

CBIO313: data mining and machine learning

Investigating the Role of miRNA Editing in Lung Adenocarcinoma Classification using Machine Learning Models



Sajda Hussien tahoun

Under supervision of:

Dr. Mohamed Elsayeh

**Table of Contents**

[**1. Introduction 1**](#_Toc168006197)

[1.1 Dataset overview 2](#_Toc168006198)

[**2. Methodology 2**](#_Toc168006199)

[2.1 Exploratory Data Analysis and Pre-processing 2](#_Toc168006200)

[2.2 Feature Engineering 3](#_Toc168006201)

[2.3 Machine Learning Model Development 4](#_Toc168006202)

[2.4 Model Evaluation and Fine-tuning 4](#_Toc168006203)

[2.5 Model Deployment 5](#_Toc168006204)

[**3. Results 5**](#_Toc168006205)

[**4. Conclusion 5**](#_Toc168006206)

[4.1 Limitations 6](#_Toc168006207)

# **1. Introduction**

Across the globe, lung cancer is the number one cause of death from cancer. Among all lung cancers, adenocarcinoma is the most frequent type, falling under the category of non-small-cell lung cancer (NSCLC) (Keita Maemura et al., 2018). MicroRNA editing, specifically the conversion of adenosine to inosine (A-to-I), has been linked to tumor characteristics across different cancer forms. Recent examinations of The Cancer Genome Atlas (TCGA) data have identified numerous microRNAs subject to A-to-I editing in human cancer samples, with some of them implicated in prognostic outcomes (Keita Maemura et al., 2018).

My main objective is to create diverse machine learning models for classifying whether an individual has cancer based on miRNA editing expressions. I kickstart this process by conducting Exploratory Data Analysis (EDA), employing data visualization techniques, and ensuring thorough data cleaning, including the removal of duplicates, and handling null values.

By employing various machine learning models—such as XGBoost, K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Decision Tree—I aim to identify the most accurate model for this classification task. After initial model training, I refine my approach by focusing on the most important features, utilizing hyperparameter tuning through GridSearch to further optimize model performance. The culmination of this project involves deploying the best-performing model using Streamlit, providing an interactive platform for real-time cancer classification.

This project not only seeks to enhance diagnostic accuracy through advanced computational techniques but also aims to contribute to the broader field of cancer research by demonstrating the practical applications of machine learning in medical diagnostics.

## **1.1 Dataset overview**

The dataset used in this project consists of 2568 different miRNA expressions collected from 38 samples. These samples are evenly divided into two groups: the first 19 samples are from diseased individuals diagnosed with cancer, and the remaining 19 samples are from control individuals without cancer. This balanced composition is crucial for ensuring a robust comparative analysis between the diseased and control groups.

Each sample's miRNA expression profile provides a comprehensive snapshot of the molecular activity within the cells, making it possible to identify patterns and signatures associated with cancer. The miRNA expressions included in the dataset have been specifically measured to capture the conversion of adenosine to inosine (A-to-I) editing, a process linked to various tumor characteristics and prognostic outcomes.

# **2. Methodology**

## **2.1 Exploratory Data Analysis and Pre-processing**

Exploratory Data Analysis (EDA) and preprocessing are foundational steps in any data science project, crucial for understanding the dataset and preparing it for modeling. EDA involves summarizing the main characteristics of the data, often using visual methods such as histograms, scatter plots, and box plots. This process helps identify patterns, anomalies, and relationships within the data. Recent studies emphasize that EDA is essential for uncovering underlying structures and generating hypotheses for further analysis.

Preprocessing, on the other hand, includes cleaning the data by handling missing values, removing duplicates, and standardizing or normalizing the features. This step ensures that the dataset is in a suitable format for machine learning algorithms, which are sensitive to the quality of the input data. Proper preprocessing can significantly enhance model performance by improving the accuracy and efficiency of the learning process.

Together, EDA and preprocessing lay the groundwork for building robust and accurate machine learning models. They ensure that the data is both understandable and usable, facilitating better decision-making and more reliable outcomes in the subsequent modeling stages.

As for our dataset, I conducted Exploratory Data Analysis (EDA) and preprocessing to understand and prepare the dataset for modeling. During EDA, I visualized the data using techniques such as histograms, scatter plots, and box plots to identify patterns, anomalies, and relationships within the dataset. For preprocessing, I handled missing values, removed duplicates, and standardized or normalized the features to ensure that the dataset was in a suitable format for machine learning algorithms. These steps are essential for gaining insights into the data's characteristics and ensuring its quality before proceeding with model development.

For the outliers, in the context of biological data, where every miRNA expression holds significance, outlier detection becomes paramount. While we aim to identify outliers, we refrain from discarding them. This approach stems from a shift in perspective on outliers within biological research. Historically, such outliers were often dismissed, despite their potential to unveil crucial insights into biological processes. Cook et al. (2021) underscore the importance of reconsidering outliers, emphasizing their potential as rare events or individuals pivotal in driving evolutionary dynamics within populations. Therefore, while we acknowledge the necessity of scaling features to handle variance, we maintain a stance of inclusivity towards outliers, recognizing their value in enhancing our understanding of biological phenomena.

## **2.2 Feature Engineering**

Feature engineering involves transforming raw data into a format that is more suitable for machine learning algorithms, with the goal of improving model performance. This process includes creating new features, selecting relevant features, and transforming existing features to extract more useful information.

