A Machine Learning Approach for Early Pancreatic Cancer Risk Stratification Using Circulating MicroRNA Profiles

Krithik Alluri

Introduction

Pancreatic cancer (PC) remains one of the most lethal forms of cancer globally, largely due to its typically late diagnosis and rapid progression. Pancreatic Ductal Adenocarcinoma (PDAC), the most common subtype, is responsible for the vast majority of PC cases and is associated with an alarmingly low five-year survival rate of around 10 to 11 percent. The asymptomatic nature of early-stage PC and the absence of effective screening tools lead to diagnosis at advanced stages, rendering curative treatment impossible for many patients. This highlights the critical need for early detection strategies or risk stratification tools that can identify high-risk individuals before clinical manifestation. In recent years, circulating microRNAs (miRNAs) have emerged as promising non-invasive biomarkers for a variety of cancers. These small non-coding RNAs are involved in post-transcriptional gene regulation and are remarkably stable in blood, making them ideal candidates for liquid biopsy-based diagnostics.

In particular, there is a major gap in pre-diagnostic screening for PC. Current blood-based biomarkers, such as CA 19-9, have limited utility for early detection in asymptomatic individuals. Circulating miRNAs, which can reflect tumor-specific changes even before macroscopic lesions are visible, offer a promising alternative.

In this study, I explored the use of circulating miRNA expression profiles in predicting pancreatic cancer risk using machine learning approaches. I utilized a publicly available dataset, GSE262260, which comprises pre-diagnostic plasma samples from individuals who either developed pancreatic cancer within five years or remained cancer-free. This nested case-control dataset includes 462 incident PC cases and 462 matched controls. By applying a series of preprocessing steps, including imputation, variance filtering, and univariate feature selection, I identified 100 relevant miRNAs to be used in model training. Three machine learning algorithms, Logistic Regression, Random Forest, and Gradient Boosting, were rigorously tuned and evaluated using stratified train-test splits and 5-fold cross-validation. The performance of these models was primarily assessed using the Area Under the Receiver Operating Characteristic Curve (ROC AUC), with secondary metrics including accuracy, precision, recall, and confusion matrices.

Literature Review

The exploration of circulating microRNAs (miRNAs) as biomarkers for cancer detection has garnered significant attention in recent years. miRNAs are small, non-coding RNA molecules that play crucial roles in regulating gene expression post-transcriptionally. Their stability in blood and involvement in various cellular processes make them attractive candidates for non-invasive cancer diagnostics.

Several studies have highlighted the potential of specific miRNAs in the context of pancreatic cancer. For instance, Schultz et al. (2014) demonstrated that certain miRNA signatures in whole blood could effectively differentiate pancreatic cancer patients from healthy controls. Similarly, Mitchell et al. (2008) emphasized the stability of circulating miRNAs in blood and their potential as reliable biomarkers for cancer detection.

Further investigations have identified specific miRNAs associated with pancreatic cancer progression. miR-147a, for example, has been reported to act as a tumor suppressor in non-small cell lung cancer by inhibiting metastasis. Conversely, miR-10a-5p has been implicated in promoting pancreatic cancer growth by regulating pathways such as BDNF/SEMA4C. The dual roles of miRNAs like miR-202-3p, which exhibit tumor-suppressive functions in certain cancers but show increased levels in pancreatic cancer cases, underscore the complexity of miRNA-mediated regulation in oncogenesis.

In the realm of computational biology, machine learning (ML) approaches have been increasingly employed to analyze high-dimensional biological data. Sidey-Gibbons and Sidey-Gibbons (2019) highlighted the efficacy of ML techniques in identifying hidden patterns within complex datasets, emphasizing their applicability in medical diagnostics. Despite these advancements, the application of ML to pre-diagnostic miRNA profiles for early pancreatic cancer risk stratification remains underexplored.

This study aims to bridge this gap by leveraging ML algorithms to analyze circulating miRNA profiles, with the objective of developing predictive models for early pancreatic cancer risk assessment.

Methods

The dataset utilized in this study, GSE262260, was obtained from the NCBI Gene Expression Omnibus (GEO) database. It comprises pre-diagnostic plasma miRNA profiles from 462 individuals who developed pancreatic cancer within five years and 462 matched controls who remained cancer-free. The dataset offers a unique opportunity to investigate miRNA signatures preceding clinical diagnosis.

