Design and Implementation of Artificial Intelligence Algorithms for Prognostic Stratification in Acute Myeloid Leukemia

A. Basterra-García[1], D. Bermejo-Peláez[5], A. Mendoza[3], D. Brau-Queralt[5], A. D. Ramos-Guerra [2,1], J .E. Ortuño[2,1], I. Hernández-Abad [1], N. Díez [5], R. Ancos-Pintado [4], R. Garcia-Vicente [4], L. Lin [5], A. Santos[1,2], M. Linares [5], J. Martínez-López [3], M. Luengo-Oroz [5], M. J. Ledesma-Carbayo[1,2]

[1] Biomedical Image Technologies (BIT), Universidad Politécnica de Madrid, Madrid, Spain

[2] CIBER de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Madrid, Spain

[3] Servicio de Hematología y Hemoterapia, Hospital Universitario 12 de Octubre, Madrid, Spain

[4] Universidad Complutense de Madrid, Madrid, Spain

[5] Spotlab, Madrid, Spain

Abstract

Acute myeloid leukemia (AML) is a neoplasm of immature myeloid cells characterized by the rapid proliferation of abnormal cells, disrupting normal hematopoiesis and leading to bone marrow failure. Predicting AML prognosis is challenging due to variability in clinical interpretation of bone marrow aspirate (BMA) images. This study aims to enhance AML prognosis by integrating microscopy images with clinical and genetic data using artificial intelligence (AI). We used image based information and clinical and genetic information to predict prognosis using the following modeling approaches: a conventional multiple instance learning (MIL) scheme compared with the clustering-restricted attention learning model (CLAM) for BMA images, and a random forest (RF) algorithm for clinical and genetic data. An ensemble learning strategy combined the best performing models based on image or clinical and genetic information to enhance prognostic accuracy. The study found that the image-based model effectively differentiated between various AML risks, and the multivariable model accurately stratified patient cohorts. The ensemble model showed superior prognostic accuracy over individual models, with improved prediction of 6-month relapse-free survival (RFS6) and risk classification according to the 2022 European Leukemia Net (ELN) system. These findings highlight the potential of AI-integrated data as a prognostic biomarker and decision support tool for oncologists.

1. Introduction

Acute myeloid leukemia (AML) is a neoplasm of immature myeloid cells, causing rapid proliferation of abnormal cells that impair normal hematopoiesis and lead to bone marrow (BM) failure. The one-year relative survival for AML, estimated at 47.2%, reflects the percentage of patients expected to survive the effects of the cancer [1].

Survival rates for hematologic malignancies are improving, but further advancements in precision diagnostics and personalized treatments will enhance patient survival and quality of life. Predicting AML prognosis remains challenging due to variability among patients, necessitating effective prognostic guidelines. Artificial intelligence (AI) has shown promise in this field, with potential applications for diagnosis, risk stratification, predicting prognosis, and treatment and drug discovery [2]. The new European Leukemia Network (ELN) 2022 guidelines categorize patients into adverse, intermediate, and favorable prognostic groups [3]. Additionally, given that relapse is the main indicator of AML progression, predicting six-month relapse-free survival (RFS6), defined as the time between diagnosis and relapse or last follow-up, is of significant clinical interest.

In this context, bone marrow aspirate (BMA) analysis is indispensable. Despite technological advances, hematologists still rely heavily on conventional optical microscopes, leading to variability in diagnoses and prognoses and requiring significant time for reliable results.

Therefore, this study aimed to implement AI techniques for AML prognostic prediction by integrating multimodal information from medical images with clinical and genetic data.

2. Materials and methods

This study employed two approaches to estimate AML prognosis: a BMA image-based model and a model integrating clinical and genetic data. An ensemble of both models was used to enhance prognosis accuracy.

2.1. Patient cohort

A total of 175 AML cases were collected from Hospital Universitario 12 de Octubre (HU12O) in Madrid, Spain. For each case, 40 different fields of view (FOVs) were digitized, and 42 clinical and genetic variables were gathered.

