Jeffrey Amankwah & Ivan Ade

Dr. Solka

BINF 702

04/23/23

Predicting Multi-Drug Resistance using in Acute Lymphoma Leukemia Patients using the ALL Dataset and Machine Learning Models

**Abstract**

Leukemia, a common blood, and bone marrow cancer, often presents treatment challenges due to the development of multidrug resistance, which reduces chemotherapy efficacy and leads to unfavorable outcomes. This study aims to predict multidrug resistance in leukemia patients with the use of machine learning techniques and gene data from the Acute Lymphoblastic Leukemia (ALL) dataset. Various machine learning approaches, including logistic regression, random forests, and support vector machines, will be applied to the ALL dataset to predict leukemia patients' multidrug resistance based on their gene expression patterns. Feature selection methods will also be used to pinpoint the most relevant genes for multidrug resistance prediction. Identifying patients prone to multidrug resistance enables healthcare professionals to customize treatment plans, ultimately enhancing patient outcomes. This study will contribute to the creation of personalized treatment strategies for leukemia patients, ultimately improving their prognosis.

**Introduction**

Leukemia is a type of cancer originating in the bone marrow, resulting in the production of abnormal blood cells, primarily white blood cells or leukocytes (American Cancer Society, 2021). The disease can be classified into four main types: Acute Lymphoblastic Leukemia (ALL), Chronic Lymphocytic Leukemia (CLL), Acute Myeloid Leukemia (AML), and Chronic Myeloid Leukemia (CML), depending on the affected cell type and the rate of disease progression (National Cancer Institute, 2021). This relatively common cancer accounted for over 60,000 new cases and nearly 24,000 deaths in the United States in 2021 alone (Siegel et al., 2021). Risk factors for leukemia include exposure to ionizing radiation, certain chemicals, genetic factors, and a family history of the disease (National Cancer Institute, 2021). Common symptoms of leukemia encompass fatigue, fever, weight loss, frequent infections, and easy bruising or bleeding (American Cancer Society, 2021).

To diagnose leukemia, healthcare professionals employ blood tests, bone marrow biopsies, and imaging studies (Mayo Clinic, 2021). Treatment options vary depending on the type and stage of the disease and may include chemotherapy, targeted therapy, radiation therapy, stem cell transplantation, and immunotherapy (American Cancer Society, 2021). There are four main types of leukemia, which can be further divided into subtypes based on specific characteristics. Each type of leukemia differs in its prevalence, symptoms, and treatment options (American Cancer Society, 2021).

Acute Lymphoblastic Leukemia (ALL) is a rapidly progressing cancer that affects immature lymphoid cells and is the most common type of leukemia in children, although it can also occur in adults (Pui et al., 2004). Chronic Lymphocytic Leukemia (CLL) is a slow-growing cancer involving mature lymphocytes and is the most common type of leukemia in adults, particularly in older individuals (Hallek et al., 2018). Acute Myeloid Leukemia (AML) is an aggressive cancer of the myeloid cells, which can affect both children and adults and has several subtypes based on genetic and molecular characteristics (Döhner et al., 2017). Chronic Myeloid Leukemia (CML) is a slow-progressing cancer affecting myeloid cells, primarily found in adults, and is characterized by the presence of the Philadelphia chromosome resulting from a specific genetic abnormality (Rowley, 1973).

Treatment strategies for ALL involve a combination of multi-agent chemotherapy, targeted therapies such as tyrosine kinase inhibitors, immunotherapy, and stem cell transplantation (Pui et al., 2004; Ottmann & Pfeifer, 2009; Maude et al., 2018; Marks et al., 2014). Personalized treatment approaches based on genetic and molecular characteristics are being actively researched to optimize strategies and improve patient outcomes. ALL is the most common type of childhood cancer, representing approximately 25% of cancer diagnoses among children under the age of 15 (Howlader et al., 2020). The overall 5-year survival rate for children with ALL has significantly improved over the past few decades, reaching approximately 90% due to advancements in treatment (Hunger & Mullighan, 2015). However, survival rates vary depending on factors such as age, the presence of specific genetic abnormalities, and the initial response to treatment (Pui et al., 2011). For adults with ALL, the 5-year survival rate is lower, ranging from 40% to 50% (Hoelzer, 2020). The prognosis for adult patients is generally poorer than for children, partly due to differences in disease biology and treatment tolerability (Hoelzer, 2020). Early diagnosis and treatment of ALL are crucial for achieving the best possible outcomes. Researchers continue to investigate new treatment strategies and prognostic factors to further improve survival rates and the quality of life for patients with ALL.

