An Ensemble Model for Acute Myeloid Leukemia Risk Stratification Recommendations by Combining Machine Learning with Clinical Guidelines

Ming-Siang Chang[1], **Xavier Cheng-Hong Tsai**[2,3,4,5], **Wen-Chien Chou**[2,7], **Hwei-Fang Tien**[2,8], **Hsin-An Hou**[2], and **Chien-Yu Chen**[1,6*]

[1] *Department of Biomechatronics Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei City 106319, Taiwan (R.O.C.)*
[2] *Division of Hematology, Department of Internal Medicine, National Taiwan University Hospital, No.1, Changde St., Zhongzheng Dist., Taipei City 100229, Taiwan (R.O.C.)*
[3] *Department of Medical Education and Research, National Taiwan University Hospital Yunlin Branch, No. 579, Sec. 2, Yunlin Rd., Douliu City, Yunlin County 640203, Taiwan (R.O.C.)*
[4] *Department of Medical Education and Research, National Taiwan University Hospital Yunlin Branch, No. 579, Sec. 2, Yunlin Rd., Douliu City, Yunlin County 640203, Taiwan (R.O.C.)*
[5] *Department of Hematological Oncology, National Taiwan University Cancer Center, National Taiwan University Hospital, No.57, Ln. 155, Sec. 3, Keelung Rd., Da'an Dist., Taipei City 106, Taiwan (R.O.C.)*
[6] *Smart Medicine and Health Informatics Program, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei City 106319, Taiwan (R.O.C.)*
[7] *Department of Laboratory Medicine, National Taiwan University Hospital, No.1, Changde St., Zhongzheng Dist., Taipei 100229, Taiwan (R.O.C.)*
[8] *Department of Internal Medicine, Far-Eastern Memorial Hospital, New Taipei City, Taiwan, No. 21, Section 2, Nanya S. Road, Banqiao District, New Taipei City, 220, Taiwan (R.O.C.)*

Abstract

Acute Myeloid Leukemia (AML) is a complex disease requiring accurate risk stratification for effective treatment planning. This study introduces an innovative ensemble machine learning model integrated with the European LeukemiaNet (ELN) 2022 recommendations to enhance AML risk stratification. The model demonstrated superior performance by utilizing a comprehensive dataset of 1,213 patients from National Taiwan University Hospital (NTUH) and an external cohort of 2,113 patients from UK-NCRI trials. On the external cohort, it improved a concordance index (c-index) from 0.61 to 0.64 and effectively distinguished three different risk levels with median hazard ratios ranging from 18% to 50% improved. Key insights were gained from the discovered significant features influencing risk prediction, including age, genetic mutations, and hematological parameters. Notably, the model identified specific cytogenetic and molecular alterations like *TP53*, *IDH2*, *SRSF2*, *STAG2*, *KIT*, *TET2*, and karyotype (-5, -7, -15, inv(16)), alongside age and platelet counts. Additionally, the study explored variations in the effectiveness of hematopoietic stem cell transplantation (HSCT) across different risk levels, offering new perspectives on treatment effects. In summary, this study develops an ensemble model based on the NTUH cohort to deliver improved performance in AML risk stratification, showcasing the potential of integrating machine learning techniques with medical guidelines to enhance patient care and personalized medicine.

Keywords: *Acute myeloid leukemia (AML), Ensemble model, Machine learning, Risk stratification, European LeukemiaNet*

*Corresponding author. *E-mail address:* chienyuchen@ntu.edu.tw

1. Introduction

Acute Myeloid Leukemia (AML) is a rapidly progressing cancer of the blood and bone marrow, necessitating precise risk stratification for personalized treatment strategies [1]. Traditional methods often relied on single or limited biomarkers for risk assessment. However, the complexity of AML shifted the focus towards integrating a variety of biomarkers for more nuanced risk evaluation.

The ELN recommendations [2,3] for AML utilize biomarkers like cytogenetics and gene mutations to categorize patients into different risk groups, acknowledging the complex interactions between these biomarkers. For example, *t(9;11)(p21.3;q23.3)* considered intermediate-risk takes precedence over rare, concurrent adverse risk gene mutations. In addition, *CEBPA* gene mutations considered favorable in ELN will lead to poor prognosis if they carry *WT1* simultaneously [4]. Therefore, developing a model that can consider multiple biomarkers is necessary.

In health machine learning (ML) models, clinician-initiated and non-clinician-initiated data play distinct roles [5]. Clinician-initiated data, although helpful, might lead models to mimic existing clinical decisions and reduce the efficacy of risk stratification models, which aim to identify patients at risk before clinical recognition. Non-clinician data, on the other hand, can provide objective insights to enhance clinical decision-making.

This study aims to enhance AML risk stratification at diagnosis by developing an ensemble machine learning model that combines predictions with the European LeukemiaNet (ELN) 2022 recommendations [2,3]. The model utilizes multiple biomarkers, such as genetics, age, gender, and hematologic data, to accurately differentiate risk levels.

2. Results

The study focused on enhancing risk stratification in Acute Myeloid Leukemia (AML) patients using the Ensemble (ML+ELN). This machine learning-based ensemble model (Ensemble (ML)) integrates predictions from the European LeukemiaNet (ELN) 2022 criteria. The model was developed using a dataset of 1,213 AML patients from the National Taiwan University Hospital (NTUH) (Dr. Cheng-Hong Tsai), established respectively in 50 experiments on the NTUH dataset using different training and validation set partitions to ensure robustness, and subsequently tested these models on the NCRI cohort [6–9] for generalizability.

The ensemble model demonstrated superior performance in differentiating risk levels on the validation set of the NTUH cohort and an external cohort of 2,113 patients from UK-NCRI trials [6–9]. The study conducted a detailed analysis of significant features (p-value < 0.05) to better understand the model's predictions. This included comparing two groups: one where both the ELN 2022 and ensemble model classified samples as adverse or favorable and another where they differed in their classification. In the adverse group, features like Age_60, *TP53*, and specific karyotypes (-5, -7, -17) led to a higher risk prediction by the ensemble model, while *IDH2*, *SRSF2*, and *STAG2* indicated a lower risk. For the favorable group, features like Plt_55, *KIT*, and inv(16) indicated a lower risk, while Age_60, *TET2*, and *SRSF2* suggested a higher risk.

Furthermore, the study assessed the impact of hematopoietic stem cell transplantation (HSCT) on patients classified into different risk levels. The findings revealed variations in HSCT effectiveness based on risk level predictions by the ELN 2022 criteria and the Ensemble (ML) model. For example, patients categorized as adverse by ELN 2022 but favorable by the Ensemble (ML) model showed reduced benefit from HSCT.

2.1. Performance on the validation set

To ensure the models' generalizability and robustness, the study conducted 50 separate experiments, each establishing the models with a distinct partition of the training and validation sets on the NTUH cohort, and subsequently tested these models on the NCRI cohort. On the validation set, the Ensemble (ML+ELN) model performed best on the weighted F1 score, concordance index (c-index) [10], distinguishing different risk levels (p-value for logrank test [10,11]) (Table 1, Fig. 1). Those patients with identical risk levels showed outstanding performance compared to all patients

(Table 2). Besides, the published models were selected through the best c-index on the validation set in 50 experiments (Fig. 2).

Table 1. The performance of each model on the validation set (n=160) with a mean and standard deviation in 50 times experiments.

Model	Accuracy	Weighted F1 score[1]	c-index [10]
ELN 2022	0.49 ± 0.03	0.49 ± 0.03	0.63 ± 0.02
Logistic Regression	0.48 ± 0.03	0.47 ± 0.03	0.64 ± 0.02
KNN	0.45 ± 0.03	0.45 ± 0.03	0.62 ± 0.02
SVM	0.48 ± 0.03	0.48 ± 0.03	0.64 ± 0.02
Random Forest	0.49 ± 0.03	0.48 ± 0.03	0.63 ± 0.02
XGBoost	0.47 ± 0.03	0.47 ± 0.03	0.62 ± 0.02
LightGBM	0.48 ± 0.03	0.48 ± 0.03	0.63 ± 0.02
1D-CNN	0.46 ± 0.04	0.45 ± 0.04	0.62 ± 0.02
Ensemble (ML)	**0.50 ± 0.03**	0.49 ± 0.03	0.64 ± 0.02
Ensemble (ML + ELN)	0.49 ± 0.03	**0.50 ± 0.03**	**0.65 ± 0.02**

[1] The weighted F1 score calculates the F1 score for each class independently and uses a weight that depends on the number of true instances for each class to average the prediction.

