, INSL5, LDHAL6B, SPAM1, TTC22, WT1, ASB4, MTSS1, RXFP3, ELL2, PPP3CC, TGM4, CACNA1G, PTPRZ1, TRA., SOX17, PVRL1, CLDN3, CRP, ZC3H7B, AVPR1A, PDGFB, IL1F6, GLP1R, PCDHB17, CCDC103, MFAP3L
50%+	*See Appendix B for code.*

Table 3: Candidate Biomarker Genes Identified by Frequency of Selection Across Top-Performing Models. *Models:* All genes listed in this table were selected by the same 9 of 10 top-performing models: `KW_k100_LogisticEN`, `KW_k100_RBF_SVM`, `S2N_k100_KNN`, `S2N_k100_LinearSVM`, `S2N_k100_RBF_SVM`, `S2N_k50_LinearSVM`, `S2N_k50_LogisticEN`, `RFVS_RF_k100`, and `RFVS_RF_k50`.

Discussion

We concluded that classification and prediction of the GBM, LUCS, and OV cancers prefer simple dividing lines from linear models by conducting a retrospective observational design comparing various supervised machine learning and gene selection methods. As linear models are mostly interpretable, this resulted in confirmation of certain blocks of well known surrogate biomarkers for each cancer and also more recently proposed biomarkers (PDGFB for GBM and AVPR1A for LUSC). Work remains to be done in collaboration with biologists to compare the candidate biomarkers, as well as the top performing models, with non-parametric permutation tests to assess whether the observed difference in gene selection and classification performance is statistically significant or simply due to chance.

A limitation of our work from a statistical perspective is that the use of linear models and univariate feature selection methods may further emphasize independent gene effects, potentially overlooking nonlinear interactions between genes that contribute to cancer development. In reality, putative genomic markers have rarely been implemented into routine clinical use because there is a low level of trust of the single disease, single genomic marker approach, as cancer is characterized by tumor and molecular heterogeneity in the presence of mutations, nucleotide polymorphisms, epistasis and pathway-level effects on pharmacological response (Coyle & Johnston, 2010).

So, based on the inherent complexity of biological systems, it is still hard not to believe the true predictive biological signal is not nonlinear, or that complex nonlinear models such as neural networks are not better suited to complete these tasks. Understanding the conditions under which linear versus nonlinear models are more appropriate remains an important open question. Heil et. all (2023) found that when removing all linear signals in omics data, the residual nonlinear signal remained, so the similarity in performance across model types was not due to the problem domains possessing solely linear signals. Consequently, extensions of this paper may involve investigating whether the apparent success of linear classifiers reflects true approximate linear separability of cancer types in gene expression space, or whether it is mainly driven by dataset-specific structure and a small number of highly discriminative genes, and in which biological scenarios would yield better performance. If future research could expand beyond transcriptomic data to incorporate multi-omics information to further capture the complex biological mechanisms underlying cancer, such as tumor imaging and epigenetic data, perhaps nonlinear performance would dominate.

References

Alharbi, F., & Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, *10*(2), 173. https://doi.org/10.3390/bioengineering10020173

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2024). *Quarto* (Version 1.4) [Computer software]. https://quarto.org

Black, P. M., Carroll, R., Glowacka, D., Riley, K., & Dashner, K. (1994). Platelet-derived growth factor expression and stimulation in human meningiomas. *Journal of Neurosurgery*, *81*(3), 388–393. https://doi.org/10.3171/jns.1994.81.3.0388

Coyle, V., & Johnston, P. G. (2010). Genomic markers for decision making: What is preventing us from using markers? *Nature Reviews Clinical Oncology*, *7*(2), 90–97. https://doi.org/10.1038/nrclinonc.2009.214

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*, 3. https://doi.org/10.1186/1471-2105-7-3

Dorsam, R. T., & Gutkind, J. S. (2007). G-protein-coupled receptors and cancer. *Nature Reviews Cancer*, *7*(2), 79–94. https://doi.org/10.1038/nrc2069

Eschrich, S., & Yeatman, T. (2004). DNA microarrays and data analysis: An overview. *Surgery*, *136*(3), 500–503. https://doi.org/10.1016/j.surg.2004.05.021