May-Grünwald-Giemsa (MGG) staining images were digitized using a 3D-printed device that allows coupling a mobile phone and aligning its camera with a conventional optical microscope's ocular lens, and a mobile app customized specifically for fast, standarized, and easy digitization of BMA microscopy images. All images were captured at a resolution of 12 megapixels. The smartphones were mounted onto the ocular of a light microscope (Nikon Eclipse 80i) using a 100x objective lens (1,000x total magnification).

ISBN: 978-84-09-67332-2

2.2. Clinical endpoints and labelling methodology

The primary endpoint of this study was the risk categorization of patients according to the ELN 2022 system. The secondary endpoint was relapse-free survival (RFS), with patients having an RFS of less than 6 months classified as having an adverse prognosis, and those with an RFS of 6 months or longer classified as having a favorable prognosis.

Initially, deep learning (DL) models were trained using all three ELN 2022 risk categories; however, the results were suboptimal due to the limited number of intermediate cases (92 adverse, 39 intermediate, and 43 favorable).The small sample size of intermediate cases, coupled with their complexity, made it difficult to achieve reliable results. Consequently, we adopted a binary classification approach, focusing on the adverse and favorable cases based on ELN categorization. The same split was used for both the ELN-labeled cases and those labeled according to RFS6. As a result, the dataset included 92 cases categorized as adverse (68.14%) and 43 as favorable (31.85%) based on ELN categorization. For the RFS6, the dataset was more balanced, comprising 74 adverse cases (54.81%) and 61 favorable cases (45.18%).

2.3. Image-based Deep-Learning Models

2.3.1. Pre-process of BMA images

The preparation of the BMA images for analysis involved segmentation, patching, and feature extraction (Fig. 1). The segmentation process involves down-sampling the image, converting it to HSV color space, and creating a binary mask by thresholding the saturation channel and using morphological closing. Patching was performed with a 50% overlap. Following this, feature extraction was conducted using two encoders. The ResNet50 encoder, pre-trained on ImageNet [4], converted each patch into a 1,024-dimensional feature vector. Additionally, and for comparative purposes, an encoder derived from a BM cell classification algorithm pre-trained on a large number of hematology images [5] was utilized to extract 256 features per patch.

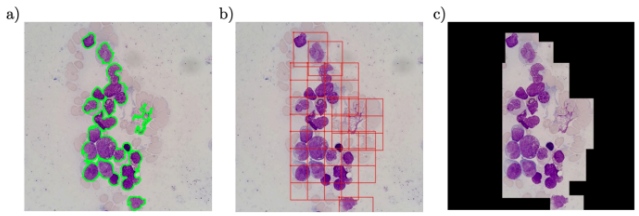

Figure 1: *Image segmentation and patching process*

2.3.2. Multiple Instance Learning (MIL) for AML Prognosis

In the present study, both the conventional Multiple Instance Learning (MIL) model and Clustering-Constrained-Attention Multiple-Instance Learning (CLAM) were employed.

In this study, where a set of patches share a common label, the use of architectures such as MIL is particularly suitable. In this context, the features extracted from the patches are subsequently combined to generate a final decision that takes into account all available information, making these architectures particularly effective for addressing this type of problem.

This approach involves dividing each slide into patches, but faces two main challenges: loss of contextual information and the need for often unavailable patch-level annotations. While weak supervision can address these issues, it sacrifices crucial spatial information. MIL overcomes these challenges by using slide-level labels for prediction tasks [6]. This learning technique is highlighted in Fig. 2.

Additionally, this study employed CLAM, a multiresolution model based on attention-guided MIL. CLAM serves as a framework for prognostic prediction, leveraging DL capabilities to tackle weakly supervised classification tasks in computational pathology. In this work, each FOV in the training set is treated as a single data point with a known slide-level label, but no class-specific information is available for any region within the slide [7].

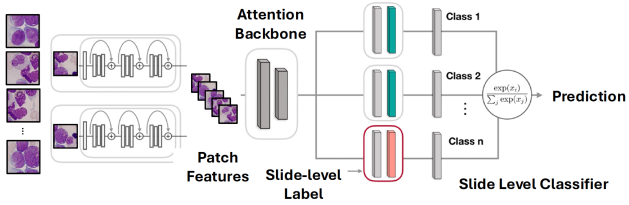

Figure 2: *High-level overview of CLAM*

During training, each slide undergoes patch selection using max-pooling, with scores input into a cross-entropy loss function. Model parameters are optimized using stochastic gradient descent (SGD) with a batch size of one and the Adam optimizer. Dropout is applied after each layer in the attention backbone for regularization to enhance generalization. A conservative learning rate is used to facilitate learning complex relationships between extracted features and class labels. Early stopping based on validation set performance is implemented to mitigate overfitting. Additionally, the network employs self-supervised learning (SL) to distinguish highly attended patches from the least-attended ones, promoting distinct cluster formation within each class during training.

