ENHANCING LEUKEMIA DETECTION WITH MACHINE LEARNING AND STATISTICAL ANALYSIS

Dr.K.S.Karunya [1], Deepaneesh R V [2] , Sanjay Saravanan P [3]
Assistant Professor[1], PG Students[2,3]
Department of Statistics, PSG College of Arts & Science,
Coimbatore, Tamil Nadu, India
karunyaks2013@gmail.com[1] , deepaneesh98@gmail.com[2] , sanjaysp026@gmail.com[3]

ABSTRACT

Leukemia is a type of blood cancer that affects both the blood and bone marrow, and it remains a significant health issue around the globe. Catching it early is really important because it can greatly improve survival rates. In this study a data-driven approach used to predicting leukemia status using a mix of statistical and machine learning techniques. Different machine learning algorithms have been applied to create predictive models and evaluate their performance using standard measures. Feature selection techniques were also utilized to pinpoint the most impactful predictors, which not only makes the models easier to understand but also helps streamline the analysis. By comparing several methods, the aim is to identify the best way to classify and predict leukemia. The findings shed light on how various clinical and lifestyle factors influence leukemia outcomes, showcasing the promise of machine learning in enhancing disease prediction. This study demonstrates the power of machine learning in improving leukemia prediction, enabling early better patient outcomes. It emphasizes the potential of advanced models in medical diagnostics.

Keywords: Leukemia, Machine Learning Algorithms, Prediction Models, Accuracy Evaluation, Model Performance

INTRODUCTION

Leukemia is a cancer that affects the blood and bone marrow, caused by the uncontrolled production of abnormal white blood cells. These cells multiply excessively, crowding out healthy blood cells necessary for proper bodily function. This global health issue impacts all ages, with certain types more common in specific age groups. Early detection and accurate diagnosis are crucial for effective treatment and better recovery. Advances in medical research have incorporated machine learning techniques in leukemia diagnosis, prognosis, and treatment planning, enhancing accuracy and efficiency.

TYPES OF LEUKEMIA

Leukemia is categorized into two main types: acute (rapid progression) and chronic (slow progression), based on the kind of blood cells involved. The four main types are Acute Lymphocytic Leukemia (ALL), Acute Myelogenous Leukemia (AML), Chronic Lymphocytic Leukemia (CLL), and Chronic Myelogenous Leukemia (CML).

SYMPTOMS OF LEUKEMIA

Symptoms vary but can include fatigue, frequent infections, easy bruising or bleeding, bone pain, swollen lymph nodes, enlarged liver or spleen, night sweats, weight loss, and fever. It's crucial to seek medical advice if these symptoms persist.

DIAGNOSTIC TESTS FOR LEUKEMIA

Diagnosis typically involves a Complete Blood Count (CBC), blood smear, and bone marrow biopsy. Additional tests may identify chromosomal abnormalities and gene mutations and use imaging to assess the leukemia's extent.

TREATMENT OPTIONS FOR LEUKEMIA

Treatment varies based on leukemia type, patient health, and stage. Chemotherapy is the primary treatment, supplemented by radiation, targeted therapy, immunotherapy, and stem cell transplantation in high-risk cases.

PROGNOSIS AND RISK FACTORS

Prognosis varies depending on leukemia type, age, health, and treatment response. Acute forms can be aggressive and potentially curable, while chronic forms progress slowly. Risk factors include genetic conditions, chemical exposure, previous radiation, smoking, and family history**.**

MACHINE LEARNING APPLICATION & TECHNIQUES

Predictive modeling in leukemia research uses machine learning to analyze patient data, improving diagnosis, prognosis, and treatment responses. By examining blood test results, genetic markers, and symptoms, these models facilitate early detection and assess disease progression. They also refine predictions using patient histories and identify those at higher risk for complications, enabling timely interventions. Overall, this approach enhances leukemia diagnosis, treatment optimization, and patient care.

In this paper, machine learning and statistical analysis methods are used to predict the leukemia status in a patient.

REVIEW OF LITERATURE

The article **"Blinatumomab Boosts Chemotherapy as Initial Treatment for Some Kids with ALL" by Carmen Phillips (2025)** highlights that adding blinatumomab to standard chemotherapy significantly improves disease-free survival in children with standard-risk B-cell acute lymphoblastic leukemia (ALL). A study with over 1,400 participants found a 96% disease-free survival rate after 2.5 years with both treatments, versus 88% with chemotherapy alone. With minimal side effects and FDA approval in June 2024, blinatumomab may become a new standard treatment, though its 28-day infusion may limit accessibility.