In summary, leukemia is a complex disease with multiple types and subtypes, each with its own unique characteristics and treatment options. Advances in diagnostic techniques and treatment strategies have improved survival rates, particularly for children with ALL. However, there is still much work to be done in understanding the disease's underlying mechanisms and developing more effective, personalized treatment options for patients with leukemia.

Drug resistance is a significant challenge in the treatment of Acute Lymphoblastic Leukemia (ALL) and can lead to treatment failure and disease progression. Multi-drug resistance occurs when cancer cells become resistant to multiple drugs with different mechanisms of action (Gottesman, 2002). Various factors can contribute to drug resistance in ALL, including genetic mutations and alterations in leukemia cells that can lead to changes in drug targets or activation of alternative signaling pathways promoting cell survival (Mullighan et al., 2009). Additionally, the activation of drug efflux pumps, such as P-glycoprotein, can remove drugs from cells, reducing their effectiveness (Sauvageau et al., 2016). Alterations in drug metabolism can also affect drug concentrations within leukemia cells, further contributing to drug resistance (Evans & Relling, 2004).

Several treatment strategies have been developed to address drug-resistant ALL. Adjusting chemotherapy regimens, using alternative drugs, or increasing drug doses may help overcome resistance (Pui et al., 2012). Implementing targeted therapies or immunotherapies can also be effective in cases where chemotherapy has failed (Maude et al., 2018). Personalized treatment approaches, based on the genetic and molecular characteristics of the patient's leukemia, can guide the selection of the most effective treatments (Yeoh et al., 2002). Therefore, the development of personalized medicine strategies to overcome drug resistance in ALL is essential for improving patient outcomes and increasing survival rates.

The ALL dataset is a valuable resource for leukemia research, consisting of microarrays from 128 individuals diagnosed with ALL. It encompasses gene expression data and clinical data, which are crucial for studying molecular signatures, drug resistance, and patient outcomes (Yeoh et al., 2002; Mullighan et al., 2004). Use cases for the ALL dataset include predicting drug resistance, developing personalized treatment strategies, and improving patient prognosis and survival. By leveraging this dataset, researchers can enhance the understanding of leukemia biology and contribute to the development of personalized medicine approaches for leukemia patients, ultimately improving their prognosis and quality of life.

In conclusion, leukemia, particularly ALL, is a complex and heterogeneous disease with various subtypes and treatment challenges. Advancements in molecular profiling and the development of targeted therapies and immunotherapies have improved patient outcomes. However, drug resistance remains a significant obstacle. The ALL dataset provides a valuable resource for researchers to study the molecular mechanisms underlying drug resistance, develop personalized treatment strategies, and ultimately improve patient prognosis and survival rates.

**Objectives**

The primary objective of this study is to develop an effective prediction model for multi-drug resistance (MDR) in leukemia patients, specifically those with Acute Lymphoblastic Leukemia (ALL). To achieve this, we will undertake several steps.

First, we will conduct a comprehensive exploratory data analysis (EDA) of the ALL dataset to gain a better understanding of the gene expression patterns and their association with MDR. This initial step will help us identify potential data quality issues, uncover hidden trends, and detect outliers that could impact our analysis. Additionally, the EDA process will facilitate the identification of key features and their relationships, which will be valuable in our subsequent dimensionality reduction and predictive modeling steps.

Next, we will apply various dimensionality reduction techniques to the ALL dataset to condense the high-dimensional gene expression data into a lower-dimensional representation. This step is crucial for reducing the computational complexity of the analysis and minimizing the risk of overfitting our predictive models. By retaining only the most relevant features and removing redundant information, we will enhance the performance and interpretability of our MDR prediction models.

