BMC Medical Informatics and Decision Making

RESEARCH **Open Access**

Identification of relevant features using SEQENS to improve supervised machine learning models predicting AML treatment outcome

Pedro Pons-Suñer[1*], François Signol[1*], Noemi Alvarez[2], Claudia Sargas[3], Sara Dorado[4], Jose Vicente Gil Ortí[3], Juan A. Delgado Sanchis[1], Marta Llop[3], Laura Arnal[1], Rafael Llobet[1], Juan-Carlos Perez-Cortes[1], Rosa Ayala[2†] and Eva Barragán[3†]

Abstract

Background and objective This study has two main objectives. First, to evaluate a feature selection methodology based on SEQENS, an algorithm for identifying relevant variables. Second, to validate machine learning models that predict the risk of complications in patients with acute myeloid leukemia (AML) using data available at diagnosis. Predictions are made at three time points: 90 days, six months, and one year post-diagnosis. These objectives represent fundamental steps toward the development of a tool to assist clinicians in therapeutic decision-making and provide insights into the risk factors associated with AML complications.

Methods A dataset of 568 patients, including demographic, clinical, genetic (VAF), and cytogenetic information, was created by combining data from Hospital 12 de Octubre (Madrid, Spain) and Instituto de Investigación Sanitaria La Fe (Valencia, Spain). Feature selection based on an enhanced version of SEQENS was conducted for each time point, followed by the comparison of four classifiers (XGBoost, Multi-Layer Perceptron, Logistic Regression and Decision Tree) to assess the impact of feature selection on model performance.

Results SEQENS identified different relevant features for each prediction horizon, with Age, TP53, – 7/7Q, and EZH2 consistently relevant across all time points. The models were evaluated using 5-fold cross-validation, XGBoost achieve the highest average ROC-AUC scores of 0.81, 0.84, and 0.82 for 90-day, 6-month, and 1-year predictions, respectively. Generally, performance remained stable or improved after applying SEQENS-based feature selection. Evaluation on an external test set of 54 patients yielded ROC-AUC scores of 0.72 (90-day), 0.75 (6-month), and 0.68 (1-year).

†Rosa Ayala and Eva Barragán are joint senior authors.

*Correspondence:
Pedro Pons-Suñer
pedropons@iti.es
François Signol
fsignol@iti.es

Full list of author information is available at the end of the article

Conclusions The models achieved performance levels that suggest they could serve as therapeutic decision support tools at different times after diagnosis. The selected variables align with the European LeukemiaNet (ELN) 2022 risk classification, and the SEQENS-based feature selection effectively reduced the feature set while maintaining prediction accuracy.

Keywords Acute myeloid leukemia, Machine learning, Patient evolution, Therapy outcome, Recurrence, Mortality, Complications risk factors, Cytogenetics, Variant Allele Frequency

Introduction

Acute myeloid leukemia (AML) is a heterogeneous and multifactorial disease characterized by a wide range of molecular alterations leading to malignant transformation of hematopoietic stem cells [1]. It is the most frequent type of acute leukemia in adults with a 5-year relative survival of 30.5% reported by the Surveillance, Epidemiology, and End Results Program (SEER) in the USA (SEER12: https://seer.cancer.gov/statfacts/html/amyl.html). The European Cancer Information System (ECIS) reported a 5-year relative survival of 46.9% for men aged 15 to 44 years versus 7.5% for men aged 65–74 years (Source: ECIS - European Cancer Information System From https://ecis.jrc.ec.europa.eu, accessed on 3/JAN/2023).

The goal of treatment is to achieve complete remission (CR) of leukemia with initial therapy, followed by consolidation and/or maintenance efforts to deepen the remission and maximize response duration.

For patients fit for intensive chemotherapy, induction therapy remains centered around one or two cycles of anthracyclines and cytarabine, as recommended by ELN 2022. According to PETHEMA protocols [2], intensive chemotherapy (IC) regimens include 3 + 7 or 2 + 5 induction (idarubicin or daunorubicin and Ara-C), liposomal cytarabine and daunorubicin (CPX-351), fludarabine, idarubicin, and Ara-C (FLAG-IDA), high-dose cytarabine (HDAC), fludarabine and high-dose cytarabine (FLA), idarubicin, cytarabine, and etoposide (ICE), mitoxantrone plus cytarabine, and clofarabine plus cytarabine. After achieving CR, consolidation ideally involves regimens incorporating intermediate-dose cytarabine [3].

For patients not eligible for intensive chemotherapy, non-intensive treatment options include low-dose cytarabine (LDAC)-based regimens, hypomethylating agent (HMA) monotherapy, venetoclax (VEN)-based regimens, clinical trial or supportive care. Non-fit intensive treatment patients should be evaluated early during the first cycle, after three cycles, and then repeated every three cycles for patients in remission [3]. In contrast, patients undergoing intensive chemotherapy should be evaluated after the first cycle and subsequently every two cycles if in remission.

Treatment failure occurs due to initial resistance and failure to achieve CR but more frequently due to the recurrence of leukemia after achieving CR. These failures are associated with an unfavourable outcome. In addition, treatment toxicity contributes to the mortality of these patients, especially in the early phases of treatment. Therefore, an early estimation of the response is crucial to correctly manage the patients.

The potential of artificial intelligence (AI) methodologies to enhance the management of patients with leukemia is presented in the reviews [4–6], demonstrating their capacity to facilitate advancements in the comprehension of leukemia mechanisms and the stratification of risks for treatment adaptation. A number of potential avenues have been proposed for the development of AI models with clinical utility. These include the expansion of the accessibility of multicentric databases, the implementation of prospective model validation, and the progressive incorporation and evaluation of machine learning (ML) models in clinical trials.

The initial fundamental step in the development of models that could support medical decisions is to identify the relevant features involved in the prediction of interest. In the context of this study, the focus is on the evolution of AML patients. Karami et al. [7] sought to identify prognostic factors for survival in patients with AML using machine learning techniques. A dataset comprising 249 patients was employed. The authors compare several feature selection algorithms in order to identify the 25 most predictive features. Subsequently, a number of machine learning models are trained, evaluated and compared. The optimal combination is RelieF [8] with Gradient Boosted Tree model, which yielded an ROC-AUC of approximately 0.93, 72% sensitivity, and 91% specificity. In [9], although developed for chronic myeloid leukemia (CML), authors present a comparative study of eight machine learning models trained with an 837-patient cohort to predict five-year survival from 12 variables identified as the more predictive ones by the minimum redundancy maximum relevance feature (mRMR) selection method [10]. The variables extracted from electronic medical records comprise demographic data, medical history, clinical manifestations, and laboratory data. The optimal model exhibits a ROC-AUC of 0.85, a sensitivity of 86%, and a specificity of 85%.

A number of studies have demonstrated the utility of AI models in predicting the evolution of leukemia in patients, thereby facilitating the personalisation of their care. In [11, 12], authors showed that AI models could tailor the AML therapeutic decisions according to genomic-clinical data. Based on survival analysis (log-additive Cox hazard modeling), the models provide the evolution of a patient from diagnosis by estimating the probability of the patient being alive with or without remission, or with or without recurrence over time. Siddiqui et al. [13] compared several ML models (Logistic-regression (LR), Decision-Tree, and Random-Forest) in predicting in-hospital mortality after AML induction treatment with patient demographics, comorbidities, and information available at admission. The models achieved ROC-AUC scores ranging from 0.70 to 0.78 and could have helped to detect 51 patients who suffered treatment-related mortality. Another comparison of 9 ML models is presented in [14] where they were trained for two purposes: predict complete remission and 2-year overall survival. In an external test cohort, models achieved ROC-AUC ranging in [0.71–0.80] and [0.65–0.75], respectively.