The study "**Leukemia Incidence Trends (1990-2017)" by Ying Dong et al. (2020) shows a rise in total leukemia cases from 354,500 to 518,500**, despite a 0.43% annual decline in the age-standardized incidence rate (ASIR). Acute lymphoblastic leukemia (ALL) cases increased, chronic lymphocytic leukemia (CLL) cases more than doubled, and acute myeloid leukemia (AML) cases rose, while chronic myeloid leukemia (CML) ASIR fell. These trends signal significant public health concerns.

Chennamadhavuni et al. (2023) describe leukemia as a blood cancer characterized by abnormal white blood cell production, classified into acute and chronic forms. Chemotherapy is the main treatment, and an interprofessional healthcare team is crucial for improving outcomes.

Davis et al. (2014) explain leukemia involves abnormal growth of blood stem cells, with ALL being more common in children. Risk factors include genetics and ionizing radiation, and symptoms are nonspecific. Diagnosis involves blood tests and bone marrow examination, and treatments focus on chemotherapy and stem cell transplantation, with long-term monitoring for complications. Survival rates are highest in younger patients and those with CML or CLL.

A Statistical Study of Mortality from Leukemia by Sacks and Seeman (1947) examines leukemia death rates from 1900 to 1944, revealing an increase from 1.9 per 100,000 in 1920 to 3.7 in 1940. Higher rates were seen among white males over 55. Factors like better diagnostic methods and more hospital resources contributed to this trend. The study notes classification challenges and underreporting, concluding that U.S. leukemia mortality trends are consistent with those in other countries, highlighting the need for further research.

SCOPE OF THE STUDY

This study focuses on predicting leukemia in patients using machine learning techniques. It analyzes the impact of medical and demographic factors on leukemia diagnosis. The research explores key risk indicators and their contribution to disease prediction. It also highlights the challenges posed by data limitations, such as selection bias and the absence of healthy individuals in the dataset. Lastly, the study aims to provide insights for improving early detection and guiding future research in leukemia prediction models.

DATA SOURCE

The dataset for this study was sourced from an open-source repository on GitHub. It was refined to include only high-quality responses, resulting in 402 records. Non-essential variables, such as 'Country' and 'Patient ID,' were removed to focus on relevant medical and demographic factors for leukemia prediction. The data was also filtered by age categories to target the most relevant patient groups. This refined dataset serves as the foundation for the machine learning models and statistical analysis

OBJECTIVES

- To develop highly optimized statistical and machine learning models by fine-tuning hyperparameters through a looping method for accurate leukemia prediction.
- To identify the most effective model by evaluating accuracy and determining the one with the highest predictive performance.
- To determine key variables influencing leukemia status by analyzing feature importance or model coefficients from the best-selected model..

RESEARCH METHODOLOGY

METHODS FOR BINARY CLASSIFICATION

This study employed various machine learning and statistical models for binary classification. Optimal hyperparameters were selected through a looping method, with the best model determined based on accuracy during tuning and final selection.

LOGISTIC REGRESSION

Logistic Regression is a statistical method for binary classification that estimates the probability of an outcome belonging to a class. It transforms a linear combination of input features to output values between 0 and 1. This simple yet effective model, commonly used as a baseline, is optimized using Maximum Likelihood Estimation to minimize binary cross-entropy loss and enhance predictive accuracy.

XGBOOST (EXTREME GRADIENT BOOSTING)

XGBoost is a gradient boosting algorithm that builds decision trees sequentially, with each tree correcting errors from previous ones. It uses regularization and second-order gradient approximation to improve efficiency and handles missing values well. Hyperparameters such

as learning rate, tree depth, and number of estimators were optimized using a looping method to select the most accurate model.

CATBOOST (CATEGORICAL BOOSTING)

CatBoost is designed for efficiently handling categorical data. It utilizes ordered boosting and target-based encoding to reduce overfitting and works well with imbalanced datasets, eliminating extensive preprocessing. Key hyperparameters were optimized iteratively, focusing on accuracy.

RANDOM FOREST

Random Forest combines multiple decision trees to enhance classification performance and reduce overfitting. Each tree is trained on a different data subset, with final predictions made through majority voting. This method is effective for high-dimensional data, and hyperparameters like the number of trees and tree depth were optimized for best accuracy.