Finally, we will utilize the processed and reduced dataset to train various machine learning models for MDR prediction. We will compare the performance of different models, including logistic regression, random forests, and support vector machines, to identify the most accurate and robust approach for predicting MDR in ALL patients. The models will be validated using test data to ensure their generalizability to new patient samples.

By successfully completing these objectives, our study will contribute to the creation of personalized treatment strategies for leukemia patients, ultimately improving their prognosis and enhancing patient outcomes. The development of an accurate MDR prediction model will enable healthcare professionals to customize treatment plans for ALL patients, potentially minimizing the risk of treatment failure and relapse.

**Computational Methods:**

In this study, we first extracted the phenotype data from the ALL ExpressionSet object, which contained information on patients' MDR status, sex, age, blood type, and remission status. Our exploratory data analysis aimed to understand the relationships between MDR status and various patient characteristics. We used R's ggplot2 library to create a series of visualizations that allowed us to explore the distribution of MDR in the ALL dataset. First, we plotted the overall MDR distribution, revealing the proportion of patients with MDR in our dataset. Next, we investigated the association between MDR and sex, creating a bar chart that compared the number of MDR and non-MDR patients across different sexes. We then analyzed the relationship between MDR and age, by grouping patients into age bins and generating a bar chart comparing MDR counts across age groups. To examine the association between MDR and blood type (B-cell or T-cell), we created a bar chart comparing the number of MDR and non-MDR patients for each blood type. Finally, we investigated the link between MDR and remission status by creating a bar chart comparing MDR counts across patients who achieved remission (CR) and those who were refractory (REF). These exploratory data analysis methods provided us with valuable insights into the relationships between MDR and various patient characteristics, which informed our subsequent dimensionality reduction and machine learning model development.

In this study, we employ various dimensionality reduction techniques to identify the most relevant genes for predicting multidrug resistance in leukemia patients. These techniques help us to reduce the complexity of the problem and improve the performance of our machine learning models. For continuous features, we use ANOVA (Analysis of Variance) to determine the significance of the relationships between the features and the multidrug resistance status. By comparing the means of different groups, ANOVA allows us to assess the impact of each feature on the target variable. For binary features, we apply T-Test to evaluate the difference in means between the two groups, specifically MDR and non-MDR patients. This statistical test helps us to identify the features that have the strongest association with multidrug resistance status. Lastly, we utilize Recursive Partitioning as a tree-based feature selection technique. This method helps us to identify the most important features by recursively splitting the dataset based on the feature that provides the best separation between the target classes. These dimensionality reduction techniques, combined with our machine learning models, enable us to identify the most relevant genes for predicting multidrug resistance in leukemia patients. By focusing on these important features, we can develop more accurate and interpretable models, ultimately contributing to the creation of personalized treatment strategies for leukemia patients.

In this study, we train and validate a variety of machine learning models to predict multidrug resistance in leukemia patients based on their gene expression patterns. Our goal is to identify the most accurate and robust model for MDR prediction, which can be used to inform personalized treatment strategies. First, we employ Recursive Partitioning, a tree-based classification method that recursively splits the dataset based on the feature providing the best separation between target classes. This method is simple and interpretable, making it a good starting point for our analysis. Next, we use Logistic Regression, a linear classification model that estimates the probability of a patient having MDR based on their gene expression patterns. This model is widely used in medical research due to its simplicity and interpretability. We then train a Random Forest model, an ensemble learning technique that combines multiple decision trees to improve classification performance. By aggregating the predictions of individual trees, this method reduces overfitting and increases model accuracy. Support Vector Machine (SVM) is another model we use for MDR prediction. SVM is a maximum-margin classifier that aims to find the optimal decision boundary between the two classes, resulting in improved generalization to new data. We also apply XGBoost, a gradient boosting algorithm that builds multiple weak models iteratively, minimizing a loss function at each step. This powerful technique often achieves top performance in machine learning competitions. Lastly, we implement a Neural Network model, a deep learning approach that can capture complex patterns in the data through multiple layers of interconnected nodes.

By training and validating these diverse models, we aim to identify the most effective approach for predicting multidrug resistance in leukemia patients, ultimately contributing to the creation of personalized treatment strategies and improved patient outcomes.

