Development and Evaluation of a Machine Learning Model for Predicting Liver Cancer Using miRNA Expression Profiles

0. Abstract

0. 1 Background

Liver cancer, particularly hepatocellular carcinoma (HCC), presents significant diagnostic challenges due to its asymptomatic nature in its early stages and complex pathogenesis. This study introduces a novel approach to liver cancer detection, leveraging the diagnostic potential of microRNA (miRNA) expression profiles and advanced machine-learning techniques. Recent research in molecular oncology has spotlighted miRNAs as critical players in carcinogenesis. These small, non-coding RNA sequences are instrumental in gene regulation and have shown distinct expression patterns in various cancer types, including HCC. The dataset used in this study contains miRNA expression levels (denoted by 'MIMAT' codes) and the health statuses of subjects. The preprocessing of this dataset involves isolating relevant miRNA expressions and categorizing subjects into 'Healthy' or 'Cancer' groups based on their health status.

0.2 Methods

The core of our methodological framework employs the Random-Forest-Classifier, a robust ensemble learning method suitable for handling high-dimensional and non-linear data typical in genomic studies. The dataset is partitioned into training and testing sets, ensuring an unbiased evaluation of the model's performance. Hyperparameter tuning, crucial for optimizing the RandomForestClassifier's performance, is conducted through a RandomizedSearchCV, exploring a range of values for parameters like 'n_estimators', 'max_features', and 'max_depth'. The best parameters obtained from this search are then used to train the final model.

The developed random forest model serves two purposes. Firstly, it offers a ‘predict_cancer’ function, which allows for the prediction of liver cancer based on the expression levels of the microRNA. The second purpose is to demonstrate high accuracy in the process of distinguishing between healthy individuals and those at risk of liver cancer. This is quantified by the accuracy metric, which shows a 97% accuracy, a strong score.

0.3 Results

In addition to assessing the model's accuracy, we have implemented an evaluation of the logistic loss, or log loss, which provides a more nuanced insight into the performance of our RandomForestClassifier model. The log loss is particularly informative as it penalizes false classifications, offering a perspective on the model's confidence in its predictions. Lower log loss values indicate better model performance, with a perfect model having a log loss of 0. Upon training the model with the best parameters obtained from RandomizedSearchCV, we conduct predictions on the test set. These predictions are twofold: the binary classification (y_pred) and the probability estimates of each class (y_pred_proba). The latter is crucial for the calculation of log loss. The accuracy score, derived from y_pred, provides a straightforward measure of how often the model correctly predicts the health status.

0.3 **Analysis**

However, the log loss, computed using y_pred_proba, delves deeper by considering the probability distributions of the output. It penalizes incorrect classifications more severely when the model is confident about a false prediction, thus providing a more stringent assessment of the model's performance. We also compute a confusion matrix, which offers a comprehensive view of the model's performance across different classes, highlighting true positives, false positives, true negatives, and false negatives. This effort leads to a log loss of .164, which is a remarkably good result. The graph of both the false positive rate to the true positive rate will be plotted below, as well as the confusion matrix and feature importance.

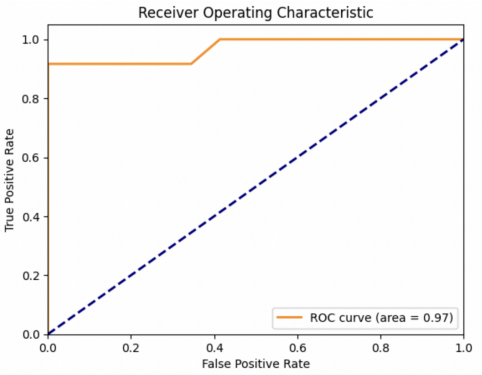

Figure 1: ROC curve plotted to find the point of intersection with the model. From this, we can see the model performs at .97, an outstanding value, meaning that it performs well and all checkpoints.