In [15], the authors propose using ML models to anticipate the blood transfusion AML patients will need during their treatment (red-blood-cell and platelet quantities). They compare five ML models: Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, LR, and Lasso-logistic. These models exhibit ROC-AUC scores in the range [0.82–0.88] for red-blood-cell prediction. The ROC-AUC score for platelets varies from 0.80 to 0.85 in the training cohort, while it ranges from [0.63–0.70] in the test cohort. Besides, a significant increase in overall survival was observed in populations receiving more blood transfusions, suggesting that optimal blood management could lead to an overall increase in patient survival. In [16], the authors compare ten ML models for predicting mortality and relapse in children with acute lymphoblastic leukemia (ALL). The dataset comprises 161 patients with 15 variables (demographic, laboratory, clinical side effects). The model with the highest accuracy to predict mortality achieves an ROC-AUC of 0.74, while the relapse prediction reaches 0.84 ROC-AUC. More recently, our previous study [17] trained an XGBoost model to predict the risk of complications at 90 days after diagnosis in AML patients, achieving a ROC-AUC of 0.85 by leave-one-out cross-validation on the training dataset and 0.7 on the external test dataset.

Table 1 Performance summary of relevant works in the literature. When available, results from external cohort validation are reported. Three metrics are shown: ROC-AUC, sensitivity (True Positive Rate or TPR), and specificity (True Negative Rate or TNR)

Study	Prediction objective	Method	ROC-AUC	TPR	TNR
[7]	AML survival rate	ReliefF + GBT	0.93	0.72	0.91
[9]	CML 5-year survival	mRMR + SVM	0.85	0.86	0.85
[13]	AML inpatient mortality	Random Forest	0.78	0.09	0.99
[14]	AML complete remission	SVM	0.80	0.77	0.51
[14]	AML 2-year survival	SVM	0.75	0.67	0.74
[15]	Red-blood-cells in AML	LassoLR	0.88	–	–
[15]	Platelets in AML	SVM	0.70	–	–
[16]	ALL mortality in children	ANN	0.74	0.80	0.68
[16]	ALL relapse in children	Boosting	0.84	0.97	0.71
[17]	90-day AML complications	XGBoost	0.7	0.57	0.79

A summary of all aforementioned key works in the literature is presented in Table 1. In light of this background, this paper presents two main contributions. The first is a feature selection process utilising an enhanced version of the SEQENS algorithm [18]. The second is a ML model that estimates the risk of complications (resistant disease, recurrence or death) in patients diagnosed with AML, based on the variables identified as relevant in the first step.

This paper introduces a three-stage feature selection methodology. Firstly, the variables are ranked according to their relation strength (importance) with the target (remission or complication). This ranking is generated using the SEQENS algorithm [18]. This ensemble algorithm computes multiple Sequential Feature Selections using multiple inductors across multiple dataset partitions. SEQENS offers several notable advantages: (a) it supports all types of tabular data, (b) it explores potential interactions between variables and does not limit itself to treating them independently, and (c) it enhances the stability and robustness of its rankings by combining multiple inductors (each partitioning the hyperspace differently), thereby benefiting from their diverse perspectives. In the second stage, the published SEQENS is improved by randomly shuffling the target variable. It is then possible to assess whether the variables provide more information than would be expected by chance. Variables that meet this criterion are considered relevant. Finally, we search for the most concise and predictive possible subset from these relevant variables using the Sequential Backward Search algorithm [19].

Consequently, applying this methodology leads to another interesting result, providing a list of the variables that are related to the target (relevant). These relevant variables can be debated, confirming or refining implications already known or opening up new research perspectives on disease mechanisms.

Figure 1 illustrates the purpose and intended application of the proposed predictive model. This suport-decision model, which uses patients' pre-induction clinico-biological variables as input, aims to estimate the risk of complications within the subsequent three, six, or twelve months (90, 180, and 365 days, respectively) following an AML diagnosis. Complications are here defined as resistant

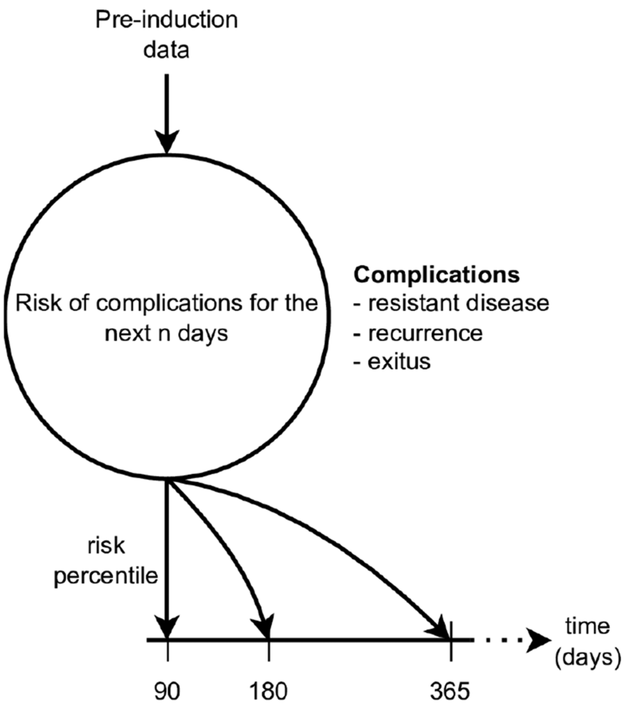

Fig. 1 Model predicting the *n*-days risk of complications for AML patients as a clinical support-decision system

disease, recurrence, or death. In response to the recommendation to utilise multicentric data, this study employs cohorts from two Spanish hospitals.

Section Materials and methods describes the dataset and how it was acquired and preprocessed. Section Results presents the relevant features obtained for the three time-window predictions and the performance exhibited by the predictive models. Section Discussion is a discussion about the main findings of this paper. Finally, the conclusions of this work are detailed in Section Conclusions.

Materials and methods

Dataset

The dataset includes newly diagnosed or relapsed/refractory AML reported to the PETHEMA AML registry (NCT02607059). Diagnosis was established according to the World Health Organization criteria [20].

The dataset is multicentric and combines two distinct AML populations: one acquired by Hospital 12 de Octubre in Madrid, consisting of 500 patients, and another by Hospital La Fe in Valencia, comprising 221 patients, for a total of 721 individuals. By integrating data from different populations, our goal is to better capture the inherent diversity of the disease, thus improving the generalisability of the lists of relevant features and the constructed models.

The first dataset was collected by Hospital 12 de Octubre (hereinafter referred to as H12O) as part of a study approved by the *Comité de Ética de la Investigación Clínica con Medicamentos* (CEIm) with registration numbers 19/434 and 20/236. This dataset initially comprised 500 patients and 56 variables. During the preprocessing phase, patients and variables that did not meet quality standards, as well as those with unclear outcomes at the selected time periods after diagnosis, were excluded. The quality criteria applied are detailed in Sections Data curation and Patient condition labelling.

The second dataset was acquired by Hospital La Fe (HUiPLaFe, Valencia, Spain), hereinafter referred to as LAFE. All adult patients (> 18 years) with newly diagnosed or relapsed/refractory AML (excluding acute promyelocytic leukemia), in accordance to the World Health Organization criteria (2016 and 2022), regardless of the treatment received, were eligible for the study. The Institutional Ethics Committee for Clinical Research of Instituto de Investigación Sanitaria La Fe (IISLaFe) approved this study with registration number 2019/0117. Written informed consent in accordance with the recommendations of the Declaration of Human Rights, the Conference of Helsinki, and institutional regulations were obtained from all patients [21, 22].

The LAFE dataset comprised 221 patients and 31 variables before the preprocessing phase. The variables in H12O that were unavailable in the LAFE dataset were automatically imputed using the same imputation methods applied to the training data. Some patients were excluded based on the quality criteria outlined in Section Patient condition labelling, resulting in a reduction of the number of usable patients and variables, depending on the specific scenario.

The variables in this study encompass various types: demographic data, clinico-biological variables, cytogenetic information, and genetic data.