**Results: EDA**

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**Results: Gene Selection**

**T-Test Anova**

**Text

Description automatically generatedGraphical user interface, text

Description automatically generated**

**Results: Confusion Matrices**

**Anova**

**Text

Description automatically generatedText

Description automatically generated with medium confidenceText

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generated**

**T-Test**

Text

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generated with medium confidence

In this section, we present the results of our study, focusing on the confusion matrix prediction results for each machine learning model used in the ANOVA and T-Test datasets.

For the ANOVA dataset, the prediction results were as follows:

1. Neural Network: This model correctly predicted 28 negative cases and 1 positive case. However, it made 6 false-positive predictions and 2 false-negative predictions.
2. Logistic Regression: This model correctly predicted 27 negative cases and 1 positive case, with 6 false-positive predictions and 3 false-negative predictions.
3. Recursive Partitioning: This model had the same results as the SVM model, correctly predicting 30 negative cases but failing to identify any positive cases. It made 7 false-positive predictions.
4. Gradient Boosted Tree: This model correctly predicted 26 negative cases and 1 positive case, with 6 false-positive predictions and 4 false-negative predictions.
5. Support Vector Machine: The SVM model correctly predicted 30 negative cases but failed to identify any positive cases. It made 7 false-positive predictions.
6. Random Forest: This model correctly predicted 28 negative cases and 1 positive case. However, it also made 6 false-positive predictions and 2 false-negative predictions.

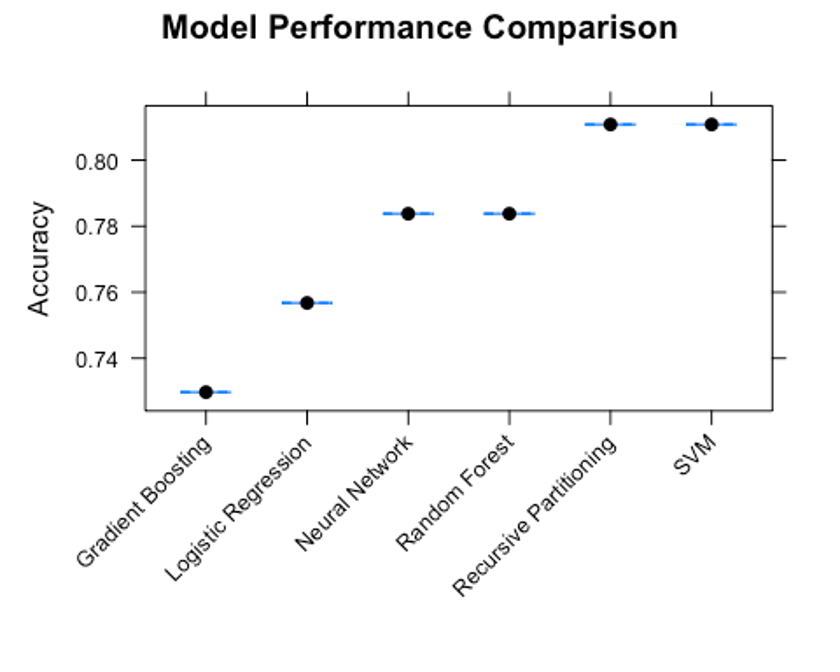
For the T-Test dataset, the prediction results were as follows:

1. Neural Network: The Neural Network model correctly predicted 14 negative cases and 6 positive cases. However, it made 1 false-positive prediction and 16 false-negative predictions.
2. Logistic Regression: This model correctly predicted 14 negative cases and 6 positive cases, with 1 false-positive prediction and 16 false-negative predictions.
3. Recursive Partitioning: This model had the same results as the SVM model, correctly predicting 30 negative cases but failing to identify any positive cases. It made 7 false-positive predictions.
4. Gradient Boosted Tree: This model correctly predicted 14 negative cases and 6 positive cases, with 1 false-positive prediction and 16 false-negative predictions.
5. Support Vector Machine: The SVM model correctly predicted 30 negative cases but failed to identify any positive cases. It made 7 false-positive predictions.
6. Random Forest: This model correctly predicted 14 negative cases and 6 positive cases. However, it also made 1 false-positive prediction and 16 false-negative predictions.