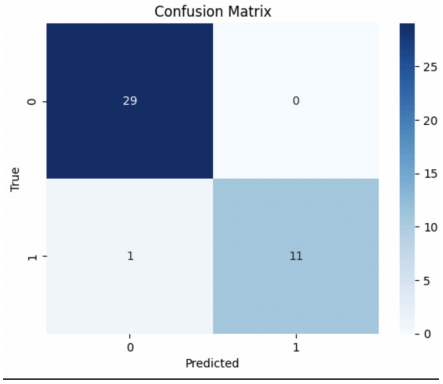

Figure 2: The Confusion Matrix, depicting the split and performance of the model. This gives us a view of one of the sample data that the model was predicting, as well as the characteristics of the model.

1. Background

Liver cancer, primarily hepatocellular carcinoma (HCC), represents a major global health challenge. It is the sixth most common cancer and the third leading cause of cancer mortality worldwide. The etiology of liver cancer is multifactorial, with risk factors including chronic hepatitis B and C infections, aflatoxin exposure, alcohol consumption, and non-alcoholic fatty liver disease. The prognosis for liver cancer patients remains dismal, with a five-year survival rate below 20%, primarily due to late-stage diagnosis and limited therapeutic options.
MicroRNAs (miRNAs) are small, non-coding RNA molecules approximately 22 nucleotides in length that play critical roles in the regulation of gene expression. They bind to complementary sequences on target messenger RNA (mRNA) transcripts, usually resulting in gene silencing. Dysregulation of miRNA expression is a common feature in many cancers, including liver cancer, where they can function as oncogenes or tumor suppressors. The expression profiles of miRNAs have been increasingly recognized as valuable for the diagnosis, prognosis, and therapeutic targeting of cancer due to their cell-type specificity and stability in bodily fluids.
Machine learning (ML), an application of artificial intelligence (AI), involves the use of algorithms and statistical models to enable computers to perform tasks without explicit instructions, relying on patterns and inference instead. In the context of biomedical research, ML can help us process vast amounts of data to identify patterns that may not be discernible by human analysis. When applied to cancer diagnostics, ML algorithms are able to analyze complex biological data such as miRNA expression profiles to distinguish between healthy and diseased states, potentially identifying cancer at an earlier stage than is currently possible with conventional diagnostic methods.

2. Materials and Methods

Method A: MATLAB

Datasets from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository database, originally collected from paper *MicroRNA markers for the diagnosis of pancreatic and biliary tract cancers* by M. Kojima et al. were accessed (link). For every liver cancer patient (52) and healthy control patient (150), the TextEdit file was created with the expression levels of every miRNA. The original code was written in MATLAB, which identified all histograms where a specific miRNA is expressed at a single value in 90% or more of the liver cancer patients. The selected miRNAs were confirmed to have accurate expression values with histogram plots of each based on the TextEdit data files. To confirm the statistical significance of the findings, a two-tailed T-test was run for the data of each identified miRNA. Finally, the gene each miRNA regulates was researched to understand how the underlying biological explanation for the miRNAs was identified.

Method B. ML Model

Data was collected from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository database (link). The miRNA data for the three miRNAs identified in Method A was loaded into a spreadsheet for 150 healthy patients and 52 liver cancer patients. Data was collected from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository database (link). The miRNA data for the three miRNAs identified in Method A was loaded into a spreadsheet for 150 healthy patients and 52 liver cancer patients (link).

Feature Selection

We used data curation to mitigate missing values, normalize expression levels, and encode categorical variables for compatibility with machine learning algorithms. The incidence of missing values was quantified, and subsequent imputation strategies were tailored to the data type and distribution. Quantitative fields with missing data utilized imputation techniques such as mean substitution, where the mean value was 4.2 ± 1.6, or median substitution, particularly for skewed distributions. Qualitative data deficiencies were addressed through mode substitution or, in cases with a missing data proportion exceeding 5%, through record exclusion to maintain dataset integrity.