2.2. Performance on the external cohort

The Ensemble (ML+ELN) model performed best on all evaluation methods on the external cohort. The model showcased its effectiveness in survival analysis, improving an average concordance index (c-index) [10] from 0.61 to 0.64 (Table 3, Fig. 3), and the median hazard ratios [12]: 1.85 to 2.2 for adverse versus intermediate groups (18% improved), 1.53 to 1.97 for intermediate versus favorable groups (29% improved), and 2.83 to 4.24 for adverse versus favorable groups (50% improved) (Fig. 4a). Furthermore, it also demonstrated its capability to differentiate between various groups. This differentiation was evident in the model's survival curves, with median p-values for logrank test [10,11] ranging from 1.37e-22 to 3.01e-41 for adverse versus intermediate groups and 8.31e-8 to 2.35e-19 for intermediate versus favorable groups (Fig. 4b).

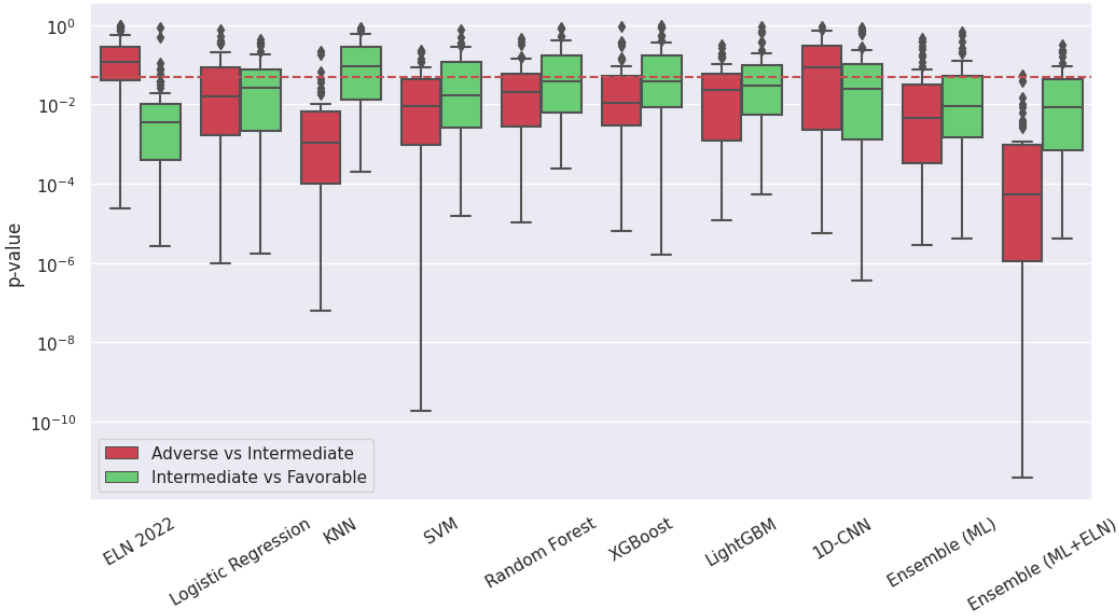

Figure 1. Distribution of p-values for logrank test between different survival curves for 50 experiments in different models on the validation set. The red dashed line has a p-value of 0.05.

Table 2. The performance of the Ensemble (ML+ELN) model for the samples where the Ensemble (ML) model and ELN 2022 predict identical risk levels. The values are mean and standard deviation in 50 times of experiments.

Model	Ratio of identical predictions	Accuracy	Weighted F1 score[1]
Ensemble (ML + ELN)	0.47 ± 0.03	0.63 ± 0.04	0.62 ± 0.05

[1] The weighted F1 score calculates the F1 score for each class independently and uses a weight that depends on the number of true instances for each class to average the prediction.

Table 3. The performance on the NCRI cohort

Model	c-index with mean and standard deviation in 50 experiments[1]	c-index for the published models [2]
ELN 2022	0.61 ± 0.00	0.612
Ensemble (ML)	0.64 ± 0.01	0.642
Ensemble (ML+ELN)	**0.64 ± 0.00**	**0.645**

[1] Each model was trained in 50 experiments on the NTUH cohort, using different training and validation partitions. The whole NCRI cohort was used to test the performance in each experiment.
[2] Published models were selected through the best c-index on the validation set in 50 experiments.

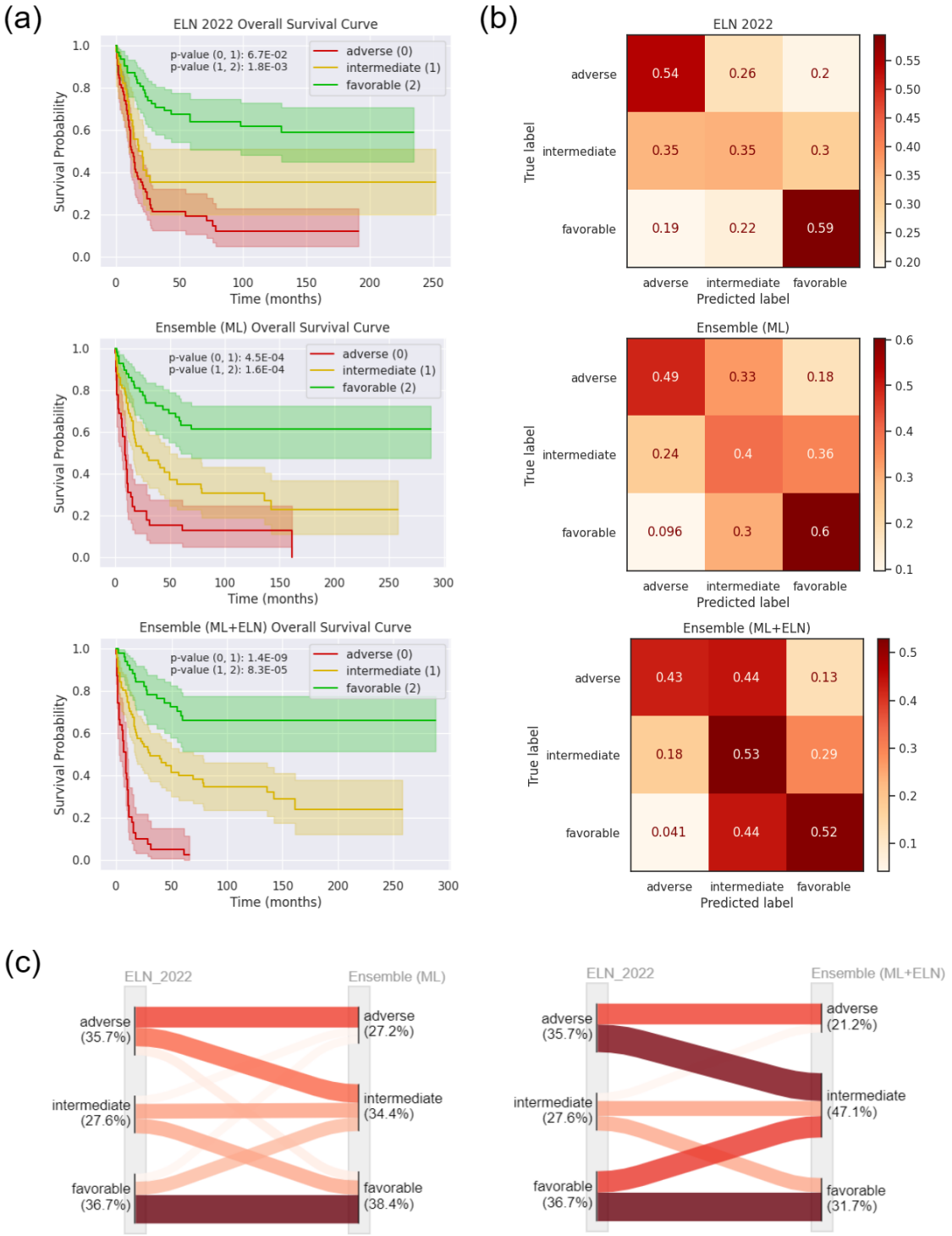

Figure 2. Evaluation matrices on the validation set. (a) Survival curves for the published models. (b) Confusion matrices from cumulative results of 50 experiments. (c) Sample distribution flows from the cumulative results of 50 experiments.

