Random Forest and Radial Basis Function Support Vector Machine for Classifying Acute Myeloid Leukemia Subtypes

Vu N. Le

British International School HCMC, 246 Nguyen Van Huong, Thao Dien, Ho Chi Minh City, 70000, Vietnam

@lelouis.lnv@gmail.com

Abstract

Acute Myeloid Leukemia (AML) is a genetically heterogeneous malignancy with multiple subtypes classified under the WHO 2022 classification system. This study intends to explore the efficacy of Random Forest (RF) and Radial Basis Function Support Vector Machine (RBF SVM) to classify AML subtypes based on probe intensity values from the GSE13159 dataset. I employ LASSO feature selection to reduce dimensionality, SMOTE to address moderate class imbalance in the dataset, and 10-fold cross-validation for hyperparameter tuning. The tuned RF model achieved a classification accuracy (CA) of 76% whilst the tuned RBF SVM reached a CA of 82%. Moreover, RF's Feature Importance Analysis identified key genes linked with AML prognosis, including BEX2, BAG2, DELE1, and PLEKHF1. The modest accuracy of both models implies that transcriptomic data alone was insufficient for precise classification, and, thus, multi-omics integration is essential for higher accuracy and potential application of ML models in clinical settings. These findings suggest that machine learning can support — but not replace — cytogenetic testing.

Keywords

Computational Biology and Bioinformatics; Genomics; Acute Myeloid Leukemia; Random Forest; Radial Basis Function Support Vector Machine.

Introduction

Literature Review

Acute myeloid leukemia (AML) is an aggressive form of blood cancer from the rapid proliferation of clonal myeloid blast cells that cannot differentiate into mature leukocytes.[1] During the cell cycle, cyclins and cyclin-dependent kinases (CDKs) regulating hematopoiesis. The disease originates from genetic and epigenetic alterations from the mutations in proto-oncogenes, which encodes proteins to regulate the growth and differentiation of myeloid blast cells. When proto-oncogenes undergo activation, which can be through either mutations or overexpression, they become oncogenes. This can give rise to further rapid and uncontrolled growth of immature myeloid cells in the bone marrow.

AML subtypes are classified under the World Health Organisation (WHO) 2022 classification system based on genetic, morphologic, and immunophenotypic characteristics. Accurate classification of AML subtypes is critical for guiding treatment decisions and improving patient outcomes. For example, subtypes like "AML with t(8;21)" and "AML with inv(16)" have unique molecular signatures that influence prognosis and therapy.[8] Other certain AML subtypes, such as those with TP53 aberrations, still present particularly poor outcomes despite recent advancements in AML treatment, including low-intensity induction therapy with hypomethylating agents and venetoclax or targeted oral therapies.[4] The high dimensionality of gene expression data and the complex relationships between different 'biomarkers' pose significant challenges for traditional statistical methods like regression models.[10] Currently, the standard testing method for classifying AML subtypes remains cytogenetic testing (Karyotyping, FISH, CGH, aCGH),[16] which achieves high accuracy in identifying key translocations and mutations associated with different subtypes.

Vu Nguyen Le

With the advent of high-throughput gene expression profiling, machine learning (ML) approaches have been explored for AML subtype classification as a potential diagnostic method. Gene expression-based ML models, such as Random Forest (RF) and Support Vector Machine (SVM) can analyse complex molecular patterns for hematologic cancers, identify prognostic "biomarkers", predict treatment response, and perform risk stratification.[3,5-11,13-14] In fact, recent studies have highlighted the utility of RF in AML research across different aspects, such as predicting complete remission (CR) outcomes in AML patients,[8] integrating microarray and RNA-seq data for multiclass AML subtype classification,[12] and cancer biomarker selection.[10] In comparison to other optimised ML methods, although some studies reported that RF was less effective compared to other ML classifiers in their studies,[12,14] this limitation was said to be attributed to dataset-specific characteristics or insufficient feature selection. Other studies demonstrate RF's robustness in handling high-dimensional datasets and identifying important biomarkers.[9-10]

Despite these advances, gaps remain in leveraging ML to classify AML subtypes using biomarkers identified from microarray datasets systematically. Some studies focus on general cancer classification or prognosis but do not address AML subtype heterogeneity explicitly.[13] Additionally, challenges like class imbalance, which hinders the scalability of predictive models, and the integration of heterogeneous datasets persist.[10] Addressing these challenges through tailored ML models could bridge this gap and enhance subtype classification capabilities.