Gene variant analyses were conducted using Next-Generation Sequencing (NGS) strategies with harmonized criteria previously established by the PETHEMA group [23]. IISLafe and H12O utilized a targeted NGS gene panel covering 32 genes frequently mutated in AML: ASXL1, BCOR, BRAF, CALR, CBL, CEBPA, CSF3R, DNMT3A, ETV6, EZH2, FLT3, GATA2, HRAS, IDH1, IDH2, JAK2, KIT, KRAS, MPL, NPM1, NRAS, PTPN11, RUNX1, SETBP1, SF3B1, SRSF2, STAG2, TET2, TP53, U2AF1, WT1, and ZRSR2. La Fe performed sequencing using the Ion Torrent Genexus System (ThermoFisher Scientific), while 12 Octubre used the Ion GeneStudio S5 System (ThermoFisher Scientific). Quality assessment criteria included uniformity > 85% and a mean read depth of 1000X.

Mutational profiling was performed via targeted NGS using panels implicated in myeloid pathology. The total number of reads obtained in each sample was two million, with an average depth of coverage > 1000 reads per nucleotide and high uniformity amongst all fragments (median 92%). Data analysis was performed with Ion Reporter v4.4 (Life Technologies, Carlsbad, CA, USA), identifying single nucleotide variants (SNV) and small insertions or deletions (InDels). Default parameters were applied, filtering out variants with a total coverage of at least 70 reads and a variant allelic coverage of at least 10 reads. Additionally, variants with a minor allele frequency (MAF) > 0.01 in the general population according to the single nucleotide polymorphism database (NCBI, dbSNP150 [24]) and/or the 5000-exome sequencing project [25] were also rejected as possible polymorphisms [26]. Remaining variants were annotated using COSMIC, ClinVar, and CKB Jackson Laboratory, allowing those variants present in some tumors to be retained, as well as those presenting dbSNP and previously identified as cancer mutations. This way, only pathogenic or probably damaging variants with VAF $\geq$1% were considered for the study. Variants that were absent from dbSNP or COSMIC but deemed deleterious due to associated protein-level functional changes, or due to their occurrence in conserved regions, were also included in the final analysis. Cytogenetic analyses were performed locally.

In our dataset, each of the included genes' Variant Allele Frequency (VAF) is represented as a continuous value between 0 and 1, corresponding to the percentage of mutated cells in a sample that carry an observed gene mutation. Cytogenetic variables are binary, where a value of 1 indicates that the chromosome is altered, while a value of 0 indicates an unaltered state.

It is important to note that, while the intended purpose of the model is to estimate outcomes for different time horizons after diagnosis (as seen in Fig. 1), all predictions are made at the time of diagnosis using only pre-induction data. No follow-up data is considered in this study, and the model does not incorporate continuous follow-up times, as its focus is on predefined target horizons.

Data curation

A data curation process was carried out to determine the variables suitable for inclusion in the predictive model. For this purpose, several criteria were considered:

Table 2 Descriptive statistics for demographic and clinical variables used in the final models. The mean, standard deviation, and 5th to 95th percentile ranges are computed from all patients in the combined H12O+LAFE dataset. Since *Gender* is a categorical, binary variable, a mean of 0.4 indicates that 60% of the patients are in category *0* (male) and 40% are in category *1* (female)

	Mean	Std	5%–95%
White-Blood-Cells (WBC)	34303.0	55150.2	1103.5–157744.5
Age	63.4	15.0	35.0–84.0
Bone-Marrow-Blasts (BM_Blasts)	54.6	26.1	15.0–94.0
Gender	0.4	0.5	0–1

- Bias criterion: variables exhibiting bias were discarded. For instance, a variable with a strong correlation with the target should be revised and, if this correlation is due to an abnormality during data acquisition, the feature must be excluded.
- Quality criterion: variables with a significant amount of missing values that cannot be accurately imputed, or variables that are quasi-constant (e.g., a binary variable where only a small proportion of patients exhibit values different from the mode), were removed.
- Expert criterion: the variables that are suspected or known to have an influence on the evolution of leukemia were included, provided that their quality was sufficient.

The results of the quality check were reviewed in collaboration with healthcare professionals. Their expertise was crucial in confirming which variables should be included in the model and which should be ultimately discarded. The variables that passed the quality criteria at any proposed time horizon (90 days, 6 months, or 1 year after diagnosis) are detailed in Table 2 (demographic and clinical variables) and Table 3 (genetic and cytogenetic variables). It is important to note that, because the set of observations included in the dataset varies across different target time periods, preprocessing steps were conducted independently for each target.

Quasi-constant variables

Quasi-constant variables are eliminated, specifically those where the most frequent value (the mode) is present in more than 95% of the patients. In other words, all variables in which fewer than 5% of the patients exhibit values different from the mode are discarded. This process is referred to as quasi-constant filtering.

These variables are excluded due to the insufficient number of distinct observations (lack of variability), which makes it difficult to determine whether any observed relationship with the target is meaningful or merely due to chance. In other words, to disambiguate these quasi-constant variables, increasing the number of obsevations (patients) would be necessary.

Missing data

Several variables with a high proportion of missing values were excluded from the study to mitigate the uncertainty and potential biases introduced by imputation. To guide this process, the IQA missing data assessment algorithm, presented in [27], was employed. IQA determined the optimal imputer for each variable by evaluating the cross-validation performance of a set of candidate methods, including univariate imputers (mean, mode, and zero imputation) and multivariate imputers (Iterative

Table 3 Description of all genetic (VAF) variables and cytogenetic abnormalities. The percentage of mutations (value > 0) of each variable is presented

	% mut.	**Mean**	**Std**		**% mut.**	**Mean**	**Std**
TET2	24.77	0.10	0.20	EPOR	3.04	0.02	0.08
NF1	6.78	0.05	0.16	ETV6	1.87	0.01	0.07
KMT2A	5.37	0.03	0.15	KDM6A	3.50	0.02	0.10
EZH2	5.37	0.03	0.15	PHF6	2.34	0.01	0.09
JAK2	11.92	0.03	0.11	PRPF40B	0.47	0.00	0.03
CBL	6.07	0.02	0.08	SETBP1	0.93	0.01	0.05
SRSF2	10.75	0.05	0.15	SF3A1	2.57	0.01	0.04
KIT	10.28	0.02	0.08	SMC1A	0.08	0.00	0.04
SF3B1	5.14	0.02	0.09	U2AF1	2.10	0.01	0.08
TP53	23.13	0.12	0.26	VHL	1.40	0.01	0.07
DNMT3A	23.13	0.10	0.19	CALR	0.16	0.00	0.03
ASXL1	9.81	0.04	0.13	MPL	0.62	0.00	0.04
IDH1	18.69	0.06	0.13	SH2B3	0.08	0.00	0.03
KRAS	9.81	0.02	0.07	FLT3_ITK	10.05	0.16	0.37
NRAS	19.86	0.05	0.12	FLT3_ITD	10.51	0.06	0.22
NPM1	14.02	0.05	0.14	−7/7q	9.58	0.11	0.32
IDH2	11.68	0.05	0.14	−17/17p	7.01	0.08	0.28
ZRSR2	2.34	0.02	0.14	−5/5q	12.38	0.15	0.36
RUNX1	16.12	0.08	0.20	t(8;21)	1.64	0.02	0.14
FLT3	23.36	0.06	0.14	inv(16)/t(16;16)	4.67	0.06	0.23
BCOR	0.55	0.05	0.19				

Random Forest and Bayesian Ridge Regression). Both the proportion of missing values and the imputation quality were assessed for each variable. Variables with more than 40% missing data that could not be satisfactorily imputed were removed from further analysis. For each data split, imputation models were trained exclusively on the training set and subsequently applied to impute missing values in both the training and test sets. This approach ensures consistency across splits and prevents overestimating model performance in later stages.

