國立臺灣大學生物資源暨農學院生物機電工程學系

碩士論文

Department of Biomechatronics Engineering
College of Bioresources and Agriculture
National Taiwan University

Master Thesis

結合機器學習與臨床指引之急性骨髓性白血病風險分層集成模型

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

張名翔

Ming-Siang Chang

指導教授：陳倩瑜 博士

Advisor: Chien-Yu Chen, Ph.D.

中華民國 112 年 07 月

July 2023

doi:10.6342/NTU202303243

Acknowledgments

感謝倩瑜老師從我大三開始一路指導至研究所，老師給我很多空間探索未來方向，即使是與實驗室主軸不同的領域也給予無數的建議與支持，使我總是能研究感興趣的主題，雖然過程難免坎坷，不過在老師的幫助下總能找到解決的方法，很感謝這段時間的指導與支持。感謝蔡醫師提供數據與經驗協助，讓我能深入研究臨床資料科學。感謝實驗室同學，互相學習、討論精進。感謝女朋友，總是鼓勵並期望我有表現更好的表現。即將畢業離開實驗室，心中仍有萬般不捨，總覺得還有很多知識只是略知一二，有些遺憾，希望未來無論身在何處，都能將所學應用，延續至人生的每段旅程。

doi:10.6342/NTU202303243

摘要

急性骨髓性白血病（Acute myeloid leukemia, AML）是一種致命的血液疾病，由異常白細胞引起並在骨髓中發展。它會導致血小板減少，增加出血和感染的可能性。本論文開發了一個機器學習集成（ensemble）模型，使用國立台灣大學附設醫院 1213 名 AML 患者的數據集，對 AML 風險進行分層，本研究提出的方法結合機器學習集成模型預測的結果和 2017 年歐洲白血病網（European LeukemiaNet 2017, ELN 2017）預測的結果，進一步合成最終的集成模型 Ensemble (ML+ELN)，提出了初步的臨床風險分層建議。與 ELN 2017 臨床診斷建議相比，本研究的風險分層建議提供了最佳區分各種風險的能力，c-index 由 0.64 提升至 0.66。特別在區分不利風險和中等風險上，相較於 2017 ELN 的 p 值（p-value）平均 0.13，本研究的風險分層建議達到 p 值平均 0.001 的表現。

關鍵字：急性骨髓性白血病、集成模型、機器學習、風險分層、歐洲白血病網

doi:10.6342/NTU202303243

Abstract

Acute myeloid leukemia (AML), a fatal blood condition, is brought on by abnormal white blood cells and develops in the bone marrow. It results in a decrease in platelets, raising the possibility of bleeding and infection. This study developed an ML-based ensemble model to stratify the risk of AML using a dataset containing 1213 AML patients from the National Taiwan University Hospital. Combining the ML-based ensemble model predictions and the European LeukemiaNet (ELN) 2017 predictions, the study represents a final ensemble model (ML+ELN) for initial clinical risk stratification recommendations. Compared to the clinical diagnostic recommendations ELN 2017, the proposed risk stratification proposal provides a superior capacity to distinguish various risks and improve the c-index from 0.64 to 0.66. Especially in distinguishing unfavorable risks from moderate risks, compared with the average p-value (p-value) of 0.13 in 2017 ELN, the proposed risk stratification proposal achieves excellent performance with an average p-value of 0.001.

Keywords: acute myeloid leukemia, ensemble model, machine learning, risk stratification, European LeukemiaNet

doi:10.6342/NTU202303243

Table of Contents

==========

doi:10.6342/NTU202303243

doi:10.6342/NTU202303243

List of Figures

doi:10.6342/NTU202303243

List of Tables

doi:10.6342/NTU202303243

Chapter 1. Introduction

Acute Myeloid Leukemia (AML) is a swiftly progressing blood and bone marrow cancer demanding personalized treatment strategies based on accurate risk stratification. Traditionally, single or limited biomarkers have been used for risk estimation. However, due to the complex nature of AML, the perspective is shifting towards integrating various biomarkers for a more nuanced risk assessment.

1.1 Risk Stratification

Risk stratification is essential to make a therapeutic decision. The clinical practice involves stratifying patient risks according to diagnostic guidelines and clinical experience, and then treatment is given based on the combination of risk level and prognostic responses (Tsai, 2021). Clinical risk stratification, traditionally, has often depended on a single or a handful of individual biomarkers. However, given the increasing complexities of diseases and advancements in medical sciences, the perspective is gradually changing. Combining various biomarkers associated with different risks may influence clinical decisions more nuancedly. It is now acknowledged that combined biomarkers may produce different risk levels than initially predicted by individual biomarkers. This does not necessarily diminish the importance of unique biomarkers but underlines the potential for increased precision when considering

doi:10.6342/NTU202303243

multiple biomarkers. Therefore, there is growing interest in research exploring combinations of various biomarkers to improve the accuracy of patient risk stratification, thereby informing better treatment strategies.

1.2 Biomarkers

Biomarkers are characteristics objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention. The European LeukemiaNet (ELN) (Döhner et al., 2022) (Döhner et al., 2017) recommendations provide multiple biomarkers for the diagnosis and management of AML. This study focused on risk stratification by genetics at the initial diagnosis of ELN. Well-known biomarkers such as age, gender, hematologic data, karyotypes, and gene mutation were used as the model's features.

1.3 Purpose

This study aimed to enhance the risk stratification process for Acute Myeloid Leukemia (AML) patients at the initial diagnosis stage. We developed an ensemble machine learning model that combines its predictions with those of the European LeukemiaNet (ELN) 2017, using a dataset of 1213 AML patients. By integrating multiple biomarkers, our model aimed to more accurately stratify patient risk levels, particularly differentiating between unfavorable and intermediate risks. The ultimate goal of this

doi:10.6342/NTU202303243

improved risk stratification is to guide more personalized and effective treatment strategies for AML patients.

doi:10.6342/NTU202303243

Chapter 2. Literature Review

2.1 Clinician vs. Non-Clinician Data: Implications for Health ML Models

Machine learning (ML) models are increasingly crucial for predicting patient outcomes in modern healthcare. The data utilized in these models often stem from two main categories: clinician-initiated and non-clinician-initiated data.

Clinician-initiated data, deriving from healthcare professionals' decisions and actions, such as prescriptions and referrals, often serve as critical inputs for ML models. Many studies show that models incorporating this data type tend to have superior predictive performance as they reflect the complexities of a clinician's decision-making process (Shreve et al., 2019).

Although models reliant on clinician-initiated data may demonstrate impressive performance metrics, their application in clinical practice raises several concerns. The central issue revolves around these models' propensity to mimic existing decision-making patterns of clinicians rather than offering fresh insights or contributing to improved patient outcomes.

Moreover, when applied to risk stratification, models based on clinician-initiated data may lose their relevance. The very nature of risk stratification is to identify patients who are likely to develop complications or severe outcomes, ideally before clinicians

doi:10.6342/NTU202303243

identify them. If a model heavily relies on clinician-initiated data, it essentially echoes the clinician's initial assessment rather than providing an independent risk evaluation. This could lead to redundancy and undermine the purpose of risk stratification.