Aim

This study aimed to evaluate the feasibility of RF and SVM with RBF kernel for multiclass classification of AML subtypes using probe-level intensity data from a microarray dataset. The study also aimed to optimize and evaluate model performance through statistical metrics and data visualization figures. Through this approach, the research study seeks to contribute to enhance hematologic cancer subtype classification and advance computational methods for cancer diagnostics. As such, the study asked the question, "How effectively can Random Forest and Radial Basis Function Support Vector Machine classify Acute Myeloid Leukemia subtypes based on microarray-derived probe intensity values?"

Methods

Search Strategy

The study searched for appropriate probe intensity values from the Gene Expression Omnibus (GEO) database. Different keywords, such as "acute myeloid leukemia", "AML", "WHO classification", and "gene expression biomarkers" were combined with Boolean operators (AND, OR) to refine results and ensure relevance to AML subtypes. Moreover, only datasets featuring human subjects with AML, well-annotated microarray probe-level intensity values, and recently updated entries were included in the filtering process. Datasets focusing on animal models, unrelated diseases, or lacking relevant clinical metadata were excluded. As such, priority was given to datasets with suitable probe intensity signals for AML subtypes under the WHO classification, adequate sample sizes, and diversity across subtypes. Ultimately, the GSE13159 dataset was chosen because it met these criteria, containing probe-level microarray data with AML subtypes under the WHO classification. Unfortunately, due to inconsistencies in subtype labelling and expression scaling across other datasets, no suitable validation set was identified. As a result, 10-fold cross-validation was employed as an alternative validation strategy. The GSE13159 dataset was processed using library package in R scripts within the RStudio environment (2024.12.0+467 version).

Data Preprocessing

GSE13159 contains probe-level expression data obtained from the Affymetrix HG-U133 Plus 2.0 microarray platform. Microarray relies on probe hybridization, so each probe has an intensity value corresponding to the abundance of a target transcript. In the dataset, rows correspond to probe IDs and columns represent samples. Standard microarray preprocessing includes background correction, quantile normalization, and summarization. However, we only performed missing value removal (and any intensity values of zero were replaced with a small constant),

Not Peer-reviewed

thresholding, sampling, and dimensionality reduction, without any normalization procedures. This is because RowMeans analysis already showed that the range of probe intensities was compressed within a defined range (min: ~0.01, max: ~0.97).

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.006667 0.139016 0.236859 0.285708 0.404571 0.967323
```

Figure 1. Summary of RowMeans analysis.



Figure 2. Box plot of probe intensity distributions from the GSE13159 microarray data.

Figure 1 represents the distribution of probe intensities before filtering. Figure 2 visualises a box plot of the distribution of intensity values across all samples, with a consistent spread observed across different samples. According to RowMeans analysis, the range of probe expression values spans from 0.006667 (minimum) to 0.967323 (maximum), with the 1st quartile (0.139016) representing the lower 25% of intensities and the 3rd quartile (0.404571) representing the upper 25%. Probes with intensity values below the 1st quartile were categorised as low-intensity probes. Probes with intensity values between the 1st and 3rd quartiles were moderately expressed. Probes with expression values above the 3rd quartile were categorised as high-intensity probes. These probes often include housekeeping genes or AML-specific upregulated transcripts.

Subsequently, a median probe intensity threshold of 0.236859 was chosen and applied to remove low-intensity probes. This value represented the central tendency of average expression levels across all the sampled probes. Low-intensity probes, which had weak hybridization signals, were more susceptible to technical noise or non-transcriptionally active genes in the studied condition.

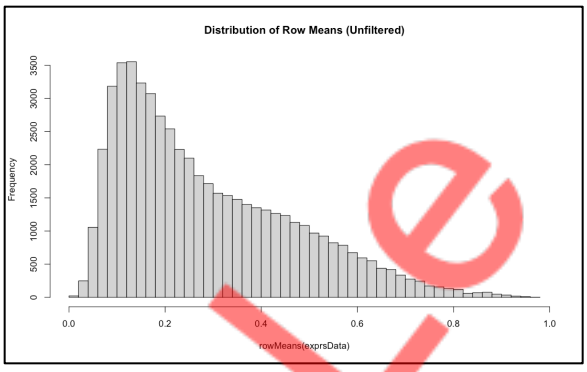


Figure 3. Distribution of row means before filtering.

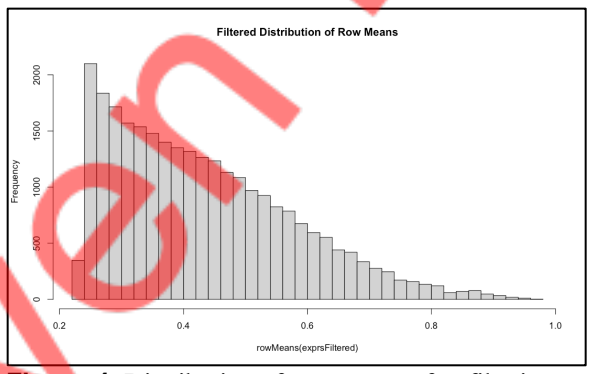


Figure 4. Distribution of row means after filtering.

Through filtering, only probes with moderate to high intensity signals, which exhibit sufficient transcriptional activity, were retained to contribute to meaningful classification outcomes. Since the dataset originally contained multiple leukemia types, we filtered it again for samples specific to only AML and no other leukemia variants. Only samples explicitly labeled as “AML” in the metadata were kept. This resulted in a final dataset of 542 remaining AML samples in the dataset, each assigned to one of the six unique subtype labels as shown in Figure 5.