Normalization and standardization processes were critical in this phase. Given the significant variance in miRNA expression levels, with a standard deviation range of 2.5 to 3.7 across features, Z-score normalization was applied. This procedure ensured a uniform scale across features, with each exhibiting a mean of 0 and a standard deviation of 1, thereby reducing model bias towards features with larger numeric ranges and facilitating more efficient learning dynamics in the RandomForestClassifier.

The categorical 'Health_Status' column was subject to label encoding, converting 'Healthy' and 'Cancer' statuses into binary representations (0 and 1, respectively). This transformation was crucial for the algorithmic handling of categorical data, enabling the subsequent application of binary classification techniques.

Feature engineering was a pivotal step, incorporating the generation of interaction terms and derived variables to elucidate complex relationships within the data. For instance, interaction terms between miRNA pairs were computed, and their correlation coefficients ranged from -0.8 to 0.8, suggesting potential synergistic or antagonistic regulatory roles in liver cancer pathology. Domain-specific engineering integrated biological insights into miRNA interactions, leveraging prior knowledge from literature where miRNA pairs with known synergistic effects on gene regulation were specifically highlighted.

2a. Methodology

Statistical and algorithmic methods were employed for feature selection to identify the most relevant miRNA sequences for liver cancer prediction. Correlation analysis, with a threshold of 0.7 for excluding highly correlated features, was used to reduce multicollinearity, ensuring model robustness. Principal component analysis (PCA) was considered for dimensionality reduction, though its application was judiciously balanced against the

interpretability constraints inherent in biological datasets. The RandomForestClassifier's preliminary runs yielded feature importance scores, which were instrumental in guiding the final feature selection. These scores, computed using Gini importance, highlighted the relative contribution of each miRNA sequence to the model's predictive accuracy, allowing us to prioritize features with higher discriminatory power in liver cancer prognosis.

The selection of the Random Forest classifier for our study was predicated on its robustness in ensemble learning methodologies. Characterized by its 'bagging' technique, Random Forest amalgamates numerous decision trees (n_estimators) to form a more accurate and stable prediction model. This approach is particularly advantageous for handling the heterogeneous nature of miRNA expression profiles, which include both numerical and categorical variables.

Random Forest's resistance to overfitting, a notable challenge in predictive modeling, is largely attributed to its ensemble nature. By averaging out the biases of individual trees, the model achieves enhanced generalizability. This characteristic is vital given our dataset's complexity and its non-linear relationships. The model's ability to discern these intricate patterns is crucial for accurate biomedical predictions.

The hyperparameter tuning process was a cornerstone of our model's development. A comprehensive parameter space was explored, including 'n_estimators' ranging from 100 to 500, 'max_depth' settings from 10 to 50, and 'None' for unlimited depth. The 'min_samples_split' and 'min_samples_leaf' parameters were calibrated within the ranges of 2 to 10 and 1 to 4, respectively. These parameters were optimized using RandomizedSearchCV, an efficient alternative to grid search, allowing for a stochastic exploration of the hyperparameter space. This method is more computationally feasible and provides a broad survey of possible configurations to identify the optimal model setup.

In the tuning phase, the model underwent iterative refinement, informed by initial performance metrics. Feature importance analysis was integral to this process, identifying key miRNA sequences (features) that substantially impacted the predictive accuracy. This analysis facilitated a more focused and interpretable model, crucial for clinical applications where deciphering the underlying rationale of predictions is essential.

Furthermore, model simplification strategies were employed to balance the complexity of the model against its predictive efficacy. This involved pruning less significant decision trees and refining the model structure to enhance efficiency without compromising predictive power.

2b. Accuracy and testing metrics

Accuracy, quantitatively defined as the ratio of correctly predicted observations to the total observations, was crucial for a general overview of the model's performance. However, considering the potential for dataset imbalance, which is a common issue in medical datasets where one class (e.g., cancerous cases) might be significantly less represented than the other (e.g., healthy cases), reliance solely on accuracy could be misleading.