Correlated variables

Correlated variables generally do not improve model performance and may obscure valuable interactions between variables when using tree-based algorithms, as well as perturbing feature importance estimations [28]. In our dataset, the maximum observed absolute Pearson correlation coefficient between any pair of features is equal or below 0.6. However, since some groups of variables in our data are coincidentally missing in the same patients, the imputation process could introduce correlations between these variables. For this reason, a post-imputation filter is applied. When a pair of variables exhibits an absolute Pearson coefficient (r) above 0.8, one of them is excluded from the study, with exceptions made for variables whose relevance in relation with AML has been firmly established. These variables are included in a pre-defined allowlist, drawn from the genetic abnormalities recognised in the 2022 European LeukemiaNet guidelines [3]. Finally, a second filter ignoring this allowlist is applied to discard one variable for each pair with correlations exceeding 0.95, as such highly correlated variables are likely to convey redundant information.

Figure 2 shows the correlation matrix of variables after imputing missing values. It is important to note that, since different numbers of patients with unknown groundtruth outcomes were excluded from the study for each time target, the resulting correlation matrices vary slightly. However, for the sake of brevity, and because the observed deviations were negligible, we present only the matrix corresponding to the 90-day target horizon. As shown, no variable pairs exceeded the predefined correlation threshold, resulting in no variables being discarded due to correlation. This was true for the three time horizons. The most highly correlated groups ($r > 0.5$) were – 5*/5q* with – 17*/17p* and *TP53_VAF*, and *FLT3_VAF* with *FLT3_ITD* and *FLT3_ITK*.

Patient condition labelling

Complete remission (CR) required the absence of extramedullary disease, < 5% blast cells in the bone marrow (BM), and absence of blasts in the peripheral blood (PB). A bone marrow blast count between 5% and 25%, along with a decrease of more than 50% from baseline, was considered partial remission (PR), regardless of PB counts. Patients who did not met the response criteria mentioned above were categorized as resistance. Patients who died before assessment for response of first cycle of induction were classified as induction death. Relapse was

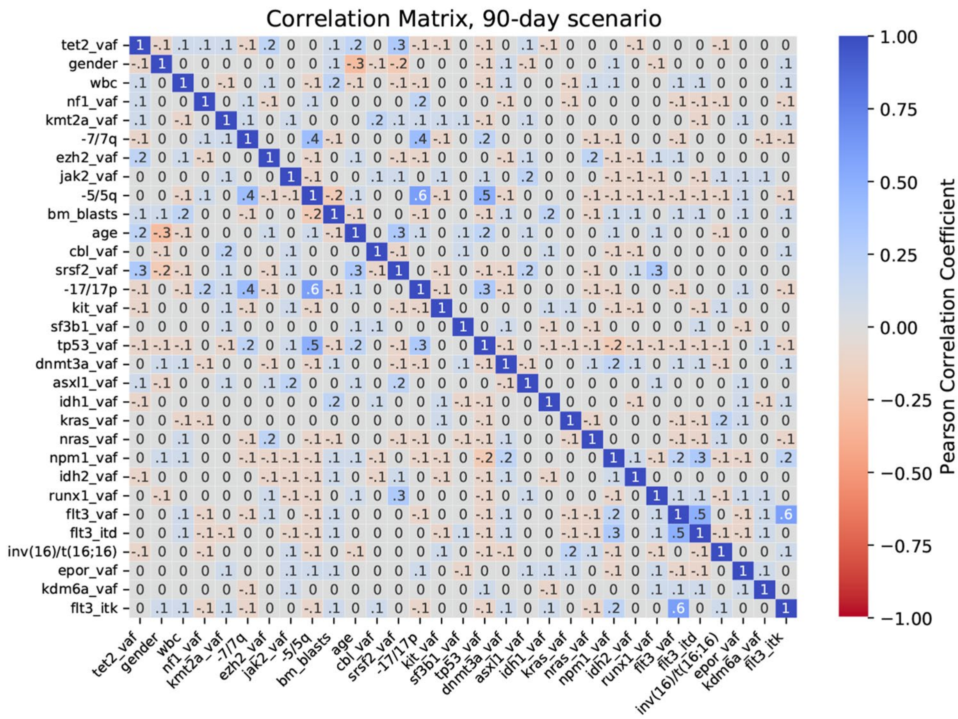

Fig. 2 Correlation matrix of the dataset in the 90-day target scenario after imputing missing values

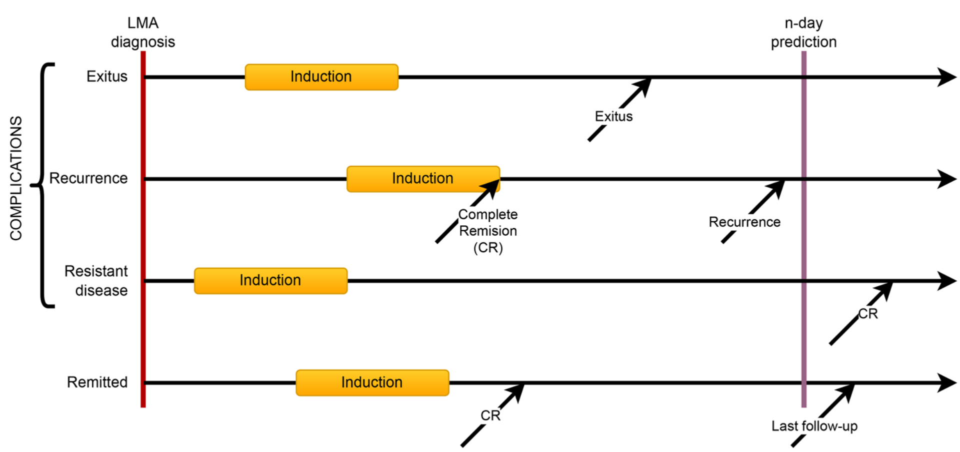

Fig. 3 n-day state computing rules. In this study, the prediction is made at 3, 6 and 12 months

defined as $\geq 5\%$ blast cells in BM or PB, or documented extramedullary AML after achieving CR [3].

The follow-up of each patient provides insight into their condition (either in remission or experiencing complications) at specific time intervals after being diagnosed with Acute Myeloid Leukemia. Consequently, the solution adopted in this paper is based on supervised learning.

The predictive model is trained using the variables associated with a patient and their condition at three distinct time periods: 90 days, 6 months, and 1 year. The patient's condition, which serves as the *target* variable, is the outcome value that the model is designed to predict. Patients with resistant disease, recurrence, or exitus are grouped into a single category labeled *with complication*. These patients are assigned a target value of 1 (positives). Patients in remission are classified under the *in remission* category and are assigned a target value of 0 (negatives). Thus, the prediction task is framed as a binary classification problem.

To determine the patient's status at a given time, the time intervals between several key dates are used: last follow-up, death, recurrence, remission, and diagnosis.

Figure 3 illustrates the criteria used to establish the n-day patient's status. Patients are classified as having complications within the first n days if they (1) die, (2)

Table 4 Number of patients in each status for the three prediction time windows. The classes balance %complications/%remitted is provided in the last row

	90 days	**6 months**	**1 year**
Remission	189	190	129
Resistant disease	32	10	3
Recurrence	61	68	104
Exitus	146	235	332
Total	428	503	568
%positive/%negative	56/44	62/38	77/23

experience a recurrence after an initial complete remission, or (3) have a resistant disease (achieving complete remission later than *n* days). For a target period of n days, patients who remain in complete remission with a last follow-up occurring after the n-day mark are considered to be in remission. However, some patients with complete remission were excluded from the analysis if their last follow-up occurred before the n-day threshold, as their remission status at the n-day mark could not be reliably confirmed.