```
                                AML complex aberrant karyotype
                                                            48
                                    AML with inv(16)/t(16;16)
                                                            28
AML with normal karyotype + other abnormalities
                                                           351
                                    AML with t(11q23)/MLL
                                                            38
                                        AML with t(15;17)
                                                            37
                                         AML with t(8;21)
                                                            40
```

Figure 5. Unique AML subtypes in the GSE13159 dataset.

Vu Nguyen Le

Following subtype filtering, the samples were aligned by matching phenotype metadata (rows) with corresponding probe intensity values-(columns). Samples without both expression data and their corresponding label were removed, so that only one-to-one correspondent records remained.

To address the high dimensionality of microarray expression data, Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied for feature selection. Multinomial cross-validation was implemented for LASSO to determine the optimal regularization parameter, selecting probes with non-zero coefficients.

To address the moderate class imbalance between subtypes, Synthetic Minority Over-Sampling Technique (SMOTE) was implemented, which artificially generated synthetic samples for underrepresented AML subtypes. In a theoretical sense, this method helps to minimise classification bias as the model trains on data that is dominated by the "normal karyotype + other abnormalities" subtype.

Moreover, top-ranked probes were identified through the RF model's Feature Importance, selected for their contribution to AML subtype classification rather than their statistical differential expression.

Train-Test Split

Following preprocessing, the dataset was split into a training set (70%) and a testing set (30%). The stratified split ensured that each AML subtype was proportionally represented in both sets. This separation allowed for independent model evaluation while maintaining sufficient training data for model learning.

Machine Learning Pipelines

To classify AML subtypes, four ML models were trained and evaluated: Pre-SMOTE RF, Post-SMOTE and Untuned RF, Post-SMOTE and Tuned RF, and Tuned SVM with RBF kernel. The Pre-SMOTE RF serves as a baseline model, handling data without any class balancing interventions. The Tuned RF (ntree=500) and Tuned RBF SVM models were fine-tuned through 10-fold cross-validation.

There are 3 hyperparameters tuned. Firstly, the mtry from the RF. This measures the number of features randomly selected at each tree split. Secondly, the cost C ($\boldsymbol{C > 0}$) in the RBF SVM. This regularization parameter controls the trade-off between maximizing the margin and minimizing classification errors. A low C cost means that there is a larger margin, more misclassifications but better generalization. Thirdly, the sigma $\boldsymbol{\sigma}$ ($\boldsymbol{0 < \sigma < 1}$) from the RBF SVM. This determines the reach of a single training instance on the decision boundary. A low value means that every training example will have a far reach.

Results and Discussion

Model Evaluation

Each model's evaluation was based on standard classification metrics, including classification accuracy (CA), confidence intervals (CI), Cohen's Kappa coefficient, P-value, and class-specific performance indicators such as Sensitivity and Specificity.

CA ($\boldsymbol{0\% \leq CA \leq 100\%}$) measures the proportion of correctly classified instances among all predictions. A higher CA score means that the model is more accurate.

