Identification of Origin for Cancer Samples With Machine Learning Models

Jackson McDonald and **Christian Rutherford**
{jamcdonald2,chrutherford}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

Accurate identification of the tissue origin for cancer is crucial in informing treatment strategies for cancer patients. This study investigates the uses of various models including random forests, gradient boosting classifiers, support vector machines, and logistic regression models to compare the effectiveness of these models for identifying the tissue of origin. The effectiveness of the models was quantified using a f1 score to determine which model was the most effective for classification in this context.

1 Introduction

Accurate determination of the tissue of origin of cancer remains a central challenge in caner care, which impacts the selection of treatment strategies to improve patient outcomes. 3-5% of caner patients are identified as having cancer of unknown origin. Follow up diagnosis of the origin presents various difficulties with computed tomography (CT) and positron emission tomography (PET) providing 20–27% and 24–40% respectively (He et al. 2020).

Advances in computational methods particularly machine learning have expanded the tools available for diagnostic decision-making, enabling the analysis of complex, high-dimensional data beyond the capacity of traditional statistical approaches.

The use of machine learning in medical diagnosis has a long history, dating back to foundational work on decision trees and rule-based systems in the late twentieth century, where Quinlan's development of the Iterative Dichotomizer 3 (ID3) algorithm and the early diagnostic applications of Michalski and Chilausky's AQ system marked a turning point in automated decision rule generation. These methods were soon applied to challenging diagnostic tasks, including oncology, where Bratko and Mulec demonstrated the potential of ID3 for complex medical classification problems (Kononenko 2001). Still today, machine learning models continue to provide massive breakthroughs in medical classification issues with models achieving far better identification than classical lab results providing f1 scores between .882-.888 for identification across 13 types of cancer.

2 Background

The dataset used in this study consists of microRNA (miRNA) expression profiles sourced from The Cancer Genome Atlas (TCGA). There are 6 different types of cancers contained in the profiles provided for analysis: Breast Invasive Carcinoma, Kidney Renal Clear Cell Carcinoma, Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, Pancreatic Adenocarcinoma, and Uveal Melanoma.

For our feature extraction, we found the miRNA sequence frequency across the patient and mapped the frequencies to their respective miRNA sequences. Next, we identified sequences with the 20 most variable frequencies across the patients and chose to use them for our features. This allowed us to reduce our feature size significantly to reduce concerns of overfitting and run time when developing the models.

3 Experiments

Our experimental setup was designed to evaluate multiple machine learning algorithms for multi-class cancer classification using miRNA expression profiles from The Cancer Genome Atlas.

Research Questions

We aimed to answer the following questions:

1. Can miRNA expression profiles accurately distinguish between six different cancer types?
2. Which machine learning algorithms perform best for this classification task?
3. How does hyperparameter tuning impact model performance?
4. How does bootstrapping affect the issue of class imbalance?

Experimental Setup

Dataset: We consolidated data from 2,895 patients across six cancer types: Breast Invasive Carcinoma (1,096 patients), Kidney Renal Clear Cell Carcinoma (544), Lung Adenocarcinoma (519), Lung Squamous Cell Carcinoma (478), Pancreatic Adenocarcinoma (178), and Uveal Melanoma (80). Each patient sample contained 1,881

miRNA expression features. **Data Preprocessing:** We applied bootstrap resampling to address class imbalance by oversampling minority classes to match the size of the largest class (1,096 samples per class). The final balanced dataset contained 6,576 samples. We performed stratified train-test split (80/20) and applied StandardScaler normalization to all features. **Models Evaluated:** We trained 4 models, each with two variants:

- Random Forest (baseline and hyperparameter-tuned)
- Gradient Boosting (baseline and hyperparameter-tuned)
- Support Vector Machine with RBF kernel (baseline and tuned)
- Logistic Regression (baseline and tuned)

Hyperparameter Tuning: 5-fold cross-validation with GridSearchCV was utilised to optimise hyperparameters. For Random Forest, we tuned n_estimators, max_depth, min_samples_split, and min_samples_leaf. For Gradient Boosting, we optimised n_estimators, learning_rate, max_depth, and min_samples_split. For SVM, we tuned C, gamma, and kernel type. For Logistic Regression, we optimised C, penalty, and solver.

Evaluation Metrics: All models were evaluated by accuracy score, F1-score (weighted), and 5-fold cross-validation accuracy on the test set.

4 Results

Our experiments demonstrate that miRNA expression profiles provide highly discriminative features for cancer type classification, with all tested models achieving exceptional performance.

Overall Model Performance

Table 1 presents the test accuracy for our eight model variants. The results are notable: all models achieved accuracy above 98.4%, with an average accuracy of 99.33% across the variants. This performance indicates that miRNA expression patterns contain strong cancer-type-specific signals.

Model	Accuracy
Random Forest (tuned)	**99.7%**
Random Forest	99.54%
Gradient Boosting	99.54%
Gradient Boosting (tuned)	99.54%
SVM (tuned)	99.32%
Logistic Regression (tuned)	99.32%
Logistic Regression	99.24%
SVM (RBF)	98.4%

Table 1: Test accuracy for all model variants. Random Forest (Tuned) achieved the best performance at 99.7% accuracy.