Patients with inconsistent date sequences (e.g., with missing dates, dates in an incorrect chronological order, or with wrong formats) were excluded from the study. Table 4 presents the number of observations available for each patient condition in the dataset, corresponding to each prediction target. Furthermore, Fig. 4 illustrates the evolution of patient outcomes (detailing remission and complications) as we consider subsequent periods after diagnosis. This figure demonstrates how some patients transition between different statuses over time with varying proportions. Due to the lack of data annotations, it also shows how some patients with initially unknown statuses enter the usable dataset as their condition becomes known at later time points. In order to provide insight into the stability of the target variable as the prediction window increases, we calculated the percentage of individuals who did not change their remission-complication status. From 90 to 180 days, 392 out of 503 patients maintained their status (78%), and from 180 to 365 days, 430 patients out of 568 retained their condition (76%). In other words, at each increment of the prediction

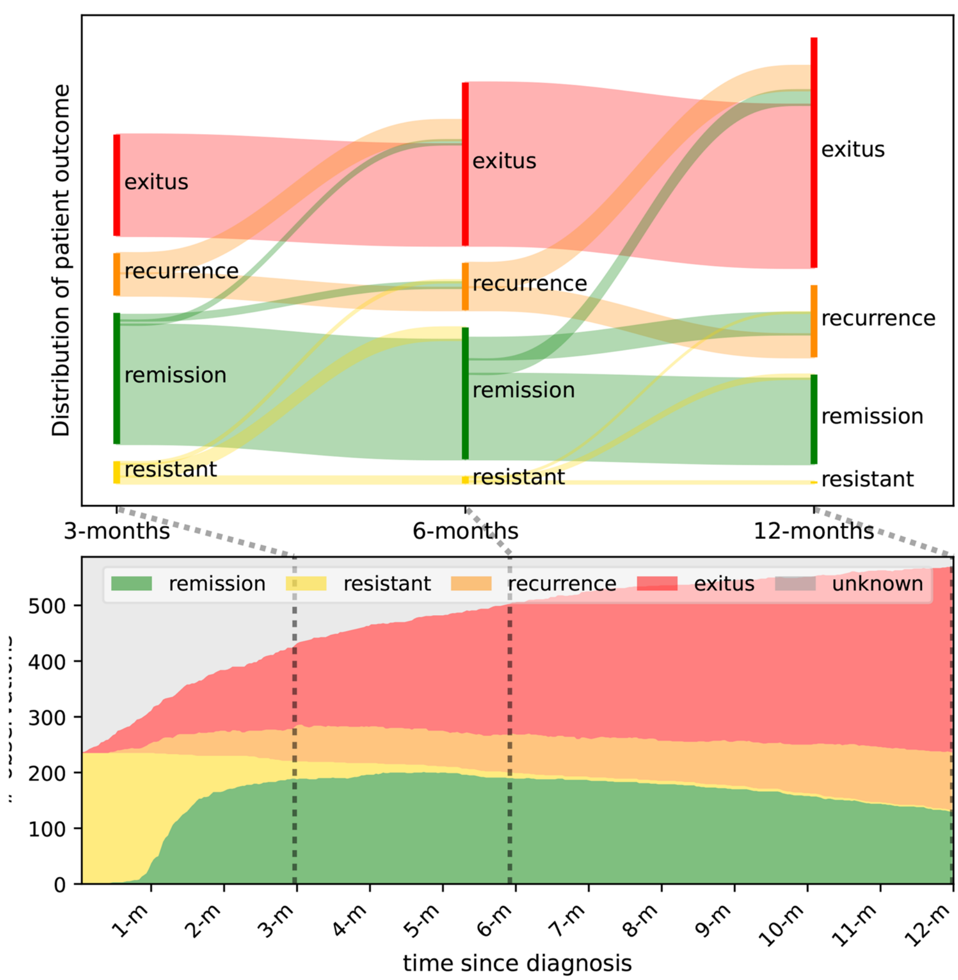

Fig. 4 Evolution and transfering of patient outcomes at different periods after diagnosis

Table 5 Features discarded during preprocessing based on various quality criteria at different target time horizons. The number of discarded features for each criterion is indicated in parentheses. The correlation filter is not included, as no variables were removed based on this criterion

Time	Missing-data filter	Quasi-constancy filter
90	(8) cebpa_vaf, bcorl1_vaf, wt1_vaf, hb, pb_blasts, ldh, stag2_vaf, platelet	(6) thpo_vaf, calr_vaf, mpl_vaf, sh2b3_vaf, epas1_vaf, rad21_vaf
180	(11) cebpa_vaf, bcorl1_vaf, wt1_vaf, hb, pb_blasts, ldh, smc1a_vaf, platelet, stag2_vaf, epas1_vaf, bcor_vaf	(3) thpo_vaf, rad21_vaf, sh2b3_vaf
365	(10) cebpa_vaf, bcorl1_vaf, wt1_vaf, pb_blasts, ldh, stag2_vaf, platelet, epas1_vaf, bcor_vaf, hp	(5) thpo_vaf, calr_vaf, rad21_vaf, smc1a_vaf, prpf40b_vaf

period, approximately 75% of the dataset remains consistent, with a trend toward an increasing number of complications.

Since we exclude the patients with unknown status at each target time horizon, the total number of available observations varies across scenarios, as it can be seen in Table 4. For this reason, the previous preprocessing steps can have different outcomes. Table 5 shows the number and sets of variables that were discarded from the analysis at each scenario.

Model building methodology

Ideally, a predictive model designed for decision support should have the following characteristics: (1) require the fewest possible input variables, (2) especially do not include features that are irrelevant to the target, and (3) achieve the highest possible performance. To meet these requirements, a three-stage methodology is proposed: (1) rank the variables based on their importance in relation to the target, (2) identify relevant and irrelevant variables (or, more precisely, those whose contribution is indistinguishable from chance), and (3) select the most predictive subset among the relevant features.

The first phase involves estimating the relevance of each variable and ranking them by descending order. To accomplish this, the SEQENS ensemble algorithm is employed [18]. Figure 5 outlines the steps and components of the method. The dataset (a) is initially divided into multiple train/test partitions (b). These partitions serve as the input to multiple feature selectors (c, blocks S_1 to S_n). SEQENS employs the sequential feature search [19], a greedy wrapper feature selection that require the use of an inductor. Five inductors are used

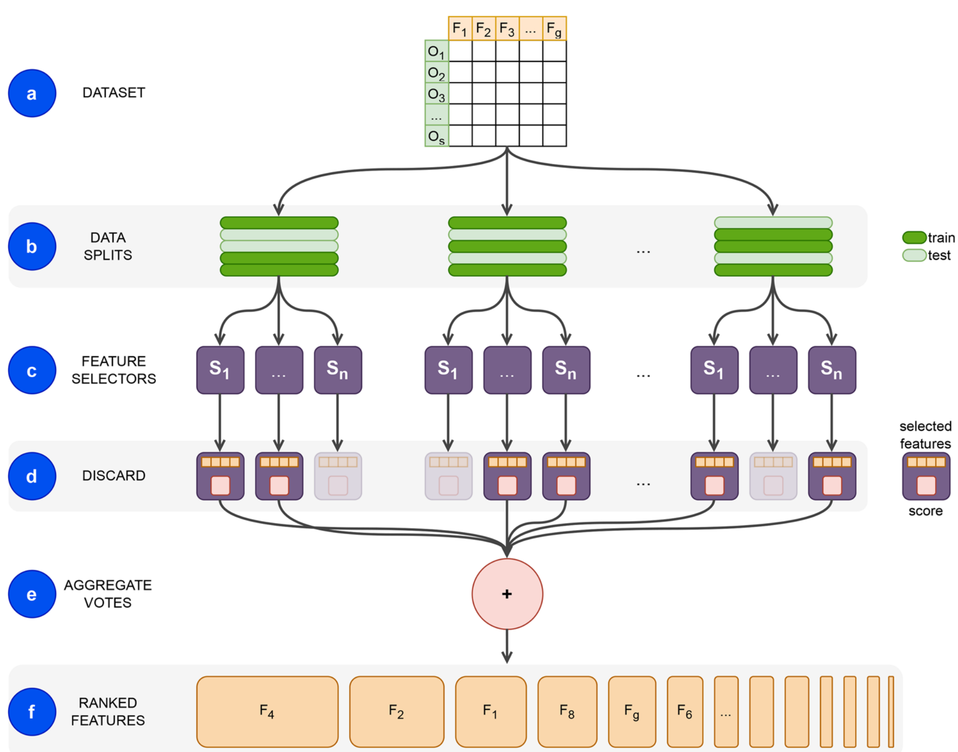

Fig. 5 Synopsis of the SEQENS components. (**a**) The dataset is comprised of observations O_i (rows) and variables or features F_j (columns). (**b**) Observations are divided into a train set and a test set. (**c**) Each partition is sent to multiple sequential feature selectors. Each selector outputs a scored subset of selected variables. (**d**) Low-scored subsets are discarded. (**e**) The result is aggregated from the selected subsets. (**f**) The output is a list of features ranked by relevance