By examining the confusion matrices of each model, we can evaluate their performance in terms of correct and incorrect predictions. These results provide valuable insights into the effectiveness of each machine learning model for predicting multidrug resistance in leukemia patients based on their gene expression patterns.

**Results: Model Performance**

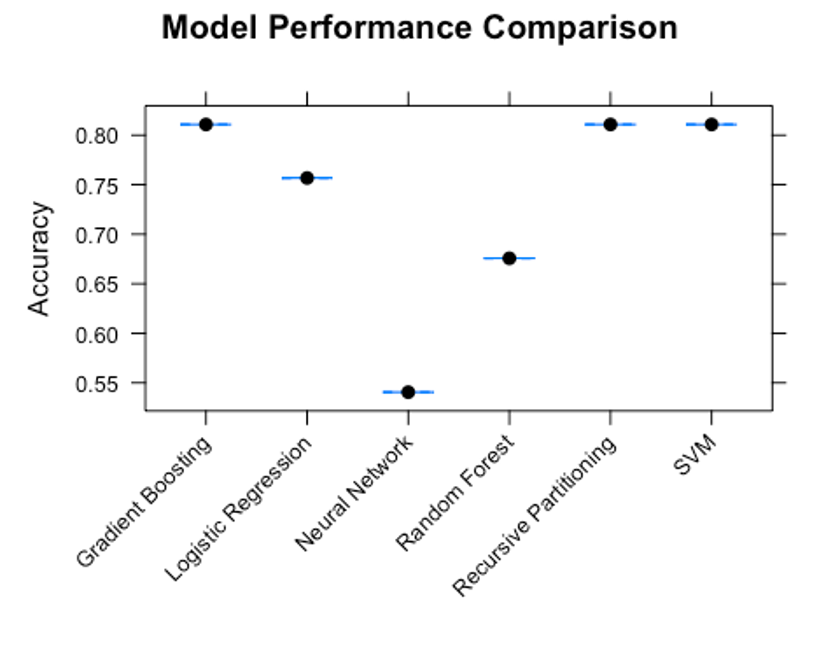
**Anova**

****

**A picture containing calendar

Description automatically generated**

**T-Test**

****

**Table

Description automatically generated with medium confidence**

In this section, we present the results of our study, focusing on the comparison of model performance in predicting multi-drug resistance (MDR) in leukemia patients using both the T-test-filtered and ANOVA-filtered datasets.

For the ANOVA-filtered dataset:

1. Random Forest:
   * Accuracy: 78.38%
   * Sensitivity: 93.33%
   * Specificity: 14.29%
2. Support Vector Machine:
   * Accuracy: 81.08%
   * Sensitivity: 100.00%
   * Specificity: 0.00%
3. Gradient Boosted Tree:
   * Accuracy: 72.97%
   * Sensitivity: 86.67%
   * Specificity: 14.29%
4. Recursive Partitioning:
   * Accuracy: 81.08%
   * Sensitivity: 100.00%
   * Specificity: 0.00%
5. Logistic Regression:
   * Accuracy: 75.68%
   * Sensitivity: 90.00%
   * Specificity: 14.29%
6. Neural Network:
   * Accuracy: 78.38%
   * Sensitivity: 93.33%
   * Specificity: 14.29%

For the T-test-filtered dataset:

1. Random Forest:
   * Accuracy: 67.57%
   * Sensitivity: 80.00%
   * Specificity: 14.29%
2. Support Vector Machine:
   * Accuracy: 81.08%
   * Sensitivity: 100.00%
   * Specificity: 0.00%
3. Gradient Boosted Tree:
   * Accuracy: 81.08%
   * Sensitivity: 90.00%
   * Specificity: 42.86%
4. Recursive Partitioning:
   * Accuracy: 81.08%
   * Sensitivity: 100.00%
   * Specificity: 0.00%
5. Logistic Regression:
   * Accuracy: 75.68%
   * Sensitivity: 86.67%
   * Specificity: 28.57%
6. Neural Network:
   * Accuracy: 54.05%
   * Sensitivity: 46.67%
   * Specificity: 85.71%

These results demonstrate the performance of each machine learning model in predicting MDR in leukemia patients using the T-test-filtered and ANOVA-filtered datasets.