Key Findings

Ensemble Methods Excel: Random Forest and Gradient Boosting models consistently outperformed other approaches. The tuned Random Forest achieved the highest accuracy (99.7%), correctly classifying 1,312 out of 1,316 test samples. **Minimal Impact of Hyperparameter Tuning:** Interestingly, hyperparameter tuning provided only marginal improvements (0.16% for Random Forest, 0% for Gradient Boosting, 0.92% for SVM, 0.08% for Logistic Regression). This suggests that default parameters are already well-suited for this problem, and that the discriminative power comes primarily from the miRNA features themselves. **Bootstrap Resampling Effectiveness:** Addressing class imbalance through bootstrap resampling was highly effective. Despite the original dataset having a 13.7:1 imbalance ratio (Breast: 1,096 vs Uveal Melanoma: 80), our balanced approach achieved near-perfect classification across all cancer types. **Feature Scaling Importance:** StandardScaler normalisation was essential for SVM and Logistic Regression performance, while tree-based methods (Random Forest, Gradient Boosting) performed well without it.

Clinical Implications

The 99.7% accuracy achieved with our best model has significant clinical implications. With only 4 misclassifications out of 1,316 test cases, this approach might serve as a reliable diagnostic tool to complement traditional histopathology. The high performance across all six cancer types suggests that miRNA profiling could aid in:

- Differential diagnosis when tissue origin is ambiguous
- Early cancer detection screening programs
- Validation of initial pathology assessments
- Research into cancer biology and miRNA biomarkers

5 Broader Impacts

Our cancer classification models have potential benefits but also raise important ethical considerations that must be addressed before clinical deployment. On the up side, achieving 99.7% accuracy in distinguishing between six cancer types could revolutionise diagnostic pathways, possibly reducing misdiagnosis rates, accelerating treatment decisions, and improving patient outcomes. For patients in underserved healthcare settings where expert pathologists might be scarce, such automated tools could equalise access to high-quality diagnostic assessments. Furthermore, the system can serve as a valuable second opinion tool, used to reduce diagnostic uncertainty and catch errors that may otherwise go undetected. The identification of discriminative miRNA biomarkers can also accelerate cancer research and lead to new therapeutic targets.

However, there are several concerns warranting a few important considerations. Firstly, dataset bias is a critical issue: The Cancer Genome Atlas primarily contains data from patients in the USA, which has potential to underrepresent genetic diversity from other populations. Cancer manifestation and miRNA expression patterns might vary across varying ethnic groups, and a model trained predominantly on one population could result in reduced accuracy or systematic biases if applied to others, potentially worsening existing healthcare disparities. Second, the risk of over-reliance on algorithmic decisions is already very real; clinicians might defer too heavily to high-accuracy AI predictions, meaning

they may overlooking atypical cases or rare presentations that fall outside the training distribution, because "the AI said otherwise." Third, the class imbalance in our original dataset (with Breast Cancer having 13.7x more samples than Uveal Melanoma) raises questions about whether bootstrap resampling truly eliminates systematic biases, or simply obfuscates them. More uncommon cancer types might still be more prone to misclassification in real-world deployments. Lastly, there are concerns of medical liability and accountability: when AI systems make diagnostic errors, assigning responsibility to the algorithm developers, healthcare institution, and/or clinician becomes convoluted. Before deployment, this system would require **extensive** prospective validation across myriad **diverse** patient populations, **clear guidelines** on appropriate use cases (i.e. diagnostic aid versus sole decision-maker), **transparency** about model limitations, and robust governance frameworks to ensure equitable access and accountability.

6 Conclusions

This study investigated the usage of machine learning algorithms for classifying 6 cancer types utilising miRNA expression profiles from The Cancer Genome Atlas. We evaluated eight model variants across four algorithm families: Random Forest, Gradient Boosting, Support Vector Machines, and Logistic Regression.

Key Findings

Our results demonstrate that miRNA expression profiles are highly discriminative for cancer type classification. The tuned Random Forest model achieved 99.7% accuracy, correctly classifying 1,312 of 1,316 test samples. Notably, all our models exceeded 98.4% accuracy, with an average of 99.33% across all variants. These results indicate that miRNA expression patterns contain strong, cancer-type-specific signatures, bootstrap resampling effectively addresses class imbalance (13.7:1 ratio), ensemble methods (Random Forest, Gradient Boosting) slightly outperform other approaches, and default hyperparameters are already well-suited for this task

Future Directions

Several avenues for future research could extend this work:

- **Feature Selection:** Investigate which specific miRNAs are most discriminative and validate their biological significance
- **Deep Learning:** Explore neural network architectures that might capture complex miRNA interaction patterns
- **Multi-modal Integration:** Combine miRNA data with other genomic markers (mRNA, DNA methylation) for even higher accuracy
- **Clinical Validation:** Test the model on prospective patient cohorts in real clinical settings
- **Explainability:** Develop interpretable models to help clinicians understand classification decisions
- **Expanded Cancer Types:** Extend classification to additional cancer types beyond these six

In conclusion, our work demonstrates that machine learning applied to miRNA expression data achieves near-perfect cancer type classification, with significant potential for clinical diagnostic applications.

7 Contributions

Christian was responsible for the data extraction and the initial version of the models while Jackson implemented the bootstrapping of data, worked on the iterations of the model leading to the final experiment. Both authors worked in conjunction on the paper.

8 Acknowledgements

The authors thank ChatGPT (OpenAI) for providing writing assistance and editorial suggestions during the preparation of this paper.

References

He, B.; Dai, C.; Lang, J.; Bing, P.; Tian, G.; Wang, B.; and Yang, J. 2020. A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on dna somatic mutation. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1866(11):165916.

Kononenko, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine* 23(1):89–109.