here: XGBoost, Gradient Boosting, Random Forest, Support Vector Machine, and K-Nearest Neighbours. Consequently, for each partition, five sequential feature selections are computed, with each selection being derived from one of the five inductors available (c). Each feature selection outputs two elements: the subset of variables that best predict the target and the prediction score (d). SEQENS finally integrates the feature selections in a voting scheme (e). When a feature is selected in one of the sequential feature selections, it receives a weighted vote through a refinement that will be introduced subsequently in Eqs. 2. A hyperparameter optimization is performed for each inductor used in the SEQENS algorithm, tuning them for the task. Table 6 provides a summary of the hyperparameter search spaces explored for each model type.

The second phase aims to calculate the threshold at which a variable can be considered relevant. In other words, the goal is to determine how many votes are necessary to ensure that a variable is not selected by chance.

Table 6 Hyperparameter search spaces for inductor models in SEQENS

Model	Parameter	Search space
KNeighbors	n_neighbors	2 — 10
	weights	[uniform, distance]
	leaf_size	10 — 50
RandomForest	n_estimators	20 — 100
	min_samples_split	2 — 20
	min_samples_leaf	1 — 10
	max_depth	1 — 10
	max_features	[sqrt, log2]
XGBoost	booster	[gbtree]
	objective	[binary:logistic]
	n_estimators	50 — 150
	learning_rate	0.001 — 1
	max_depth	1 — 10
	grow_policy	[depthwise, lossguide]
	subsample	0.8 — 1
	colsample_bytree	0.5 — 1
	colsample_bylevel	0.5 — 1
	colsample_bynode	0.5 — 1
	reg_alpha	0.0001 — 1
	reg_lambda	0.0001 — 1
GradientBoosting	loss	[log_loss]
	n_estimators	20 — 100
	learning_rate	0.01 — 1
	max_depth	1 — 8
	min_samples_leaf	1 — 10
	min_samples_split	2 — 20
	subsample	0.8 — 1
	max_features	[sqrt, log2]
SupportVector	C	0.1 — 10
	kernel	[linear, rbf]
	gamma	[scale, auto]

For this purpose, the target is randomly shuffled, thereby breaking any potential relationships between the features and the target. It is important to note that only the target variable is shuffled, meaning that the relationships between predictive variables remain intact, preserving any potential interactions between them. After shuffling the target, SEQENS is reapplied using the same configuration as the original setup, ensuring that the sole source of randomness is the shuffling process. The ranking of relevant features is then recalculated. This procedure is repeated multiple times, shuffling the target recurrently to enhance the statistical robustness of the results. This method provides an estimate of the number of votes each variable can accumulate when the target is randomized (i.e., without any genuine information for solving the task). As a result, a different vote threshold is calculated for each feature. The greater the difference between the votes a variable receives with the original target versus the shuffled target, the more relevant the variable is deemed to be. It is important to emphasise that this stage represents a significant improvement to the original SEQENS algorithm as published in [18].

In order to identify the subset of relevant features, a weighted score is computed by integrating the number of votes each feature receives and the overall performance when the feature is included. Let $x_1, x_2, \ldots, x_n$ be the n covariates in the dataset, with $v_1, v_2, \ldots, v_n$ denoting their respective number of votes in the ensemble. Additionally, let $\beta_1^i, \beta_2^i, \ldots, \beta_{v_i}^i$ represent the scores of the v_i sequential feature selections where a variable x_i is selected, and let S be the total number of feature selections. The weighted score for each variable is calculated using Equation 1. In the undeveloped form of the formula, the left component represents the percentage of sequential feature selectors that vote for a variable, while the right component is equivalent to the mean score of those sequential feature selectors. Since β ranges between [0,1], the weighted score is also constrained within this range. A weighted score of zero indicates either that a variable was never selected or, if selected, its predictive power (possibly in combination with other variables) was no better than random chance. Conversely, a score of 1 indicates that the variable was consistently selected and also perfectly predicts the target (again, possibly in combination with other variables).

$$\text{weighted\_score}(x_i) = \left(\frac{v_i}{S}\right)\left(\frac{1}{v_i}\sum_{s=1}^{v_i}\beta_s^i\right) = \frac{1}{S}\cdot\sum_{s=1}^{v_i}\beta_s^i \quad (1)$$

Each score β_s^i, which has been defined as the score of each sequential feature selector where the variable x_i has been voted, is calculated based on its ROC-AUC. The ROC-AUC is adjusted to ensure that a naive model (with predictions equivalent to random chance) has no influence on the weighting process (i.e., a ROC-AUC of 0.5 is adjusted to 0, while 1 remains 1). For a given variable x_i and a specific selector s, this adjustment follows 2.

In the case of the shuffled target scenario, a weighted score is calculated for each instance of target shuffling. This allows the calculation of a confidence interval around the mean score for each feature. Features whose original-target weighted score (as defined by Equation 1) exceeds the upper bound of the 95% confidence interval of the shuffled-target weighted score are considered relevant. This approach is designed to eliminate from consideration any variables whose contribution to the task may be due to chance.

$$\beta_s^i = \frac{|\mathrm{ROCAUC}(s) - 0.5|}{1 - 0.5} \quad (2)$$

The third phase involves feature selection, where the goal is to identify the subset of relevant variables with the highest predictive power. To achieve this, the Backward Sequential Feature Search (BSFS) algorithm is employed, following a greedy approach [19]. This iterative process eliminates the variable that contributes least to predictive power at each step. At each iteration, a subset of size L is used as a seed, from which all possible subsets with one fewer variable (size $L - 1$) are generated and evaluated. For example, a subset of three variables (v1,v2,v3) generates three subsets of two variables: (v1,v2), (v2,v3), (v1,v3). The subset that performs best is then used as the seed for the next iteration. The process continues until no features remain, with the best-performing subset from all iterations being selected as the final solution.

Predictive models

This study compares four types of classifiers that estimate the probability that a patient belongs to either the "with complications" or "in remission" class. The classifiers are: shallow decision trees (DT) with a maximum depth of 5, logistic regression (LR), multi-layer perceptrons (MLP) as artificial neural networks [29], and extreme gradient boosting (XGBoost) as an ensemble tree-based classifier [30].

The use of XGBoost is motivated by its widespread adoption in clinical decision support systems across a range of diseases, including COVID-19 [31, 32], coronary diseases [33], and osteosarcoma [34]. In a comparative empirical study involving 11 well-established classifiers across 71 datasets [35], found that the Stochastic Gradient Boosting Trees (GBDT) algorithm achieved the highest mean ROC-AUC. Meanwhile [36], found that all considered implementations of GBDT perform exceptionally well across various scenarios. The model operates by training a series of decision trees sequentially, where each tree aims to reduce the prediction error by improving on the failures of the previous tree. For more detailed insights into this model, readers can refer to [37, 38]. Practically, XGBoost is particularly notable for its comprehensive documentation, faster performance, enhanced regularization compared to other GBDT implementations [30] and is generally easier to optimize for tabular data [39].