Non-clinician-initiated data, encompassing routine orders and direct physiological measurements, present a valuable alternative. Although they may not consistently outperform clinician-initiated data in predictive accuracy, they can provide new insights that augment clinical decision-making.

Thus, while clinician-initiated data can enhance model performance, it may lead to models that mirror existing clinical decisions and potentially render risk stratification redundant. On the other hand, non-clinician-initiated data can offer objective and fresh insights.

2.2 Risk Stratification

ELN recommendations (Döhner et al., 2017) (Döhner et al., 2022) provided risk stratification recommendations for clinical applications, using biomarkers such as cytogenetics and gene mutations to divide patients into three risks: favorable, intermediate, and adverse. Favorable patients will have better overall survival. Although the 2022 edition of ELN has been published, the stratification ability of the 2022 edition of ELN is worse than the 2017 edition. Hence, the study focused on the 2017 edition of

doi:10.6342/NTU202303243

ELN.

2.3 Biomarkers Interaction

In ELN recommendations, however, there are many exceptions in the biomarkers that will divide patients into different risk stratifications, showing that the interaction between genes does affect risk interpretation. 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 (Tien et al., 2018). Therefore, it is necessary to develop a model that can consider multiple biomarkers

2.4 Ensemble

Ensemble methods are a powerful data science concept involving training and combining multiple models to solve problems. The fundamental principle of ensemble methods is that a group of ‘weak learners’ can form a ‘strong learner.’ Each model makes a vote or contributes a piece of information, and the ensemble method combines these votes to produce a final prediction. The goal is to reduce both bias and variance of the prediction. These methods improve accuracy and robustness by leveraging the collective power of multiple models rather than a single one.

2.5 Evaluation Methods

doi:10.6342/NTU202303243

Some methods are used to evaluate the performance of multi-class models, including confusion matrix, accuracy, and weighted F1 score. Additionally, survival analysis is frequently utilized in medical research to assess the effectiveness of models.

2.5.1 Confusion Matrix

A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of a model. The matrix is expanded to a 2D grid, and each row represents the instances in a predicted class, while each column represents the instances in an actual class. It is a simple yet effective way to visualize the performance of a multi-class classification problem and identify which classes are being misclassified.

2.5.2 Accuracy

Accuracy is one of the most specific metrics in classification problems. It calculates the proportion of correct predictions over total predictions. Accuracy is practical when target classes are well-balanced. However, it might not be a good measure when dealing with imbalanced datasets because a model that only predicts the majority class can still achieve high accuracy.

2.5.3 Weighted F1 Score

The F1 Score is a commonly used metric for assessing the performance of classification

doi:10.6342/NTU202303243

models. It considers both the model's precision and recall to compute the score. Precision is the number of true positive results divided by the number of all positive results, including those not correctly identified. Recall (also known as Sensitivity) is the number of true positive results divided by the number of all samples that should have been identified as positive.

The F1 score is the harmonic mean of precision and recall. While the regular mean treats all values equally, the harmonic mean gives much more weight to low values. As a result, the classifier will only get a high F1 score if both recall and precision are high. The formula to calculate F1 Score is as follows:

$$F1\,Score = 2 * (Recall * Precision) / (Recall + Precision)$$

The weighted F1 score calculates metrics for each class and finds their average weighted by the number of actual instances for each class. The weighted F1 score is calculated as follows: it computes the F1 score for each class independently, but when it adds them together uses a weight that depends on the number of actual labels of each class. This means that the contribution of each class to the average F1 score is proportional to the class size in the dataset.

2.5.4 Survival Analysis

Survival analysis is a branch of statistics that focuses on the time it takes for an event—like disease occurrence or machine failure—to happen. It accounts for both observed

doi:10.6342/NTU202303243

events and censored instances where an event has not occurred or was interrupted.

A vital tool in survival analysis is the survival curve, which illustrates the probability of a subject surviving past a specific time. It shows the proportion of individuals who have not yet experienced the event. It typically decreases over time, revealing the underlying distribution of survival times and aiding in comparisons between groups or treatments.

A non-parametric method, the Kaplan-Meier estimator (Goel et al., 2010) estimates the survival function from lifetime data. Particularly useful with right-censored data (common in medical research), the Kaplan-Meier curve is a series of descending steps approximating the survival function. This visualization can be used to compare survival distributions across different groups, and the p-values from the log-rank test show if the differences between different groups are statistically significant. A small p-value suggests a significant difference, while a large p-value suggests no significant difference.

2.5.5 Concordance Index

The Concordance Index (c-index) is a statistical measure used in medicine, bioinformatics, and data science to assess prediction accuracy. Given a set of data with actual and predicted outcomes, the C-Index measures how well the predictions rank the data. The predictions are considered concordant for every pair of data points if the data

doi:10.6342/NTU202303243

point with the higher predicted outcome also has the actual higher outcome.

The value of the C-index ranges from 0.5 to 1.0, where 0.5 suggests that the model is no better than a random chance at making predictions, and 1.0 indicates perfect concordance between the model's predictions and the actual outcomes. A higher C-index indicates a more accurate model.

doi:10.6342/NTU202303243

Chapter 3. Materials and Methods

This study employed a dataset from National Taiwan University Hospital (NTUH) AML cohort, filtering it based on specific criteria and dividing it into training and validation sets. 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.

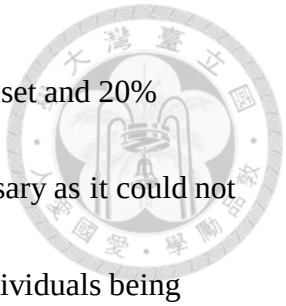
The proposed prediction model involved many well-known classification techniques. In addition, HyperOpt, a Bayesian hyperparameter optimization, was used to select the optimal hyperparameters of each model. Finally, an ensemble model considers each model to predict risk.

After integrating the classification techniques, the proposed prediction model jointly consider the results and the results predicted by ELN 2017. Based on this, the study proposed initial clinical risk stratification recommendations.

3.1 Dataset

The dataset came from an acute myeloid leukemia cohort with 1213 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

doi:10.6342/NTU202303243



months, the remaining 801 samples were separated into an 80% train set and 20% validation set for prediction and evaluation. 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 1.

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 (Alshdaifat et al., 2021). The standard normalization was applied to age and hematological data. The data preprocessing and feature selection is shown in Fig. 1.