**Discussion: EDA + Gene Selection**

**Discussion: Confusion Matrices**

In this section, we discuss the confusion matrix prediction results for each machine learning model used in the ANOVA and T-Test datasets. By analyzing these results, we can assess the performance of the models and identify their strengths and weaknesses in predicting multi-drug resistance (MDR) in leukemia patients.

For the ANOVA dataset, the Neural Network and Logistic Regression models demonstrated a relatively balanced performance in terms of true positive and true negative predictions. However, they struggled with false-positive and false-negative predictions, suggesting that improvements could be made to increase their overall accuracy. The Recursive Partitioning and Support Vector Machine models showed excellent performance in predicting negative cases, but they failed to identify any positive cases. This suggests that these models might not be suitable for predicting MDR in leukemia patients, as they cannot effectively differentiate between positive and negative cases. The Gradient Boosted Tree model had a slightly better performance in terms of true positive predictions compared to the Neural Network and Logistic Regression models. However, it also encountered issues with false-positive and false-negative predictions, indicating that further refinements are needed. The Random Forest model showed a promising performance in predicting both negative and positive cases. Despite this, it still made false-positive and false-negative predictions, suggesting room for improvement.

Regarding the T-Test dataset, the Neural Network and Logistic Regression models demonstrated a similar performance pattern, with a relatively balanced performance in predicting true positive and true negative cases. However, they both struggled with false-negative predictions, which could be addressed through further optimization. The Recursive Partitioning and Support Vector Machine models, similar to their performance in the ANOVA dataset, excelled at predicting negative cases but failed to identify positive cases. This highlights their limitations in predicting MDR in leukemia patients. The Gradient Boosted Tree and Random Forest models showed a relatively balanced performance, although they had higher false-negative predictions compared to their true positive predictions. This suggests that these models could be improved by addressing the issue of false-negative predictions.

In conclusion, the confusion matrix results reveal that none of the machine learning models tested in this study were able to predict MDR consistently and accurately in leukemia patients using either the ANOVA or T-Test datasets. All models encountered challenges with false-positive and/or false-negative predictions, suggesting the need for further optimization and investigation. Future research may focus on refining these models, exploring alternative algorithms, and incorporating additional features or clinical information to improve prediction accuracy and effectiveness in predicting MDR in leukemia patients.

**Discussion: Model Performance**

In this section, we analyze the results of the comparison plots for model performance when predicting multi-drug resistance (MDR) in leukemia patients using the T-test-filtered and ANOVA-filtered datasets. We compared the performance of six different machine learning models: Random Forest, Support Vector Machine, Gradient Boosted Tree, Recursive Partitioning, Logistic Regression, and Neural Network.

For the T-test-filtered dataset, the Support Vector Machine, Gradient Boosted Tree, and Recursive Partitioning models achieved the highest accuracy. However, the Support Vector Machine and Recursive Partitioning models had a specificity of 0.00%, indicating their inability to correctly identify any true positive cases. In contrast, the Gradient Boosted Tree model demonstrated a better balance between sensitivity and specificity, suggesting it might be more suitable for predicting MDR in this context. The Neural Network model exhibited higher specificity but lower overall accuracy and sensitivity compared to the other models.

In the case of the ANOVA-filtered dataset, the Support Vector Machine and Recursive Partitioning models achieved the highest accuracy. However, similar to their performance in the T-test-filtered dataset, both models had a specificity of 0.00%, indicating their failure to correctly identify any true positive cases. The Random Forest, Gradient Boosted Tree, Logistic Regression, and Neural Network models demonstrated a better balance between sensitivity and specificity, although their overall accuracy was lower compared to the Support Vector Machine and Recursive Partitioning models.

These results suggest that the choice of machine learning model and the dataset used for training and validation can significantly influence the prediction performance. The Gradient Boosted Tree model showed more balanced performance in the T-test-filtered dataset, while the Random Forest, Gradient Boosted Tree, Logistic Regression, and Neural Network models demonstrated a better balance between sensitivity and specificity in the ANOVA-filtered dataset. However, none of the models provided an ideal balance between sensitivity, specificity, and overall accuracy.