2.4. Multivariable Model

RF was selected for classification owing to its superior performance with imbalanced data relative to other machine learning (ML) classifiers [8]. RF hyperparameter tuning was performed by cross-validated grid-search. Missing numerical values were imputed with the column median, and a new category was added for missing categorical data before one-hot encoding.

A total of 42 variables – demographic (e.g., age at diagnosis, sex), clinical (e.g., hemoglobin, LDH, ECOG, Auer rods), and genetic (e.g., NPM1, TP53, RUNX1) – were integrated into the multivariable model, selected through a literature review and hematologist consultation for prognostic relevance.. The SHAP (or SHapley Additive exPlanations) [9] was employed to visualize each feature's contribution to producing the final prediction of the RF model.

ISBN: 978-84-09-67332-2

2.5. Integration of imaging and clinical data

After obtaining the results from the image-based models and the RF, an ensemble model based on the integration of imaging information and genetic and clinical data was constructed to integrate information from both models. This approach was implemented as the mean value of the predictions of the imaging and clinical models alone.

2.6. Statistical analysis

Stratified five-fold cross-validation was used to train all models and optimize RF hyperparameters. In each fold, the dataset was divided into a training set and an independent test set. To evaluate model performance, 27 patients (20%) with baseline imaging and clinical data were randomly selected to constitute the independent test set, while the remaining 108 (80%) were used for training.. Model performance was assessed using the area under the receiver operating characteristic (ROC) curve (AUC), and the corresponding 95% confidence interval (CI) was estimated through a bootstrap resampling method with 1000 iterations. Kaplan-Meier (KM) survival analysis was conducted to stratify patients according to the model's predictions, utilizing a threshold of 0.5. The log-rank test was utilized to assess the significance of differences between survival curves, considering p-values less than 0.05 as significant. The Cox proportional hazards models were employed to calculate the concordance index. Python (version 3.12.3) was employed for the statistical analysis and model implementation.

3. Results and discussion

In AML, traditional prognosis methods based on clinician interpretation of BMA images face significant variability [5]. Thus, integrating multimodal information through AI techniques could offer more accurate and consistent prognostic predictions.

3.1. Image-based Models

In this study, models trained with features extracted from the BM cell classification encoder exhibited superior performance in the independent test cohort compared to those utilizing ResNet50 features, likely due to the BM cell classification encoder's specialized training on hematology images. Among the models trained using the ELN 2022 system, the CLAM model, leveraging BM cell features, outperformed the conventional MIL model, achieving an AUC of 0.738 (95% CI: 0.661 – 0.817) compared to the MIL model's AUC of 0.661 (95% CI: 0.573 – 0.742). Conversely, for cases classified according to RFS6, the CLAM model again demonstrated superior performance with an AUC of 0.633 (95% CI: 0.417 – 0.853), compared to the standard MIL model, which achieved an AUC of 0.556 (95% CI: 0.475 – 0.634). This result suggests certain relationship between BMA image information and RFS6, which is an interesting finding.

3.2. Multivariable models

The RF model, which used a total of 42 clinical and genetic variables, achieved an exceptional AUC of 0.885 (95% CI: 0.796 – 0.960), when trained with cases labeled according to the ELN 2022 system, incorporating both ELN variables and patient-specific factors such as age at diagnosis and sex. In contrast, when the RF model was trained with cases labeled according to RFS6, it achieved an AUC of 0.774 (95% CI: 0.686 – 0.844), indicating a strong association between RFS6 and both clinical and genetic data.

3.3. Ensemble model

The ensemble approach resulted in an AUC of 0.944 (95% CI: 0.848 – 1.000) in the case of ELN 2022 stratification, surpassing the performance of individually trained models. This highlights the potential of integrating multimodal information to enhance prognostic prediction in AML. Similarly, when the ensemble model was generated with cases labeled according to RFS6, it demonstrated an AUC of 0.783 (95% CI: 0.596 – 0.954), once again outperforming both separately trained models.