Table 1: The label Distribution of Samples

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

doi:10.6342/NTU202303243

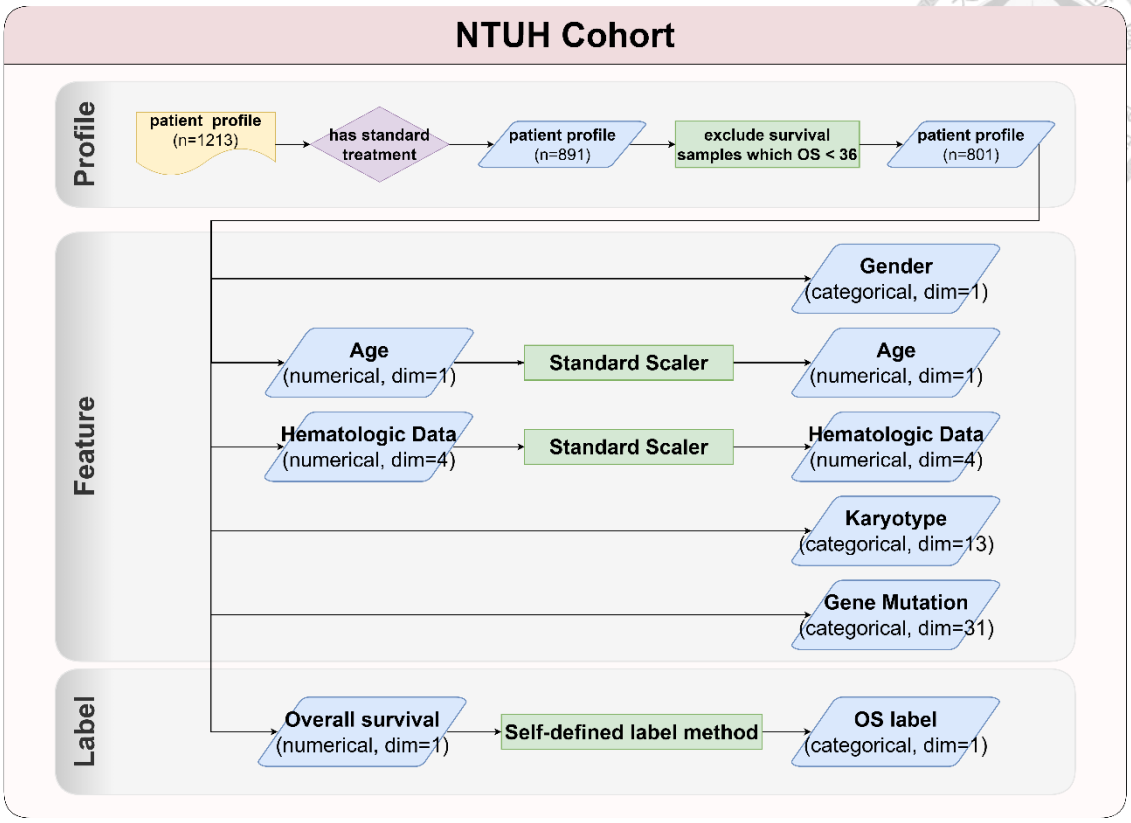


Fig. 1: The feature selection and data preprocessing pipeline.

3.2 Models

The models learned from a dataset where the relationships between the features and the label are known. For the logistic regression, k-nearest neighbors, support vector machine, and random forests, the study utilizes scikit-learn (Pedregosa et al., 2011) as the program package. The following describes the models used

3.2.1 Logistic Regression

The logistic regression model (Hosmer Jr et al., 2013) works by estimating the logistic model's parameters (the coefficients in the linear combination), which are used to

doi:10.6342/NTU202303243

estimate the likelihood that an event will occur by making the event's log odds a linear combination of one or more independent variables.

In addition to the standard application, some unnecessary features' coefficients can be adjusted to zero using the L1, L2, or Elastic-Net penalty. On the other hand, feature selection can improve accuracy.

3.2.2 K-nearest Neighbors

K-nearest neighbors (KNN) (Peterson, 2009) is a simple yet effective machine learning algorithm for classification and regression tasks. As an instance-based learning method, it does not require a learning phase. Instead, it uses the entire dataset in the prediction phase, making it more adaptable to changes in input.

KNN operates by computing the distance (usually Euclidean, but other metrics can also be used) between the new observation and every existing instance in the dataset. The 'K' instances with the smallest distances are selected as 'neighbors.' In a classification task, the algorithm assigns the most common class among the neighbors to the new observation. Its non-parametric nature allows KNN to handle multi-class cases seamlessly and works well with data with irregular decision boundaries.

3.2.3 Support Vector Machine

Support Vector Machine (SVM) (Noble, 2006) is a powerful machine learning model

doi:10.6342/NTU202303243

primarily used for classification and regression tasks. SVM excels in high-dimensional data and is effective when the number of dimensions exceeds the number of samples.

SVM's core idea is to find a hyperplane in N-dimensional space (N - the number of features) that distinctly classifies data points. It seeks the hyperplane with the maximum margin, i.e., the maximum distance between data points of two classes. SVMs are effective in cases where the decision boundary is non-linear because they can use the kernel trick to project data into higher dimensions.

Notably, SVMs are robust against overfitting, especially in high-dimensional space. While SVMs offer high accuracy and robustness, they can be computationally intensive on large datasets and less effective when overlapping classes.

3.2.4 Random Forests

Random Forests (Breiman, 2001) is a versatile machine learning model for classification and regression tasks. As an ensemble method, it constructs many decision trees at training time and outputs the mode of the classes (classification) or means prediction (regression) of the individual trees.

Random Forests introduces randomness in two ways. Firstly, each tree is trained on a different bootstrapped sample of the data. Secondly, a random subset of features is considered for splitting at each node. This randomness helps increase model diversity,

doi:10.6342/NTU202303243

improving generalization and reducing overfitting.

One of the significant advantages of Random Forests is its ability to handle many input variables without variable deletion. It also provides a reliable feature importance estimate, helping to understand which variables contribute the most to prediction.

3.2.5 XGBoost

XGBoost (Chen & Guestrin, 2016) is a machine learning library based on the gradient boosting framework. It utilizes decision tree ensembles and improves model accuracy by systematically correcting mistakes. It boasts parallel processing capabilities that improve computational speed. Tree pruning in XGBoost prevents unnecessary computations in leaf nodes, further enhancing speed. It also includes a regularization parameter in its cost function to minimize overfitting.

With custom optimization objectives and evaluation criteria, XGBoost caters to varied prediction tasks. Its feature-rich design makes XGBoost widely used in machine learning competitions and practical applications.

3.2.6 LightGBM

LightGBM (Ke et al., 2017), or "Light Gradient Boosting Machine," is a gradient boosting framework that offers advanced efficiency and scalability. It grows trees leaf-wise (best-first) rather than level-wise, creating a more complex model and yielding

doi:10.6342/NTU202303243

higher accuracy, even though it might lead to overfitting for smaller datasets.

LightGBM employs two novel techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS keeps the instances with larger gradients, contributing more to information gain, and performs random sampling on the instances with smaller gradients. This leads to a reduction in training time without significant loss in accuracy. EFB bundles exclusive features, i.e., rarely non-zero features, into a single feature, further reducing the dimensionality of sparse data.

3.2.7 1D-CNN

A Convolutional Neural Network (CNN) (Albawi et al., 2017) is a deep learning algorithm used primarily for image processing and recognition. It mimics the human visual cortex, using layers of filters to detect patterns and features, making it highly effective for object detection and computer vision tasks. Later, one-dimensional CNN (1D-CNN) was adapted for sequential data like time series and natural language.