CI is a range of values that estimate the true CA of the model with a certain probability. For example, (95% CI: 68.41% - 82.17%) means the accuracy is likely between 68.41% and 82.17% with 95% confidence. This is used to measure uncertainty in performance.

The Cohen's Kappa Coefficient ($\boldsymbol{-1 \leq \kappa \leq 1}$), or Kappa score for short, measures the level of agreement between the model's predictions and actual labels while accounting for agreement occurring by chance. A score $\boldsymbol{\kappa = 1}$ means perfect agreement, $\boldsymbol{\kappa = 0}$ means random agreement, and $\boldsymbol{\kappa < 0}$ means that the performance is worse than random chance.

The P-value ($\boldsymbol{0 \leq P \leq 1}$) measures the probability that the observed accuracy is due to random chance with respect to the No Information Rate (NIR). The NIR ($\boldsymbol{0 \leq NIR \leq 1}$) is the accuracy a naive classifier would achieve by always predicting the most frequent class. The No Information Rate (NIR) for the testing dataset

Uploaded by Linh Nguyen Le

was 0.6522. A score $\boldsymbol{P < 0.05}$ means that the model performance is statistically significant, whilst $\boldsymbol{P > 0.05}$ means that the model would not be better than random chance.

Sensitivity ($\boldsymbol{0 \leq x \leq 1}$) is the ability of the model to correctly identify True Positive cases. Specificity ($\boldsymbol{0 \leq x \leq 1}$) is the ability of the model to correctly identify True Negative cases. For both Sensitivity and Specificity, a score of 1 means that all the True cases were detected whilst a score of 0 means that none of the True cases were detected.

The evaluation for all 4 models were as followed. Firstly, the Pre-SMOTE RF model achieved an overall CA of 75.78% (95% CI: 68.41% - 82.17%), with a Kappa score of 0.4062, indicating moderate agreement beyond chance. The model held meaningful predictive capability beyond baselines guessing (P=0.002563). However, the model particularly struggled with “AML with t(11q23)/MLL” and “AML with t(8;21)”, both of which had sensitivity of zero, whilst there is a strong bias toward the predominant subtype “AML with normal karyotype + other abnormalities” (Sensitivity = 1.00, but Specificity = 0.3036).

The Post-SMOTE RF (Untuned) model improved in CA to 77.02% (95% CI: 69.74% - 83.27%) and exhibited an increase in Kappa score to 0.4468 and a significant P-value (P = 0.0008079), indicating better agreement. Notably, the sensitivity of minority classes improved, such as an increase from 0.00 from the Pre-SMOTE RF to 0.25 for the “AML with inv(16)/t(16;16)” subtype. Yet, specificity remained at 1.00 for minority classes, plausibly due to overfitting on the resampled data.

The Tuned RF (Post-SMOTE), with mtry = 5, achieved a CA of only 75.78% — the same as the Pre-SMOTE RF. This tuned model also held a slightly higher Kappa score of 0.4064 but the same P-value as the Pre-SMOTE RF. While Sensitivity and Specificity varied slightly, the model struggled with the subtype “AML with t(8;21)” (Sensitivity = 0.0833). This suggests that hyperparameter tuning did not improve classification performance, plausibly due to limitations in the feature set.

The Tuned RBF SVM model, with cost C = 8 and sigma $\boldsymbol{\sigma}$ = 0.03125, achieved the highest CA at 81.99% (95% CI: 75.17% - 87.59%) with a Kappa score of 0.6165 and a P-value (P=2.001e-06), indicating moderate to substantial agreement. The sensitivity for “AML with t(15;17) also increased to 81.82%, and even “AML with t(8;21) improved to 33.33% sensitivity, outperforming all RF models in minority class recognition. However, the sensitivity for “AML with inv(16)/t(16;16)” remained at 25.00%. This suggests there is a persistent challenge across all four models to classify this subtype.

Data Visualization