GLMNET (GENERALIZED LINEAR MODEL WITH REGULARIZATION)

GLMNET incorporates Lasso (L1) and Ridge (L2) penalties to improve feature selection and prevent overfitting. Lasso automatically selects features by shrinking coefficients, while Ridge stabilizes predictions by distributing weights. The model was fine-tuned iteratively based on regularization strength and mixing parameters, again focusing on accuracy.

All models were fine-tuned to achieve optimal performance, with the final model selected based on accuracy during tuning and evaluation.

ANALYSIS AND INTERPRETAION

Following the analysis, five machine learning models were developed using the optimal hyperparameters to achieve the best accuracy. These models were evaluated based on their training accuracy, test accuracy, and kappa values to assess their reliability and generalizability. The performance comparison allows us to identify well-fitted models while also addressing potential overfitting concerns. A summary of the model performances is provided in the table below:

	Training		Testing		
Model	**Accuracy**	**Kappa value**	**Accuracy**	**Kappa value**	**Model type**
Logistic Regression	0.879	0.879	0.843	0.6477	Good Fitted
CatBoost	1	1	0.8099	0.5686	Over Fitted
XG Boost	0.8577	0.6804	0.8347	0.6312	Good Fitted
Random Forest	1	1	0.8099	0.5783	Over Fitted
Glmnet	0.879	0.7284	0.8512	0.6643	Good Fitted

Table 1

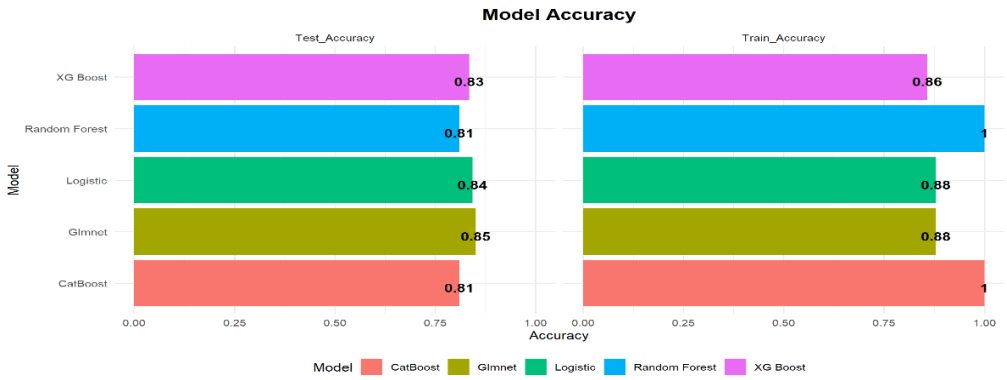

Figure 1

From the above (Table 1 and Figure 1), compares the performance of different machine learning models in predicting leukemia presence. It includes training and testing accuracy, kappa values (a measure of agreement), and an assessment of model fitting.

Logistic Regression achieved a training accuracy of 87.90% and a testing accuracy of 84.30%, with a kappa value of 0.6477, indicating good generalization. CatBoost and Random Forest obtained perfect training accuracy (100%) but had lower testing accuracy at 80.99%, which suggests they may be overfitting the data. XGBoost and Glmnet demonstrated slightly lower testing accuracy compared to Logistic Regression but still performed well overall. Given its highest testing accuracy of 84.30%, Logistic Regression is considered the best model, as it effectively balances accuracy and generalization.

LOGISTIC REGRESSION

CONFUSION MATRIX FOR TRAINING SET

		Reference	
		Negative	**Positive**
Prediction	**Negative**	168	18
	Positive	16	79

Table 2

Accuracy	0.8790
95% CI (Lower)	0.8350
95% CI (Upper)	0.9147
No Information Rate	0.6548
P-Value (Acc > NIR)	0.0000
Kappa	0.7310

Table 3

This confusion matrix shows the training performance of Logistic Regression

Accuracy: 87.90% (indicating a well-performing model). Kappa: 0.7310 (reflecting substantial agreement). Confidence Interval: 83.50% - 91.47% (demonstrating consistent accuracy). P-Value: 0.0000 (indicating statistical significance). Logistic Regression effectively distinguishes between positive and negative cases.