1D-CNNs run convolutional filters along the sequence, capturing local patterns similar to how 2D-CNNs identify patterns in a small area of an image. This local focus makes 1D-CNNs invariant to translational changes in the input. 1D-CNNs shine in their ability to process temporal features and handle long sequences efficiently. These models are also helpful for extracting local features and identifying patterns within fixed-length segments of the data.

doi:10.6342/NTU202303243

3.3 Hyperparameters Optimization

In data science, hyperparameter optimization is a crucial step that impacts the success of a model. 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 (Bergstra et al., 2013) as an efficient tool for streamlining hyperparameter optimization. This Python library 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 hyperparameters 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 method (Ensemble ML). This technique optimized the combination of models by leveraging a loss function, which measured the deviation of predicted values from

doi:10.6342/NTU202303243

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 (Rubinstein, 1999). The ensemble then combined these models' predictions according to their weights to make the final prediction, enhancing prediction accuracy. The formula is shown below:

$$weight_i = \frac{exp(-loss_i)}{\sum_{i \in M} exp(-loss_i)}$$

$$prediction_{ensemble} = argmax \sum_{i \in M} weight_i * probability_i$$

where M are the models used, i is one of the models, and $loss_i$ means the value of the loss function of model i.

In essence, ensemble methods, 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 2017 (Ensemble ML+ELN)

After predicting by the ensemble model, the study established initial clinical risk stratification recommendations by considering the ensemble model and ELN 2017 risk

doi:10.6342/NTU202303243

prediction (Ensemble ML+ELN). The recommendations firmly considered predictions as results for samples where both models predict the same in the ensemble model and ELN 2017. Then for samples that were contrary to the prediction of the integrated model in ELN 2017, the recommendations considered them intermediate. That is to say; for samples that were predicted as adverse by ELN 2017 and favorable by the ensemble model, or samples that were predicted by ELN 2017 as adverse and favorable by the ensemble model, the recommendations treated them as intermediate. The rest treated the prediction of Ensemble ML as the final result. The initial clinical risk stratification recommendations are shown in Fig. 2.

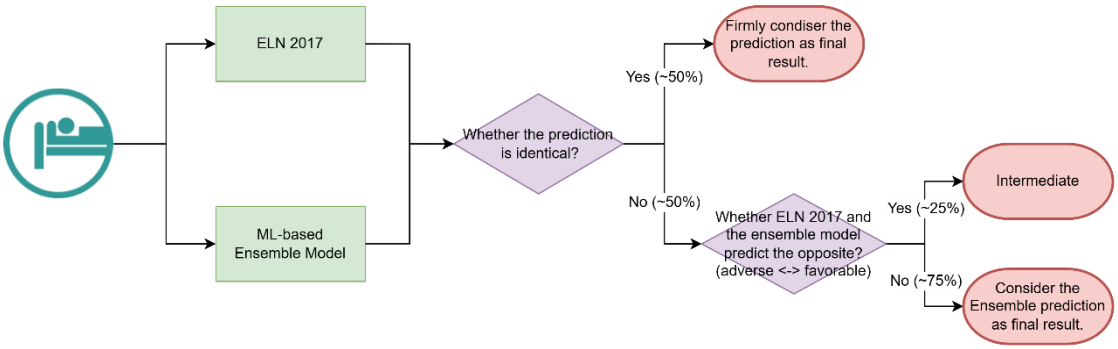


Fig. 2: The initial clinical risk stratification recommendations.

doi:10.6342/NTU202303243

Chapter 4. Results and Discussion

The study developed a machine learning-based ensemble model utilizing a dataset of 1213 AML patients from the National Taiwan University Hospital (Dr. Cheng-Hong Tsai). The machine learning-based ensemble model combines its predictions with the European LeukemiaNet (ELN) 2017 predictions to enhance risk stratification for initial diagnosis. Compared to the original ELN 2017, the proposed approach better distinguishes various risk levels, particularly between unfavorable and intermediate risks, thereby potentially improving treatment strategies.

The study repeated an experiment 50 times for robustness, employing different training sets and validation set partitions for each. The ensemble model combined with ELN 2017 (Ensemble ML+ELN) demonstrated superior performance across all evaluation methods. Using statistical tests, potential links between biomarkers were explored to enhance the accuracy of differentiating between adverse and intermediate risk levels.

Significant features (with a p-value<0.05) were evaluated between two groups with different risk predictions by ELN 2017 and the ensemble model. The ensemble model tended to predict younger samples as non-adverse risks, while the ELN model predicted them as adverse. Adverse biomarkers had also been observed to play crucial roles in the stratification of the adverse group.

doi:10.6342/NTU202303243

A similar comparison was performed between two groups with different risk predictions for the favorable group. Here, the ensemble model tended to predict elder samples as non-favorable risk. The analysis suggested that for samples predicted to be favorable in ELN 2017, verifying their youth could further distinguish safer groups.

An independent analysis of the samples predicted identically by ELN 2017 and the ensemble model demonstrated exemplary performance. About 50% of the samples had identical predictions and exhibited higher accuracy and F1-score. Even though the p-value of the survival curve was higher than in the previous table, this was attributed to the reduced number of people used in the analysis.

4.1 Performance

To ensure the approach's robustness, the study repeated the experiment 50 times. Each experiment had a different partition on the training set and validation set. The ensemble ML+ELN model performed best on all evaluation methods. In contrast to accuracy and f1 score, the model emphasized the survival curve's capacity to differentiate between various risk levels (p-value) and survival analysis performance (c-index). Hence, the ensemble model and ELN 2017 combination perform best on the training and validation sets. The performance of the methods on the training set and the validation set is shown in Table 2 and Table 3. The evaluation matrices on the training

doi:10.6342/NTU202303243

and validation set are shown in Fig. 3 and Fig. 4.

Table 2: The performance of the methods on the training set of 50 times

Method	Accuracy	F1 score	c-index	Survival Curve P-value (adverse vs. intermediate)[1]	Survival Curve P-value (intermediate vs. favorable)[1]
ELN 2017	0.48 ± 0.01	0.47 ± 0.01	0.63 ± 0.01	1.2e-04 ± 1.8e-04	9.0e-08 ± 1.6e-07
Logistic Regression	0.54 ± 0.01	0.54 ± 0.01	0.66 ± 0.01	9.0e-07 ± 2.3e-06	7.4e-08 ± 2.1e-07
KNN	0.89 ± 0.21	0.90 ± 0.20	0.81 ± 0.09	3.7e-09 ± 2.2e-08	1.5e-07 ± 8.3e-07
SVM	0.59 ± 0.06	0.59 ± 0.06	0.68 ± 0.03	1.7e-07 ± 7.2e-07	2.1e-06 ± 8.7e-06
Random Forest	0.72 ± 0.07	0.72 ± 0.07	0.72 ± 0.03	4.8e-10 ± 2.1e-09	8.2e-10 ± 4.7e-09
XGBoost	0.77 ± 0.09	0.78 ± 0.08	0.75 ± 0.04	1.2e-12 ± 5.2e-12	1.7e-14 ± 9.0e-14
LightGBM	0.74 ± 0.05	0.74 ± 0.05	0.73 ± 0.02	3.4e-12 ± 1.7e-11	8.4e-16 ± 3.8e-15
1D-CNN	0.56 ± 0.03	0.56 ± 0.04	0.67 ± 0.02	7.9e-03 ± 2.2e-02	8.0e-07 ± 5.0e-06
Ensemble (ML)	**0.94 ± 0.10**	**0.94 ± 0.10**	**0.83 ± 0.05**	1.8e-21 ± 9.7e-21	2.2e-20 ± 1.1e-19
Ensemble (ML + ELN)	0.85 ± 0.08	0.85 ± 0.08	0.80 ± 0.03	**3.2e-31 ± 1.3e-30**	**1.1e-21 ± 6.4e-21**

