Heliyon 11 (2025) e42030

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A novel perspective on survival prediction for AML patients: Integration of machine learning in SEER database applications☆

Zheng-yi Jia [a], Maierbiya Abulimiti [a], Yun Wu [b], Li-na Ma [a], Xiao-yu Li [a], Jie Wang [c,*]

[a] School of Pharmacy, Xinjiang Medical University, Urumqi, 830011, China
[b] Department of General Medicine, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, 830011, China
[c] Department of Pharmacy, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, 830011, China

ARTICLE INFO

Keywords:
Acute myeloid leukemia
Machine learning
SEER database
Epidemiological characteristics
Prognosis prediction

ABSTRACT

Objective: The purpose of this study is to explore the epidemiological characteristics of acute myeloid leukemia (AML) and establish a more accurate model for predicting the prognosis of AML patients based on machine learning.

Methods: We obtained clinical data of a total of 87,090 AML patients between 1975 and 2019 from the SEER database. First, we used Kaplan-Meier analysis to examine the prognosis of patients in different strata. Then, we discussed the independent factors that influenced the overall survival (OS) of AML patients, using univariate and multivariate Cox regression analysis. Finally, we used 11 machine learning algorithms to predict the survival rate of AML patients at 1, 2, and 3 years, respectively. We also used five-fold cross-validation with 20 cycles to obtain the optimal parameters for each model, in order to improve the accuracy of predictions.

Results: The Kaplan-Meier analysis showed that the survival rate of patients diagnosed after 2010 was significantly higher than that of those diagnosed before. In addition, older age, male gender, and non-black race were associated with poor prognosis. Among the FAB subtypes, M3 AML had a better prognosis than other subtypes, and among the WHO subtypes, AML associated with Down syndrome had the best prognosis, followed by AML with eosinophilic abnormalities. The Cox regression analysis demonstrated that gender, age, race, and family income were significantly related to the survival of AML patients. Among the 11 machine learning models, the random forest classifier performed best on multiple evaluation metrics in predicting survival at 1, 2, and 3 years. In addition, both the XGBoost classifier and the neural network classifier showed high accuracy and reliability at each prediction stage.

Conclusion: Through in-depth analysis, this study provides a deeper understanding of the epidemiological characteristics of AML and successfully establishes a prediction model based on machine learning, which demonstrates good accuracy and reliability in predicting the prognosis of AML patients.

* Corresponding author.
E-mail address: JieW629@163.com (J. Wang).

https://doi.org/10.1016/j.heliyon.2025.e42030
Received 13 March 2024; Received in revised form 30 July 2024; Accepted 15 January 2025
Available online 19 January 2025

1. Introduction

Acute myeloid leukemia (AML) manifests as a malignant neoplasm originating from the bone marrow, encompassing over 50 % of all adult leukemias, thus rendering it the most prevalent form of acute leukemia in adults [1]. Despite the ongoing research dedicated to AML over the past few decades, providing us with greater insights into its incidence, clinical presentations, and therapeutic modalities, the overall survival rates remain dishearteningly low, with a mere 20 %–30 % 5-year survival rate [2]. This holds particularly true for high-risk and elderly patients, among whom the prospects of survival are worrisome. Consequently, offering accurate survival prognoses and devising effective personalized treatment strategies for AML patients remains a formidable challenge. With the continual emergence of tumor-associated biomarker discoveries and survival analysis methodologies, various survival analysis techniques, such as Kaplan-Meier curves and Cox proportional hazard models, have been employed to widen the scope of survival prediction research. However, the intricacy, heterogeneity, and individual variability inherent in AML present limitations to the application of these conventional approaches in survival prediction. Confronted with voluminous and complex clinical data, encompassing biochemical and genetic profiles, researchers must seek novel perspectives and methodologies to handle such data and achieve heightened accuracy in survival prognoses. In recent years, machine learning techniques have demonstrated substantial potential in the realm of medical healthcare, particularly in leveraging vast heterogeneous datasets for disease prediction, treatment selection, and patient risk stratification [3]. By incorporating machine learning models, it becomes possible to analyze multifaceted numerical, structured, and unstructured data, ultimately constructing prediction models of greater precision [4]. Hence, integrating complex survival analysis methodologies with machine learning models holds practical significance in the realm of patient survival prediction research [5].

The present study aims to utilize the large-scale AML patient data from the SEER database, firstly performing traditional survival analysis methods such as Kaplan-Meier and Cox regression to explore the relationship between clinical characteristics and survival time. On this basis, we apply 11 widely used machine learning models, including decision trees, support vector machines, random forests, and neural networks, for AML patient survival prediction to improve the accuracy of 1-, 2-, and 3-year survival status prediction. The study flow of this research showed in Supplementary Figure 1 By comparing the prediction models of traditional survival analysis methods with those of machine learning methods, we identify key clinical features and risk factors in predicting the survival results, providing more targeted suggestions for clinicians to develop personalized treatment strategies. By integrating existing clinical feature data, traditional survival analysis, and machine learning techniques, this study brings a new perspective to the future of AML patient survival prediction research, with the aim of improving the survival rate and quality of life for AML patients. The study aims to promote the development of AML patient survival prediction and precision medicine, providing strong support for clinical practice.

2. Materials and methods

2.1. Data source

The data used in this study was sourced from the Surveillance, Epidemiology, and End Results (SEER) database (SEER*Stat version 8.4.1) of the National Cancer Institute in the United States. The SEER database is an epidemiological resource that covers the entire nation and aims to collect and report incidence and survival data for various cancer diseases. This study analyzed data of patients with Acute Myeloid Leukemia (AML) between 1975 and 2019. The data collected by the SEER database was provided by multiple registries located in different geographical regions. This study covers data from 18 SEER registries, including regions such as California, Kentucky, Missouri, New Jersey, Connecticut, Florida, Georgia, Illinois, North Carolina, and Washington state, which ensures a representative sample for our study.

We retrieved data from six SEER datasets, obtaining a total of 499,809 registered records of AML patients. Through deduplication using unique patient IDs, we obtained information for 295,624 patients. (Supplementary Table 1). This information included gender, age, race, year of diagnosis, number of occurrences, primary site, survival status, cause of death, survival time, average annual household income, residential city size, and ICD-O-3 classification code. To ensure the accuracy and consistency of the study, we performed further data filtering. We included only patients with a primary malignancy occurring once and limited the primary site to bone marrow. Additionally, we excluded patients with unknown race, incomplete survival information, unclear cause of death, and those diagnosed with a condition other than AML. Finally, our study sample consisted of 87,090 patient records.

We utilized the following histologic codes from the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) to identify patients: 9840 (Acute erythroid leukemia), 9861 (Acute myeloid leukemia, not otherwise specified), 9865 (Acute myeloid leukemia with t(6; 9)(p23; q34), DEK-NUP214), 9866 (Acute promyelocytic leukemia with t(15; 17)(q22; q11-12), PML/RARA), 9867 (Acute monocytic leukemia), 9869 (Acute myeloid leukemia with inv(3)(q21q26.2) or t(3; 3)(q21; q26.2), RPN1-EVI1), 9871 (Acute myelomonocytic leukemia with abnormal eosinophils), 9872 (Acute myeloid leukemia, minimally differentiated), 9873 (Acute myeloid leukemia, without maturation), 9874 (Acute myeloid leukemia, with maturation), 9891 (Acute monoblastic leukemia), 9895 (Acute myeloid leukemia with multilineage dysplasia), 9896 (Acute myeloid leukemia with t(8; 21)(q22; q22)), 9897 (Acute myeloid leukemia with 11q23 abnormalities), 9898 (Myeloid leukemia associated with Down syndrome), and 9910 (Acute megakaryoblastic leukemia).

2.2. Data analysis

We utilized univariate and multivariate COX regression analyses to compute hazard ratios and confidence intervals in order to

ascertain the relationship between variables and survival outcomes. In statistical terms, we define a p-value less than 0.05 as being indicative of significant difference. To visualize the COX proportional hazards model, we employed the survival package [6] for model construction and generated a nomogram using the regplot package.