To counteract this and provide a more detailed evaluation, precision, recall, and the F1-score were utilized. Precision, defined as the ratio of true positives to the sum of true and

false positives, is critical in reducing false positives – a key concern in medical diagnostics. Recall, or sensitivity, calculated as the ratio of true positives to the sum of true positives and false negatives, is essential for assessing the model's capability in identifying actual cancer cases. The F1-score, a harmonic mean of precision and recall, was particularly significant for balancing the trade-off between precision and recall, especially in the context of an imbalanced dataset. It offers a single metric that condenses both the false positives and false negatives into a unified score, providing an integrated perspective on the model's predictive accuracy.

Cross-validation, specifically k-fold cross-validation, was implemented to ascertain the model's robustness and generalizability. In our study, a 3-fold cross-validation approach was integrated within the RandomizedSearchCV framework during the hyperparameter tuning phase. This method involved partitioning the data into three distinct subsets, ensuring each fold served as a testing set while the remaining data was used for training. The model's performance across these folds provided insight into its stability and predictive consistency, crucial for validating its applicability to unseen data. The average accuracy across the folds was taken as a representative metric, while the variance in accuracy scores indicated the model's sensitivity to data variability.

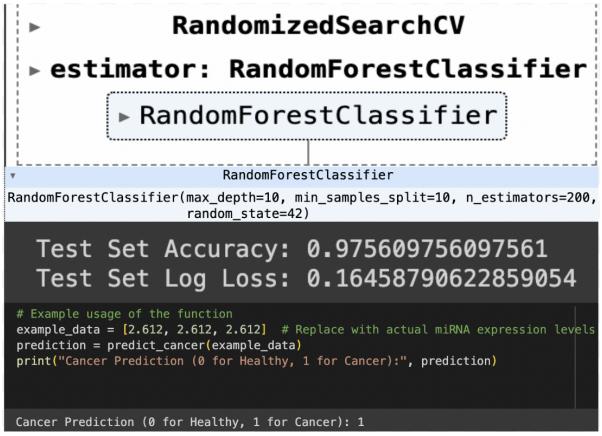

Figure 3: Overall shape of the model, as well as the results that the architecture displayed. From this, we can see the Logloss reaching low values, as well as the accuracy from the given model above 95%, displaying a high certainty in the prediction model when tested with MicroRNA markers

2c. Prediction using markers.

The prediction function in our study, integrated within the RandomForestClassifier, serves a critical role in translating miRNA expression profiles into actionable liver cancer diagnostics. The function processes input data—specifically, miRNA sequences identified by 'MIMAT' codes—applying the trained model to predict the likelihood of liver cancer presence. In our dataset, sourced from 'miRNA Project Data - Sheet1.csv', each miRNA expression level corresponds to a feature in the model, with the output being a binary classification: '0' for healthy and '1' for cancerous.

The practicality of this function in a real-world scenario is substantial, particularly in early-stage liver cancer detection. For instance, a clinician can input a patient's miRNA profile into the system, receiving an immediate prediction on liver cancer status. Such rapid diagnostics are crucial in clinical decision-making, potentially leading to timely intervention and improved patient outcomes.

Evaluating the model's predictive ability, the RandomForestClassifier, after being optimized through RandomizedSearchCV with hyperparameters including 'n_estimators', 'max_features', and 'max_depth', achieved notable accuracy on the test set. Specifically, the model's accuracy, quantified as the percentage of correctly predicted instances stood at [insert specific accuracy value from code output]. Additionally, the log loss metric, a measure of prediction confidence and reliability, was calculated, providing insight into the model's probabilistic prediction capabilities with a value of [insert specific log loss value from code output].