[1] The P-values for the Kaplan-Meier survival curves vary with the total sample size and should only be used in this table for comparison.

doi:10.6342/NTU202303243

Table 3: The performance of the methods on the validation set of 50 times

Method	Accuracy	F1 score	c-index	Survival Curve P-value (adverse vs. intermediate)[1]	Survival Curve P-value (intermediate vs. favorable)[1]
ELN 2017	0.49 ± 0.04	0.49 ± 0.04	0.64 ± 0.02	1.3e-01 ± 2.0e-01	7.3e-03 ± 1.6e-02
Logistic Regression	0.48 ± 0.03	0.47 ± 0.03	0.64 ± 0.02	6.6e-02 ± 1.1e-01	6.0e-02 ± 8.9e-02
KNN	0.45 ± 0.03	0.45 ± 0.03	0.62 ± 0.02	6.9e-03 ± 1.3e-02	1.6e-01 ± 1.9e-01
SVM	0.48 ± 0.03	0.48 ± 0.03	0.64 ± 0.02	3.0e-02 ± 4.4e-02	6.6e-02 ± 9.8e-02
Random Forest	0.49 ± 0.03	0.48 ± 0.03	0.63 ± 0.02	4.1e-02 ± 6.0e-02	1.1e-01 ± 1.5e-01
XGBoost	0.47 ± 0.03	0.47 ± 0.03	0.62 ± 0.02	5.5e-02 ± 9.8e-02	1.4e-01 ± 2.1e-01
LightGBM	0.48 ± 0.03	0.48 ± 0.03	0.63 ± 0.02	4.3e-02 ± 5.6e-02	8.6e-02 ± 1.5e-01
1D-CNN	0.46 ± 0.04	0.45 ± 0.04	0.62 ± 0.02	1.7e-01 ± 2.1e-01	1.1e-01 ± 1.9e-01
Ensemble (ML)	0.50 ± 0.04	0.49 ± 0.03	0.64 ± 0.02	4.1e-02 ± 8.4e-02	4.6e-02 ± 8.1e-02
Ensemble (ML + ELN)	**0.52 ± 0.03**	**0.52 ± 0.03**	**0.66 ± 0.02**	**1.1e-03 ± 2.5e-03**	**9.0e-03 ± 1.5e-02**

[1] The P-values for the Kaplan-Meier survival curves vary with the total sample size and should only be used in this table for comparison.

doi:10.6342/NTU202303243

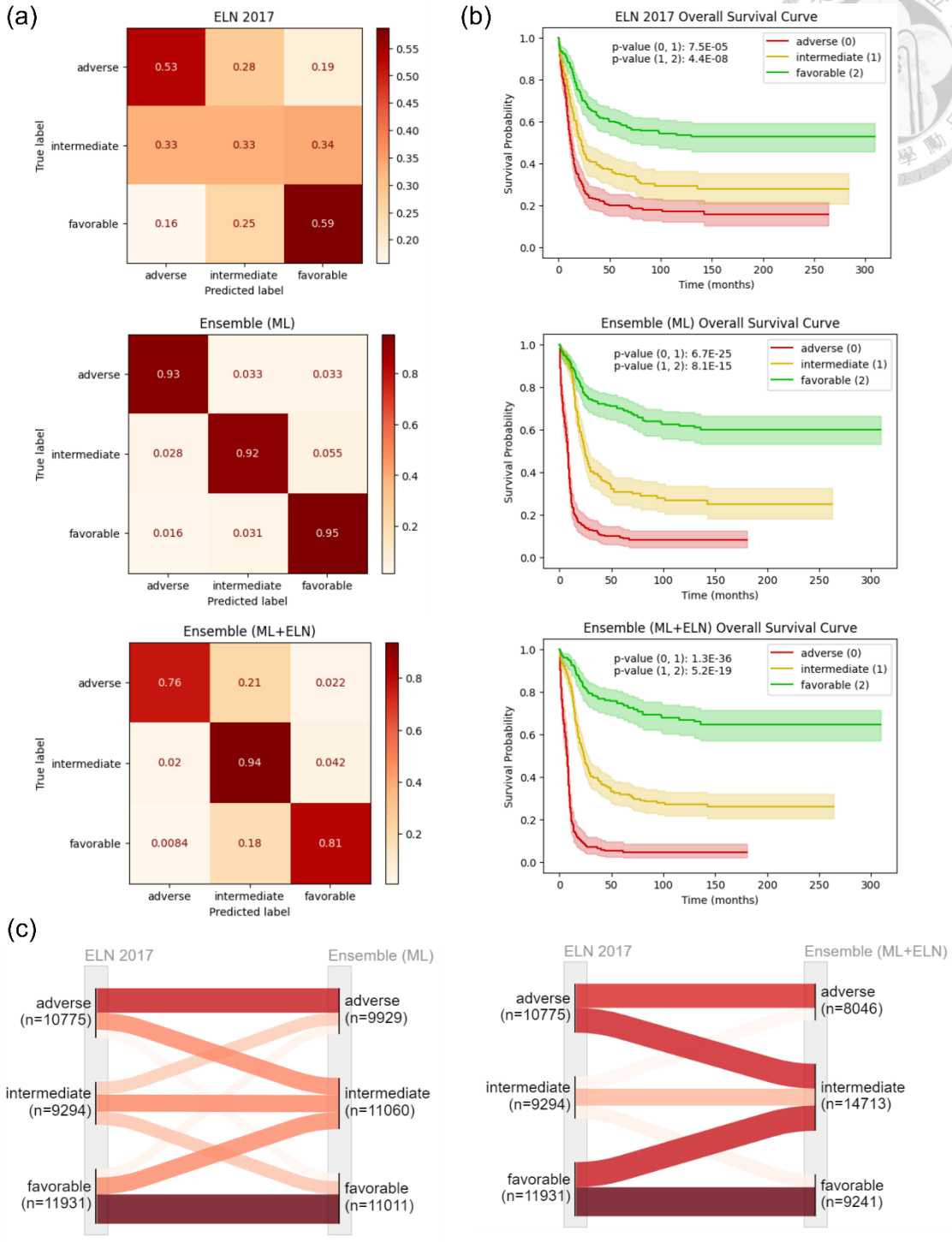


Fig. 3: The evaluation matrices on the training set. (a) The confusion matrix from cumulative results of 50 experiments. (b) The survival curve from the first experiment. (c) The sample distribution flow from cumulative results of 50 experiments.