In the process of data manipulation, we categorized the age variable into the following groups: 0–18 years, 19–46 years, 46–64 years, 65–80 years, and over 80 years of age. The diagnostic year variable was divided into the following groups: 1975–1989, 1990–1999, 2000–2009, and 2010–2019. The variable of annual family income was classified into the following groups: less than 50,000, 50,000–70,000 more than 70,000 and unknown. The size of the residential city was classified into the following groups: over 25 million inhabitants, adjacent to a city, non-adjacent to a city, and unknown. AML patients were categorized into those with WHO classification, FAB classification, and those with unknown classification. We analyzed whether there were differences in survival status among the different groups through Kaplan-Meier curves.

We selected variables such as gender, age, diagnostic year, race, residential city size, annual family income, and ICDO3 code as predictive factors and constructed a machine learning model to predict 1-year, 2-year, and 3-year survival rates for patients. We divided all patients into training and test sets in a ratio of 7:3. In the process of model construction, we utilized 11 models including XGBoost classifier [7], cross-validated glmnet classifier [8], K-Nearest Neighbor classifier [9], Linear Discriminant Analysis classifier [10], Logistic Regression classifier [11], Naive Bayes classifier [12], Neural Network classifier [13], Quadratic Discriminant Analysis classifier [14], Random Forest classifier [15], Decision Tree classifier [16] and Support Vector Machine classifier [17]. We employed

Table 1
Patients' baseline characteristics.

	Year of Dignosis	1975–1989 (N = 8310)	1990–1999 (N = 10624)	2000–2009 (N = 30248)	2010–2019 (N = 39528)	Overall (N = 88710)
Sex	Female	3697 (44.5 %)	4733 (44.6 %)	13625 (45.0 %)	18368 (46.5 %)	40423 (45.6 %)
	Male	4613 (55.5 %)	5891 (55.5 %)	16623 (55.0 %)	21160 (53.5 %)	48287 (54.4 %)
Age	Mean (SD)	60.4 (20.3)	61.0 (20.8)	62.9 (19.8)	59.0 (21.5)	60.7 (20.8)
	Median [Min, Max]	65.0 [0, 85.0]	67.0 [0, 85.0]	68.0 [0, 85.0]	64.0 [0, 85.0]	66.0 [0, 85.0]
Race	Black	794 (9.6 %)	859 (8.1 %)	2828 (9.3 %)	4020 (10.2 %)	8501 (9.6 %)
	Other (American Indian/AK Native, Asian/ Pacific Islander)	511 (6.1 %)	1201 (11.3 %)	2513 (8.3 %)	4067 (10.3 %)	8292 (9.3 %)
	White	7005 (84.3 %)	8564 (80.6 %)	24907 (82.3 %)	31441 (79.5 %)	71917 (81.1 %)
Vital_status	Alive	25 (0.3 %)	114 (1.1 %)	1187 (3.9 %)	13737 (34.8 %)	15063 (17.0 %)
	Dead	8285 (99.7 %)	10510 (98.9 %)	29061 (96.1 %)	25791 (65.2 %)	73647 (83.0 %)
Cause_of_death	Dead (attributable to causes other than this cancer dx)	820 (9.9 %)	1037 (9.8 %)	2556 (8.5 %)	2636 (6.7 %)	7049 (7.9 %)
	Alive or dead due to cancer	7490 (90.1 %)	9587 (90.2 %)	27692 (91.5 %)	36892 (93.3 %)	81661 (92.1 %)
Type	FAB Classification	1751 (21.1 %)	3077 (29.0 %)	11477 (37.9 %)	11359 (28.7 %)	27664 (31.2 %)
	unclear	6559 (78.9 %)	7529 (70.9 %)	16154 (53.4 %)	23031 (58.3 %)	53273 (60.1 %)
	WHO Classification	0 (0 %)	18 (0.2 %)	2617 (8.7 %)	5138 (13.0 %)	7773 (8.8 %)
City of residence	Counties in metropolitan areas ge 1 million pop	0 (0 %)	7768 (73.1 %)	18593 (61.5 %)	23220 (58.7 %)	49581 (55.9 %)
	Counties in metropolitan areas of 250,000 to 1 million pop	0 (0 %)	1526 (14.4 %)	5798 (19.2 %)	8095 (20.5 %)	15419 (17.4 %)
	Counties in metropolitan areas of lt 250 thousand pop	0 (0 %)	106 (1.0 %)	2210 (7.3 %)	3202 (8.1 %)	5518 (6.2 %)
	Nonmetropolitan counties adjacent to a metropolitan area	0 (0 %)	366 (3.4 %)	2176 (7.2 %)	2744 (6.9 %)	5286 (6.0 %)
	Nonmetropolitan counties not adjacent to a metropolitan area	0 (0 %)	268 (2.5 %)	1456 (4.8 %)	2266 (5.7 %)	3990 (4.5 %)
	Unknown/missing/no match/Not 1990–2018	8310 (100 %)	590 (5.6 %)	15 (0.0 %)	1 (0.0 %)	8916 (10.1 %)
Household_income	median_household_income<50000$	0 (0 %)	356 (3.4 %)	3936 (13.0 %)	6474 (16.4 %)	10766 (12.1 %)
	median_household_income50000-70000$	0 (0 %)	3355 (31.6 %)	13033 (43.1 %)	18786 (47.5 %)	35174 (39.7 %)
	median_household_income>70000$	0 (0 %)	6907 (65.0 %)	13264 (43.9 %)	14267 (36.1 %)	34438 (38.8 %)
	Unknown	8310 (100 %)	6 (0.1 %)	15 (0.0 %)	1 (0.0 %)	8332 (9.4 %)

5-fold cross-validation with 20 repetitions to select the optimal parameters for each model and used these optimal parameters to construct the prediction model. Subsequently, we evaluated model performance on the test set, including metrics such as accuracy, area under the curve, balanced accuracy, cross-entropy, logarithmic loss, precision, and recall. All of the aforementioned data manipulation, model construction, validation and evaluation were performed using R (version 4.3.0).

3. Results

3.1. Baseline characteristics

Table 1 presents the baseline clinical characteristics of the patients. A total of 87,090 eligible patients were included in our study.