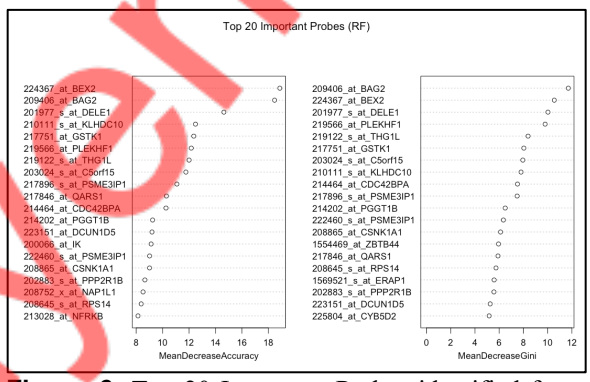


Figure 6. Top 20 Important Probes identified from Variable Importance Plot of Post-SMOTE RF’s Feature Importance using Mean Decrease Accuracy and Mean Decrease Gini.

The Mean Decrease Gini (MDG) measures the impurity reduction in decision trees. The Mean Decrease Accuracy (MDA) measures the impact of excluding each gene on model accuracy. Both metrics suggest the top 20 probes that contribute substantially to classification. The Probe's ID was correspondently mapped to a specific gene symbol using the Affymetrix HG-U133_Plus_2 Array annotation data (chip hgu133plus2). Among the most critical genes identified are BEX2, BAG2, DELE1, PLEKHF1, and PSME3.

Confusion Matrix diagrams were compiled and row-normalized for each model. Each row sums to one using row-wise proportion scaling to account for class imbalance. This allows for proportionate scaling to see underrepresented classes more clearly.

Version 1.0

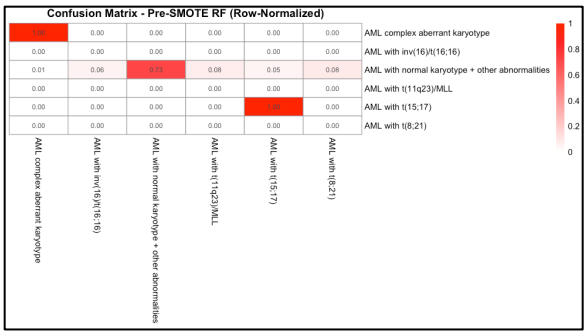


Figure 7. Row-Normalized Confusion Matrix for baseline RF model before SMOTE balancing. Overall accuracy: 75.78% (95% CI: 68.41% – 82.17%), Kappa: 0.4062, NIR: 65.22%, p-value (Acc>NIR): 0.002563.

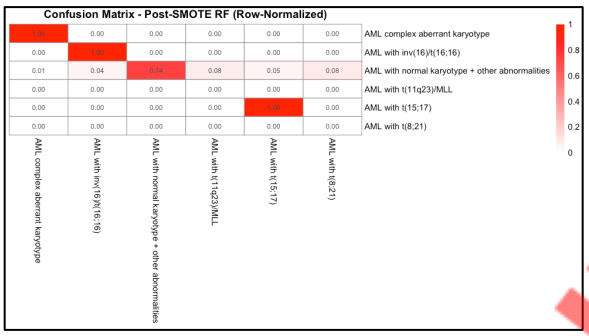


Figure 8. Row-Normalized Confusion Matrix for untuned RF model after SMOTE balancing. Overall accuracy: 77.02% (95% CI: 69.74% - 83.27%), Kappa: 0.4468, NIR: 65.22%, p-value (Acc>NIR): 0.0008079.

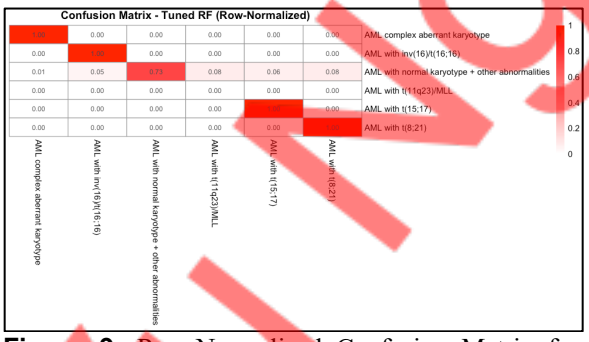


Figure 9. Row-Normalized Confusion Matrix for tuned RF model after SMOTE balancing. Overall accuracy: 75.78% (95% CI: 68.41% – 82.17%), Kappa: 0.4064, NIR: 65.22%, p-value (Acc>NIR): 0.002563. Tuned parameters: mtry =5.

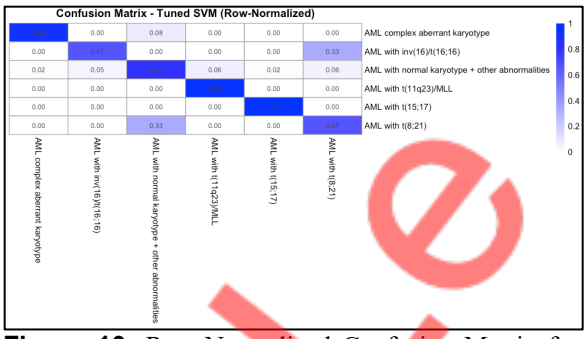


Figure 10. Row-Normalized Confusion Matrix for tuned Post-SMOTE RBF SVM model. Overall accuracy: 81.99% (95% CI: 75.17% - 87.59%), Kappa: 0.6165, NIR: 65.22%, p-value (Acc>NIR): 2.001e-06. Tuned parameters: C = 8, σ = 0.03125.

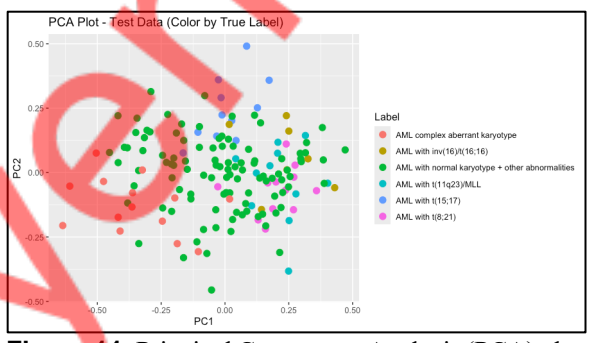


Figure 11. Principal Component Analysis (PCA) plot on the testing dataset. PC1 and PC2 retain most variance in probe expression values. Data points colored by true AML subtype labels. No clear class separation.

Table 1. Sensitivity of Pre-SMOTE RF, Post-SMOTE and Untuned RF, Post-SMOTE and Tuned RF, and Tuned RBF SVM across AML subtypes. Data samples were rounded to three decimal places.

	Sensitivity			
	Pre-SMOTE RF	Post-SMOTE RF	Tuned RF	Tuned RBF SVM
AML complex aberrant karyotype	0.929	0.929	0.929	0.857
AML with inv(16)/t(16;16)	0.000	0.250	0.125	0.250
AML with normal karyotype + other abnormalities	1.000	1.000	1.000	0.971
AML with t(11q23)/MLL	0.000	0.000	0.000	0.273
AML with t(15;17)	0.364	0.364	0.182	0.818
AML with t(8;21)	0.000	0.000	0.083	0.333

Vu Nguyen

Table 2. Specificity of Pre-SMOTE RF, Post-SMOTE and Untuned RF, Post-SMOTE and Tuned RF, and Tuned RBF SVM across AML subtypes. Data samples were rounded to three decimal places.

	Specificity			
	Pre-SMOTE RF	Post-SMOTE RF	Tuned RF	Tuned RBF SVM
AML complex aberrant karyotype	1.000	1.000	1.000	0.993
AML with inv(16)/t(16;16)	1.000	1.000	1.000	0.993
AML with normal karyotype + other abnormalities	0.304	0.339	0.304	0.554
AML with t(11q23)/MLL	1.000	1.000	1.000	1.000
AML with t(15;17)	1.000	1.000	1.000	1.000
AML with t(8;21)	1.000	1.000	1.000	0.987

Data Discussion

There are several observations that can be made from the evaluation results and visualization.

Firstly, Pre-SMOTE RF showed strong overall CA but was biased toward the dominant subtype, often failing in minority subtype detection, such as “AML with t(11q23)/MLL” and “AML with t(8;21))” with near-zero detection rates.