doi:10.6342/NTU202303243

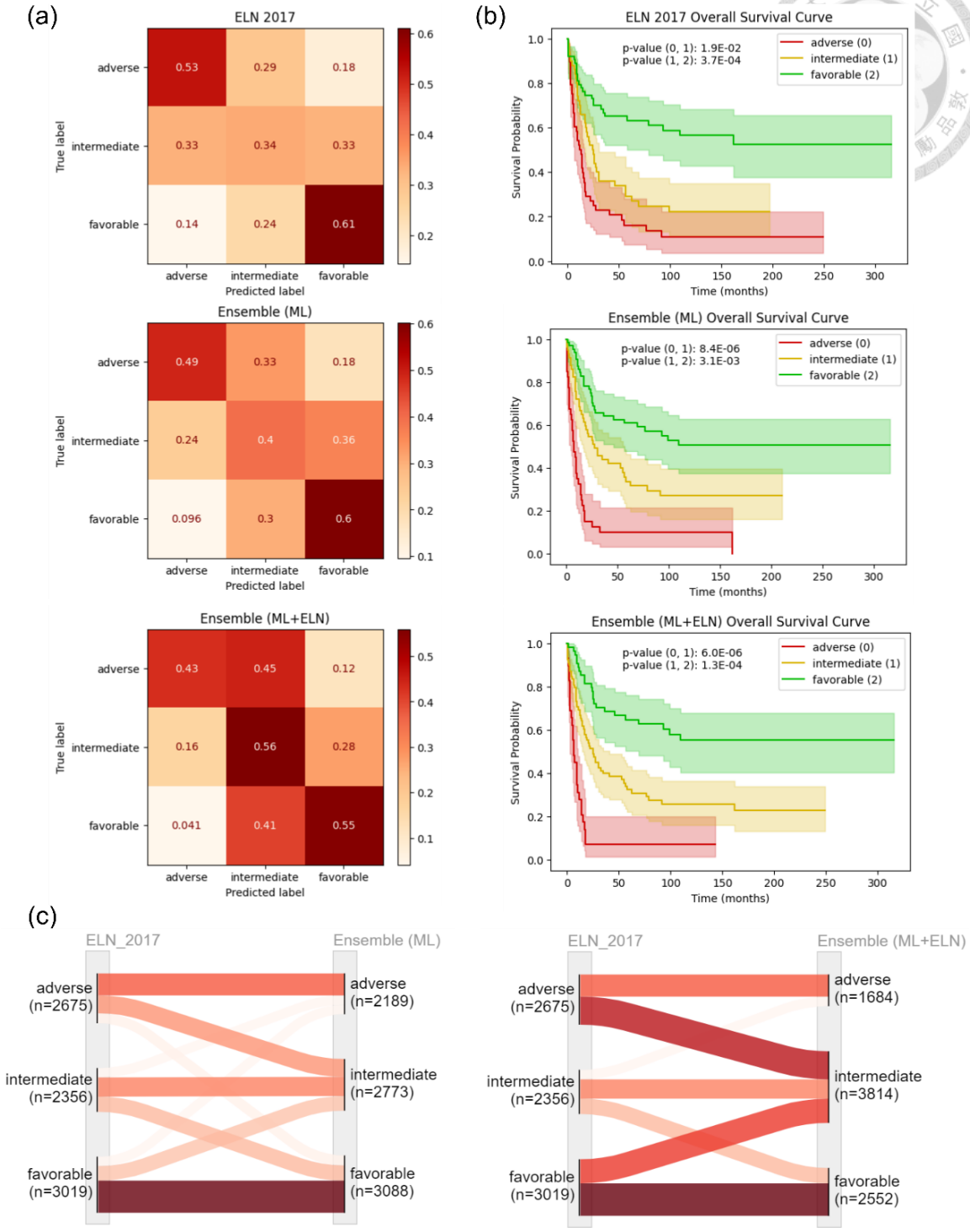
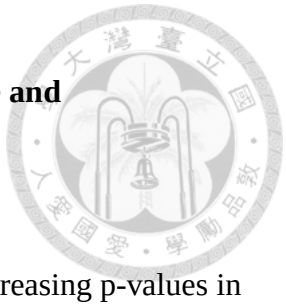


Fig. 4: The evaluation matrices on the validation set. (a) The confusion matrix from cumulative results of 50 experiments. (b) The survival curve from the first experiment. (c) The sample distribution flow from cumulative results of 50 experiments.

doi:10.6342/NTU202303243



4.2 Revealing Biomarker Interactions to Distinguish the Adverse and Intermediate of Survival Curves

Statistical tests were performed for groups with decreasing or not decreasing p-values in the survival curve between adverse and intermediate groups. T-test (Kim, 2015) is applied to numerical features, and Fisher's exact test (Upton, 1992) is applied to categorical features because of the small number of each categorical feature. However, none of the significant features consistently occur in 50 experiments. It potentially revealed links between biomarkers, improving accuracy in distinguishing between adverse and intermediate risk levels. The method for determining the samples that lower the p-value is documented in Appendix B.

4.3 Insights for patients predicted as adverse on ELN 2017

The study evaluated the two groups' significant features (p-value<0.05) using the statistical test in Chapter 4.2. One group was samples predicted as adverse by ELN 2017 and the ensemble model. The other group was samples predicted as adverse by ELN 2017 with non-adverse by the ensemble model. The groups can also be identified from Fig. 4c. For the left image of Fig. 4c, one group is the same as the line at the top of the ELN 2017 adverse section, and the other is the same as the lines at the bottom and middle of ELN 2017 adverse section. Alternatively, for the right image of Fig. 4c, one

doi:10.6342/NTU202303243

group is the same as the line at the top of the ELN 2017 adverse section, and the other is the same as the line at the bottom of the ELN 2017 adverse section.