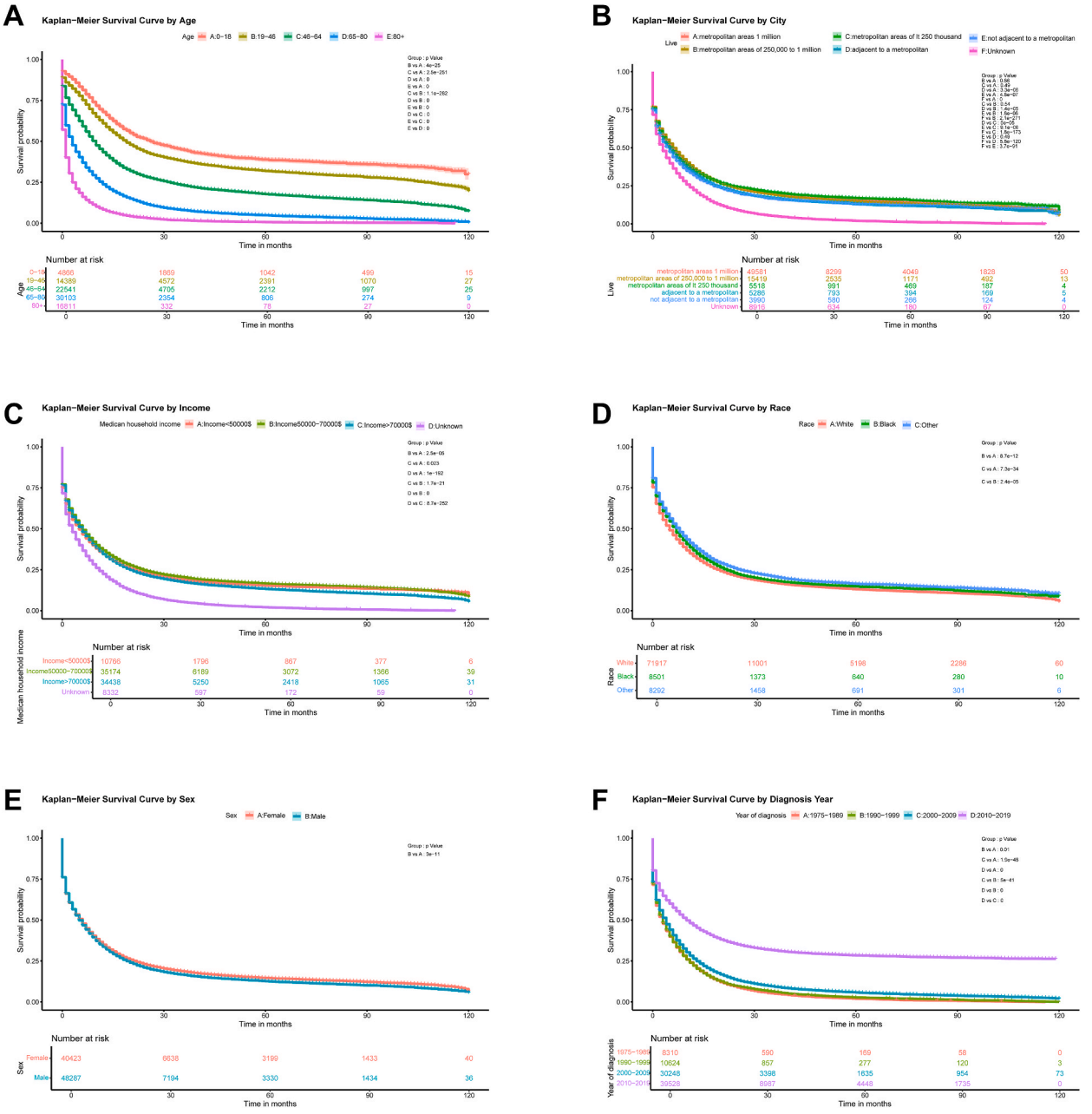

Fig. 1. KM curve analysis displaying survival rate trends in patients. The graph illustrates the impact of clinical baseline characteristics on survival rate for all patients. ABCDEF correspond to age, gender, diagnosis year, race, city size of residence, and household income, respectively. The survival status of patients is visualized through Kaplan Meier (KM) curves.

The data reveals a relatively balanced distribution of gender, with females accounting for 45.6 % (39,696) of the total population. The median age of the patients was 66 years, ranging from birth to ≥85 years. The median ages for AML diagnosis in the years 1975–1989, 1990–1999, 2000–2009, and 2010–2019 were 65, 67, 68, and 64, respectively. The majority of the patients were of White ethnicity (81.1 %), while the remaining consisted of Black individuals (9.6 %), American Indian/Alaska Natives, Asian/Pacific Islanders (9.4 %). A total of 78.3 % of the patients had a household annual income above 50,000,with nearly half of them reaching 70,000. Of the patients, 48,505 (55.7 %) resided in metropolitan areas with a population greater than 1 million, while only 3931 patients (4.5 %) lived in rural areas far from cities. Additionally, it should be noted that the AML subtypic information for all patients in the SEER database was not explicitly recorded. Out of the total, 27,664 patients (31.2 %) were classified according to the FAB classification system, while 7773 patients (8.8 %) were classified according to the WHO classification system. However, more than half of the patients (60.1 %) were simply identified as AML without providing further detailed subtype information.

3.2. Survival analysis

To eliminate the survival time disparities caused by the year of diagnosis, the follow-up period in this study was capped at 120 months. The overall survival based on patients' clinical characteristics is depicted in Fig. 1 using the Kaplan-Meier curve. It is observed that the older the age at diagnosis, the worse the prognosis, suggesting age as an important prognostic factor. Male patients exhibit poorer outcomes compared to females, hinting at gender as a potential prognostic factor. Patients diagnosed between 2010 and 2019 demonstrate significantly better prognosis compared to those diagnosed earlier, indicating a potential marked improvement in the treatment of AML after 2010. American Indian/Alaska Native and Asian/Pacific Islander individuals display more favorable prognoses than Black individuals, while White individuals have the poorest prognosis. This suggests a potential association between AML prognosis and race. Interestingly, the size of the patient's residential city and the annual household income seem unrelated to prognosis.

After performing KM analysis on two types of AML classification (FAB and WHO), we observed some clear trends, as shown in Fig. 2. Within the FAB classification, Acute Promyelocytic Leukemia (M3) exhibited the best prognosis, followed by M7 subtype, while prognosis in other FAB subtypes of AML remained similar. According to reports, M3 AML is typically associated with lower white blood cell counts, fewer sequential associations, and a better response to treatment with all-trans retinoic acid [18]. Acute Megakaryoblastic Leukemia (M7) is also believed to have peculiar treatment response mechanisms and disease traits [19]. In the WHO classification, we observed that AML associated with Down syndrome had the best prognosis, followed by AML with eosinophilia and then by t(8; 21)(q22; q22), t(1; 22)(p13; q13), t(6; 9)(p23; q34), 11q23 abnormalities, t(3; 3)(q21; q26.2), and AML with multilineage dysplasia. These specific molecular alterations may be linked to varying treatment responses and survival rates. With regard to Down syndrome, individuals with Down syndrome have an increased risk of developing leukemia from birth, but those with AML associated with Down syndrome exhibit higher survival rates compared to other AML subtypes. This may be attributed to certain genetic features and epigenetic changes that render individuals with Down syndrome more sensitive to treatment [20]. Additionally, AML patients with eosinophilia abnormalities may have a distinct prognosis, which could be related to gene rearrangement and signaling abnormalities in these patients [21].

Single-factor Cox regression analysis was performed on sex, age, year of diagnosis, race, family income, and size of residential city. Variables with P-values less than 0.05 were included in the multivariable Cox regression analysis, and the results were presented in a forest plot shown in Fig. 3. The results indicate that compared to females, males have a higher risk of death (OS, HR = 1.03, P < 0.001, 95%CI: 1.01–1.04). Patients diagnosed at an older age have a higher risk of death than those diagnosed at a younger age (OS, HR = 1.03, P < 0.001, 95%CI: 1.028–1.029). Patients diagnosed in 1990–1999 (OS, HR = 0.30, P < 0.001, 95%CI: 0.17–0.53), 2000–2009 (OS, HR = 0.25, P < 0.001, 95%CI: 0.14–0.43), and 2010–2019 (OS, HR = 0.14, P < 0.001, 95%CI: 0.08–0.25) have significantly lower

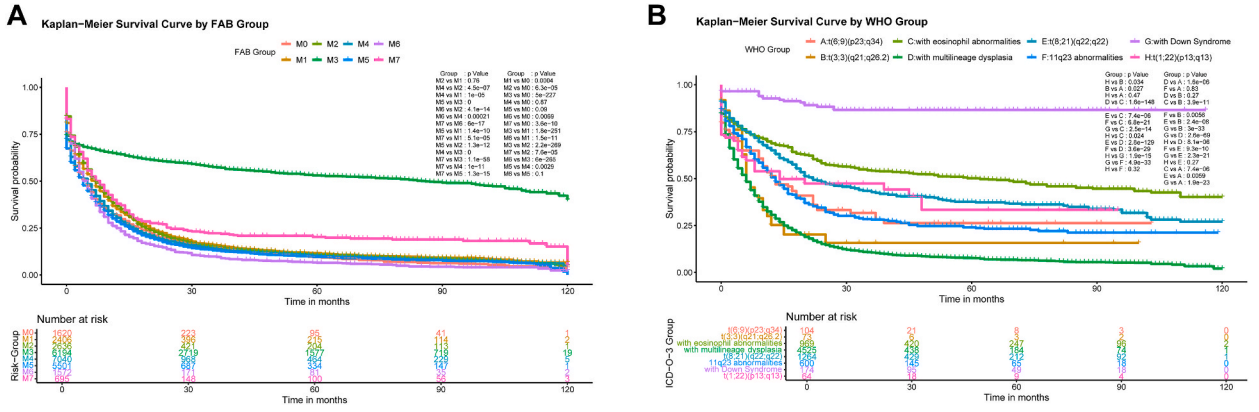

Fig. 2. KM curve analysis based on different AML classification criteria. This graph presents KM curve analysis based on different classification criteria. A is based on the FAB classification and B is based on the WHO classification. The impact of different classification criteria on survival rate is compared through Kaplan Meier (KM) curves.