The KM survival curves (Fig. 3 and Fig. 4) for overall survival (OS) in the independent test set revealed that the ensemble models effectively stratified patients into adverse and favorable risk groups for both ELN and RFS6 (p-value < 0.05), with strong correlations to OS (ELN model: C-index 0.73; RFS6 model: C-index 0.75). Notably, in the case of RFS6, the survival curves indicate that there is a 20% probability that adverse cases will survive beyond 6-7 months.

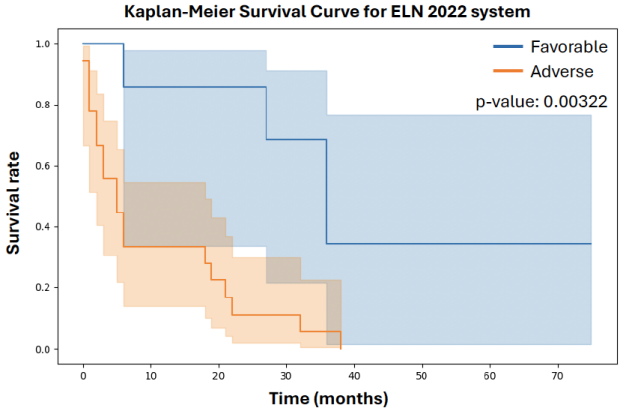

Figure 3: *Kaplan-Meier survival curves for the ensemble model trained for endpoint ELN 2022*

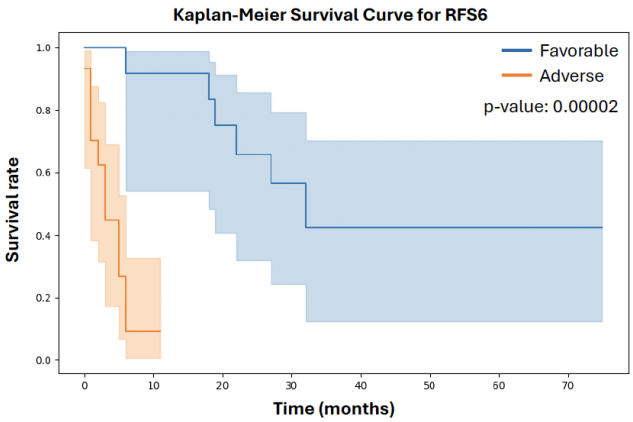

Figure 4: *Kaplan-Meier survival curves for the ensemble model trained for endpoint RFS6*

ISBN: 978-84-09-67332-2

The SHAP analysis revealed that the primary contributors to the predicted outcomes in the RF model for the ELN 2022 system were the NPM1 mutation, monocyte count, and karyotype complexity. Positive values of the NPM1 mutation were strongly associated with a favorable prognosis, consistent with the literature and the definition of the ELN 2022, where the NPM1 mutation is considered a prognostic factor for a favorable outcome [1]. For cases classified according to the RFS6, the most significant clinical and genetic variables were age at diagnosis, karyotype complexity, and the presence of chronic heart disease. Advanced age was correlated with an adverse prognosis, and increased karyotype complexity also indicated a worse prognosis, in agreement with existing research [1].