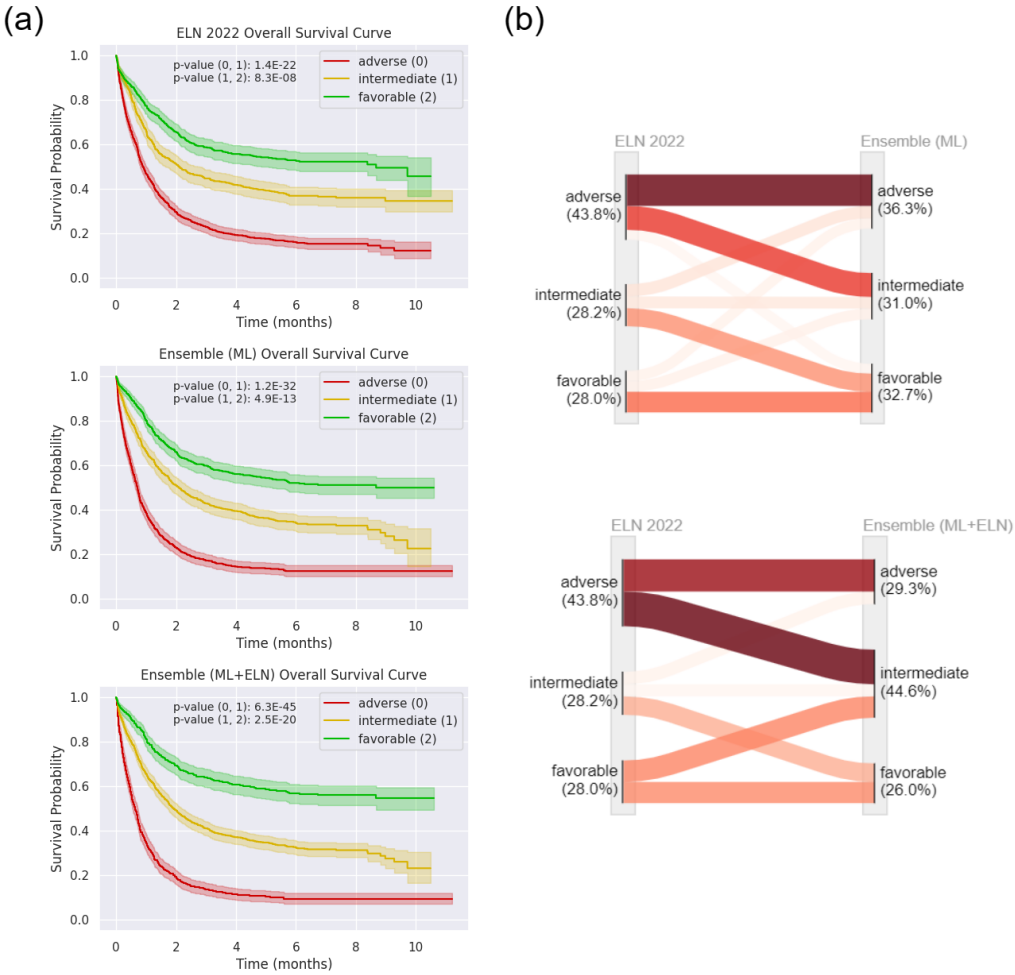

Figure 3. Evaluation matrices on the NCRI cohort. (a) Survival curves for the published models. (b) Sample distribution flows from the cumulative results of 50 experiments.

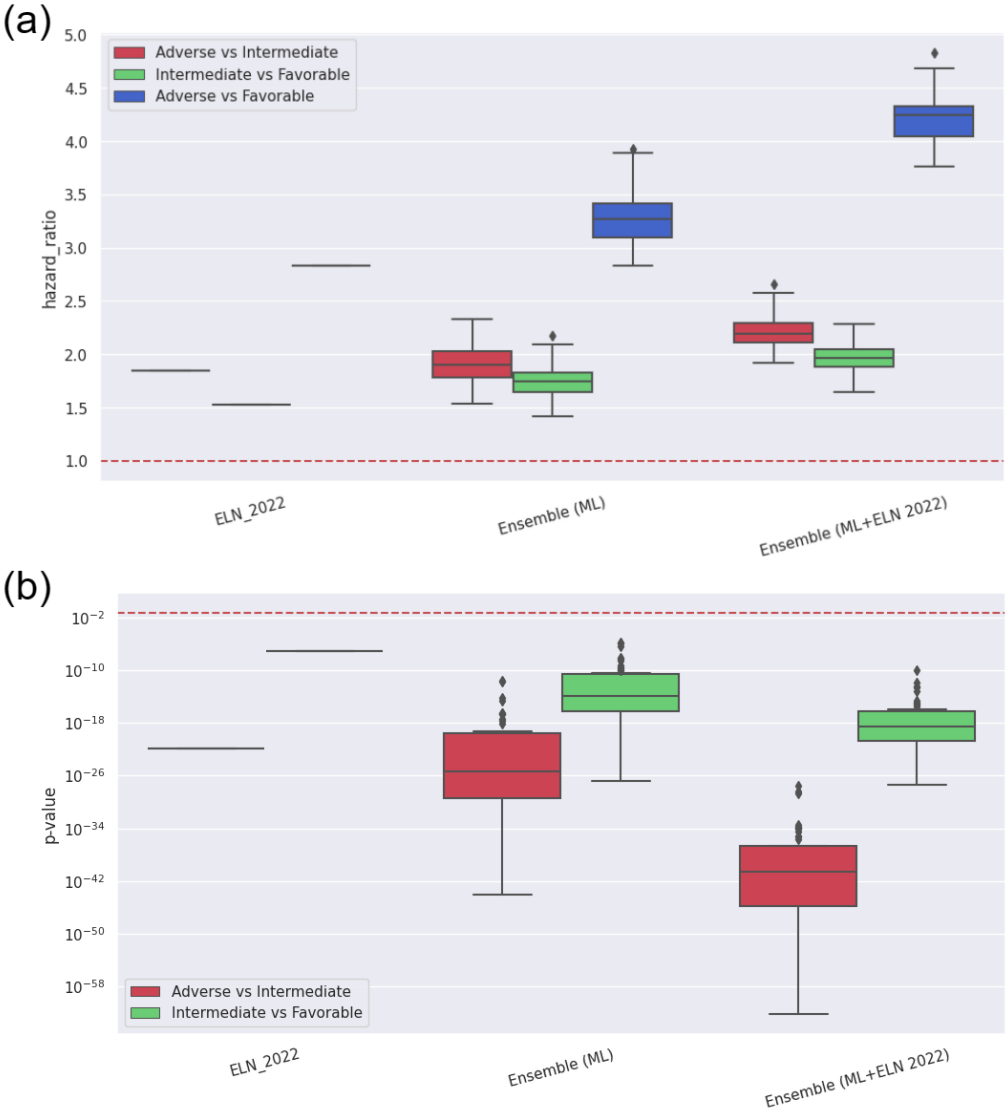

Figure 4. Effect sizes in different models on the external cohort. (a) Distribution of hazard ratios for different risk levels for 50 experiments in different models. The red dashed line has a hazard ratio of 1. (b) Distribution of p-values for logrank test between different survival curves for 50 experiments in different models on the external cohort. The red dashed line has a p-value of 0.05.

2.3. Insights for patients predicted as adverse by ELN 2022

The study assessed significant features (p-value < 0.05) in two groups using different statistical tests: T-tests [13] for numerical features and chi-square tests [14] for categorical ones. One group included samples classified as adverse by both the ELN 2022 and ensemble model, while the other consisted of samples identified as adverse by ELN 2022 but not by the ensemble model. The survival curves for these groups are illustrated in Fig. 5a.

In 50 experiments, certain significant features repeatedly appeared in an external cohort (Fig. 5c), and distribution of p-values for logrank test [10,11] between the survival curves of the two groups suggested that the Ensemble (ML+ELN) model was more effective in identifying patients with higher risk (Fig. 5b).

An odds ratio was used to demonstrate the impact of these significant features on risk prediction (Fig. 5d). The feature 'age' was categorized with a threshold of 60 years (Age_60). Features like Age_60, *TP53*, and specific karyotypes (-5, -7, -17) led the Ensemble (ML+ELN) model to categorize patients as adverse. In contrast, features such as *IDH2*, *SRSF2*, and *STAG2* led to a non-adverse prediction. Therefore, for patients identified as adverse in ELN 2022, determining if they are older than 60 or carry *TP53*, or the aforementioned karyotypes can help further identify a more adverse risk group while checking for *IDH2*, *SRSF2*, or *STAG2* can help identify a less adverse group.