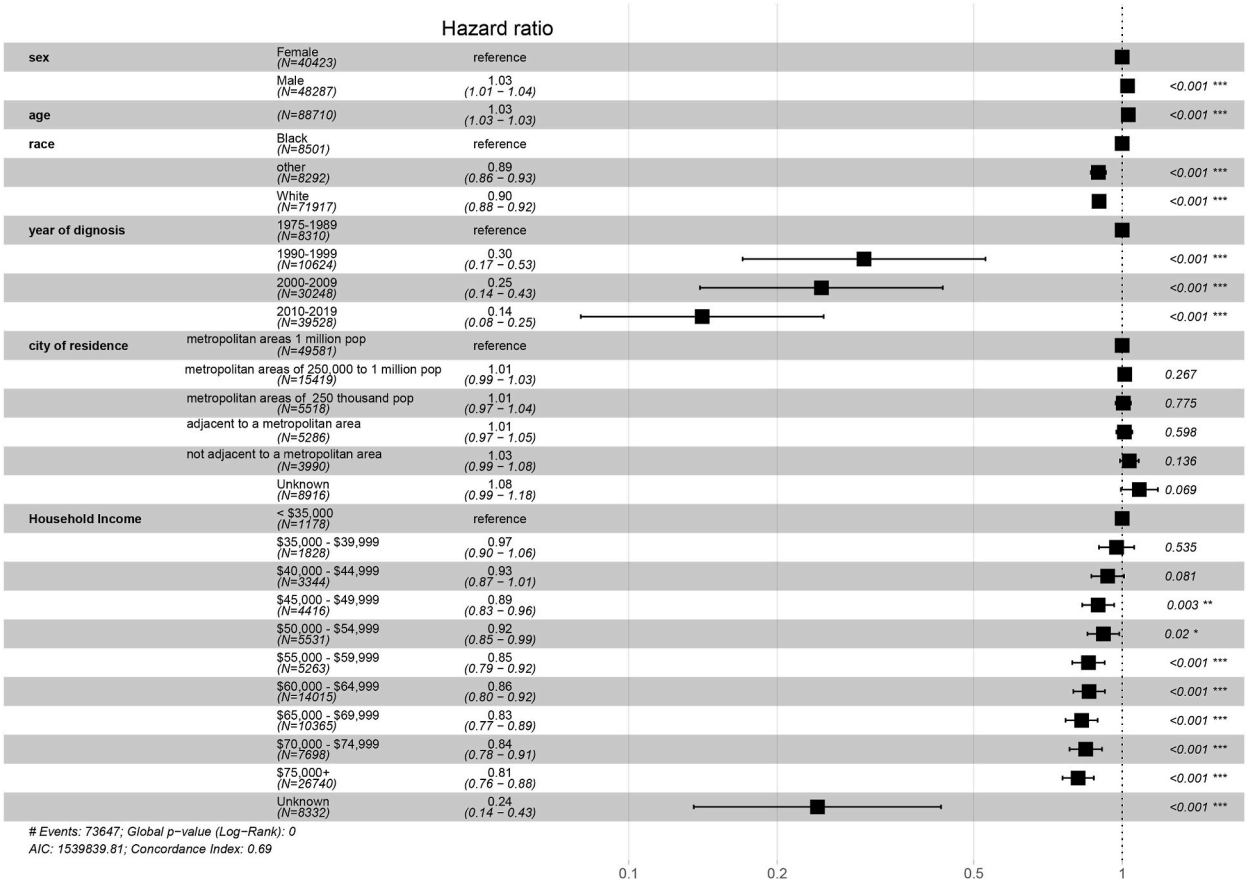

Fig. 3. Forest plot illustrating the association between clinical baseline characteristics and survival rate in patients. This figure is a forest plot showing the correlation between clinical baseline characteristics and survival rate in a survival analysis. Each square represents a Cox proportional hazards regression model, with the horizontal axis representing the effect size of the variable and the vertical line representing the confidence interval of the model.

risk of death compared to those diagnosed in 1975–1989. Patients with a family income above 45,000 per year have a significantly lower risk of death compared to those with a family income below 35,000 per year, and the risk of death decreases more significantly with an increase in family income. In contrast, the size of the residential city where the patient lives does not significantly correlate with survival. A column chart based on the COX regression model was developed to facilitate the model's clinical utilization, and scores for each variable were calculated based on its weight to predict a patient's survival probability, as shown in Fig. 4. In summary, according to the COX regression results, sex, age, race, and family income are significantly associated with the survival of leukemia patients, while the size of the residential city where the patient lives is not significantly associated with survival.

3.3. Machine learning

In this study, we utilized a dataset of 60,963 patients to train 11 machine learning models for predicting survival status at 1, 2, and 3 years after AML diagnosis. Hyperparameter tuning results for model training are documented in Supplementary Table 2. We then validated and evaluated the predictive performance of these models using the remaining dataset of 26,127 patients. Performance metrics were obtained for each model, and visualized the accuracy of the models (Fig. 5), the AUC (Fig. 6, Supplementary Fig. 2), as well as the confusion matrix (Supplementary Fig. 3). we predicted the survival status of patients at 1-year, 2-year, and 3-year time points using 11 different models. The evaluation metrics for each model are presented in Tables 2–4, respectively.

For the 1-year survival prediction, the Random Forest Classifier had the highest AUC value (0.766) and classification accuracy (0.741) among all models, indicating a high predictive accuracy for 1-year survival prediction. However, it's worth noting that the XGBoost Classifier, Neural Network Classifier, and Linear Discriminant Analysis Classifier also demonstrated good performance in most evaluation metrics. On the other hand, the Naive Bayes Classifier, K-Nearest Neighbor Classifier, and Decision Tree Classifier showed relatively poor performance, particularly the K-Nearest Neighbor Classifier, whose logloss value was much higher than that of other models.For the 2-year survival prediction, the Random Forest Classifier still exhibited high AUC value (0.793) and classification accuracy (0.831). The XGBoost Classifier and Neural Network Classifier also showed good performance in all evaluation metrics.

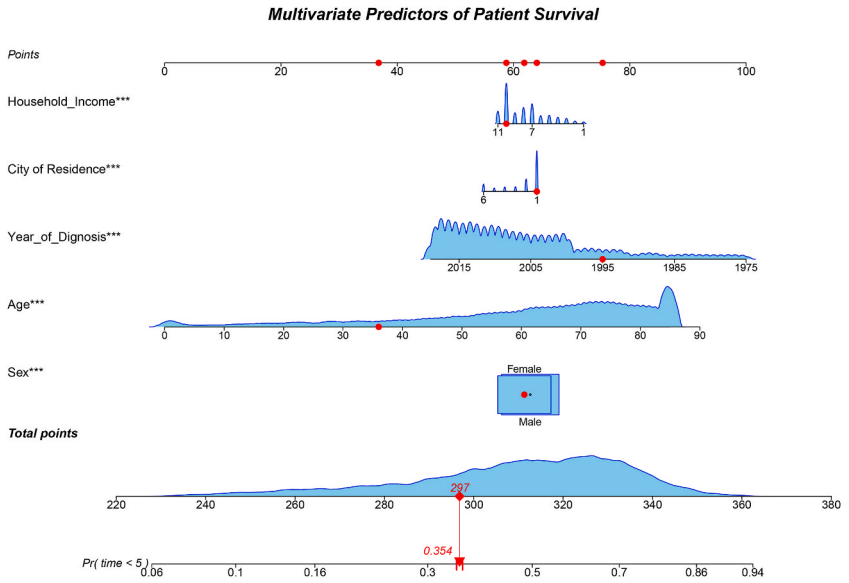

Fig. 4. Column chart analysis based on COX model for predicting patient survival rate. This figure shows an analysis of important variables for predicting patient survival rate based on the COX model. The red dot in the graph is a sample used to illustrate the contribution of a variable in the COX model. Each variable also has a score assigned to it to calculate the total score of the patient for predicting their survival rate.