Neural networks constitute another state-of-the-art approaches. The use of MLP has been motivated by a recent study [40] comparing XGBoost, MLP and a naive classifier (returning the majority class) to estimate the overall survival of AML patients, in which the MLP exhibited the best predictive power.

DT and LR are simple, interpretable models used here to measure the improvement brought by the use of more sophisticated but harder to explain non-linear models, such as XGBoost and MLP. While DT serves as a baseline non-linear machine-learning method, LR is often treated as a statistical method, adjusting its coefficients which can be interpreted as odds ratio. When complex models are unable to improve on the performance obtained by an explainable baseline model, the latter should be preferred.

The hyperparameters of each model (with the exception of DT) are tuned using Tree Parzen Estimator (TPE) [41], a versatile Bayesian optimization method. Following each division of the dataset into training and evaluation sets, various combinations of hyperparameters are evaluated through cross-validation conducted exclusively on the training set. The combination that yields the highest ROC-AUC is selected for building the final model. The hyperparameter search spaces explored for these predictive models are detailed in Table 7.

Evaluation method and metrics

Cross-validation

The proposed models are evaluated using 5-fold cross-validation. In each of the five data splits, 80% of the individuals are used as the training set, while the remaining 20% serve as the test set. For each fold, hyperparameter tuning is performed on additional splits of the training set as outlined in Section Predictive models. The imputation methods are fitted on the training set and subsequently applied to both the training and test sets. Thereafter, two predictive models are trained on the preprocessed training data: the first utilising all available variables, and the second employing solely the corresponding selected variables from feature selection. Using each model, the probability of complications for every patient in the test set is calculated. Performance metrics for each model (across all folds) are detailed in the subsequent sections and computed by comparing these predictions to the ground truth.

Table 7 Hyperparameter search spaces for evaluated models

Model	Parameter	Search space
XGBoost	booster	[gbtree]
	objective	[binary:logistic]
	n_estimators	50 — 150
	learning_rate	0.001 — 1
	max_depth	1 — 10
	grow_policy	[depthwise, lossguide]
	subsample	0.8 — 1
	colsample_bytree	0.5 — 1
	colsample_bylevel	0.5 — 1
	colsample_bynode	0.5 — 1
	reg_alpha	0.0001 — 1
	reg_lambda	0.0001 — 1
MultilayerPerceptron	n_layers	1 — 10
	n_neurons/layer	5 — 50
	activation	[tanh, relu]
	solver	[sgd, adam]
	alpha	[0.0001, 1]
	batch_size	10 — 100
	learning_rate	[constant, invscaling, adaptive]
	learning_rate_init	0.001 — 0.1
	max_iter	100 — 300
LogisticRegression	C	0.001 — 100
	penalty	[l1, l2, elasticnet]
	solver	[lbfgs, liblinear, saga]
	l1_ratio	0 — 1
	max_iter	100 — 300

ROC curve

The *Receiver Operating Characteristic* (ROC) curve is generated for each fold by varying the threshold on the probability of complications and computing the sensitivity (True Positive Rate, TPR) and specificity (True Negative Rate, TNR) at each threshold step. The *Area Under the Curve* (AUC) is also computed for each fold, providing an overall measure of the model's performance in terms of its balance between sensitivity and specificity. To summarize the performance across folds, the ROC curves are aggregated by binning the TPR values and averaging the corresponding TNR values at each step. Similarly, the average AUC is computed. In both cases, the 95% confidence interval of the mean is calculated, offering insight into the stability of the model's performance.

Operating threshold

Throughout the ROC curve, an operating point (i.e., the probability decision threshold) can be set by the user of the predictive model. By setting this point, the user adjusts the model's behaviour to balance the trade-off between sensitivity and specificity.

In this study, a candidate operating point is set at the threshold which maximizes Youden's index [42], calculated as shown in Equation 3. This point maximises the sum of sensitivity and specificity (or equivalently minimises the sum of errors). In addition to this threshold, thresholds corresponding to the 25th, 50th, and 75th risk percentiles are also used to generate confusion matrices comparing predictions with ground truth of each fold. Five well-established performance metrics suited for imbalanced case scenarios (True Positive Rate or TPR, True Negative Rate or TNR, Positive Predictive Value or PPV, Negative Predictive Value or NPV, and F1-Score), along with Youden's index (J), are calculated at each of these decision thresholds.

$$J = \text{sensitivity} + \text{specificity} - 1 \tag{3}$$

External test

To further assess the presented models' performance, a final test was performed using test data from patients that were unseen during all previous processes. This separate dataset consists of 54 patients from LAFE diagnosed with AML, whose status was periodically monitored after they began treatment. Although these patients were treated in the same hospitals as those in the main dataset, their records were obtained in more recent batches, and were not derived from a random split of a larger original set. While some variables were provided in different ranges, they were linearly mapped through consistent transformations, such as converting percentages (in the range 0–100) to decimals (0–1), ensuring compatibility across datasets.

Three models, each for a specific time horizon, were built using all records from the H12O+LAFE dataset as a training set. The variables used for each model were those selected following the feature selection presented in Section Feature selection using the H12O+LAFE dataset. The external test results in Section External test present thepredictive power of the three models.

Imputers described in Section Missing data are trained to fill-in missing values in the training set, which are then used to impute missing values in the test set. The external test set is complete and contains no missing variables or values; therefore, imputation was not required.

For each time horizon, the model's hyperparameters were tuned, and the model was trained using the entire training dataset, interchanging the target for each specific prediction interval. These models were subsequently used to generate predictions for the 54 patients in the test set. The groundtruth labels for these patients were finally compared with the model predictions, and the same performance metrics used in prior experiments were calculated for this evaluation.

Results

Relevant features

Following the steps described in Section Model building methodology, variables were ranked by their relevance to the prediction task across all scenarios using SEQENS. In this section, we present the ranking results for the combined dataset. Figure 6 illustrates the relevant features identified in each scenario, with their scores detailed and available for side-by-side comparison in Table 8. These relevant features, identified for each dataset and time threshold, serve as the starting point for the subsequent feature selection task, th results of which are presented in Section Feature selection.

As shown in Table 9, nine variables are consistently relevant across all three prediction windows: Age, TP53, SRSF2, – 7/7q, EZH2, KIT, ASXL1, – 5/5q, and NPM1, although their importance varies over time. IDH1, JAK2, White-Blood-Cells, and U2AF1, are relevant at earlier stages but cease to be relevant at one year. Bone-Marrow-Blasts, TET2, FLT3-ITD, and EPOR, which are not relevant at three months, become relevant at six months and one year. – 17/17p, BCOR and RUNX1 are only relevant at the 3-month prediction window, while MPL is relevant only at 180 days. Gender, SF3B1, and KMT2A become significant only at the 1-year prediction window.