COEFFICIENT OF LOGISTIC REGRESSION MODEL

Variable	Estimate	Std. Error	z value	Pr(>|z|)
Age	-0.0032	0.0206	-0.1560	0.8759
Gender-Female	-11.9300	3.1646	-3.7700	0.0000***
Gender-Male	-12.1174	3.1921	-3.7960	0.0000***
WBC_Count	0.0005	0.0001	4.5140	0.0000***
RBC_Count	0.7706	0.3191	2.4150	0.0157*
Platelet_Count	-0.0007	0.0027	-0.2500	0.8024
Hemoglobin_Level	-0.2485	0.1301	-1.9100	0.0561
Bone_Marrow_Blasts	0.6445	0.1136	5.6740	0.0000***
Genetic_Mutation-Present	3.6621	0.5820	6.2920	0.0000***
Family_History-yes	3.0620	0.5448	5.6200	0.0000***
Smoking_Status-Smoker	1.3663	0.4839	2.8240	0.0047**
Alcohol_Consumption-yes	-0.6764	0.4813	-1.4060	0.1598
Radiation_Exposure-yes	2.0859	0.5924	3.5210	0.0004***
Infection_History--yes	0.5496	0.4695	1.1700	0.2418
BMI	0.0004	0.0521	0.0080	0.9939
Chronic_Illness-yes	-0.1834	0.4655	-0.3940	0.6936
Immune_Disorders-yes	0.0246	0.5291	0.0460	0.9630
Socioeconomic_Status-Low	-0.6729	0.6307	-1.0670	0.2860
Socioeconomic_Status-Medium	-1.4117	0.6492	-2.1740	0.0297
Urban_Rural-Urban	0.1174	0.4969	0.2360	0.8132

Table 4

Higher white blood cell (WBC) counts (+0.0005, $p < 0.0001$) and the presence of bone marrow blasts (+0.6445, $p < 0.0001$) significantly increase leukemia risk. Genetic mutations (+3.6621, $p < 0.0001$) and family history (+3.0620, $p < 0.0001$) are also highly significant risk factors. Additionally, smoking (+1.3663, $p = 0.0047$) and radiation exposure (+2.0859, $p = 0.0004$) contribute to this risk. In contrast, a medium socioeconomic status (-1.4117, $p = 0.0297$) is associated with a lower risk of leukemia. Other factors such as age, body mass index (BMI), platelet count, hemoglobin, infection history, alcohol consumption, chronic illness, immune disorders, and urban-rural status do not show statistical significance. Overall, key risk factors for leukemia include genetic factors, bone marrow condition, smoking, radiation exposure, and family history.

CONFUSION MATRIX FOR TESTING SET

		Reference	
		Negative	**Positive**
Prediction	**Negative**	71	11
	Positive	8	31

Table 5

Accuracy	0.8430
95% CI (Lower)	0.7657
95% CI (Upper)	0.9027
No Information Rate	0.6529
P-Value (Acc > NIR)	0.0000
Kappa	0.6477

Table 6

The confusion matrix reveals that the model accurately predicted 71 negatives and 31 positives, misclassifying 8 positives and 11 negatives. With an accuracy of 84.30% and a confidence interval of 76.57% to 90.27%, the model shows reliable performance. The No Information Rate (NIR) is 65.29%, indicating the model is significantly better than random guessing. The p-value is 0.0000, confirming statistical significance, and the Kappa value is 0.6477, indicating moderate to substantial agreement. Overall, the Logistic Regression model performs well in predicting leukemia.

Subsequently, GLMNET is applied to select important variables and reduce overfitting. By combining Lasso (L1) and Ridge (L2) regularization, GLMNET eliminates irrelevant features and shrinks coefficient values, enhancing the model's generalizability for unseen data.

GENERALIZED LINEAR MODEL WITH REGULARIZATION

glmnet(x = x_train, y = y_train)
family = "binomial"
alpha = 0.1
lambda = 0.01303

Table 7

Using a looping method like cross-validation or grid search, optimal values for alpha and lambda are determined to enhance model performance. This involves testing combinations of alpha (balance between Lasso and Ridge) and lambda (regularization strength) to reduce overfitting while maintaining accuracy. The selected alpha = 0.1 and lambda = 0.01303 indicate a 10% Lasso and 90% Ridge mix, offering an optimal balance between feature selection and coefficient shrinkage. For binary classification, setting family = "binomial" applies logistic

regression, modeling the probability of one class. Each categorical variable's reference level acts as a baseline for comparison.

Gender: Female is the reference, and the coefficient indicates the effect of being Male. For Genetic Mutation, Family History, Smoking, Alcohol, Radiation, Infection, Chronic Illness, and Immune Disorders, "No" is the reference, so coefficients show the impact of having these conditions. In terms of Socioeconomic Status, Low is the baseline, with coefficients measuring the effect of moving to Medium or High. Urban/Rural comparison has Rural as the reference, with coefficients showing the effect of urban living.