However, the Decision Tree Classifier performed poorly in 2-year survival prediction, especially in terms of AUC and balanced accuracy. The typical Naive Bayes Classifier and Support Vector Machine Classifier had low values in several metrics, such as AUC and balanced accuracy, indicating poor performance.Regarding the 3-year survival prediction, the Random Forest Classifier once again had the highest AUC value (0.808). The Neural Network Classifier demonstrated satisfactory performance in most evaluation metrics. However, the Decision Tree Classifier, Naive Bayes Classifier, and Support Vector Machine Classifier all showed poor performance in metrics such as AUC and balanced accuracy.

Based on the above analysis, the Random Forest Classifier exhibits the overall best performance in multiple evaluation metrics for 1-year, 2-year, and 3-year survival predictions. In addition, the XGBoost Classifier and Neural Network Classifier also demonstrate high accuracy and reliability in all prediction phases. These models demonstrate strong abilities in handling survival prediction tasks with mixed factors, non-linear relationships, and high-dimensional data. Conversely, the K-Nearest Neighbor Classifier and Decision Tree Classifier based on distance metrics have relatively lower accuracy and stability in survival prediction.In further analysis, we calculated the contribution of each variable based on the best-performing XGBoost and Random Forest models, representing their importance as shown in Fig. 7. The results show that the ICDO3 classification, age, and household annual income are the most important variables in both models. Firstly, the ICDO3 classification refers to the pathological classification of tumors, which displays higher importance in both XGBoost and Random Forest models. This indicates that the specific type of tumor has a significant impact on survival predictions. Different types of tumors may have different biological characteristics and prognoses, and therefore, ICDO3 classification can provide critical information about the nature of the tumor. Secondly, age has been identified as an important predictive factor in both models, which is consistent with trends found in clinical practice and research. Age may be related to a patient's overall health status, immune function, and tolerance to treatment, and therefore plays a significant role in predicting patient survival rates. In addition, household annual income has also been identified as an important variable in both models. Household annual income may reflect the patient's economic status and social support level, which are associated with survival rates and treatment outcomes. Higher-income households may have better access to healthcare and support, which positively affects survival rates.It should be noted that the results of the contribution of variable importance are based on specific XGBoost and Random Forest models and may be influenced by specific conditions and samples of the research data. Therefore, when using these models for survival prediction, it is essential to consider the influence of these important variables.

In summary, for survival prediction in leukemia patients, using models such as Random Forest, XGBoost, and Neural Network may lead to improved prediction performance. However, in practical applications, other factors such as computational resources and model complexity may need to be considered. Suitable models can be chosen based on specific needs to obtain the best prediction results. These models can support clinicians in developing personalized treatment plans, determining patient follow-up strategies, and evaluating treatment outcomes, thereby improving the survival rate and quality of life for AML patients. In future research, these models can be further optimized to enhance prediction performance. Additionally, technological developments and new discoveries from the field of biomedical research may provide more possibilities for improving existing models and developing new methods.

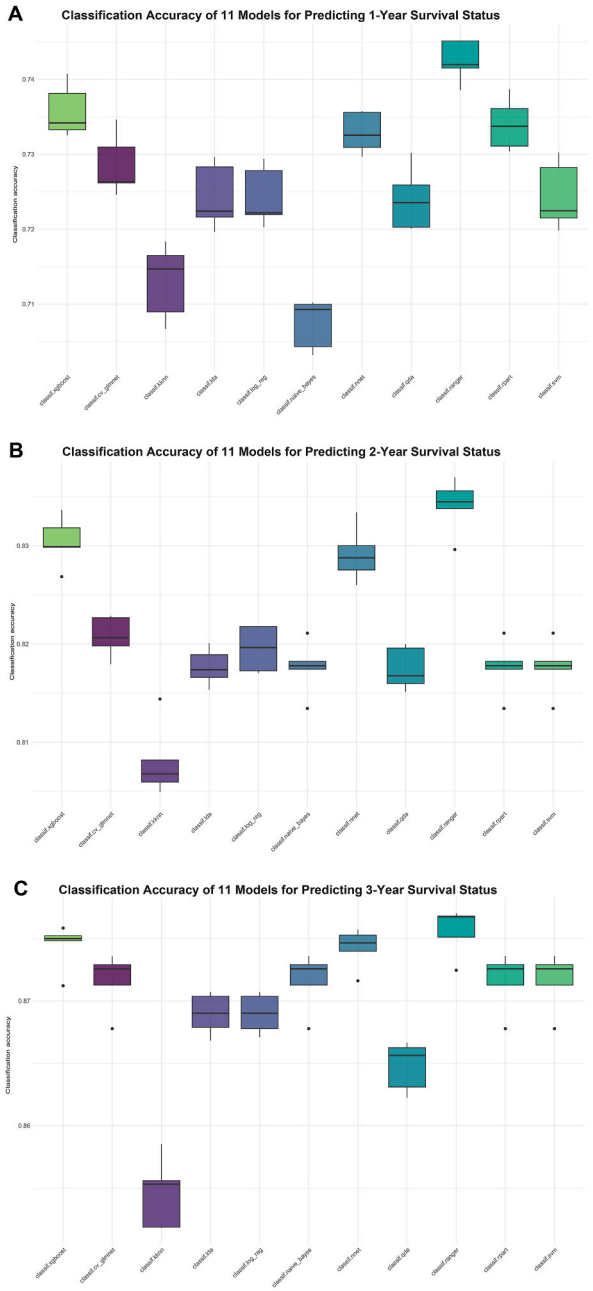

Fig. 5. Comparative boxplot analysis of predicting patient survival rate based on 11 machine learning models. This boxplot is based on 11 different machine learning methods for survival prediction models evaluating patient survival rates for 1, 2, and 3 years. These models include the XGBoost classifier, cross-validated glmnet classifier, K-nearest neighbor classifier, linear discriminant analysis classifier, logistic regression classifier, naive Bayes classifier, neural network classifier, quadratic discriminant analysis classifier, random forest classifier, decision tree classifier, and support vector machine classifier. Each box represents the median value of five non-overlapping five-fold cross-validation results, with the top and bottom edges representing the 75th and 25th percentile, respectively. The median value reflects the accuracy of classification, which indicates the prediction accuracy of different models. ABC depict the corresponding prediction results for 1, 2, and 3 years, respectively.