SMOTE re-sampling did improved Sensitivity for underrepresented subtypes (Table 1), particularly improving Sensitivity from 0.000 to 0.250 for “AML with inv(16)/t(16;16)” and an increase in Cohen’s Kappa coefficient to 0.4468 (Figure 8). However, SMOTE also introduced overfitting, evident by the poorer performance of the Tuned RF model with a lower CA of 75.78% and almost identical statistics compared to the Pre-SMOTE RF model except for a slightly higher Cohen’s Kappa coefficient.

This also suggests that hyperparameter tuning (mtry optimization) in Tuned RF yield negligible gains. In essence, the RF’s classification ability plateaued due to reliance on feature redundancy, making it less sensitive to tuning in the GSE13159, which was fundamentally imbalanced. In contrast, the Tuned RBF SVM model outperformed all RF models with 81.99% accuracy, demonstrating superior class separation from the kernel-based classifier. Yet, it can’t be ignored that, despite gains in “AML with t(15;17)”, where the Sensitivity increases from 0.182 to 0.818 (Table 1), rare subtypes, such as “AML with t(11q23)/MLL” and “AML with inv(16)/t(16;16)” remained challenging.

Additionally, all the models’ performance can also be interpreted form the confusion matrices (Figures 7-10). It can be seen that there is a struggle from all models with overlapping transcriptional signatures between A”ML with inv(16)/t(16;16)” and "AML with normal karyotype + other abnormalities". The ideal diagonal line (spanning from top-left corner to bottom-right corner) is not achieved. Regardless, the RBF SVM model demonstrated the highest diagonal structure fidelity, evident by the more concentrated blue colour.

Moreover, it should also be noted that the RF models appear to be extremely conservative in misclassifying True Negative cases as their specificity scores remain consistently 1.000 according to Table 2. This is also probably why the RF models became more sensitive for the minor classes, including “AML with t(8;21)”, “AML with t(11q23)/MLL”, "AML with inv(16)/t(16;16)", and "AML complex aberrant karyotype", where they all scored a high specificity score on. Regarding "AML with normal karyotype + other abnormalities", all three RF models averaged low Specificity scores but high Sensitivity scores. This is most likely a case of imbalance in the decision boundaries, where the RF classifier is too cautious in calling a sample a minority subtype, leading to high Specificity but low Sensitivity for those classes. This can also be confirmed with the PCA plot in Figure 11, where there is no apparent class separation between "AML with t(8;21)" (pink label) and "AML with t(11q23)/MLL" (cyan label), as opposed to a clearer identification of vector space for "AML complex aberrant karyotype" (red label).

Secondly, SMOTE improved sensitivity for underrepresented classes but resulted in overfitting, particularly with the Untuned RF model. Thirdly, hyperparameter tuning for RF did not significantly boost accuracy or class-specific performance, indicating that RF’s performance plateaued regardless of optimization. Fourthly, the tuned SVM outperformed all RF models, achieving the highest accuracy (81.99%) and better class-wise Sensitivity, although some minority subtypes still had low detection rates.

Unsupervised

On the other hand, Feature Importance Analysis, through both MDG and MDA metrics, identified BEX2, BAG2, DELE1, PSME3IP1, and PLEKHF1 as key mapped genes to the most active probe signatures.

BEX2 (Brain Expressed X-Linked 2) has been connected to AML chemoresistance, particularly in relation to the expression of the Mixed Lineage Leukemia. Research indicates with BEX2 correlates with MLL translocations by suppressing apoptosis through BCL-2 regulation.

BAG2 (Bcl2-associated Athanogene 2), which functions as a member of Hsp70 as a co-chaperone, has been shown to contribute to AML cell survival by inhibiting proteasomal degradation of oncogenic proteins.