In my project, I employed feature engineering techniques to enhance the predictive power of the machine learning models for cancer classification based on miRNA editing expressions. I performed feature selection to choose the most relevant features and removed any redundant or irrelevant ones. Additionally, I standardized or normalized the features to ensure that they were on a similar scale, which is important for certain machine learning algorithms. By optimizing the feature set in this way, I aimed to improve the accuracy and efficiency of the models in predicting cancer status based on miRNA expressions.

A screenshot of a computer

Description automatically generated

## **2.3 Machine Learning Model Development**

In the machine learning model development phase, I utilized various algorithms to build predictive models for cancer classification based on miRNA editing expressions. These algorithms included XGBoost, K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Decision Tree. I began by splitting the dataset into training and testing sets to train the models on a subset of the data and evaluate their performance on unseen data. Each model was trained using the training data, and its performance was assessed using relevant metrics such as accuracy, precision, recall, and F1-score. This iterative process allowed me to compare the performance of different algorithms and select the most promising ones for further refinement.

## **2.4 Model Evaluation and Fine-tuning**

During the model fine-tuning phase, I focused on optimizing the performance of the Random Forest (RF) algorithm, which emerged as the best-performing model during the initial evaluation on the selected features. Fine-tuning RF involved systematically exploring different combinations of hyperparameters to identify the configuration that maximized predictive performance. Utilizing techniques such as GridSearch, I searched over a predefined hyperparameter space, adjusting parameters such as the number of trees in the forest, the maximum depth of the trees, and the minimum number of samples required to split a node. By evaluating the performance of RF with varying hyperparameter values on a validation set, I identified the optimal combination that yielded the highest accuracy and robustness in cancer classification.

## **2.5 Model Deployment**

Once the RF model was fine-tuned and optimized for performance, I proceeded to deploy it using Streamlit, an open-source framework for building interactive web applications. The deployed model leveraged the optimized hyperparameters and feature set developed during the earlier phases of the project. Through the Streamlit interface, users could input miRNA editing expressions and receive real-time predictions regarding cancer classification. This deployment provided an accessible platform for users to utilize the predictive capabilities of the RF model without requiring extensive programming knowledge. By deploying the RF model on Streamlit, I facilitated the practical application of the developed model for cancer classification based on miRNA expressions.

# **3. Results**

The Random Forest (RF) algorithm emerged as the most accurate model for this dataset when deployed on the selected features, achieving a high accuracy of 97%. Despite efforts to fine-tune its hyperparameters, including the number of trees and maximum depth, hyperparameter tuning did not yield further improvements in model accuracy. This suggests that the RF model had already reached its maximum predictive capability, with little room for enhancement through hyperparameter optimization. However, it's important to note that while hyperparameter tuning did not lead to a higher accuracy, the RF model's robust performance underscores its effectiveness in handling the intricacies of the dataset and accurately classifying cancer based on miRNA editing expressions.

Feature selection also played a crucial role in optimizing the performance of the machine learning models. By identifying and selecting the most relevant features, we aimed to reduce dimensionality, improve model interpretability, and potentially enhance predictive accuracy. Through techniques such as feature importance ranking, we identified a subset of features that had the most significant impact on predicting lung adenocarcinoma status based on miRNA editing expressions. This process not only simplified the model complexity but also helped mitigate the risk of overfitting by focusing on the most informative features.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# **4. Conclusion**

In conclusion, this study highlights the efficacy of machine learning algorithms in cancer classification based on miRNA editing expressions. Through rigorous model development and evaluation, the Random Forest (RF) algorithm demonstrated superior accuracy when deployed on selected features. This success underscores the potential of RF in effectively handling the complexities of biological data and capturing nuanced patterns for accurate classification. However, it's important to acknowledge that the choice of RF as the best-performing model is specific to this dataset and may not generalize to all scenarios. Future research should continue exploring alternative algorithms and techniques to further enhance predictive performance and uncover additional insights into cancer biology.

## **4.1 Limitations**

Despite the promising results, this study has several limitations. Firstly, the dataset used for model development consists of a relatively small number of samples, with only 19 diseased and 19 control individuals. This limited sample size may not fully capture the variability present in miRNA expression patterns across different cancer types and patient populations, potentially limiting the generalizability of the findings. Additionally, the small sample size increases the risk of overfitting, where the models may learn patterns specific to the training data but fail to generalize well to unseen data. Future research with larger and more diverse datasets will be essential for validating the robustness and reliability of the developed models. Addressing these limitations through the inclusion of more samples and rigorous cross-validation techniques will be critical for advancing the field of cancer classification using machine learning approaches.

**References**

Cook, C. N., Freeman, A. R., Liao, J. C., & Mangiamele, L. A. (2021). The Philosophy of Outliers: Reintegrating Rare Events Into Biological Science. *Integrative and Comparative Biology*, *61*(6), 2191–2198. https://doi.org/10.1093/icb/icab166

Guo, S., Mao, C., Peng, J., Xie, S., Yang, J., Xie, W., Li, W., Yang, H., Guo, H., Zhu, Z., & Zheng, Y. (2024). Improved lung cancer classification by employing diverse molecular features of microRNAs. *Heliyon*, e26081–e26081. https://doi.org/10.1016/j.heliyon.2024.e26081

Keita Maemura, Watanabe, K., Ando, T., Hiyama, N., Toshio Sakatani, Amano, Y., Kage, H., Nakajima, J., Yutaka Yatomi, Nagase, T., & Takai, D. (2018). Altered editing level of microRNAs is a potential biomarker in lung adenocarcinoma. *Cancer Science*, *109*(10), 3326–3335. https://doi.org/10.1111/cas.13742