4. Discussion

Acute myeloid leukemia (AML) stands as one of the most formidable subtypes of leukemia worldwide. Approximately 30 % of leukemia-related deaths globally can be attributed to AML, despite accounting for only 40 % of leukemia-related mortality cases. Moreover, the incidence of AML has shown a gradual increase throughout the past decade, surpassing chronic lymphocytic leukemia (CLL), acute lymphoblastic leukemia (ALL), and chronic myeloid leukemia (CML) in frequency [22]. In most countries, adult AML

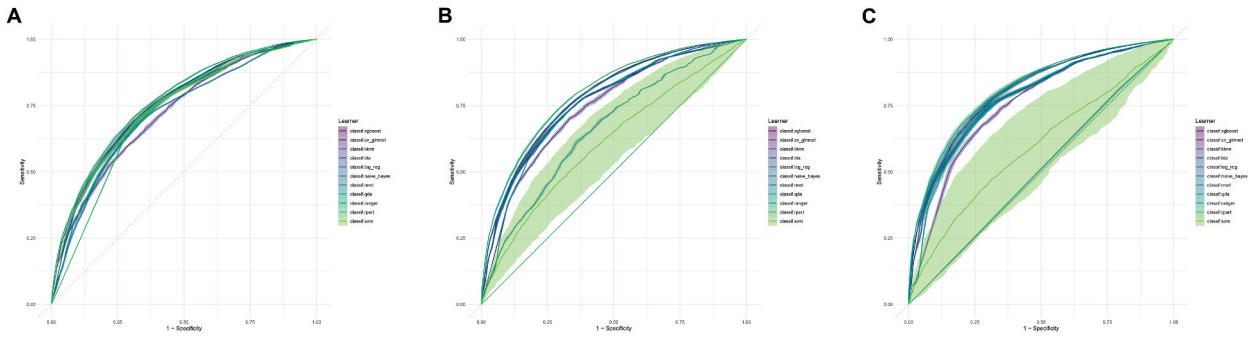

Fig. 6. Comparative ROC curve analysis of predicting patient survival rate based on 11 machine learning models. This figure demonstrates the ROC curves based on 11 different machine learning methods for survival prediction models for 1, 2, and 3 years. Each model's ROC curve depicts the average of five non-overlapping five-fold cross-validation results. The ROC curve is plotted by setting the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR), also known as sensitivity, as the vertical axis, describing the performance of the classifier under different threshold settings. AUC (area under curve) is commonly used as an evaluation index. In this figure, ABC represent the corresponding ROC curves for 1, 2, and 3 years, respectively.

Table 2
Evaluation metrics for one-year forecasting models.

	Classification accuracy	Area Under the Curve (AUC)	Balanced accuracy	Cross-entropy	Logarithmic loss	Precision	Recall
XGBoost Classifier	0.7362	0.7567	0.6289	0.2638	0.5282	0.7531	0.9162
glmnet Classifier with Cross-Validation	0.7285	0.7457	0.6044	0.2715	0.5457	0.7383	0.9366
k-Nearest Neighbor Classifier	0.7056	0.5933	0.5429	0.2944	7.2589	0.7062	0.9784
Linear Discriminant Analysis (LDA) Classifier	0.7243	0.7469	0.6206	0.2757	0.5386	0.7498	0.8984
Logistic Regression Classifier	0.7243	0.7467	0.6173	0.2757	0.5383	0.7475	0.9037
Naive Bayes Classifier	0.7062	0.7205	0.5865	0.2938	0.5584	0.7304	0.9069
Neural Network Classifier	0.7338	0.7547	0.6245	0.2662	0.5303	0.7506	0.9172
Quadratic Discriminant Analysis (QDA) Classifier	0.7230	0.7352	0.6315	0.2770	0.5513	0.7580	0.8765
Random Forest Classifier	0.7414	0.7662	0.6423	0.2586	0.5202	0.7616	0.9076
Decision Tree Classifier	0.7334	0.7193	0.6173	0.2666	0.5461	0.7458	0.9281
Support Vector Machine (SVM) Classifier	0.7245	0.7463	0.6217	0.2755	0.5396	0.7506	0.8968

exhibits a slight male predominance over females. For the period between 2000 and 2003, the age-adjusted incidence rate of AML in the United States was recorded as 3.7 per 100,000 population, with rates of 4.6 per 100,000 population for males and 3.0 per 100,000 population for females. AML incidence also displays variation among different racial and ethnic groups. Historically, AML was more prevalent in the black population before 2000; however, subsequent records reveal a lower incidence rate in black individuals compared to their white counterparts [23]. The present findings align closely with the baseline data collected in this study, which indicates a higher proportion of males (55.4 %) diagnosed with AML compared to females (45.6 %), and an increasing trend in AML diagnoses among the black population. This suggests that males may have a higher susceptibility to AML due to physiological or lifestyle differences, while notable disparities in incidence rates persist among different racial and ethnic groups.AML constitutes a malignant neoplasm of the hematopoietic system, characterized by aggressive behavior. The survival rates for AML patients are influenced by a multitude of factors. Notably, a study has demonstrated a significant relationship between age, sex, race, and patient survival rates [22]. Through our analysis of different patient subgroups, we have observed a close association between age and patient survival, with older individuals experiencing progressively lower survival rates. One plausible reason for the lower survival rates among older patients is their unsuitability for bone marrow transplantation, which can be attributed to the presence of comorbidities and compromised physical condition that make them unable to tolerate the high-intensity chemotherapy associated with this

Table 3
Evaluation metrics for two-year forecasting models.

	Classification accuracy	Area Under the Curve (AUC)	Balanced accuracy	Cross-entropy	Logarithmic loss	Precision	Recall
XGBoost Classifier	0.8293	0.7815	0.5898	0.1707	0.3949	0.8463	0.9668
glmnet Classifier with Cross-Validation	0.8215	0.7638	0.5361	0.1785	0.4099	0.8287	0.9853
k-Nearest Neighbor Classifier	0.8228	0.6114	0.5294	0.1772	4.3368	0.8266	0.9913
Linear Discriminant Analysis (LDA) Classifier	0.8182	0.7648	0.5555	0.1818	0.4070	0.8350	0.9690
Logistic Regression Classifier	0.8206	0.7636	0.5395	0.1794	0.4068	0.8298	0.9820
Naive Bayes Classifier	0.8176	0.6608	0.5000	0.1824	0.4601	0.8176	1.0000
Neural Network Classifier	0.8291	0.7785	0.5896	0.1709	0.3969	0.8462	0.9666
Quadratic Discriminant Analysis (QDA) Classifier	0.8189	0.7527	0.6042	0.1811	0.4231	0.8520	0.9422
Random Forest Classifier	0.8310	0.7929	0.6066	0.1690	0.3888	0.8521	0.9599
Decision Tree Classifier	0.8176	0.5000	0.5000	0.1824	0.4750	0.8176	1.0000
Support Vector Machine (SVM) Classifier	0.8176	0.3652	0.5000	0.1824	0.4988	0.8176	1.0000

Table 4
Evaluation metrics for three-year forecasting models.

	Classification accuracy	Area Under the Curve (AUC)	Balanced accuracy	Cross-entropy	Logarithmic loss	Precision	Recall
XGBoost Classifier	0.8730	0.7985	0.5445	0.1270	0.3154	0.8818	0.9865
glmnet Classifier with Cross-Validation	0.8716	0.5000	0.5000	0.1284	0.3833	0.8716	1.0000
k-Nearest Neighbor Classifier	0.8722	0.6041	0.5170	0.1278	3.2001	0.8755	0.9949
Linear Discriminant Analysis (LDA) Classifier	0.8680	0.7783	0.5387	0.1320	0.3275	0.8805	0.9818
Logistic Regression Classifier	0.8678	0.7774	0.5375	0.1322	0.3269	0.8802	0.9820
Naive Bayes Classifier	0.8716	0.5076	0.5000	0.1284	0.4334	0.8716	1.0000
Neural Network Classifier	0.8730	0.7982	0.5613	0.1270	0.3154	0.8858	0.9807
Quadratic Discriminant Analysis (QDA) Classifier	0.8621	0.7700	0.5612	0.1379	0.3455	0.8860	0.9660
Random Forest Classifier	0.8743	0.8083	0.5537	0.1257	0.3101	0.8840	0.9851
Decision Tree Classifier	0.8716	0.5000	0.5000	0.1284	0.3833	0.8716	1.0000
Support Vector Machine (SVM) Classifier	0.8716	0.5072	0.5000	0.1284	0.3833	0.8716	1.0000

procedure. However, it should be noted that some older patients may still derive benefits from intensive chemotherapy [24]. Additionally, the KM analysis results indicate significantly lower survival rates in males compared to females. Researchers have uncovered disparities in survival rates between sexes among both pediatric and adult AML patients [25,26]. One possible explanation is that male patients may often present with more advanced disease stages at the time of diagnosis or may have a higher propensity for malignant transformation of AML, leading to poorer treatment outcomes [26]. Furthermore, variations in survival rates among AML patients of different racial and ethnic backgrounds have also been observed. These differences may stem from substantial genetic variations between racial groups, potentially influencing treatment responses [27]. Moreover, in this study, it was found that patients with lower household incomes exhibited significantly lower survival rates compared to those with higher incomes. This suggests that AML survival rates are influenced not only by physical status, physiological indicators, and genetic backgrounds but also by differences in socio-economic status and income disparities, which can result in varying levels of treatment accessibility and care [28]. In conclusion, these preliminary findings highlight the importance of stratifying future investigations of AML's biology and epidemiology according to age, sex, race, and socio-economic background.