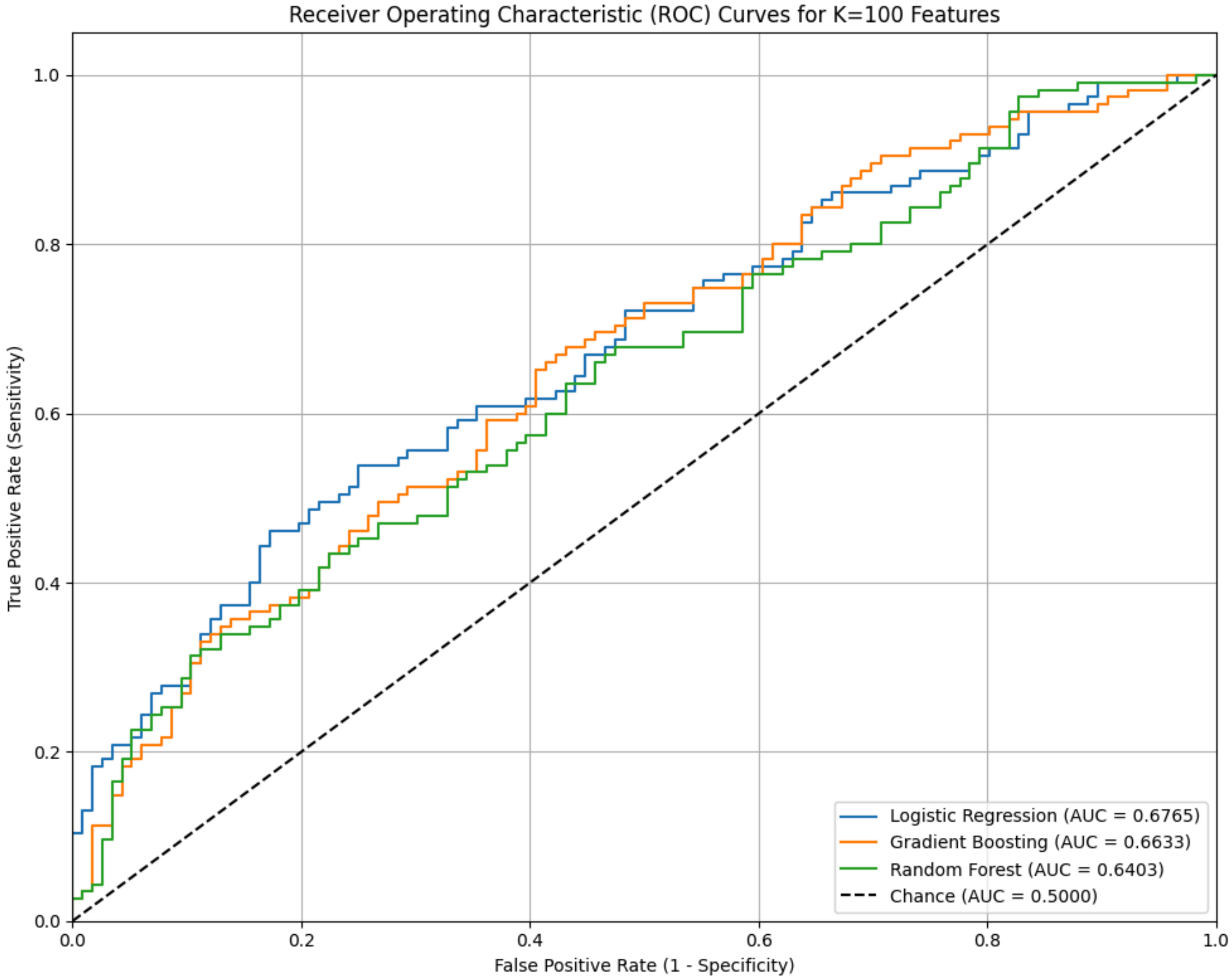
To prepare the data for analysis, several preprocessing steps were undertaken. Missing values within the dataset were addressed using median imputation, ensuring that the central tendency of the data was preserved without introducing significant bias. Features exhibiting low variance (threshold set at 0.01) were removed to eliminate non-informative variables that could potentially hinder model performance. Subsequently, univariate feature selection was performed using the ANOVA F-test via the SelectKBest method, retaining the top 100 miRNAs most relevant to the classification task. Finally, the data was standardized using the StandardScaler to ensure that each feature contributed equally to the model training process.

The preprocessed dataset was partitioned into training and testing sets using a stratified 75%-25% split to maintain the original class distribution. Three machine learning algorithms were selected for model development: Logistic Regression, Random Forest, and Gradient Boosting. Each model underwent hyperparameter tuning through 5-fold cross-validation using GridSearchCV, optimizing for the ROC AUC metric to identify the best-performing configurations.