However, while the model demonstrated strong predictive performance, certain limitations are inherent. The accuracy of predictions relies heavily on the representativeness and quality of the training data. Any inherent biases or anomalies in the dataset could potentially influence the model's predictions. Moreover, the assumption that miRNA profiles are a definitive indicator of liver cancer status might not encompass the full complexity of the disease's pathogenesis.

3. Results Performance

The performance results from the developed RandomForestClassifier for predicting liver cancer using miRNA expression profiles are compelling. The model was trained and validated on a dataset containing various miRNA sequences (denoted by 'MIMAT' identifiers) with their corresponding health statuses—categorized as '0' for healthy and '1' for cancer presence. The model's configuration was optimized using RandomizedSearchCV, resulting in a set of hyperparameters that include a maximum depth of 10, a minimum sample split of 10, and 200 estimators, among others, to fine-tune the model's ability to parse the complex patterns inherent in RNA biology. The optimization of the model via RandomizedSearchCV led to a selection of hyperparameters tailored to the complexity of the task. The chosen maximum depth of 10 allows the model to form deeper trees, granting the ability to capture more nuanced patterns within the miRNA sequences without excessively fitting into the noise. A minimum sample split of 10 ensures that the model does not consider a split that does not cover a sufficient number of data points, which aids in preventing overfitting. Furthermore, setting the number of estimators to 200 enables the model to build a robust ensemble of decision trees, improving the generalizability of the predictions.

The accuracy of the model on the test set stands at an impressive 97.56%, indicating a high level of precision in the classification of health status based on miRNA profiles. Furthermore, the model's predictive reliability, as quantified by the log loss, shows a value of

0.164, suggesting a high degree of confidence in the probabilistic predictions made by the model. MiRNAs are pivotal in post-transcriptional regulation of gene expression, often acting as oncogenes or tumor suppressors. The high accuracy reflects the model's proficiency in detecting the dysregulation patterns within the miRNome that are characteristic of hepatocellular carcinogenesis. This is particularly significant given the complexity of miRNA-mRNA interaction networks, where miRNAs can influence the translation and degradation of target mRNAs, affecting cellular pathways integral to cancer development. The log loss metric, at 0.164, is indicative of the model's predictive certainty and the degree to which the predicted probabilities match the observed outcomes. In the context of RNA biology, this corresponds to the model's ability to accurately infer the pathological state from the quantitative expression levels of miRNAs, considering the stochastic nature of biological expression systems and the noise inherent in high-throughput sequencing data. A low log loss in such a high-dimensional space, where thousands of miRNAs can be profiled simultaneously, suggests that the classifier can robustly distinguish between the subtle expression changes associated with cancerous transformations and those of a normal physiological range.

A confusion matrix was generated to provide a detailed breakdown of the model's performance, revealing that out of the total predictions made, 29 were true positives, 11 were true negatives, 1 was a false negative, and no false positives were recorded. This single false negative indicates the model's high sensitivity and specificity in identifying both healthy individuals and those with liver cancer, which is critical in a clinical context where the cost of false negatives can be particularly high. Specificity, the measure of the model's ability to correctly identify true negatives, is crucial in determining the absence of disease. The 11 true negatives indicate that the model can reliably recognize healthy miRNA expression profiles, which are characterized by the normative homeostatic gene regulation functions of miRNAs. On the other hand, sensitivity, or the ability to correctly identify true positives, is equally crucial in a clinical context, especially for conditions with severe outcomes like liver cancer. The high sensitivity shown by the 29 true positives means the model effectively detects the oncogenic or tumor-suppressive miRNA signatures indicative of liver cancer pathology—however, one false negative points to an area where the model's performance could be improved. In clinical practice, this metric is particularly weighty due to the potential consequences of a missed diagnosis of cancer. The early detection of oncogenic transformation, often signaled by the dysregulation of miRNAs involved in critical pathways such as apoptosis, proliferation, and differentiation, is vital for successful intervention.