The potential of machine learning in the field of cancer, specifically in the case of AML, a type of aggressive hematologic malignancy, is significant. In early diagnosis of AML, machine learning can utilize vast amounts of clinical and molecular data to construct predictive models that enhance the accuracy and sensitivity of early detection [29]. For instance, the machine learning model

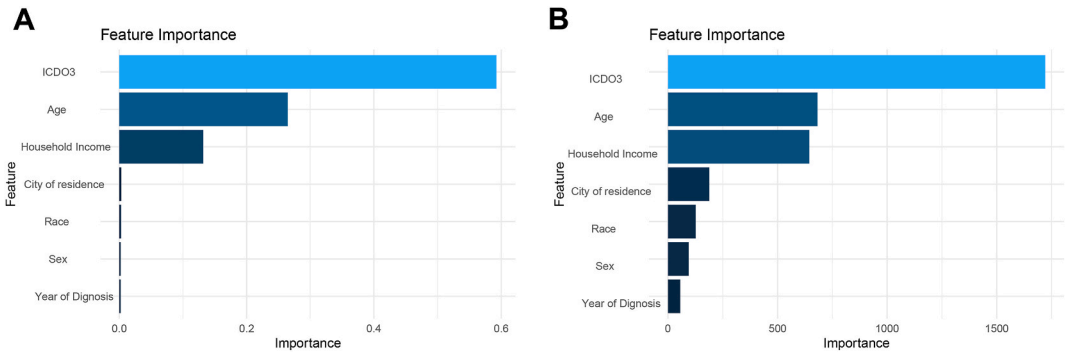

Fig. 7. Variable importance analysis based on XGBoost and random forest models. This graph displays a variable importance analysis based on XGBoost and random forest models. Variables are ranked in descending order according to their importance and compared between the two models. A and B represent the contributions of variables in the XGBoost and random forest models, respectively. These contributions are measured by calculating the relative weights of the parameters, with higher weights indicating greater contribution of variables in the model.

developed by Liu et al. [30]. For automatic classification of AML subtypes via bone marrow images has demonstrated remarkably high levels of accuracy, hastening the diagnostic and treatment processes. Moreover, through the analysis of patient gene expression, mutation information, and clinical characteristics, machine learning can identify potential AML risk factors and biomarkers. Additionally, based on patient biological and disease characteristics, machine learning can construct predictive models to forecast patient survival rates and treatment responses [31]. Eckardt et al. [32] gathered clinical and genetic information from patients and employed machine learning to construct a prognostic model that predicts patient outcomes. This model exhibited favorable performance in both internal and external validations. Furthermore, other researchers have utilized bioinformatics analysis to identify genes associated with ferroptosis [33,34], oxidative stress [35], aging [36], cuproptosis [37], and immune microenvironment [38]. These genes were used as variables to construct prognostic models for predicting patient outcomes, yielding promising results as well. However, most of these models were analyzed and validated using openly available datasets, necessitating further verification and evaluation of their application value and real-world performance. Various methods, including COX regression [39], support vector machines [29], random forests [31], and LASSO regression [40], have been employed to establish predictive models, but there is limited research on the differences and characteristics among these models. Therefore, this study utilized 11 machine learning models to construct predictive models for the survival status of AML patients at 1, 2, and 3 years. The aim of this research was to explore and compare the characteristics of these models, seeking beneficial guidance for constructing AML survival prediction models in the future. We evaluated the performance indicators of these models, including accuracy, recall rate, and the area under the AUC curve, comparing and discussing the performance differences among the models as well as the reasons behind such differences. These findings can reveal the advantages and limitations of each model in predicting the survival status of AML patients, providing guidance for the future development of efficient AML survival prediction models.

Among all the predictive models, the Random Forest classifier outperformed the others in several metrics for 1-year, 2-year, and 3-year survival prediction. Its robust performance stems from ensemble learning, which combines multiple different decision trees and aggregates their results through voting to enhance prediction accuracy. Additionally, Random Forest is capable of handling highly complex and nonlinear relationships within data, thereby exhibiting strong predictive ability in this study. The XGBoost classifier also demonstrated good performance in survival prediction. It is a gradient boosting tree model that achieves optimal predictions by weighted accumulation of outputs from a series of weak models. XGBoost can capture intricate nonlinear relationships and perform feature selection, making it highly accurate in predicting the survival status of AML patients. The neural network classifier exhibited satisfactory performance in AML patient survival prediction as well. Generally, neural networks possess strong capabilities in processing nonlinear data and can adapt to various data scales and structures. Through parameter tuning and a combination of internal weights and activation functions, this model can learn effective feature representations from vast and intricate biomedical data. On the other hand, the Naive Bayes classifier, Decision Tree classifier, and k-Nearest Neighbors (k-NN) classifier exhibited lower accuracy and stability in survival prediction compared to other models. The Naive Bayes classifier is based on Bayes' theorem and assumes independence among input features. This assumption may oversimplify the real-world relationships, making it difficult for the model to capture complex information within the input data. The k-NN classifier is based on distance measurement, and in high-dimensional data, distance metrics can become ineffective, leading to decreased predictive performance. The Decision Tree classifier is susceptible to data noise and overfitting, which can hinder its predictive performance.In summary, when dealing with complex biomedical data to predict the survival status of AML patients, models such as Random Forest, XGBoost, and neural networks may exhibit higher predictive accuracy and stability. These models leverage their capabilities in handling nonlinear data and fitting high-dimensional data to uncover hidden structural information within the data and predict patient survival status. However, in practical applications, factors such as computational resources and model complexity need to be carefully considered to select the optimal model.

Compared with traditional methods of survival analysis, machine learning methods can handle larger datasets and uncover more subtle survival data features, improving the accuracy and robustness of models. Machine learning models can overcome noise interference in datasets, deal with nonlinear correlations, and help researchers better differentiate patient survival times and predict

risks. In large datasets such as the SEER database, there may be thousands of features that influence survival, but traditional survival analysis methods may not be able to capture interactions among these features. Machine learning methods can adaptively process these complex datasets and construct more accurate and robust models to predict survival time. Machine learning methods have higher flexibility and scalability, which allows for retraining and updating models to handle new data and new problems. This flexibility helps researchers navigate heterogeneity across different laboratories and data collection processes, and allows the models to adapt to small feature or dataset changes.

While machine learning methods have advantages in large-scale datasets, they still face several challenges. Compared to traditional survival analysis methods, the models produced by machine learning methods often lack clear understanding of the specific factors that influence the outcome, resulting in a loss of interpretability and difficulty in explaining the results generated by the models. This may lead to biologically implausible findings in the medical field, making it challenging to trust machine learning models. Machine learning methods require a substantial amount of data to train the models, and these data not only need to be abundant but also of high quality. The quality of data is a crucial factor in the success of machine learning models, and manually obtained data quality is often more susceptible to biases compared to naturally occurring diverse biological data. Machine learning methods also demand higher computational resources and time compared to traditional survival analysis methods. Enhancements in computing hardware and algorithm optimization are necessary to avoid issues such as excessively long training times and unreliable models.