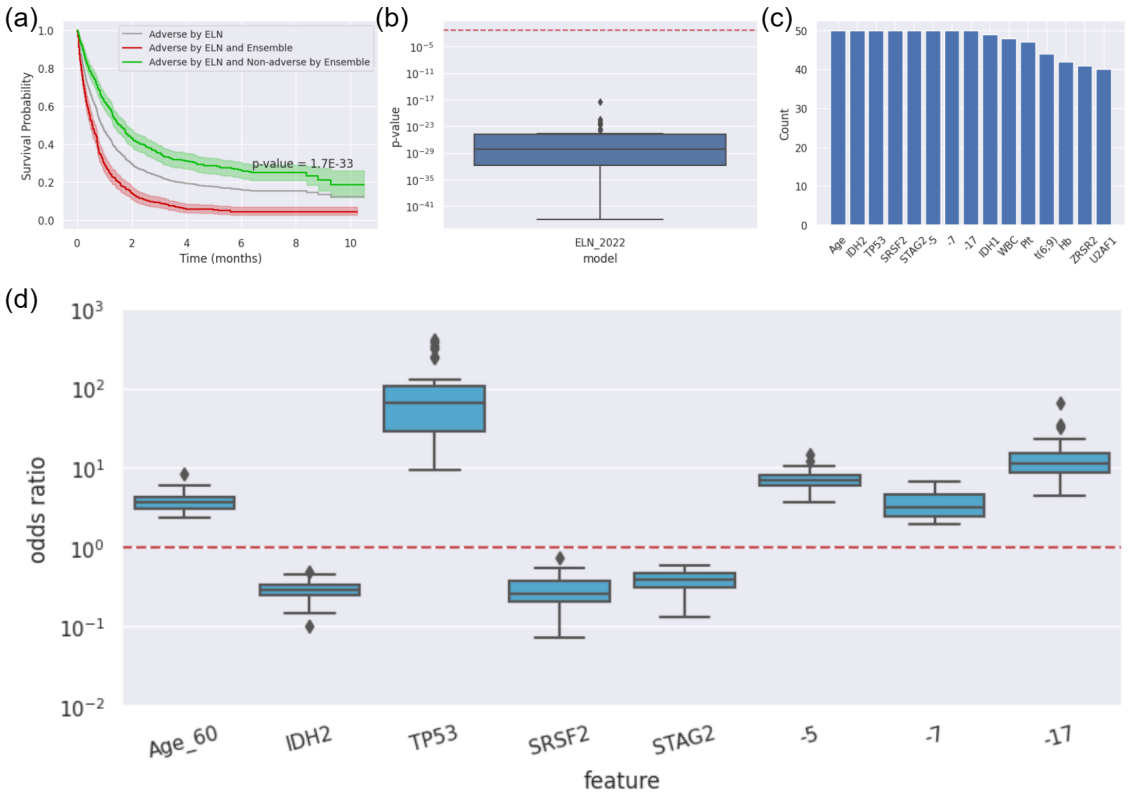

Figure 5. Biomarkers for further distinguishing patients identified as adverse according to the ELN 2022 guideline from the external cohort. (a) Survival curves for the published models where patients were classified as adverse by ELN 2022. The p-value results from the logrank test between the red and green survival curves. (b) Distribution of p-values for logrank test between the survival curves of two patient groups in 50 experiments, where one group is classified as adverse by both ELN 2022 and an Ensemble (ELN+ML) model, and the other is classified as adverse by ELN 2022 but non-adverse by the Ensemble (ELN+ML) model. The red dashed line has a p-value of 0.05. (c) The Significant biomarkers (p-value$<$0.05) were observed in over 40 of 50 experiments. (d) Analyzing the effect of consistently occurred significant biomarkers by odds ratio [15]. The red dashed line has an odds ratio of 1.

Integrating the findings from the previous literature with the current study offers a comprehensive understanding of risk stratification in Acute Myeloid Leukemia (AML) patients. Studies have

shown that older AML patients exhibit distinct genetic alterations compared to younger patients, with mutations like *TP53* being poor prognostic factors [16,17]. In addition, the *TP53* mutation, known for its stability during AML evolution, is associated with poor prognosis, especially in patients with complex karyotypes [17]. Contrarily, mutations in the cohesin complex gene *STAG2* have contributed to the disease's complexity and showed a trend toward improved long-term outcomes [18]. Moreover, the earlier research highlighted the prognostic significance of various genetic alterations and cytogenetic profiles, such as the favorable impact of *IDH2* mutations and the negative implications of an unfavorable karyotype (-5, -7, -17) on treatment outcomes and survival rates in AML. [19].

2.4. Insights for patients predicted as favorable by ELN 2022

The study assessed significant features (p-value < 0.05) in two groups using the statistical test in Chapter 2.3. One group included samples classified as favorable by both the ELN 2022 and ensemble model, while the other consisted of samples identified as favorable by ELN 2022 but not by the ensemble model. The survival curves for these groups are illustrated in Fig. 6a.

In 50 experiments, certain significant features repeatedly appeared in an external cohort (Fig. 6c), and distribution of p-value for logrank test [10,11] between the survival curves of the two groups suggested that the Ensemble (ML+ELN) model was more effective in identifying patients with low risk (Fig. 6b).

An odds ratio was used to demonstrate the impact of these significant features on risk prediction (Fig. 6d). The feature 'age' was categorized with a threshold of 60 years (Age_60), and 'Plt' (Platelets) was categorized with a threshold of 55 thousand per microliter (Plt_55). Features like Plt_55, *KIT*, and inv(16) led the Ensemble (ML+ELN) model to categorize patients as favorable. In contrast, features such as Age_60, *TET2*, and *SRSF2* led to a non-favorable prediction. Therefore, for patients identified as favorable in ELN 2022, determining if they are younger than 60, or the concentration of platelets more than 55 thousand per microliter, or carry *KIT* or inv(16) can help further identify a more favorable risk group while checking for *TET2*, *SRSF2* can help identify a less favorable group.

The recent studies on acute myeloid leukemia (AML) have underscored the importance of genetic alterations and clinical features in patient risk stratification. Older AML patients exhibit distinct genetic profiles with mutations like *DNMT3A* and *TP53* being poor prognostic indicators [16]. Platelet count at diagnosis is also a critical factor influencing treatment outcomes [20]. The findings regarding *TET2* mutation in acute myeloid leukemia (AML) patients, particularly those with intermediate-risk cytogenetics, underscore its role as an unfavorable prognostic factor [21]. Moreover, myelodysplasia-related mutations like *SRSF2* indicate poorer outcomes across all age groups, underscoring the need for intensive treatment [22]. Another earlier research highlighted the prognostic importance of favorable karyotypes inv(16) in AML [19].

2.5. Variations in Hematopoietic Stem Cell Transplantation Effectiveness Across Different Risk Levels

The study assessed the impact of hematopoietic stem cell transplantation (HSCT) on patients classified into different risk levels as part of the NCRI cohort study. Of the 2,113 patients in the cohort, 527 underwent HSCT. The study utilized hazard ratios [12] to analyze outcomes based on risk assessments made by two different models: the ELN 2022 criteria and the Ensemble (ML) model (Fig. 7).

The findings revealed notable variations in the effectiveness of HSCT depending on the risk level predictions by these two models. Specifically:

1. For patients categorized as adverse according to ELN 2022 but favorable by the Ensemble (ML) model, HSCT showed a reduced benefit. This HSCT effectiveness was comparable to, or even less than, that observed in patients deemed favorable by ELN 2022. The survival curve from the published models for these patients is shown in Fig. 7b.