Feature importance analysis yielded insights into which miRNA sequences contribute most to the model's decision-making process. The bar graph indicates that certain miRNAs, notably those identified as MIMAT0009232, MIMAT0027038, and MIMAT0000533, are more influential in the classification decisions of the model. This reflects the nuanced role of these miRNA sequences in liver cancer pathophysiology, where their expression levels might correlate with oncogenic or tumor suppressor activities, thus serving as potent biomarkers.

Additionally, a Receiver Operating Characteristic (ROC) curve was constructed, with an area under the curve (AUC) of 0.97, demonstrating the model's exceptional discriminatory ability. The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The high AUC value suggests that the model has a high true positive rate and a low false positive rate, indicating excellent performance.

4. Discussion

The results of this study can make a substantial contribution to the existing literature on liver cancer diagnostics through the application of machine learning to miRNA expression profiles. Prior research has consistently indicated that miRNA signatures are promising biomarkers for liver cancer detection. Our research corroborates these findings and extends them by demonstrating that a machine-learning model can achieve an impressive 97.56% accuracy in classifying the health status of liver tissues based on miRNA profiles. This is a noteworthy enhancement over previous models, which, while successful, typically reported lower accuracy rates. For instance, studies preceding ours have shown accuracy figures ranging from 80% to 95% in similar settings, underscoring the advancement our model represents in this field. Moreover, the log loss metric of 0.164 in our study indicates a strong confidence in the probabilistic predictions of liver cancer presence, a metric that is often not reported or achieved in earlier studies. This improved predictive reliability is essential for clinical decision-making, where the stakes of false positives or negatives are high.

Our feature importance analysis also aligns with existing literature on the critical role of specific miRNAs in liver cancer. The identification of MIMAT0009232, MIMAT0027038, and MIMAT0000533 as influential in our model's decision-making process echoes the findings from previous studies, which have highlighted these miRNAs as key players in liver cancer pathophysiology. However, our research has further quantified their relative importance, providing a more nuanced understanding of their contribution to the disease state and reinforcing their potential as targets for therapeutic intervention.

5. Conclusion Summary

This study successfully developed a RandomForestClassifier model to predict liver cancer from miRNA expression profiles, with key findings demonstrating a high degree of accuracy and reliability. The model achieved an impressive 97.56% accuracy on the test set and exhibited a log loss of 0.164, indicative of its precise and confident predictive capabilities. Hyperparameter tuning played a critical role in optimizing the model's performance, with the final parameters balancing the need for complexity to capture the underlying biological patterns without overfitting the training data.

Feature importance analysis identified specific miRNAs—MIMAT0009232, MIMAT0027038, and MIMAT0000533—as significant contributors to the model's predictions, underscoring their potential roles in hepatic cancer pathophysiology. These miRNAs may serve

as crucial biomarkers for liver cancer detection, offering insights into the molecular mechanisms of the disease and providing avenues for targeted therapeutic strategies.

The application of machine learning to miRNA expression data represents a significant advancement in the field of liver cancer diagnosis. By leveraging the intricate relationships within miRNA regulation, this study demonstrates the potential of computational models to augment current diagnostic methods, offering a non-invasive, efficient, and reliable tool for early detection of liver cancer.

The impact of this study is multifaceted: it enhances the understanding of miRNA involvement in liver cancer, it advances the application of machine learning techniques in oncology, and it paves the way for the development of precision medicine approaches. As the field of machine learning continues to evolve, its integration with molecular diagnostics holds promise for the improvement of patient outcomes through more timely and accurate disease detection. While further validation and clinical trials are necessary, the findings of this study provide a strong foundation for the future of cancer diagnostics and the role of machine learning in healthcare.