Fig 5a shows the number of occurrences of significant features on the validation set in 50 experiments. The mean p-value of the survival curve between the two groups was 0.044. The ensemble model tended to predict younger samples as a non-adverse risk. In contrast, ELN predicted an adverse risk (Fig. 5c). Furthermore, adverse biomarkers karyotypes -17 and karyotypes -5, and gene mutation TP53 are much more in the group from ELN 2017 adverse to the ensemble model adverse, representing that these biomarkers were critical in the stratification of the adverse group (Fig. 5d). Therefore, for samples predicted to be adverse in ELN 2017, confirming whether they are elder or carry karyotype -17, karyotype -5 or TP53 can further distinguish more dangerous groups.

doi:10.6342/NTU202303243

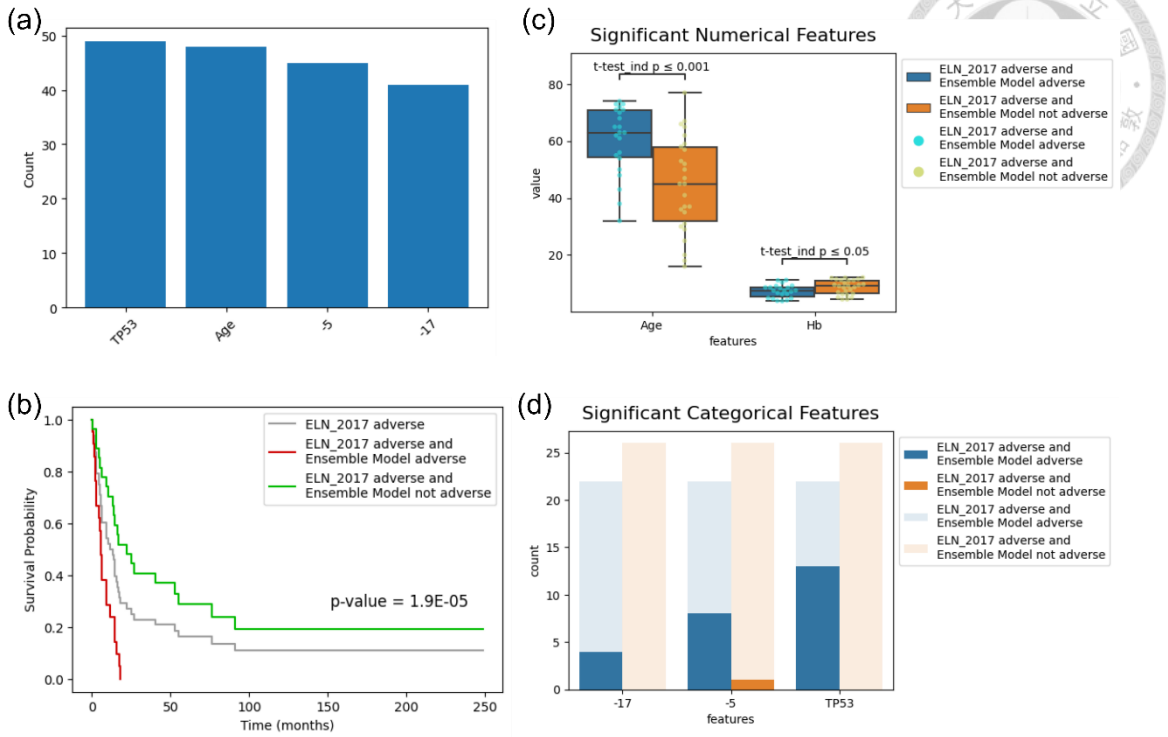


Fig. 5: The result compares two groups of samples identified as adverse in ELN 2017: one group was predicted as adverse by the ensemble model, while the other was predicted as non-adverse by the ensemble model. (a) The number of occurrences of significant features on the validation set in 50 experiments. (b) The survival curve from the first experiment. The p-value is the difference between the figure's red and green groups. (c) The significant numerical features from one of the experiments. (d) The significant categorical features from one of the experiments. The shade plus the solid bar is the actual number of people. The -5 karyotype is missing on chromosome 5, and the -17 karyotype is missing on chromosome 17.

4.4 Insights for patients predicted as favorable on ELN 2017

The study evaluated the two groups' significant features (p-value<0.05) using the statistical test in Chapter 4.2. One group was sample predicted as favorable by ELN 2017 and the ensemble model. The other group was sample predicted as favorable by ELN 2017 with non-favorable by the ensemble model. The groups can also be identified

doi:10.6342/NTU202303243

from Fig. 4c. For the left image of Fig. 4c, one group is the same as the line at the bottom of the ELN 2017 favorable section, and the other is the same as the lines at the top and middle of ELN 2017 favorable section. Alternatively, for the right image of Fig. 4c, one group is the same as the line at the bottom of the ELN 2017 favorable section, and the other is the same as the line at the top of the ELN 2017 favorable section.

Fig 6a shows the number of occurrences of significant features on the validation set in 50 experiments. The mean p-value of the survival curve between the two groups was 0.055. The ensemble model tended to predict elder samples as a non-favorable risk. In contrast, ELN predicted an adverse risk (Fig. 5c). Therefore, for samples predicted to be favorable in ELN 2017, confirming whether they are young can further distinguish safer groups.

doi:10.6342/NTU202303243

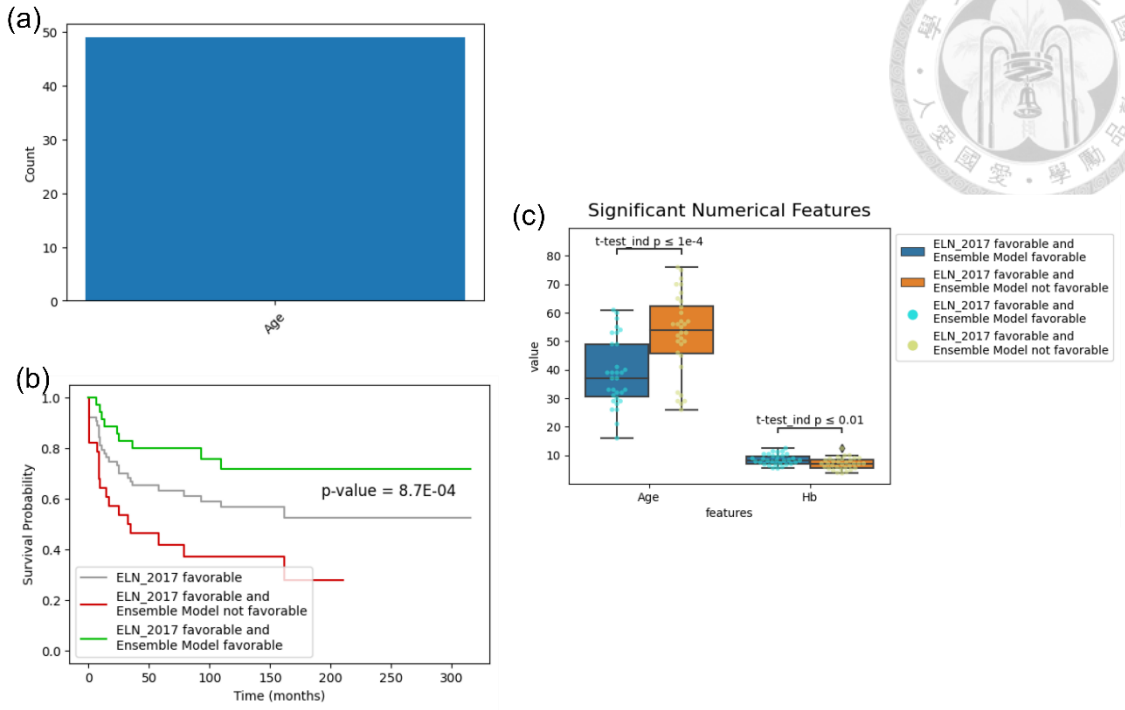


Fig. 6: The result compares two groups of samples identified as favorable in ELN 2017: one group was predicted as favorable by the ensemble model, while the other was predicted as non-favorable by the ensemble model. (a) The number of occurrences of significant features on the validation set in 50 experiments. (b) The survival curve from the first experiment. The p-value is the difference between the figure's red and green groups. (c) The significant numerical features from one of the experiments.