Reference categories are baselines for comparison, allowing us to measure impacts of other variable levels. Positive coefficients indicate higher likelihood or risk of the outcome, while negative coefficients suggest lower impact.

CONFUSION MATRIX FOR TRAINING SET

		Reference	
		Negative	**Positive**
Prediction	**Negative**	170	20
	Positive	14	77

Table 8

Accuracy	0.879
95% CI (Lower)	0.835
95% CI (Upper)	0.9147
No Information Rate	0.6548
P-Value (Acc > NIR)	0.0000
Kappa	0.7284

Table 9

The training set confusion matrix for GLMNET shows an accuracy of 87.9%, with a Kappa value of 0.7284, indicating substantial agreement. The model correctly classified 170 negatives and 77 positives, while misclassifying 20 negatives and 14 positives. The 95% confidence interval (83.5% – 91.47%) confirms reliability, and the p-value (0.0000) shows statistical significance. The No Information Rate (65.48%) highlights that the model performs significantly better than random guessing.

CO-EFFICIENT OF GENERALIZED LINEAR MODEL

Variable	Estimate		Variable	Estimate
Age	-0.002		Smoking_Status	1.0599
Gender	-0.1094		Alcohol_Consumption	-0.3555
WBC_Count	0.0003		Radiation_Exposure	1.2889
RBC_Count	0.4918		Infection_History	0.3804
Platelet_Count	0.0005		BMI	0
Hemoglobin_Level	-0.147		Chronic_Illness	-0.1575
Bone_Marrow_Blasts	0.4253		Immune_Disorders	-0.0507
Genetic_Mutation	2.5637		Socioeconomic_Status	-0.4009
Family_History	2.0862		Urban_Rural	0.0022

Table:10

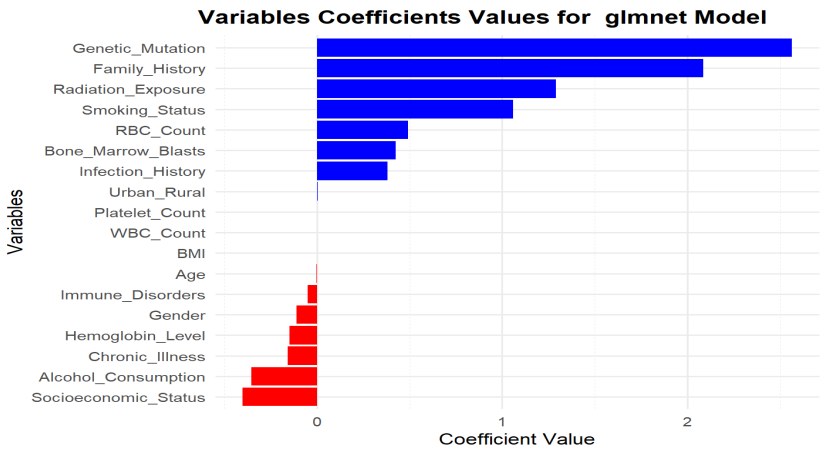

Figure 2

This table & figure shows the estimated coefficients from a GLMNET logistic regression model, illustrating how each variable affects the likelihood of leukemia presence. Positive estimates (e.g., Genetic Mutation = 2.5637, Family History = 2.0862) increase the probability of the outcome, while negative estimates (e.g., Gender = -0.1094, Hemoglobin Level = -0.1470) decrease it. Some variables have minimal impact (e.g., BMI = 0.0000, Urban/Rural = 0.0022). Strong predictors include Genetic Mutation, Family History, Smoking Status, and Radiation Exposure, all with high positive estimates.

CONFUSION MATRIX FOR TESTING SET

		Reference	
		Negative	**Positive**
Prediction	**Negative**	72	11
	Positive	7	31

Table 11

Accuracy	0.8512
95% CI (Lower)	0.7751
95% CI (Upper)	0.9094
No Information Rate	0.6529
P-Value (Acc > NIR)	0.0000
Kappa	0.6643

Table 12

The glmnet model's confusion matrix reveals an accuracy of 85.12%, with a 95% confidence interval of 77.51% to 90.94%, demonstrating strong generalization. It outperforms the No Information Rate (NIR) of 65.29%, and the Kappa score of 0.6643 indicates good agreement beyond chance.