DELE1 (DAP3 Binding Cell Death Enhancer 1), is a protein coding gene that interacts with HRI kinase, which, regarding AML blasts, can activate mitochondrial stress responses that are crucial for apoptosis evasion.

PSME3IP1 (Proteasome Activator Subunit 3 Interacting Protein 1), a proteasome activator, is associated with alterations in protein degradation pathways in regard to AML progression.

PLEKHF1 (Pleckstrin Homology And FYVE Domain Containing 1), a lipid signalling gene, is implicated in tumorigenesis gene mutations and endosomal trafficking, which may contribute to AML progression and AML's microenvironment interactions.

Limitations

There were critical challenges faced for this study — several methodological constraints impact modal generalizability. The class imbalance remains a significant issue as it affects the minority subtype predictions for all the RF models. Although SMOTE sampling did enhanced sensitivity, it introduced overfitting, evidenced by the Tuned RF's specificity drop (Table 2).

Secondly, RF's majority-voting mechanism with 500 decision trees inherently favours common subtypes, explaining why tuning failed to improve performance, even after optimizing mtry.

Additionally, feature selection via LASSO regression, while effective in redundancy elimination, have also excluded functionally relevant genes as it filtered down to less than 50 selected genes. Different approaches like Gene set enrichment analysis (GSEA) should also be attempted to provide more clinically meaningful attributes.

Finally, generalization is still uncertain as external validation was not performed due to problems with selecting datasets. Of course, 10-fold cross validation serves as a well-suited alternative. However, this still implies that generalization is uncertain as there has not been independent dataset testing.

Conclusion

This study demonstrates that ML models can undertake AML multiclass subtype classification, with the RBF SVM model performing the best and achieving a classification accuracy of 81.99%.

Feature importance analysis also revealed genes associated with apoptosis resistance, proteasomal degradation, and lipid signaling, which may prove useful for AML prognosis. This supports another role of ML models in contributing to biomarker discovery. However, as stated before, limitations in data imbalance, overfitting, and external validation restrict immediate clinical translation.

These results suggest that while ML models can provide meaningful insights into AML subtype classification, they do not replace cytogenetic and molecular testing. Instead, they may can be used as complementary tools, particularly for exploratory biomarker identification and potential prognostic applications. However, challenges such as hyperparameter tuning, feature selection, and class imbalance remain key obstacles in achieving reliable clinical applicability. Of course, only ML models that can achieved CA scores of 95%+ should be considered for clinical application, which would require a lot of feedback training. Yet, it can't be denied that ML remains a promising tool for AML subtype classification.

Vu Nguyen Le

Acknowledgements

I express gratitude to my mentor, Vo Tram Anh (University of Toronto '28), for her guidance in helping me write this research paper. The source code in R scripts can be found on the public GitHub repository under the name "lnv-louis/AML-Subtypes-Classification-ML."

References

1. Dowling, P.; Tierney, C.; Dunphy, K.; Miettinen, J. J.; Heckman, C. A.; Bazou, D.; O'Gorman, P. Identification of Protein Biomarker Signatures for Acute Myeloid Leukemia (AML) Using Both Nontargeted and Targeted Approaches. Proteomes 2021, 9 (4), 42. https://doi.org/10.3390/proteomes9040042.