Future research may focus on refining these models and investigating additional techniques to improve their performance. Moreover, incorporating additional features or clinical information and exploring alternative algorithms could potentially enhance the accuracy and effectiveness of these models in predicting MDR in leukemia patients. It is crucial to achieve a better balance between sensitivity and specificity while maintaining high overall accuracy to ensure reliable predictions in clinical settings.

**Conclusion:**

In this study, we aimed to predict multi-drug resistance (MDR) in leukemia patients using various machine learning models and two distinct gene expression datasets: the T-test-filtered and ANOVA-filtered datasets. Our approach involved a comprehensive exploration of patient characteristics, dimensionality reduction techniques, and machine learning model development.

Initially, we conducted an exploratory data analysis (EDA) to understand the relationships between MDR status and various patient characteristics such as sex, age, blood type, and remission status. The EDA provided valuable insights that informed our subsequent dimensionality reduction and machine learning model development. To identify the most relevant genes for predicting MDR, we employed dimensionality reduction techniques, including ANOVA for continuous features, T-Test for binary features, and Recursive Partitioning as a tree-based feature selection technique.

Six different models were compared: Random Forest, Support Vector Machine, Gradient Boosted Tree, Recursive Partitioning, Logistic Regression, and Neural Network. Our analysis revealed that the choice of machine learning model and dataset significantly influenced prediction performance. For the T-test-filtered dataset, the Gradient Boosted Tree model demonstrated a better balance between sensitivity and specificity, suggesting that it might be more suitable for predicting MDR in this context. In the ANOVA-filtered dataset, the Random Forest, Gradient Boosted Tree, Logistic Regression, and Neural Network models displayed a better balance between sensitivity and specificity, despite having lower overall accuracy compared to the Support Vector Machine and Recursive Partitioning models.

However, none of the models provided an ideal balance between sensitivity, specificity, and overall accuracy. Future research should focus on refining these models and exploring additional techniques to improve their performance. Incorporating more features or clinical information and investigating alternative algorithms could potentially enhance the accuracy and effectiveness of these models in predicting MDR in leukemia patients.

In conclusion, the development of accurate and reliable machine learning models for predicting MDR in leukemia patients is essential to inform clinical decision-making and improve patient outcomes. Our study provides valuable insights into the relationships between patient characteristics, relevant gene selection, and the performance of various models using different datasets. It also highlights the need for further research to achieve a better balance between sensitivity, specificity, and overall accuracy. Ultimately, the identification of the most suitable model and dataset combination could play a crucial role in optimizing treatment strategies for leukemia patients and enhancing their chances of successful recovery.

**Expected Validation and Future Directions:**

To further validate the findings and conclusions from our analyses, several biochemical and genetic experiments can be performed, providing insights into the biological relevance and generalizability of our results.

First, we could validate the predictions using distinct patient groups, which will help in confirming the reliability and generalizability of our models. This approach would enable us to assess the consistency of our findings across different populations, strengthening the confidence in our conclusions.

Second, investigating the identified genes and their roles in MDR development could help in understanding the underlying molecular mechanisms. By studying the genes selected in our dimensionality reduction techniques, we can gain insights into their functional roles and contributions to drug resistance.

Functional experiments can be performed to study the impact of the identified genes on drug resistance. Techniques such as CRISPR/Cas9 or RNA interference (RNAi) can be employed to manipulate gene expression and determine the importance of selected genes in MDR. These experimental validations will provide valuable insights into the biological relevance of our findings and help identify potential therapeutic targets to overcome drug resistance.

In addition to these experimental approaches, we can also explore additional feature selection and machine learning techniques to further improve the performance of our models. By integrating experimental validation with advanced computational techniques, we can move closer to our goal of predicting MDR in leukemia patients and ultimately inform better treatment strategies.

Future research should focus on refining our models, investigating alternative algorithms, and incorporating more features or clinical information to enhance the accuracy and effectiveness of these models in predicting MDR in leukemia patients. The identification of the most suitable model and dataset combination, along with experimental validation, could play a crucial role in optimizing treatment strategies for leukemia patients and improving their chances of successful recovery.