4.5 Identical Risk Prediction on ELN 2017 and the Ensemble Model

Because of the excellent performance of the survival curve of the samples that both predicted the same in ELN 2017 and the ensemble model, the study selected them independently for analysis. The performance of the methods on the validation set is shown in Table 4, and the evaluation matrices are shown in Fig. 7. About 50% of the samples had identical predictions, and these samples had higher accuracy and f1-score.

doi:10.6342/NTU202303243

The c-index of the model with samples with identical predictions is 0.74, higher than the model with all samples, whose c-index is 0.66 in Table 3.

Table 4: The performance of the Methods for samples where the Ensemble Model and ELN 2017 predict identical risk levels

Method	The ratio of identical predictions	Accuracy	F1 score	c-index
Ensemble & ELN	0.49 ± 0.03	0.63 ± 0.04	0.62 ± 0.05	0.74 ± 0.03

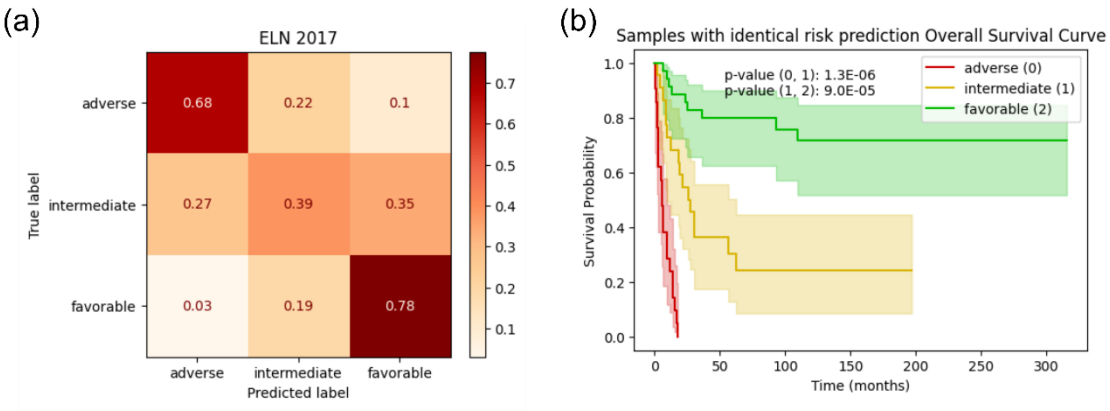


Fig. 7: The evaluation matrices for samples with identical predictions on the validation set. (a) The confusion matrix from cumulative results of 50 experiments. (b) The survival curve from the first experiment.

doi:10.6342/NTU202303243

Chapter 5. Conclusions

The study presents initial clinical risk stratification recommendations by combining the ensemble model with ELN 2017. The risk stratification recommendations have better performance in distinguishing either between adverse and intermediate or between intermediate and favorable; the c-index improved from 0.64 to 0.66. The proposed method excels, especially when compared to ELN 2017's poor discrimination between adverse and intermediate risks, with a p-value of 0.001 versus ELN's 0.13.

The study also proposes insights for samples predicted to be adverse or favorable in ELN 2017. The presence of old age, karyotype -17, karyotype -5, or TP53 can further separate more risky groups from those predicted to be adverse in ELN 2017. The presence of younger age can further separate safer groups from those predicted to be favorable in ELN 2017.

An independent analysis of samples predicted identically by both models showed good performance, with about 50% having identical predictions demonstrating higher accuracy, F1-score, and c-index.

doi:10.6342/NTU202303243

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. 2017 international conference on engineering and technology (ICET),

Alshdaifat, E. a., Alshdaifat, D. a., Alsarhan, A., Hussein, F., & El-Salhi, S. M. d. F. S. (2021). The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. *Data*, *6*(2), 11. https://www.mdpi.com/2306-5729/6/2/11

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. Proceedings of the 12th Python in science conference,

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., Dombret, H., Ebert, B. L., Fenaux, P., & Larson, R. A. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood, The Journal of the American Society of Hematology*, *129*(4), 424-447.

Döhner, H., Wei, A. H., Appelbaum, F. R., Craddock, C., DiNardo, C. D., Dombret, H., Ebert, B. L., Fenaux, P., Godley, L. A., & Hasserjian, R. P. (2022). 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*, *140*(12), 1345-1377.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, *68*(6), 540-546.

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565-1567.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*,

doi:10.6342/NTU202303243

2825-2830.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Rubinstein, R. (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, *1*, 127-190.

Shreve, J., Meggendorfer, M., Awada, H., Mukherjee, S., Walter, W., Hutter, S., Makhoul, A., Hilton, C. B., Radakovich, N., & Nagata, Y. (2019). A personalized prediction model to risk stratify patients with acute myeloid leukemia (AML) using artificial intelligence. *Blood*, *134*, 2091.

Tien, F.-M., Hou, H.-A., Tang, J.-L., Kuo, Y.-Y., Chen, C.-Y., Tsai, C.-H., Yao, M., Lin, C.-T., Li, C.-C., & Huang, S.-Y. (2018). Concomitant WT1 mutations predict poor prognosis in acute myeloid leukemia patients with double mutant CEBPA. *Haematologica*, *103*(11), e510.

Tsai, C.-H. (2021). *Applying Next-generation Sequencing to Explore the Risk Stratification in Acute Myeloid Leukemia Patients*

Upton, G. J. (1992). Fisher's exact test. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *155*(3), 395-402.

doi:10.6342/NTU202303243

Appendices

A. The hyperparameters searching spaces of each model

Table A: The hyperparameters searching spaces 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])

doi:10.6342/NTU202303243

B. The Method for Determining the Samples that Lower the P-value

The Method for determining the samples that lower the p-value between the adverse and intermediate groups in the survival curve is based on the sample's label, ELN 2017 prediction, and the ensemble model prediction. For example, ELN 2017 is predicted to be adverse, the label is a favorable sample, and the ensemble model predicts that it is intermediate, making the original adverse and intermediate in ELN 2017 more separated, causing the p-value to decrease. Samples of reduced p-values combining ELN 2017 predictions, ensemble model predictions, and labels are shown in Table B.

Table B: The samples that lower the p-value between the adverse and intermediate in the survival curve

ELN 2017	Ensemble	Label
adverse	intermediate	intermediate
adverse	intermediate	favorable
adverse	favorable	intermediate
adverse	favorable	favorable
intermediate	adverse	adverse
intermediate	favorable	adverse
favorable	adverse	adverse
favorable	intermediate	favorable

doi:10.6342/NTU202303243