2. Braoudaki, M.; Lambrou, G. I.; Vougas, K.; Karamolegou, K.; Tsangaris, G. T.; Tzortzatou-Stathopoulou, F. Protein Biomarkers Distinguish between High- and Low-Risk Pediatric Acute Lymphoblastic Leukemia in a Tissue-Specific Manner. J. Hematol. Oncol. 2013, 6 (1), 52. https://doi.org/10.1186/1756-8722-6-52.

3. Cheng, Y.; Yang, X.; Wang, Y.; Li, Q.; Chen, W.; Dai, R.; Zhang, C. Multiple Machine-Learning Tools Identifying Prognostic Biomarkers for Acute Myeloid Leukemia. BMC Med. Inform. Decis. Mak. 2024, 24 (1). https://doi.org/10.1186/s12911-023-02408-9.

4. Bhansali, R. S.; Pratz, K. W.; Lai, C. Recent Advances in Targeted Therapies in Acute Myeloid Leukemia. J. Hematol. Oncol. 2023, 16 (1), 29. https://doi.org/10.1186/s13045-023-01424-6.

5. Zhang, Y.; Liu, D.; Li, F.; Zhao, Z.; Liu, X.; Gao, D.; Zhang, Y.; Li, H. Identification of Biomarkers for Acute Leukemia via Machine Learning-Based Stemness Index. Gene 2021, 804, 145903. https://doi.org/10.1016/j.gene.2021.145903.

6. Zhong, F.-M.; Yao, F.-Y.; Yang, Y.-L.; Liu, J.; Li, M.-Y.; Jiang, J.-Y.; Zhang, N.; Xu, Y.-M.; Li, S.-Q.; Cheng, Y.; Xu, S.; Huang, B.; Wang, X.-Z. Molecular Subtypes Predict Therapeutic Responses and Identifying and Validating Diagnostic Signatures Based on Machine Learning in Chronic Myeloid Leukemia. Cancer Cell Int. 2023, 23 (1). https://doi.org/10.1186/s12935-023-02905-x.

7. Al-Tashi, Q.; et al. Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review. Int. J. Mol. Sci. 2023, 24 (9), 7781. https://doi.org/10.3390/ijms24097781.

8. Gal, O.; Auslander, N.; Fan, Y. Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression. *Cancer Inform.* **2019**, *18*. https://doi.org/10.1177/1176935119835544.

9. Yan, Z.; Li, J.; Xiong, Y.; Xu, W.; Zheng, G. Identification of Candidate Colon Cancer Biomarkers by Applying a Random Forest Approach on Microarray Data. Oncol. Rep. 2012, 28 (4), 2173–2180. https://doi.org/10.3892/or.2012.1891.

10. Najafi, A.; Shakeri, M. T. Classification and Biomarker Gene Selection for Cancer Gene Expression Data Using Random Forest. Iran. J. Pathol. 2017, 12 (4), 326–334. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5844678/.

11. Wei, Q. Application of Machine Learning Techniques to Acute Myeloid Leukemia. Master's Thesis, University of Washington, 2016. https://digital.lib.washington.edu/researchworks/handle/1773/35503.

12. Castillo, D.; Galvez, J. M.; Herrera, L. J.; Rojas, F. Leukemia Multiclass Assessment and Classification from Microarray and RNA-Seq Technologies Integration at Gene Expression Level. PLoS One 2019, 14 (7). https://doi.org/10.1371/journal.pone.0212127.

13. Qin, Y.; Pu, X.; Hu, D.; Yang, M. Machine Learning-Based Biomarker Screening for Acute Myeloid Leukemia Prognosis and Therapy from Diverse Cell-Death Patterns. Sci. Rep. 2024, 14 (1). https://doi.org/10.1038/s41598-024-68755-3.

14. Sakib, S.; Tanzeem, A. K.; Tasawar, I. K.; Rahman, M. Blood Cancer Recognition Based on Discriminant Gene Expressions: A Comparative Analysis of Optimized Machine Learning Algorithms. Proc. IEEE Int. Conf. Bioinf. Comput. Biol. 2021, 214–221.

Vo Nguyen Le

https://doi.org/10.1109/BICB52429.2021.9623210.

15. Hua, J.; Ding, T.; Shao, Y. A Transient Receptor Potential Channel-Related Model Based on Machine Learning for Evaluating Tumor Microenvironment and Immunotherapeutic Strategies in Acute Myeloid Leukemia. Frontiers in immunology 2022, 13. https://doi.org/10.3389/fimmu.2022.1040661.

16. Smith, E. Cytogenetic testing | DermNet NZ. dermnetnz.org. https://dermnetnz.org/topics/cytogenetic-testing.

Authors

Le Nguyen Vu (Louis), at the time of writing, is a senior at the British International School HCMC. He is interested in Artificial Intelligence, Decision-Making, and their interdisciplinary applications. He developed this paper as part of his engagement with the HEMA Future Foundation — a non-profit organization that supports Vietnamese children with leukemia and their families.

Vu Nguyen Le