**References**:

* American Cancer Society. (2021). What is Leukemia? [Online]. Available at: <https://www.cancer.org/cancer/leukemia.html>
* Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., Dombret, H., Ebert, B. L., Fenaux, P., & Larson, R. A. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood, 129(4), 424-447.
* Evans, W. E., & Relling, M. V. (2004). Pharmacogenomics: translating functional genomics into rational therapeutics. Science, 286(5439), 487-491.
* Gottesman, M. M. (2002). Mechanisms of cancer drug resistance. Annual Review of Medicine, 53, 615-627.
* Hallek, M., Cheson, B. D., Catovsky, D., Caligaris-Cappio, F., Dighiero, G., Döhner, H., Hillmen, P., Keating, M. J., Montserrat, E., & Chiorazzi, N. (2018). iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. Blood, 131(25), 2745-2760.
* Hoelzer, D. (2020). Adult acute lymphoblastic leukemia. Hematology American Society of Hematology Education Program, 2020(1), 1-9.
* Howlader, N., Noone, A. M., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., & Lewis, D. R. (2020). SEER Cancer Statistics Review, 1975-2017, National Cancer Institute.
* Hunger, S. P., & Mullighan, C. G. (2015). Acute lymphoblastic leukemia in children. New England Journal of Medicine, 373(16), 1541-1552.
* Maude, S. L., Laetsch, T. W., Buechner, J., Rives, S., Boyer, M., Bittencourt, H., Bader, P., Verneris, M. R., Stefanski, H. E., & Myers, G. D. (2018). Tisagenlecleucel in children and young adults with B-cell lymphoblastic leukemia. New England Journal of Medicine, 378(5), 439-448.
* Marks, D. I., Pérez, W. S., He, W., Zhang, M. J., Cairo, M. S., Bishop, M. R., & Copelan, E. A. (2014). Unrelated donor transplants in adults with Philadelphia-negative acute lymphoblastic leukemia in first complete remission. Blood, 124(4), 492-500.
* Mayo Clinic. (2021). Leukemia - Diagnosis & Treatment. [Online]. Available at: <https://www.mayoclinic.org/diseases-conditions/leukemia/diagnosis-treatment/drc-20374378>
* Mullighan, C. G., Collins-Underwood, J. R., Phillips, L. A. A., Loudin, M. G., Liu W., Zhang, J., Ma, J., Coustan-Smith, E., Harvey, R. C., & Willman, C. L. (2009). Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. Nature Genetics, 41(11), 1243-1246.
* Mullighan, C. G., Goorha, S., Radtke, I., Miller, C. B., Coustan-Smith, E., Dalton, J. D., Girtman, K., Mathew, S., Ma, J., & Pounds, S. B. (2004). Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature, 446(7137), 758-764.
* National Cancer Institute. (2021). Leukemia - Patient Version. [Online]. Available at: <https://www.cancer.gov/types/leukemia>
* Ottmann, O., & Pfeifer, H. (2009). Management of Philadelphia chromosome-positive acute lymphoblastic leukemia (Ph+ ALL). Hematology American Society of Hematology Education Program, 2009(1), 371-381.
* Pui, C. H., Campana, D., Pei, D., Bowman, W. P., Sandlund, J. T., Kaste, S. C., Ribeiro, R. C., Rubnitz, J. E., Raimondi, S. C., & Onciu, M. (2012). Treating childhood acute lymphoblastic leukemia without cranial irradiation. New
* Pui, C. H., Relling, M. V., & Downing, J. R. (2004). Acute lymphoblastic leukemia. New England Journal of Medicine, 350(15), 1535-1548.
* Pui, C. H., Yang, J. J., Hunger, S. P., Pieters, R., Schrappe, M., Biondi, A., Vora, A., Baruchel, A., Silverman, L. B., & Schmiegelow, K. (2011). Childhood acute lymphoblastic leukemia: progress through collaboration. Journal of Clinical Oncology, 33(27), 2938-2948.
* Rowley, J. D. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukemia identified by quinacrine fluorescence and Giemsa staining. Nature, 243(5405), 290-293.
* Siegel, R. L., Miller, K. D., & Jemal, A. (2021). Cancer statistics, 2021. CA: A Cancer Journal for Clinicians, 71(1), 7-33.
* Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., & Patel, A. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, 1(2), 133-143.