Model performance was assessed using several evaluation metrics. The primary metric was the Area Under the Receiver Operating Characteristic Curve (ROC AUC), which provides a measure of the model's ability to distinguish between classes. Secondary metrics included accuracy, precision, recall, F1-score, and confusion matrices, offering a comprehensive evaluation of each model's predictive capabilities.

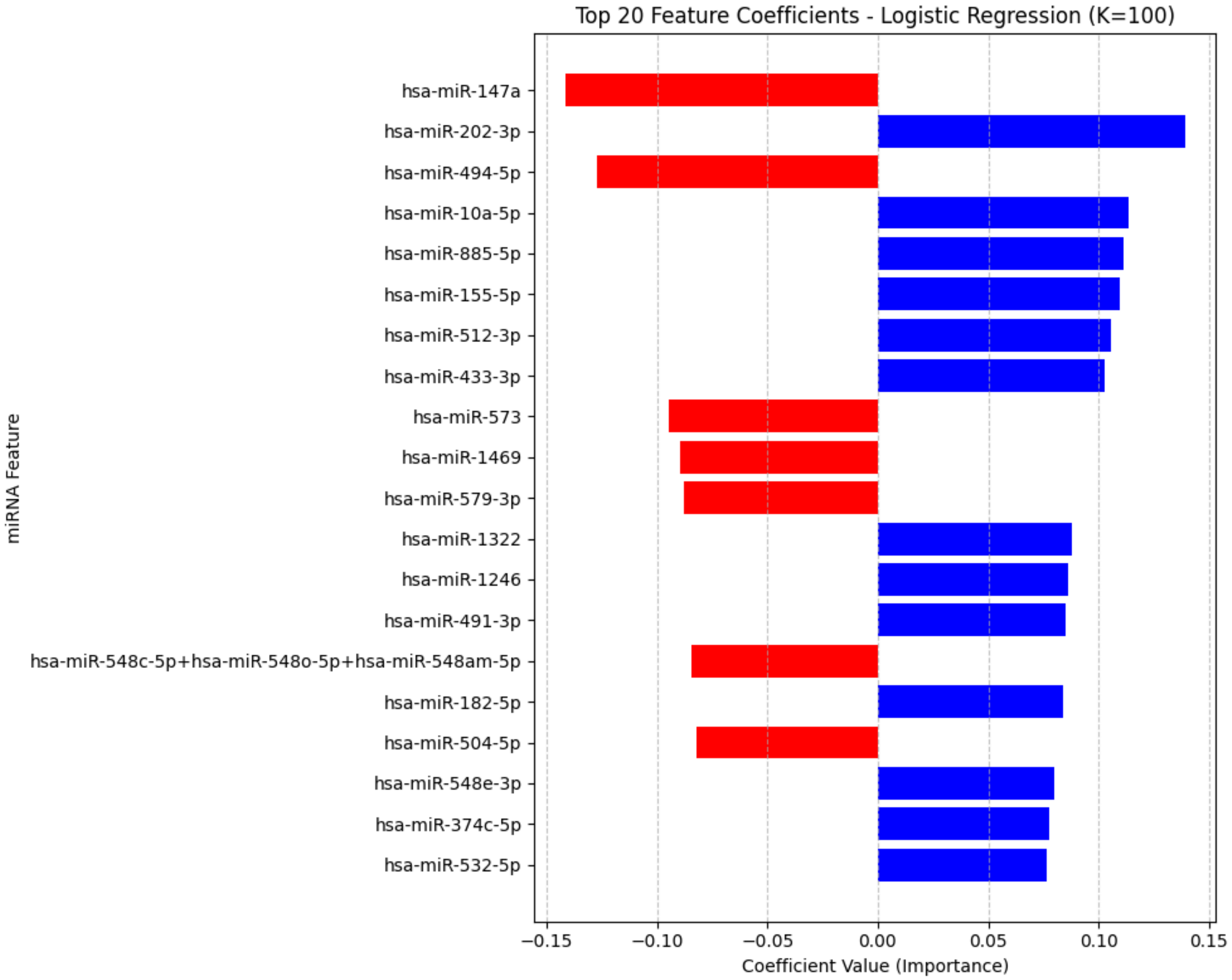
Results

Among the three machine learning models tested, the Logistic Regression classifier demonstrated the highest performance. After tuning with L2 regularization and a low regularization parameter (C=0.01), the model achieved a test ROC AUC of 0.6765 and a test accuracy of 60.17%. This was followed closely by the Gradient Boosting model, which achieved a ROC AUC of 0.6633 with comparable accuracy. The Random Forest classifier, while still effective, exhibited slightly lower performance with a ROC AUC of 0.6403. These results suggest that the relationship between the selected miRNA features and future pancreatic cancer development is likely linear or at least well captured by linear models.