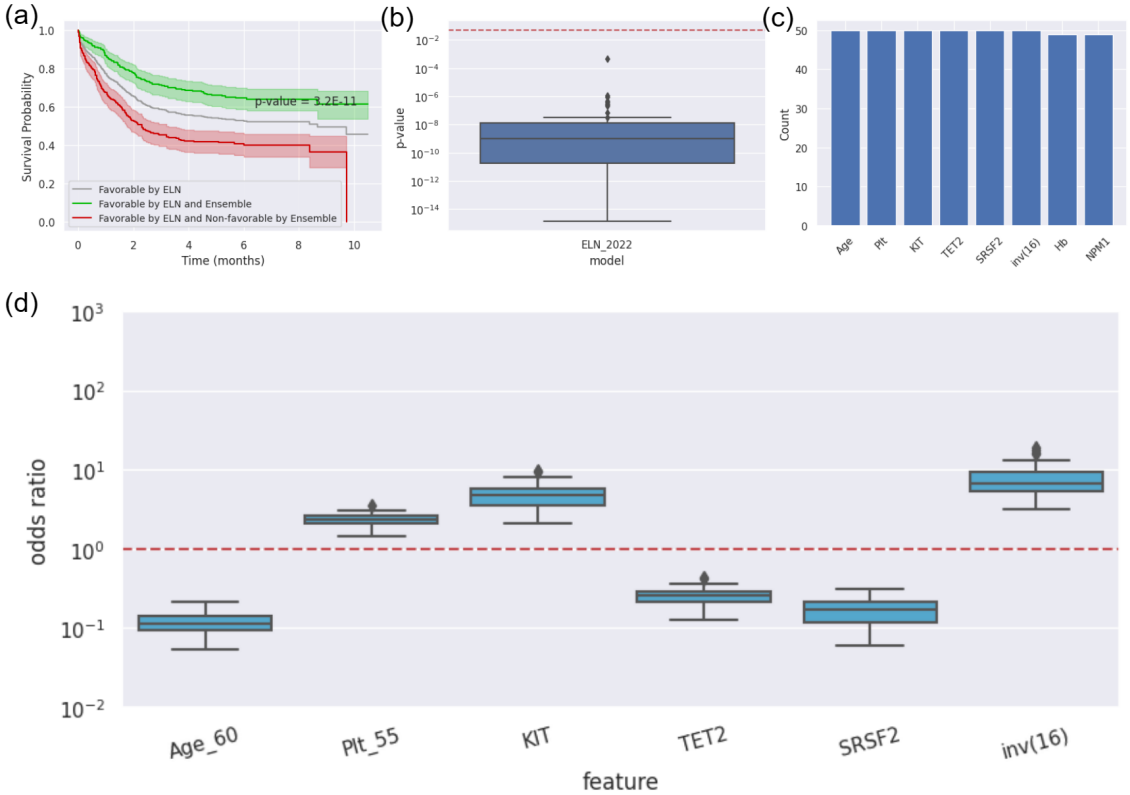

Figure 6. Biomarkers for further distinguishing patients identified as favorable according to the ELN 2022 guidelines from the external cohort. (a) Survival curves for the published models where patients were classified as favorable by ELN 2022. The p-value results from the logrank test between the red and green survival curves. (b) Distribution of p-values for logrank test between the survival curves of two patient groups in 50 experiments, where one group is classified as favorable by both ELN 2022 and an Ensemble (ELN+ML) model, and the other is classified as favorable by ELN 2022 but non-favorable by the Ensemble (ELN+ML) model. The red dashed line has a p-value of 0.05. (c) The Significant biomarkers (p-value<0.05) were observed in over 40 of 50 experiments. (d) Analyzing the effect of consistently occurred significant biomarkers by odds ratio [15]. The red dashed line has an odds ratio of 1.

2. Conversely, patients identified as favorable by ELN 2022 but adverse by the Ensemble (ML) model experienced a significantly enhanced benefit from HSCT. This effectiveness was similar to or greater than that seen in patients classified as adverse by ELN 2022. The survival curve from the published models for these patients is shown in Fig. 7d.

3. Additionally, patients rated as intermediate-risk by ELN 2022 and adverse by the Ensemble (ML) model showed a level of HSCT effectiveness similar to part of patients categorized as adverse by ELN 2022. The survival curve from the published models for these patients is shown in Fig. 7c.

Our study expands on prior research by redefining risk groups in HSCT outcomes, incorporating both ELN 2022 criteria and Ensemble (ML) model predictions. This approach offers a more detailed risk assessment, particularly identifying a subset within the NCRI cohort that responds variably to HSCT. While these findings are significant, they should be interpreted cautiously due to potential selection biases in retrospective studies. Prospective studies are needed to validate and further explore these insights.

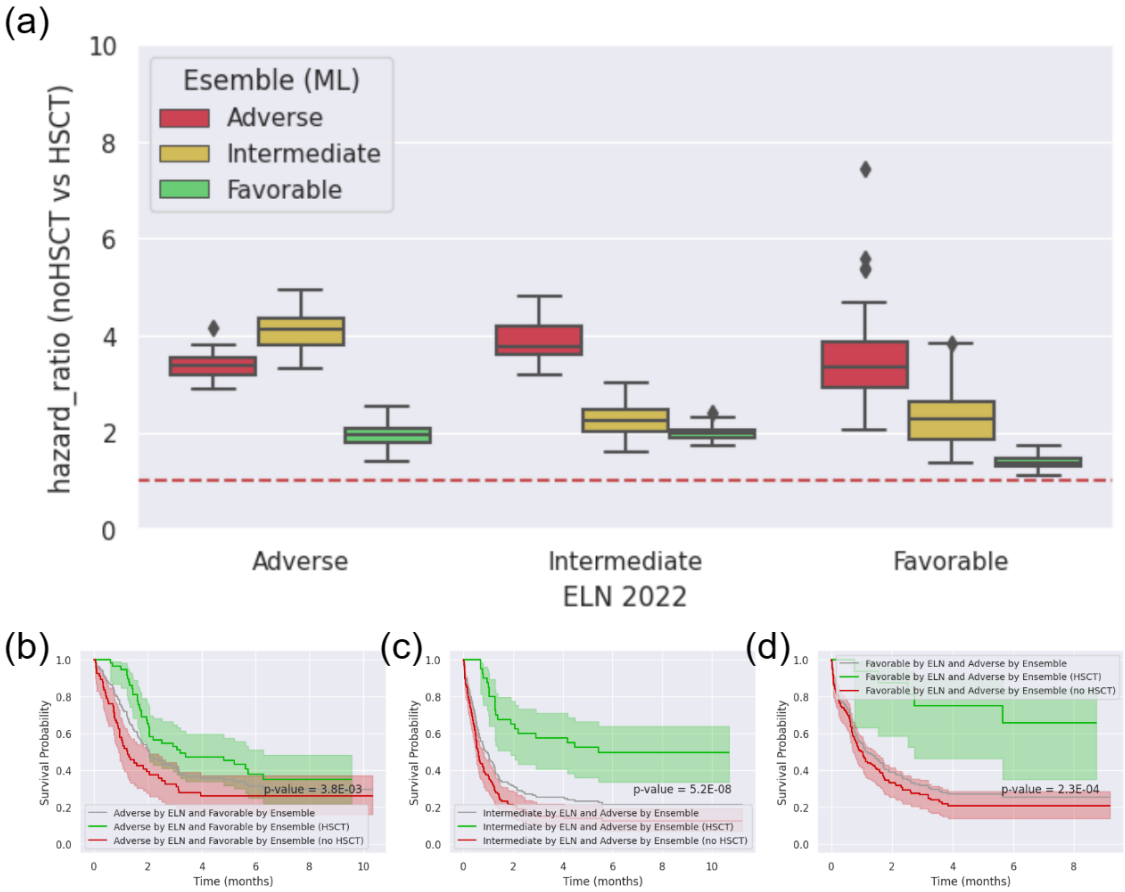

Figure 7. Comparative analysis of the effectiveness of HSCT according to ELN 2022 and the Ensemble (ML) model in predicting different risk levels on the external cohort. (a) Overall HSCT effectiveness on the NCRI cohort. (b) Survival curves for patients predicted as adverse by ELN 2022 but favorable by the published Ensemble (ML) model. The p-value results from the logrank test between the red and green survival curves. (c) Survival curves for patients predicted as intermediate by ELN 2022 and adverse by the published Ensemble (ML) model. The p-value results from the logrank test between the red and green survival curves. (d) Survival curves for patients predicted as favorable by ELN 2022 but adverse by the published Ensemble (ML) model. The p-value results from the logrank test between the red and green survival curves.

3. Methods

This study employed datasets from 1,213 patients from the National Taiwan University Hospital (NTUH) AML cohort and another from 2,113 patients from three UK-NCRI trials [6–9] as an external cohort. The NTUH cohort was filtered based on specific criteria and divided into training and validation sets to establish the proposed models. Patient survival duration determined labels, and various patient data, including karyotypes and gene mutations, were used as predictive features. The data were then normalized and preprocessed.