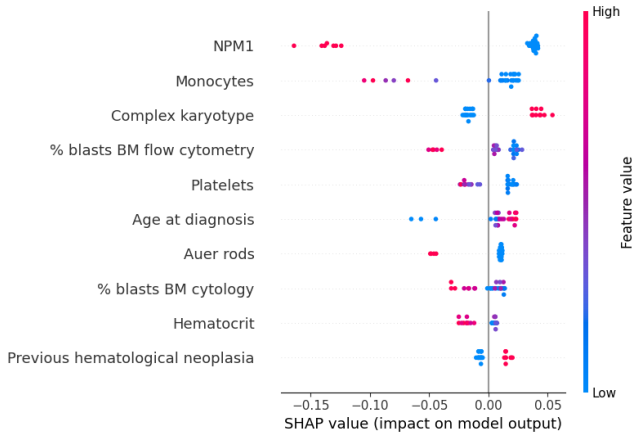

Figure 5: *Multivariable model interpretation using SHAP for ELN label*

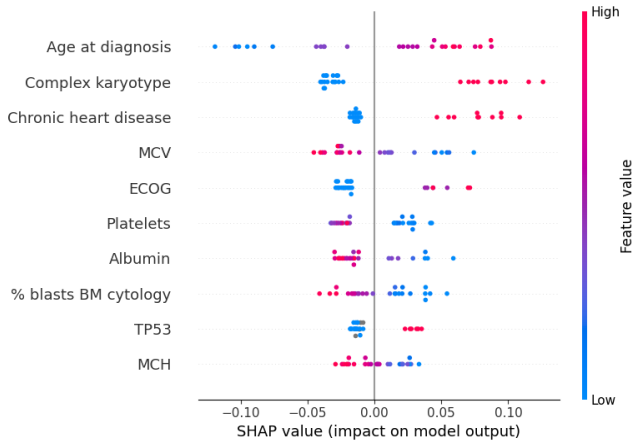

Figure 6: *Multivariable model interpretation using SHAP for RFS6 label*

4. Conclusions

This study underscores the potential of AI in improving prognosis for hematological diseases such as AML, offering a valuable tool for aiding oncologists in identifying high-risk relapse patients and guiding timely interventions. Remarkably, by using only BMA images captured with a standard smartphone, the proposed image-based AI model achieved an AUC of 0.738 for ELN 2022 risk stratification and 0.633 for predicting RFS6. When integrating clinical and genetic variables, the predictive power increased significantly, with AUCs of 0.944 for ELN2022 and 0.783 for RFS6, highlighting the crucial role of combining image-based data with clinical information for more accurate prognosis and optimized therapeutic decisions.. Additionally, our findings indicate that age at diagnosis, karyotype complexity, and the presence of chronic heart disease were the most significant variables contributing to RFS. Nonetheless, the study faced limitations, including the use of a single-center dataset, the scarcity of intermediate cases classified by the ELN system, and the necessity of incorporating potential prognostic biomarkers for more accurate model classification.

Acknowledgements

This project received partial funding from the European Union - NextGenerationEU under Spain's "Plan de Recuperación, Transformación y Resiliencia", the Instituto de Salud Carlos III (ISCIII) through grants PMPTA22/00169, PMPTA22/00088, PMPTA22/00041, and the Centro para el Desarrollo Tecnológico y la Innovación (CDTI) under grant EXP 00156466/IDI - 20230066.

References

[1] S. Shimony, M. Stahl, and R. M. Stone, "Acute myeloid leukemia: 2023 update on diagnosis, risk-stratification, and management," *American Journal of Hematology*, vol. 98, no. 3, pp. 502–526, 2023.

[2] A. Alhajahjeh and A. Nazha, "Unlocking the potential of artificial intelligence in acute myeloid leukemia and myelodysplastic syndromes," *Current Hematologic Malignancy Reports*, vol. 19, no. 1, pp. 9–17, 2024.

[3] G. Garcia-Manero, "Myelodysplastic syndromes: 2023 update on diagnosis, risk-stratification, and management," *American journal of hematology*, vol. 98, no. 8, pp. 1307–1325, 2023.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[5] D. Bermejo-Peláez, S. Rueda Charro, M. García Roa, R. Trelles-Martínez, A. Bobes-Fernández, M. Hidalgo Soto, R. García-Vicente, M. L. Morales, A. Rodríguez-García, A. Ortiz-Ruiz, *et al.*, "Digital microscopy augmented by artificial intelligence to interpret bone marrow samples for hematological diseases," *Microscopy and Microanalysis*, vol. 30, no. 1, pp. 151–159, 2024.

[6] N. Tsiknakis, E. Tzoras, I. Zerdes, G. C. Manikis, B. Acs, J. Hartman, T. Hatschek, T. Foukakis, and K. Marias, "Multiresolution self-supervised feature integration via attention multiple instance learning for histopathology analysis," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1–4, IEEE, 2023.

[7] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[8] S. Dasariraju, M. Huo, and S. McCalla, "Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm," *Bioengineering*, vol. 7, no. 4, p. 120, 2020.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.