5. Conclusions

Through in-depth analysis, this study provides a deeper understanding of the epidemiological characteristics of AML and successfully establishes a prediction model based on machine learning, which demonstrates good accuracy and reliability in predicting the prognosis of AML patients.

CRediT authorship contribution statement

Zheng-yi Jia: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Maierbiya Abulimiti:** Validation, Software, Methodology, Formal analysis, Data curation. **Yun Wu:** Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition. **Li-na Ma:** Validation, Methodology, Formal analysis, Data curation. **Xiao-yu Li:** Visualization, Software, Methodology, Formal analysis, Data curation. **Jie Wang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

The data used in this study was sourced from the Surveillance, Epidemiology, and End Results (SEER) database (SEER*Stat version 8.4.1) of the National Cancer Institute in the United States.

Ethics approval

Not applicable.

Funding

This study was supported by Xinjiang Uygur Autonomous Region Distinguished Young Scientists Fund Project (Founding NO.:2022D01E72) and Xinjiang Uygur Autonomous Region Youth Science and Technology Top Talent Project-Youth Science and Technology Innovation Talent Training (2022TSYCCX0027).

Declaration of competing interest

The authors claim to have no conflicts of interest.

Acknowledgments

We want to thank Jian-ping Hao, director of the First Affiliated Hospital of Xinjiang Medical University in Urumqi, China, for providing crucial assistance with fundraising and research projects.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2025.e42030.

References

[1] H. Döhner, D.J. Weisdorf, C.D. Bloomfield, Acute myeloid leukemia, N. Engl. J. Med. 373 (12) (2015) 1136–1152.
[2] C.E. DeSantis, C.C. Lin, A.B. Mariotto, et al., Cancer treatment and survivorship statistics, 2014, CA Cancer J Clin 64 (4) (2014) 252–271.
[3] R.C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.
[4] Z. Obermeyer, E.J. Emanuel, Predicting the future - big data, machine learning, and clinical medicine, N. Engl. J. Med. 375 (13) (2016) 1216–1219.
[5] K.H. Metzeler, T. Herold, M. Rothenberg-Thurley, et al., Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia, Blood 128 (5) (2016) 686–698.
[6] Terry M. Therneau, Patricia M. Grambsch, Modeling Survival Data: Extending the Cox Model, Springer, New York, 2000. ISBN 0-387-98784-3.
[7] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794.
[8] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Software 33 (1) (2010) 1–22.
[9] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1) (1967) 21–27.
[10] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (2) (1936) 179–188.
[11] D.R. Cox, Regression models and life-tables, J. Roy. Stat. Soc. B 34 (2) (1972) 187–220.
[12] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons, 1973.
[13] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
[14] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer Science & Business Media, 2009.
[15] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[16] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[17] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[18] C.C. Coombs, M.S. Tallman, R.L. Levine, Molecular therapy for acute myeloid leukaemia, Nat. Rev. Clin. Oncol. 13 (5) (2016) 305–318.
[19] H. Kuusanmäki, O. Dufva, M. Vähä-Koskela, et al., Erythroid/megakaryocytic differentiation confers BCL-XL dependency and venetoclax resistance in acute myeloid leukemia, Blood 141 (13) (2023) 1610–1625.
[20] M.M. O'Brien, X. Cao, S. Pounds, et al., Prognostic features in acute megakaryoblastic leukemia in children without Down syndrome: a report from the AML02 multicenter trial and the Children's Oncology Group Study POG 9421, Leukemia 27 (3) (2013) 731–734.
[21] A. Görgens, S. Radtke, M. Möllmann, et al., Revision of the human hematopoietic tree: granulocyte subtypes derive from distinct hematopoietic lineages, Cell Rep. 3 (5) (2013) 1539–1552.
[22] X. Song, Y. Peng, X. Wang, et al., Incidence, survival, and risk factors for adults with acute myeloid leukemia not otherwise specified and acute myeloid leukemia with recurrent genetic abnormalities: analysis of the surveillance, epidemiology, and End results (SEER) database, 2001-2013, Acta Haematol. 139 (2) (2018) 115–127.
[23] B. Deschler, M. Lübbert, Acute myeloid leukemia: epidemiology and etiology, Cancer 107 (9) (2006) 2099–2107.
[24] N.J. Short, M.E. Rytting, J.E. Cortes, Acute myeloid leukaemia2018, Lancet 392 (10147) (2018) 593–606.
[25] M.J. Hossain, L. Xie, Sex disparity in childhood and young adult acute myeloid leukemia (AML) survival: evidence from US population data, Cancer Epidemiol 39 (6) (2015) 892–900.
[26] M. Yi, A. Li, L. Zhou, Q. Chu, Y. Song, K. Wu, The global burden and attributable risk factor analysis of acute myeloid leukemia in 195 countries and territories from 1990 to 2017: estimates based on the global burden of disease study 2017, J. Hematol. Oncol. 13 (1) (2020) 72.
[27] H. Weng, F. Huang, Z. Yu, et al., The m6A reader IGF2BP2 regulates glutamine metabolism and represents a therapeutic target in acute myeloid leukemia, Cancer Cell 40 (12) (2022) 1566–1582.e10.
[28] H. Döhner, D.J. Weisdorf, C.D. Bloomfield, Acute myeloid leukemia, in: D.L. Longo (Ed.), N Engl J Med, vol. 373, 2015, pp. 1136–1152, 12.
[29] J.N. Eckardt, M. Bornhäuser, K. Wendt, J.M. Middeke, Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects, Blood Adv 4 (23) (2020) 6077–6085.
[30] K. Liu, J. Hu, Classification of acute myeloid leukemia M1 and M2 subtypes using machine learning, Comput. Biol. Med. 147 (2022) 105741.
[31] H. Awada, A. Durmaz, C. Gurnari, et al., Machine learning integrates genomic signatures for subclassification beyond primary and secondary acute myeloid leukemia, Blood 138 (19) (2021) 1885–1895.
[32] J.N. Eckardt, C. Röllig, K. Metzeler, et al., Prediction of complete remission and survival in acute myeloid leukemia using supervised machine learning, Haematologica 108 (3) (2023) 690–704.
[33] Y. Song, S. Tian, P. Zhang, N. Zhang, Y. Shen, J. Deng, Construction and validationof a novel ferroptosis-related prognostic model for acute myeloid leukemia, Front. Genet. 12 (2022 Jan 17) 708699.
[34] R. Shao, H. Wang, W. Liu, J. Wang, S. Lu, H. Tang, Y. Lu, Establishment of aprognostic ferroptosis-related gene profile in acute myeloid leukaemia, J. Cell Mol. Med. 25 (23) (2021 Dec) 10950–10960.
[35] C. Dong, N. Zhang, L. Zhang, The multi-omic prognostic model of oxidative stress-related genes in acute myeloid leukemia, Front. Genet. 12 (2021 Sep 30) 722064.
[36] H. Shi, L. Gao, W. Zhang, M. Jiang, Identification and validation of a siglec- based and aging-related 9-gene signature for predicting prognosis in acute myeloid leukemia patients, BMC Bioinf. 23 (1) (2022 Jul 19) 284.
[37] P. Li, J. Li, F. Wen, et al., A novel cuproptosis-related LncRNA signature: prognostic and therapeutic value for acute myeloid leukemia, Front. Oncol. 12 (2022) 966920.
[38] F. Zhong, F. Yao, Y. Cheng, et al., m6A-related lncRNAs predict prognosis and indicate immune microenvironment in acute myeloid leukemia, Sci. Rep. 12 (1) (2022) 1759.
[39] R. Shouval, M. Labopin, N.C. Gorin, et al., Individualized prediction of leukemia-free survival after autologous stem cell transplantation in acute myeloid leukemia, Cancer 125 (20) (2019) 3566–3573.
[40] S. Abelson, G. Collord, S.W.K. Ng, et al., Prediction of acute myeloid leukaemia risk in healthy individuals, Nature 559 (7714) (2018) 400–404.