The proposed prediction model involved many well-known classification techniques. In addition, HyperOpt [23], a Bayesian hyperparameter optimization, was used to select the optimal hyperparameters of each model. Finally, an ensemble (ML) model considers each model to predict risk.

After integrating the classification techniques, the proposed prediction model jointly considers the results and the results predicted by ELN 2022. Based on this, the study proposed initial clinical risk stratification recommendations.

3.1. Preprocessing the dataset for establishing models

The dataset for establishing models came from an acute myeloid leukemia cohort with 1,213 samples provided by Dr. Cheng-Hong Tsai, Department of Hematology, NTUH, including clinical, hematological, karyotype, and gene mutation data. After excluding those who did not get standard therapy and those who survived with follow-up time that did not surpass 36 months, the remaining 801 samples were separated into an 80% train set and 20% validation set for prediction and evaluation (Fig. 8). The exclusion was necessary as it could not confirm whether the short follow-up period may have led to these individuals being categorized into a poorer risk group.

The label of samples was defined by overall survival; less than 12 months is an adverse risk, between 12 months to 60 months is an intermediate risk, and more than 60 months is a favorable risk. The label distribution is shown in Table 4.

The 13 types of karyotypes on the ELN recommendations for initial diagnosis were selected as features. In addition, age, gender, four types of hematological data, and 31 types of gene mutations were also included. A previous study showed that the preprocessing techniques employed have no significant effect when none of the features have more than 10% missing values [24]. The standard normalization [25] was applied to age and hematological data. The data preprocessing and feature selection are shown in Fig. 8.

Table 4. The distribution of overall survival label for 801 samples

Overall survival label	Number (n=801)	Percentage (%)
Adverse (< 12 months)	254	31.71
Intermediate	282	35.21
Favorable (> 60 months)	265	33.08

3.2. Models

The study employs various machine learning models [25] with unique characteristics, as learned from a dataset where the relationships between features and labels are known. Logistic Regression [26] estimates the logistic model's parameters to predict the likelihood of an event by making its log odds a linear combination of independent variables. It allows feature selection and adjustment of coefficients for unnecessary features using penalties like L1, L2, or Elastic-Net.

K-nearest Neighbors (KNN) [27] is a simple, instance-based learning model that computes the distance between a new observation and existing instances, using the 'K' closest instances for prediction. It's adaptable to input changes and works well with multi-class cases and irregular decision boundaries.

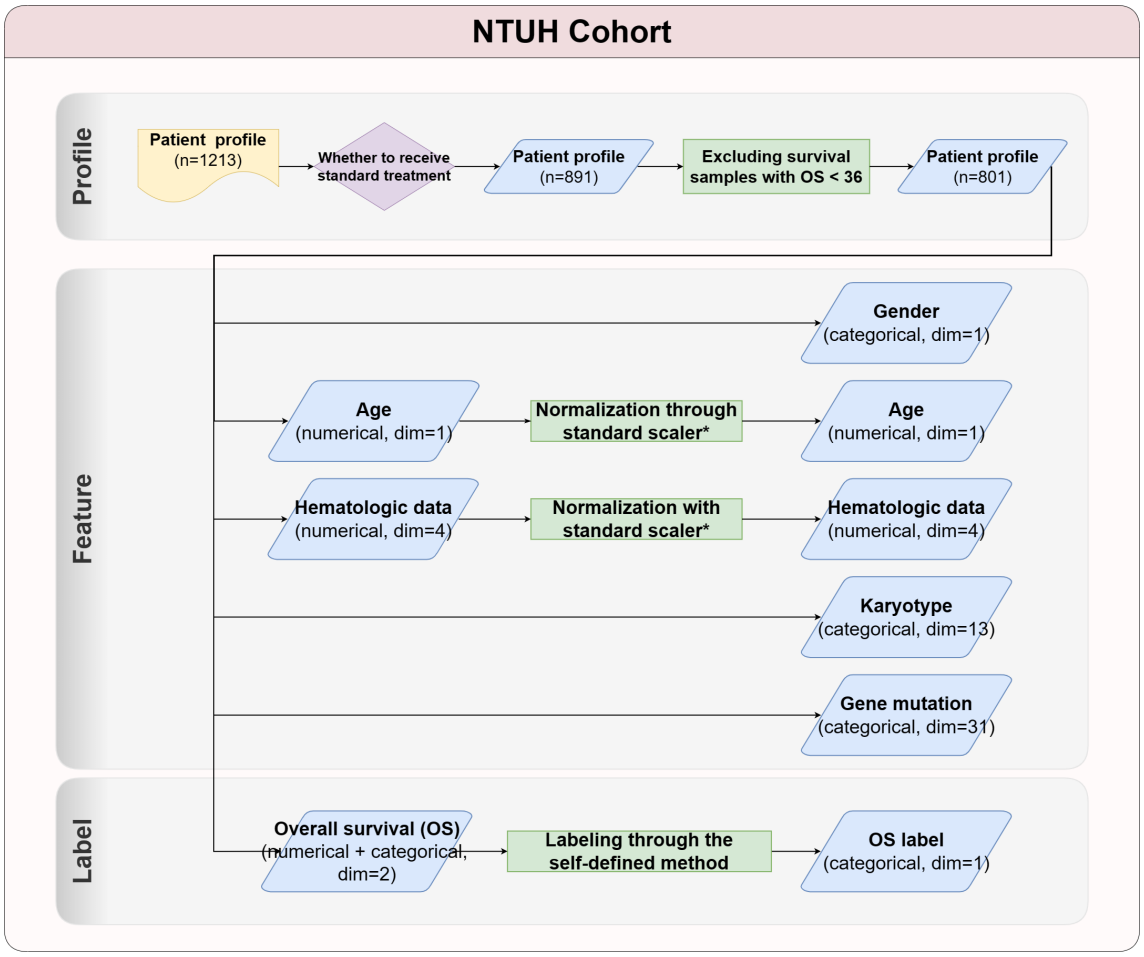

Figure 8. Feature selection and data preprocessing on the NTUH cohort. *Standard scaler is a normalization method in scikit-learn [25].

Support Vector Machine (SVM) [28] is effective in high-dimensional spaces, seeking a hyperplane with the maximum margin for classification. It uses the kernel trick for non-linear decision boundaries and is robust against overfitting but can be computationally intensive.

Random Forests [29], an ensemble model, constructs many decision trees and outputs the mode or mean prediction of these trees. It introduces randomness in training and considers a random subset of features for splitting, making it effective against overfitting and capable of handling many input variables.

XGBoost [30], based on gradient boosting, uses decision tree ensembles and systematically corrects errors, offering fast computation and a regularization parameter to minimize overfitting. It's adaptable to various prediction tasks with custom optimization objectives.

LightGBM [31] grows trees leaf-wise and employs Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) for efficiency and scalability. It's more accurate but can overfit on smaller datasets.

Lastly, 1D-CNN, adapted from Convolutional Neural Networks [32] for sequential data like time series and natural language, captures local patterns in sequences, handles long sequences efficiently, and extracts local features effectively.

3.3. Hyperparameters Optimization

Hyperparameter optimization is a crucial step that impacts the success of a model in machine learning. Hyperparameters are set before the training process and significantly influence a model's learning effectiveness. The optimization process involves searching for the best hyperparameter configurations, which can be computationally intensive and time-consuming.

This study used Hyperopt [23] as an efficient tool for streamlining hyperparameter optimization. Hyperopt uses the Tree-structured Parzen Estimator (TPE) method, which constructs a probabilistic model of the objective function to guide the search, leading to a more efficient process and often superior results. Hyperopt can be easily integrated into machine learning workflows, significantly reducing the resources and time required for optimization and improving the model's overall effectiveness. Table A in the Appendix shows the hyperparameter searching spaces of each model.

3.4. Machine Learning-based Ensemble Model (Ensemble (ML))

This study employed the loss function-based approach as the machine learning-based ensemble model (Ensemble (ML)). This technique optimized the combination of models by leveraging a loss function, which measured the deviation of predicted values from actual values.