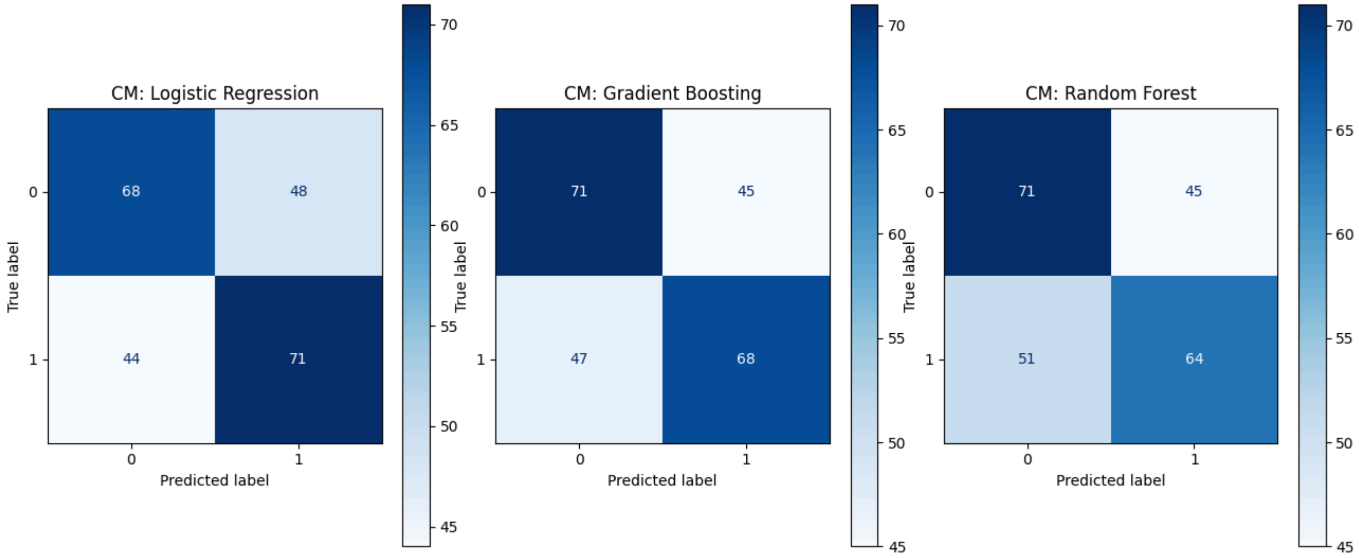
(Figure 1: ROC Curves Comparing Classifiers)

To further probe the biological interpretability of the model, I examined the top 20 most influential miRNAs based on the absolute values of their coefficients in the Logistic Regression model. These miRNAs include hsa-miR-147a and hsa-miR-494-5p, both of which had strong negative coefficients, suggesting a protective association, and hsa-miR-202-3p and hsa-miR-10a-5p, which had positive coefficients, indicating a potential oncogenic role.



(Figure 2: Top 20 Most Predictive miRNAs)

Complementing the quantitative results above, I visualized confusion matrices for all models. These matrices highlight the trade-offs each classifier made in balancing sensitivity and specificity. The Logistic Regression model correctly identified 71 of the 115 positive cases, misclassifying 44, while the Gradient Boosting and Random Forest models had similar distributions of true and false positives and negatives.



(Figure 3: Confusion Matrices Across Models)

Discussion

The results of this study suggest that circulating miRNA profiles possess moderate potential for early risk stratification of pancreatic cancer when analyzed using optimized machine learning techniques. The Logistic Regression model, despite its simplicity compared to tree-based methods, outperformed other classifiers, which may imply that the predictive relationships among the top 100 miRNAs are largely linear in nature. The top-ranked miRNAs identified in this study have been previously implicated in various cancers. For instance, miR-147a has been reported to suppress tumor proliferation in NSCLC, while miR-202-3p, though tumor-suppressive in some cancers, showed a positive association with PC risk in this model. MiR-494-5p, a less studied variant of the miR-494 family, emerged as a potentially novel protective biomarker.