FINDINGS

- Key Risk Factors: Genetic mutations, family history, bone marrow blasts, smoking, and radiation exposure significantly increase the likelihood of leukemia.
- White Blood Cell (WBC) Count: A higher WBC count is strongly associated with leukemia, indicating abnormal cell production.
- Protective Factors: Individuals from a medium socioeconomic background showed a lower risk compared to those from a low-income background.
- Non-Significant Factors: Variables such as age, BMI, platelet count, hemoglobin levels, infection history, alcohol consumption, chronic illness, and immune disorders did not show a significant direct impact on leukemia prediction.
- Classification Performance: The model effectively distinguishes between leukemia-positive and negative cases with high accuracy and reliability.

CONCLUSION

Leukemia risk is influenced by genetic, lifestyle, and environmental factors. Genetic mutations and a family history of leukemia are strong predictors. High white blood cell counts and increased bone marrow blasts indicate blood and bone marrow abnormalities. Lifestyle factors like smoking and radiation exposure also elevate risk, while middle-income individuals may experience a protective effect, though the reasons are unclear.

Despite analyzing factors like BMI, platelet count, and alcohol consumption, most did not show significant impacts on leukemia prediction. Early identification through screenings and genetic testing can improve disease management and survival rates. A comprehensive approach, including medical monitoring and lifestyle changes, is essential to reduce leukemia incidence and enhance patient outcomes.

RECOMMENDATIONS

To reduce leukemia risk and improve early detection, individuals with a family history or genetic predisposition should have regular screenings. Lifestyle changes like quitting smoking and minimizing radiation exposure can lower risk. Those with abnormal WBC counts and bone marrow changes should be closely monitored. Public health programs should raise awareness of risk factors and promote preventive healthcare. More research is needed on the effects of socioeconomic status and environmental influences on leukemia development.

LIMITATION OF THE STUDY

- The dataset was sourced from an open repository and refined to 402 records, which may introduce selection bias and limit the applicability of results to a broader leukemia patient population.
- Filtering by age categories may have excluded certain groups with different leukemia risk patterns, reducing the study's ability to generalize findings across all age groups.
- The removal of variables like ‘Country’ helped focus on medical factors but may have eliminated useful contextual information, such as regional healthcare access and environmental influences.
- The dataset consists of records from patients visiting the hospital, meaning it primarily includes individuals already experiencing health issues. As a result, healthy individuals who have not sought medical attention are not represented, limiting comparisons with a control group.
- The absence of time-event data prevents tracking leukemia progression, making it difficult to assess how risk factors change over time and limiting early intervention strategies.
- There is a lower representation of individuals from low socioeconomic backgrounds, which may impact the study's ability to assess leukemia risk factors and healthcare accessibility across different economic groups.

REFERENCE

Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics, 28(4)*, 213–225.

Chennamadhavuni, A., Lyengar, V., & Al-Kali, A. (2023). Leukemia. *StatPearls Publishing.*

Davis, A. S., Viera, A. J., & Mead, M. D. (2014). Leukemia: An overview for primary care. *American Family Physician, 89(9)*, 731-738.

Dong, Y., Shi, O., Zeng, Q., Lu, X., Wang, W., Li, Y., & Zhang, L. (2020). Leukemia incidence trends at the global, regional, and national level between 1990 and 2017. *Frontiers in Oncology, 10*, 598072.

Hedegaard, M., Matthäus, C., Hassing, S., Krafft, C., Diem, M., & Popp, J. (2011). Spectral unmixing and clustering algorithms for assessment of single cells by Raman.

H., & Mullighan, C. G. (2020). Pediatric acute lymphoblastic leukemia. *Haematologica, 105(11)*, 2524–2539.

Inaba, K. (Year missing). Microscopic imaging. *Theoretical Chemistry Accounts, 130(4–6)*, 1249–1260.

Malard, F., & Mohty, M. (2020). Acute lymphoblastic leukaemia. *The Lancet, 395(10230)*, 1146–1162.

Phillips, C. (2025). Blinatumomab boosts chemotherapy as initial treatment for some kids with ALL. *National Cancer Institute.*

Roberts, K. G., & Mullighan, C. G. (2020). The biology of B-progenitor acute lymphoblastic leukemia. *Cold Spring Harbor Perspectives in Medicine, 10(4)*, 1–22.

Sacks, B., & Seeman, I. (1947). A statistical study of mortality from leukemia. *Blood, 2(6)*, 695-709.

Xu, J., Yu, T., Zois, C. E., Cheng, J. X., Tang, Y., Harris, A. L., & Huang, W. E. (2021). Unveiling cancer metabolism through spontaneous and coherent Raman spectroscopy and stable isotope probing. *Cancers, 13(7)*, 1718.