In the approach, each base model in the ensemble was assigned a weight based on its performance as measured by the cross-entropy loss function [33]. The ensemble then combined these models' predictions according to their weights to make the final prediction, enhancing prediction accuracy. The formulas Eq. 1 and Eq. 2 are shown below:

$$\text{weight}_i = \frac{\exp(-\text{loss}_i)}{\sum_{k \in M} \exp(-\text{loss}_k)} \tag{1}$$

$$\text{prediction}_{\text{ensemble}} = \arg\max_j \left(\sum_{i \in M} \text{weight}_i \cdot \text{probability}_{i,j} \right) \tag{2}$$

where M represents the set of models used in the ensemble, i denotes an individual model within this ensemble, weight_i is the weight assigned to model i, $\text{probability}_{i,j}$ is the probability assigned to class j by model i, and j represents the class. The function $\arg\max_j$ selects the class j for which the sum of weighted probabilities across all models is maximized.

In essence, ensemble models, particularly loss function-based ones, utilize a mechanism to fine-tune the combination of multiple models. This bridges the gap between simple voting schemes and a

more optimized, intelligent way of ensemble learning, thus improving prediction performance across various tasks.

3.5. Clinical Risk Stratification Recommendations by the Combination of Ensemble Model and ELN 2022 (Ensemble (ML+ELN))

After predicting by the Ensemble (ML) model, the study established initial clinical risk stratification recommendations by considering the Ensemble (ML) model and ELN 2022 risk prediction (Ensemble (ML+ELN)). The proposed clinical risk stratification recommendations are shown in Fig. 9.

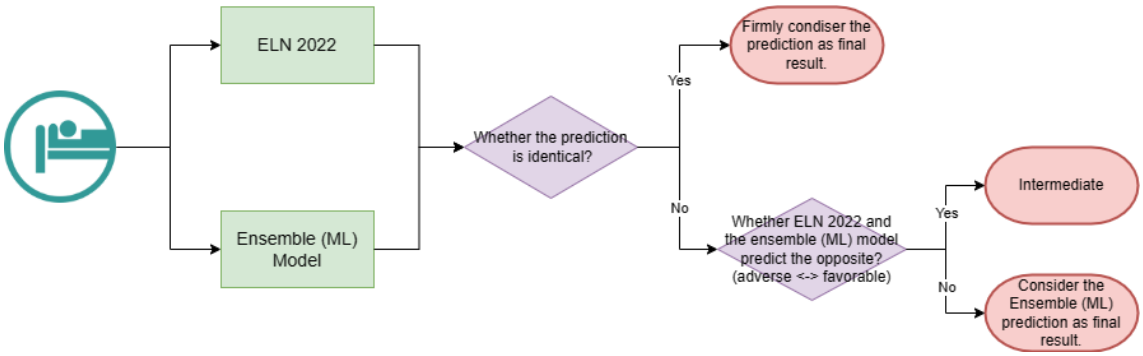

Figure 9. The proposed clinical risk stratification used in the Ensemble (ML+ELN) model.

The Ensemble (ML+ELN) model considered patients' risk levels by the following step.

1. The ELN 2022 and Ensemble (ML) models predict patient risk levels, and the Ensemble (ML+ELN) model then confidently predicts those patients with the same risk level.
2. For samples contrary to the prediction, the Ensemble (ML+ELN) model considered them intermediate. That is to say; for samples predicted as adverse by ELN 2022 and favorable by the Ensemble (ML) model or those predicted by ELN 2022 as adverse and favorable by the Ensemble (ML) model, the Ensemble (ML+ELN) model treated them as intermediate.
3. The rest treated the Ensemble (ML) prediction as the final result.

4. Conclusions

This study presents a significant advancement in the risk stratification of Acute Myeloid Leukemia (AML) patients through the development and validation of an ensemble machine learning model. This model, integrated with the European LeukemiaNet (ELN) 2022 recommendations, has shown remarkable efficacy in accurately categorizing AML patients into distinct risk groups. The application of this model promises to refine the current prognostic assessments and aid in personalized treatment planning.

The use of a comprehensive dataset, including 1,213 patients from the National Taiwan University Hospital and an external cohort of 2,113 patients from UK-NCRI trials, underscores the robustness and generalizability of the model. Including a diverse range of biomarkers such as karyotypes, gene mutations, and hematologic data enhances the model's precision in risk prediction, demonstrating its superiority over traditional single-biomarker approaches.

The study successfully delineated key genetic and clinical features influencing AML prognosis through detailed analysis. In patients predicted as adverse by ELN 2022, factors such as age above 60, *TP53* mutations, and specific karyotypes (e.g., -5, -7, -17) were linked with higher risk, while mutations like *IDH2*, *SRSF2*, and *STAG2* indicated a lower risk. Conversely, for patients deemed favorable by ELN 2022, features like a platelet count above 55,000 per microliter, *KIT* mutations, and inv(16) karyotype were associated with favorable outcomes, whereas older age and mutations

such as *TET2* and *SRSF2* suggested a less favorable prognosis. These insights underscore the nuanced nature of AML risk stratification and the model's capacity to capture these complexities.

The analysis of hematopoietic stem cell transplantation (HSCT) effectiveness across different risk levels offers novel perspectives on treatment effects. The study's findings highlight the model's potential to identify subsets of patients who might benefit more or less from HSCT, thus guiding treatment choices and improving outcomes.

In conclusion, this thesis represents a significant leap forward in AML risk stratification, providing a powerful tool for personalized medicine. Integrating advanced machine learning techniques with established medical guidelines opens new avenues for enhancing patient care in AML and potentially other complex diseases. The study underscores the importance of leveraging technology in healthcare to achieve more precise and effective patient management.

References

[1] C.-H. Tsai, "Applying next-generation sequencing to explore the risk stratification in acute myeloid leukemia patients," Thesis, 2021.

[2] H. Döhner, E. Estey, D. Grimwade, S. Amadori, F. R. Appelbaum, T. Büchner, H. Dombret, B. L. Ebert, P. Fenaux, and R. A. Larson, "Diagnosis and management of aml in adults: 2017 eln recommendations from an international expert panel," *Blood, The Journal of the American Society of Hematology*, vol. 129, no. 4, pp. 424–447, 2017.

[3] H. Döhner, A. H. Wei, F. R. Appelbaum, C. Craddock, C. D. DiNardo, H. Dombret, B. L. Ebert, P. Fenaux, L. A. Godley, and R. P. Hasserjian, "Diagnosis and management of aml in adults: 2022 recommendations from an international expert panel on behalf of the eln," *Blood, The Journal of the American Society of Hematology*, vol. 140, no. 12, pp. 1345–1377, 2022.

[4] F.-M. Tien, H.-A. Hou, J.-L. Tang, Y.-Y. Kuo, C.-Y. Chen, C.-H. Tsai, M. Yao, C.-T. Lin, C.-C. Li, and S.-Y. Huang, "Concomitant wt1 mutations predict poor prognosis in acute myeloid leukemia patients with double mutant cebpa," *Haematologica*, vol. 103, no. 11, p. e510, 2018.

[5] B. K. Beaulieu-Jones, W. Yuan, G. A. Brat, A. L. Beam, G. Weber, M. Ruffin, and I. S. Kohane, "Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?" *NPJ digital medicine*, vol. 4, no. 1, p. 62, 2021.

[6] Y. Tazi, J. E. Arango-Ossa, Y. Zhou, E. Bernard, I. Thomas, A. Gilkes, S. Freeman, Y. Pradat, S. J. Johnson, and R. Hills, "Unified classification and risk-stratification in acute myeloid leukemia," *Nature Communications*, vol. 13, no. 1, p. 4622, 2022.

[7] A. Burd, R. L. Levine, A. S. Ruppert, A. S. Mims, U. Borate, E. M. Stein, P. Patel, M. R. Baer, W. Stock, and M. Deininger, "Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the beat aml master trial," *Nature medicine*, vol. 26, no. 12, pp. 1852–1858, 2020.

[8] A. K. Burnett, N. H. Russell, R. K. Hills, J. Kell, J. Cavenagh, L. Kjeldsen, M.-F. McMullin, P. Cahalin, M. Dennis, and L. Friis, "A randomized comparison of daunorubicin 90 mg/m2 vs 60 mg/m2 in aml induction: results from the uk ncri aml17 trial in 1206 patients," *Blood, The Journal of the American Society of Hematology*, vol. 125, no. 25, pp. 3878–3885, 2015.