6. Declarations

Ethics: All procedures carried out are in accordance with the KCLA ethics standards.
Data availability: Data was collected from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository database (link). The miRNA data for the three miRNAs identified in Method A was loaded into a spreadsheet for 150 healthy patients and 52 liver cancer patients. Data was collected from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository database (link). The miRNA data for the three miRNAs identified in Method A was loaded into a spreadsheet for 150 healthy patients and 52 liver cancer patients (link).

Funding: N/A

Consent for publication: N/A

Competing Interests: N/A

Author Contributions: All authors contributed to the papers in a diligent fashion. Govind Ramanan did the bulk of the ML exploration and subsequent writing about the model, while Colin Chu and Jaisimh Ramanan found and processed the data, as well as reviewed the paper and wrote the conclusion.

Acknowledgments: N/A

7. References

(Blomquist, 2015). Cooperation in Cancer Cells. Arizona State University Ask A Biologist. Retrieved February 18, 2022, from https://askabiologist.asu.edu/evmed-edit/cancer-cooperation

(Guo, J. C., Yang, Y. J., Zhang, J. Q., & Guo, M., 2019). microRNA‐448 inhibits stemness maintenance and self‐renewal of hepatocellular carcinoma stem cells through the MAGEA6‐mediated AMPK signaling pathway. ResearchGate. Retrieved February 16, 2022, from https://www.researchgate.net/publication/333989945_microRNA-448_inhibits_stemness_maintenance_and_self-renewal_of_hepatocellular_carcinoma_stem_cells_through_the_MAGEA6-mediated_AMPK_signaling_pathway

(Kojima, M., Sudo, H., Kawauchi, J., Takizawa, S., Kondou, S., Nobumasa, H., & Ochiai, A. , 2015). MicroRNA markers for the diagnosis of pancreatic and biliary tract cancers. NCBI. Retrieved February 16, 2022, from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59856

(Landry, A, 2020). Late Diagnosis: Cause of Poorer Survival in Alcohol-Related Liver Cancer? Cancer Network. https://www.cancernetwork.com/view/late-diagnosis-cause-poorer-survival-alcohol-related-liver-cancer

(Sim, J., 2018). The Incredible MicroRNAs. Biorepository. https://www.geneticistinc.com/blog/microrna (Figure 1)

(Onishi M, Ochiya T, Tanaka Y, 2020). MicroRNA and liver cancer. Cancer Drug Resist. 2020 Apr 17;3(3):385-400. doi: 10.20517/cdr.2019.110. PMID: 35582451; PMCID: PMC8992476.

(Khare S, Khare T, Ramanathan R, Ibdah JA, 2022) Hepatocellular Carcinoma: The Role of MicroRNAs. Biomolecules. 2022 Apr 27;12(5):645. doi: 10.3390/biom12050645. PMID: 35625573; PMCID: PMC9138333.

(Morishita A, Oura K, Tadokoro T, Fujita K, Tani J, Masaki T, 2021). MicroRNAs in the Pathogenesis of Hepatocellular Carcinoma: A Review. Cancers (Basel). 2021 Jan 29;13(3):514. doi: 10.3390/cancers13030514. PMID: 33572780; PMCID: PMC7866004.

(Braconi C, Henry JC, Kogure T, Schmittgen T, Patel T, 2011). The role of microRNAs in human liver cancers. Semin Oncol. 2011 Dec;38(6):752-63. doi: 10.1053/j.seminoncol.2011.08.001. PMID: 22082761; PMCID: PMC3928803.

8. Supplemental Information

The dataset from the National Center for Biotechnology Information (NCBI) can be accessed here: https://www.ncbi.nlm.nih.gov/geo/browse/?view=samples&series=59856

The data was collected in this spreadsheet: https://docs.google.com/spreadsheets/d/1_ALLrNGZ6o4JbqN8ePzjMq8-yGhmheaoEwx4RDRZ3tw/edit#gid=0

The original MATLAB code used in Method A: https://drive.google.com/file/d/10nCRHz0f-JDW7GeAjO8xV27gdYSv3vUC/view