A plausible explanation for the predictive utility of circulating miRNAs lies in their role as mediators of intercellular communication. Tumor cells are known to release miRNAs into circulation via exosomes and other vesicles, which may reflect early tumorigenic changes before a clinical tumor is detectable. These miRNAs can also be released by immune and stromal cells responding to microenvironmental changes associated with incipient malignancy.

It is essential to recognize the limitations of this study. First, the analysis was confined to a single dataset, which limits the generalizability of the findings. Second, although the model performance was statistically significant, the ROC AUC values indicate that there is considerable room for improvement before clinical application is viable. Additionally, while the data preprocessing pipeline was rigorous, other feature selection methods or deep learning architectures could potentially yield improved results. The biological relevance of the identified miRNAs also requires validation in experimental settings to move from correlation to causation.

Future research should focus on validating these findings in independent cohorts and exploring the integration of miRNA data with other omics modalities or known clinical risk factors. Longitudinal studies tracking miRNA expression over time could also shed light on temporal dynamics and improve model accuracy. Functional studies investigating the mechanistic roles of top-ranked miRNAs in pancreatic tumorigenesis are equally critical to translating computational discoveries into clinical insights.

Conclusion

This study demonstrates that machine learning models trained on circulating pre-diagnostic miRNA expression data can moderately predict the 5-year risk of developing pancreatic cancer. The Logistic Regression model, utilizing 100 selected features, achieved the best performance and highlighted a panel of miRNAs that may serve as early biomarkers. Although not yet ready for clinical implementation, these findings open promising avenues for future multi-omic, computational, and experimental research to develop non-invasive tools for early cancer risk assessment.

References

Bartel, D. P. (2004). MicroRNAs. *Cell*, *116*(2), 281–297. https://doi.org/10.1016/s0092-8674(04)00045-5

Chen, P., Zhang, W., Chen, Y., Zheng, X., & Yang, D. (2020). Comprehensive analysis of aberrantly expressed long non‑coding RNAs, microRNAs, and mRNAs associated with the competitive endogenous RNA network in cervical cancer. *Molecular Medicine Reports*, *22*(1), 405–415. https://doi.org/10.3892/mmr.2020.11120

Li, C., Wang, X., & Song, Q. (2020). MicroRNA 885-5p Inhibits Hepatocellular Carcinoma Metastasis by Repressing AEG1; *OncoTargets and Therapy*, *Volume 13*, 981–988. https://doi.org/10.2147/ott.s228576

Lu, Y., & Luan, X. R. (2019). miR-147a suppresses the metastasis of non-small-cell lung cancer by targeting CCL5. *Journal of International Medical Research*, *48*(4). https://doi.org/10.1177/0300060519883098

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., & Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, *105*(30), 10513–10518. https://doi.org/10.1073/pnas.0804549105

O'Neill, K., Syed, N., Crook, T., Dubey, S., Potharaju, M., Limaye, S., Ranade, A., Anichini, G., Patil, D., Datta, V., & Datar, R. (2023). Profiling of circulating glial cells for accurate blood‑based diagnosis of glial malignancies. *International Journal of Cancer*, *154*(7), 1298–1308. https://doi.org/10.1002/ijc.34827

Rachagani, S., Macha, M. A., Heimann, N., Seshacharyulu, P., Haridas, D., Chugh, S., & Batra, S. K. (2014). Clinical implications of miRNAs in the pathogenesis, diagnosis and therapy of pancreatic cancer. *Advanced Drug Delivery Reviews*, *81*, 16–33. https://doi.org/10.1016/j.addr.2014.10.020

Schultz, N. A., Dehlendorff, C., Jensen, B. V., Bjerregaard, J. K., Nielsen, K. R., Bojesen, S. E., Calatayud, D., Nielsen, S. E., Yilmaz, M., Holländer, N. H., Andersen, K. K., & Johansen, J. S. (2014). MicroRNA biomarkers in whole blood for detection of pancreatic cancer. *JAMA*, *311*(4), 392. https://doi.org/10.1001/jama.2013.284664

Sidey-Gibbons, J. a. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0681-4

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA a Cancer Journal for Clinicians*, *73*(1), 17–48. https://doi.org/10.3322/caac.21763

Wang, J., Tao, W., Chen, X., Farokhzad, O. C., & Liu, G. (2017). Emerging Advances in Nanotheranostics with Intelligent Bioresponsive Systems. *Theranostics*, *7*(16), 3915–3919. https://doi.org/10.7150/thno.21317