[9] A. K. Burnett, R. K. Hills, A. Hunter, D. Milligan, W. Kell, K. Wheatley, J. Yin, M. McMullin, H. Dignum, and D. Bowen, "The addition of gemtuzumab ozogamicin to low-dose ara-c improves remission rate but does not significantly prolong survival in older patients with acute myeloid leukaemia: results from the lrf aml14 and ncri aml16 pick-a-winner comparison," *Leukemia*, vol. 27, no. 1, pp. 75–81, 2013.

[10] C. Davidson-Pilon, "lifelines: survival analysis in python," *Journal of Open Source Software*, vol. 4, no. 40, p. 1317, 2019.

[11] J. M. Bland and D. G. Altman, "The logrank test," *Bmj*, vol. 328, no. 7447, p. 1073, 2004.

[12] S. L. Spruance, J. E. Reid, M. Grace, and M. Samore, "Hazard ratio in clinical trials," *Antimicrobial agents and chemotherapy*, vol. 48, no. 8, pp. 2787–2792, 2004.

[13] T. K. Kim, "T test as a parametric statistic," *Korean journal of anesthesiology*, vol. 68, no. 6, pp. 540–546, 2015.

[14] T. M. Franke, T. Ho, and C. A. Christie, "The chi-square test: Often used and more often misinterpreted," *American journal of evaluation*, vol. 33, no. 3, pp. 448–458, 2012.

[15] J. M. Bland and D. G. Altman, "The odds ratio," *Bmj*, vol. 320, no. 7247, p. 1468, 2000.

[16] C.-H. Tsai, H.-A. Hou, J. Tang, C. Liu, C. Lin, W. Chou, M. Tseng, Y. Chiang, Y. Kuo, and M. Liu, "Genetic alterations and their clinical implications in older patients with acute myeloid leukemia," *Leukemia*, vol. 30, no. 7, pp. 1485–1492, 2016.

[17] H.-A. Hou, W. Chou, Y. Kuo, C. Liu, L. Lin, M. Tseng, Y. Chiang, M. Liu, C. Liu, and J. Tang, "Tp53 mutations in de novo acute myeloid leukemia patients: longitudinal follow-ups show the mutation is stable during disease evolution," *Blood cancer journal*, vol. 5, no. 7, pp. e331–e331, 2015.

[18] C.-H. Tsai, H.-A. Hou, J.-L. Tang, Y.-Y. Kuo, Y.-C. Chiu, C.-C. Lin, C.-Y. Liu, M.-H. Tseng, T.-Y. Lin, and M.-C. Liu, "Prognostic impacts and dynamic changes of cohesin complex gene mutations in de novo acute myeloid leukemia," *Blood cancer journal*, vol. 7, no. 12, p. 663, 2017.

[19] H. Hou, C. Lin, W. Chou, C. Liu, C. Chen, J. Tang, Y. Lai, M. Tseng, C. Huang, and Y. Chiang, "Integration of cytogenetic and molecular alterations in risk stratification of 318 patients with de novo non-m3 acute myeloid leukemia," *Leukemia*, vol. 28, no. 1, pp. 50–58, 2014.

[20] Y. Zhang, Q. Wu, B. Yuan, Y. Huang, L. Jiang, F. Liu, P. Yan, Y. Jiang, J. Ye, and X. Jiang, "Influence on therapeutic outcome of platelet count at diagnosis in patients with de novo non-apl acute myeloid leukemia," *BMC cancer*, vol. 23, no. 1, p. 1030, 2023.

[21] W.-C. Chou, S.-C. Chou, C.-Y. Liu, C.-Y. Chen, H.-A. Hou, Y.-Y. Kuo, M.-C. Lee, B.-S. Ko, J.-L. Tang, and M. Yao, "Tet2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics," *Blood, The Journal of the American Society of Hematology*, vol. 118, no. 14, pp. 3803–3810, 2011.

[22] X. C.-H. Tsai, K.-J. Sun, M.-Y. Lo, F.-M. Tien, Y.-Y. Kuo, M.-H. Tseng, Y.-L. Peng, Y.-K. Chuang, B.-S. Ko, and J.-L. Tang, "Poor prognostic implications of myelodysplasia-related mutations in both older and younger patients with de novo aml," *Blood Cancer Journal*, vol. 13, no. 1, p. 4, 2023.

[23] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in science conference*, vol. 13. Citeseer, Conference Proceedings, p. 20.

[24] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance," *Data*, vol. 6, no. 2, p. 11, 2021. [Online]. Available: https://www.mdpi.com/2306-5729/6/2/11

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[26] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley Sons, 2013, vol. 398.

[27] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[28] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[29] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[30] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, Conference Proceedings, pp. 785–794.

[31] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[32] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*. Ieee, Conference Proceedings, pp. 1–6.

[33] R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodology and computing in applied probability*, vol. 1, pp. 127–190, 1999.

Table A. The hyperparameters searching space of each model

Model	Hyperparameter Searching Space
Logistic Regression	'C': hp.loguniform('C', -5, 0), 'penalty': hp.choice('penalty', ['l1', 'l2']), 'solver': hp.choice('solver', ['liblinear']), 'tol': hp.loguniform('tol', -4, -2), 'max_iter': hp.choice('max_iter', [1000]), 'random_state': hp.choice('random_state', [0])
KNN	'n_neighbors': hp.choice('n_neighbors', [5, 10, 15, 20, 25, 30]), 'weights': hp.choice('weights', ['uniform', 'distance']), 'algorithm': hp.choice('algorithm', ['auto', 'ball_tree']), 'leaf_size': hp.choice('leaf_size', [20, 30, 40]), 'p': hp.choice('p', [1, 2])
Support Vector Machine	'C': hp.loguniform('C', -5, 2), 'kernel': hp.choice('kernel', ['poly', 'rbf', 'sigmoid']), 'gamma': hp.choice('gamma', ['scale', 'auto']), 'coef0': hp.uniform('coef0', 0, 1), 'tol': hp.loguniform('tol', -4, -2), 'cache_size': hp.choice('cache_size', [2000]), 'shrinking': hp.choice('shrinking', [True, False]), 'break_ties': hp.choice('break_ties', [False, True]), 'class_weight': hp.choice('class_weight', [None, 'balanced']), 'probability': hp.choice('probability', [True])
Random Forest	'n_estimators': hp.choice('n_estimators', np.arange(100, 1001, 100, dtype=int)), 'criterion': hp.choice('criterion', ['gini']), 'max_depth': hp.choice('max_depth', np.arange(5, 20, dtype=int)), 'min_samples_split': hp.choice('min_samples_split', np.arange(2, 11, dtype=int)), 'min_samples_leaf': hp.choice('min_samples_leaf', np.arange(1, 11, dtype=int)), 'min_weight_fraction_leaf': hp.uniform('min_weight_fraction_leaf', 0, 0.5), 'min_impurity_decrease': hp.uniform('min_impurity_decrease', 0, 0.5), 'class_weight': hp.choice('class_weight', [None, 'balanced']), 'n_jobs': hp.choice('n_jobs', [32]), 'max_features': hp.choice('max_features', ['sqrt'])
XGBoost	'eta': hp.loguniform('eta', -7, 0), 'max_depth': hp.choice('max_depth', np.arange(1, 11, dtype=int)), 'subsample': hp.uniform('subsample', 0.2, 1), 'colsample_bytree': hp.uniform('colsample_bytree', 0.2, 1), 'colsample_bylevel': hp.uniform('colsample_bylevel', 0.2, 1), 'min_child_weight': hp.loguniform('min_child_weight', -16, 2), 'alpha': hp.uniform('alpha', 0, 1), 'lambda': hp.uniform('lambda', 0, 1), 'gamma': hp.uniform('gamma', 0, 1)
LightGBM	'learning_rate': hp.loguniform('learning_rate', -5, -2), 'max_depth': hp.choice('max_depth', np.arange(3, 11, dtype=int)), 'num_leaves': hp.choice('num_leaves', np.arange(8, 51, dtype=int)), 'min_data_in_leaf': hp.choice('min_data_in_leaf', np.arange(10, 20, dtype=int)), 'verbose': hp.choice('verbose', [-1])
1D-CNN	'hidden_size': hp.choice('hidden_size', [32, 64, 128]), 'optimizer': hp.choice('optimizer', [torch.optim.Adam, torch.optim.AdamW]), 'learning_rate': hp.loguniform('learning_rate', -5, -2), 'batch_size': hp.choice('batch_size', [32, 64, 128, 256, 512])