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

Goto, N., Westcott, P. M. K., Goto, S., & al., et. (2024). SOX17 enables immune evasion of early colorectal adenomas and cancers. *Nature*, *627*, 636–645. https://doi.org/10.1038/s41586-024-07135-3

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1-3), 389–422. https://doi.org/10.1023/A:1012487302797

Heil, B. J., Crawford, J., & Greene, C. S. (2023). The effect of non-linear signal in classification problems using gene expression. *PLoS Computational Biology*, *19*(3), e1010984. https://doi.org/10.1371/journal.pcbi.1010984

Kaggle. (2025). *Classification of cancer types*. Kaggle Competition. https://www.kaggle.com/competitions/classification-of-cancer-types/data

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1), 5416. https://doi.org/10.1038/s41467-019-13056-x

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605. https://jmlr.org/papers/v9/vandermaaten08a.html

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv Preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426

Motiwala, T., & Jacob, S. T. (2006). Role of protein tyrosine phosphatases in cancer. *Progress in Nucleic Acid Research and Molecular Biology*, *81*, 297–329. https://doi.org/10.1016/S0079-6603(06)81008-1

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., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Quinn, T. P., Erb, I., Richardson, M. F., & Crowley, T. M. (2018). Understanding sequencing data as compositions: An outlook and review. *Bioinformatics*, *34*(16), 2870–2878. https://doi.org/10.1093/bioinformatics/bty175

scikit-learn developers. (2026). *Sklearn.decomposition.KernelPCA*. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.KernelPCA.html.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive

evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, *21*(5), 631–643. https://doi.org/10.1093/bioinformatics/bti033

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, *9*, 319. https://doi.org/10.1186/1471-2105-9-319

Tang, N., Lai, X., Wen, W., Wu, Y., & Xiong, X. (2026). MLANA is a potential prognostic biomarker and correlated with immune infiltration in skin cutaneous melanoma. *Clinical and Experimental Medicine*, *26*(1), 156. https://doi.org/10.1007/s10238-026-02078-7

Tian, Y., Wang, Z., Sun, M., Li, J., Zheng, W., Yang, F., & Zhang, Z. (2025). Olig1/2 drive astrocytic glioblastoma proliferation through transcriptional co-regulation of various cyclins. *Genes*, *16*(5), 573. https://doi.org/10.3390/genes16050573

UCSC Xena. (2026). *TCGA GBM gene expression dataset (HT HG-U133A)*. https://xenabrowser.net/datapages/?dataset=TCGA.GBM.sampleMap/HT_HG-U133A.

Xing, E. P., Jordan, M. I., & Karp, R. M. (2001). Feature selection for high-dimensional genomic microarray data. *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 601–608. https://doi.org/https://www.researchgate.net/publication/2371787_Feature_Selection_for_High-Dimensional_Genomic_Microarray_Data

Yen, H.-Y., Jazayeri, A., & Robinson, C. V. (2023). G protein-coupled receptor pharmacology—insights from mass spectrometry. *Pharmacological Reviews*, *75*(3), 397–415. https://doi.org/https://doi.org/10.1124/pharmrev.120.000237

Zhao, N., Peacock, S. O., Lo, C. H., Heidman, L. M., Rice, M. A., Fahrenholtz, C. D., Greene, A. M., Magani, F., Copello, V. A., Martinez, M. J., Zhang, Y., Daaka, Y., Lynch, C. C., & Burnstein, K. L. (2019). Arginine vasopressin receptor 1a is a therapeutic target for castration-resistant prostate cancer. *Science Translational Medicine*, *11*(498), eaaw4636. https://doi.org/10.1126/scitranslmed.aaw4636

Appendix A

ChatGPT and Claude were used to generate code comments for readability and to format entries in the `.bib` file. They were not used to generate results or make methodological decisions. All analyses, model design, and interpretations were independently developed and validated by the authors. Finally, this report was typeset using Quarto and the LaTeX engine is lualatex.

Appendix B

Model training/computation was implemented with the `scikit-learn` library in Python. All trained models are saved under a model registry with `.joblib` and `.json` files. Code is available upon request only, as the final Github repository remains to be organized and the raw files are impossible to attach here.