Feature selection

After identifying the relevant features in each scenario, a subset of these was selected using the Backward Sequential Feature Search algorithm, as explained in Section Model building methodology (third phase). Figure 7 shows the selection process and final result for each of the three prediction scenarios using the combined dataset. The candidates remaining after each iteration of the

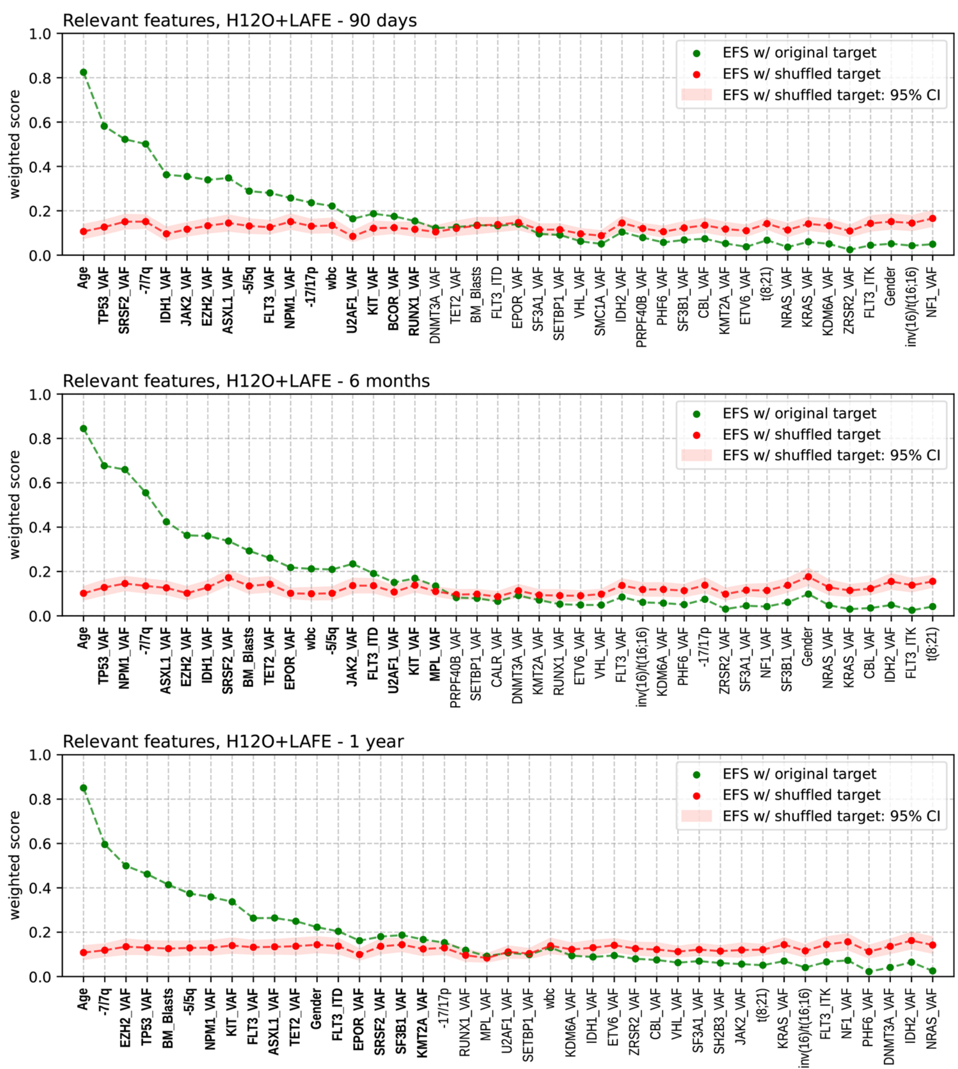

Fig. 6 Features importance ranking for the combined H12O+LAFE dataset. (top) 90-day-prediction. (middle) 6-month prediction. (down) 1-year prediction. Variables are ranked by their difference between their weighted score with the original target and their mean weighted score with shuffled targets. Relevant features are highlighted in bold font

Table 8 Relevance evolution over time of the variables identified as relevant in two or more time-periods

	90 days	6 months	1 year	evolution
Age	0.72	0.74	0.74	–
TP53_VAF	0.45	0.55	0.33	∧
−7/7q	0.35	0.42	0.48	/
NPM1_VAF	0.11	0.51	0.23	∧
EZH2_VAF	0.21	0.26	0.36	/
ASXL1_VAF	0.20	0.30	0.13	∧
SRSF2_VAF	0.37	0.17	0.04	\
−5/5q	0.16	0.11	0.25	∨
IDH1_VAF	0.27	0.23		\
Bone-Marrow-Blasts		0.16	0.29	/
JAK2_VAF	0.24	0.10		\
KIT_VAF	0.07	0.03	0.20	∨
FLT3_VAF	0.15		0.13	–
TET2_VAF		0.12	0.11	–
White-Blood-Cells	0.09	0.11		–
EPOR_VAF		0.12	0.06	\
U2AF1_VAF	0.08	0.04		\
FLT3_ITD		0.05	0.07	–

Table 9 Relevant features found by SEQENS for the H12O+LAFE dataset for each target period. Variables in bold appear as common features in at least two time targets

90 days	6 months	1 year
Age	**Age**	**Age**
TP53_VAF	**TP53_VAF**	**TP53_VAF**
SRSF2_VAF	**SRSF2_VAF**	**SRSF2_VAF**
−7/7q	**−7/7q**	**−7/7q**
EZH2_VAF	**EZH2_VAF**	**EZH2_VAF**
KIT_VAF	**KIT_VAF**	**KIT_VAF**
ASXL1_VAF	**ASXL1_VAF**	**ASXL1_VAF**
−5/5q	**−5/5q**	**−5/5q**
NPM1_VAF	**NPM1_VAF**	**NPM1_VAF**
IDH1_VAF	**IDH1_VAF**	Gender
−17/17p	**BM_Blasts**	**BM_Blasts**
JAK2_VAF	**JAK2_VAF**	SF3B1_VAF
BCOR_VAF	**TET2_VAF**	**TET2_VAF**
RUNX1_VAF	**FLT3_ITD**	**FLT3_ITD**
FLT3_VAF	MPL_VAF	**FLT3_VAF**
WBC	**WBC**	KMT2A_VAF
U2AF1_VAF	**U2AF1_VAF**	
	EPOR_VAF	**EPOR_VAF**

aforementioned feature selection process are shown, with each iteration progressively removing one feature from the initial set of relevant features.

While the detailed process for each of the nine scenarios is not presented here for the sake of brevity, the complete feature selection results across the three datasets and three target periods can be found in Table 10.

Model cross-validation performance

In the nine scenarios (comprising three target periods and three datasets), models were optimised, trained, and evaluated using both the complete set of variables and the selected subset. Table 11 shows the ROC-AUC results for the XGBoost classifiers, the multilayer perceptron classifiers, decision trees, and logistic regression. Figure 8 presents the ROC curves for the XGBoost models corresponding to the combined dataset models, along with their respective AUC values.

Furthermore, in this section, we present confusion matrices for each of the final XGBoost models, which achieved the best performance among the three classifier types, trained using the selected variables of the combined dataset. Each matrix includes five well-known metrics suited for imbalanced scenarios, along with Youden's index (J). The a-posteriori probability that maximizes J is selected as the optimal cut-point, which varies depending on the scenario: 90 days (Table 12), 6 months (Table 13) and 1 year (Table 14).

Lastly, while only the ROC curves for the combined dataset have been presented, the complete ROC-AUC results for all nine scenarios can be found in Table 15. This table summarizes the change in ROC-AUC when the training and validation datasets are limited to the selected variables.

External test

Finally, we present the performance metrics of three XGBoost models trained on the entire H12O+LAFE dataset, each one designed to predict complications at a specified time horizon and evaluated using a blind test set as explained in Section External test. Figure 9 displays the ROC curves and AUC values for each model, while Tables 16, 17, and 18 summarize various metrics when setting the operating point at different risk thresholds.

Discussion

To the best of our knowledge, no extensive work has been conducted on machine learning models specifically designed to predict treatment outcomes in acute myeloid leukemia (AML) at different time intervals post-diagnosis. This study also aims to provide insights into the genetic and clinico-biological variables that may serve as valuable predictors for estimating the risk of complications at these time intervals.

In [18], ten feature selection algorithms were compared, demonstrating that, in average, SEQENS identifies relevant variables more effectively than other state-of-the-art algorithms. Consequently, we have proposed a feature selection method based on SEQENS, incorporating enhancements to the original version such as the combination of five inductors and target shuffling (see Section Model building methodology). Nevertheless, it would also be worthwhile to apply alternative feature selection methods, especially those that have not been compared in [18].

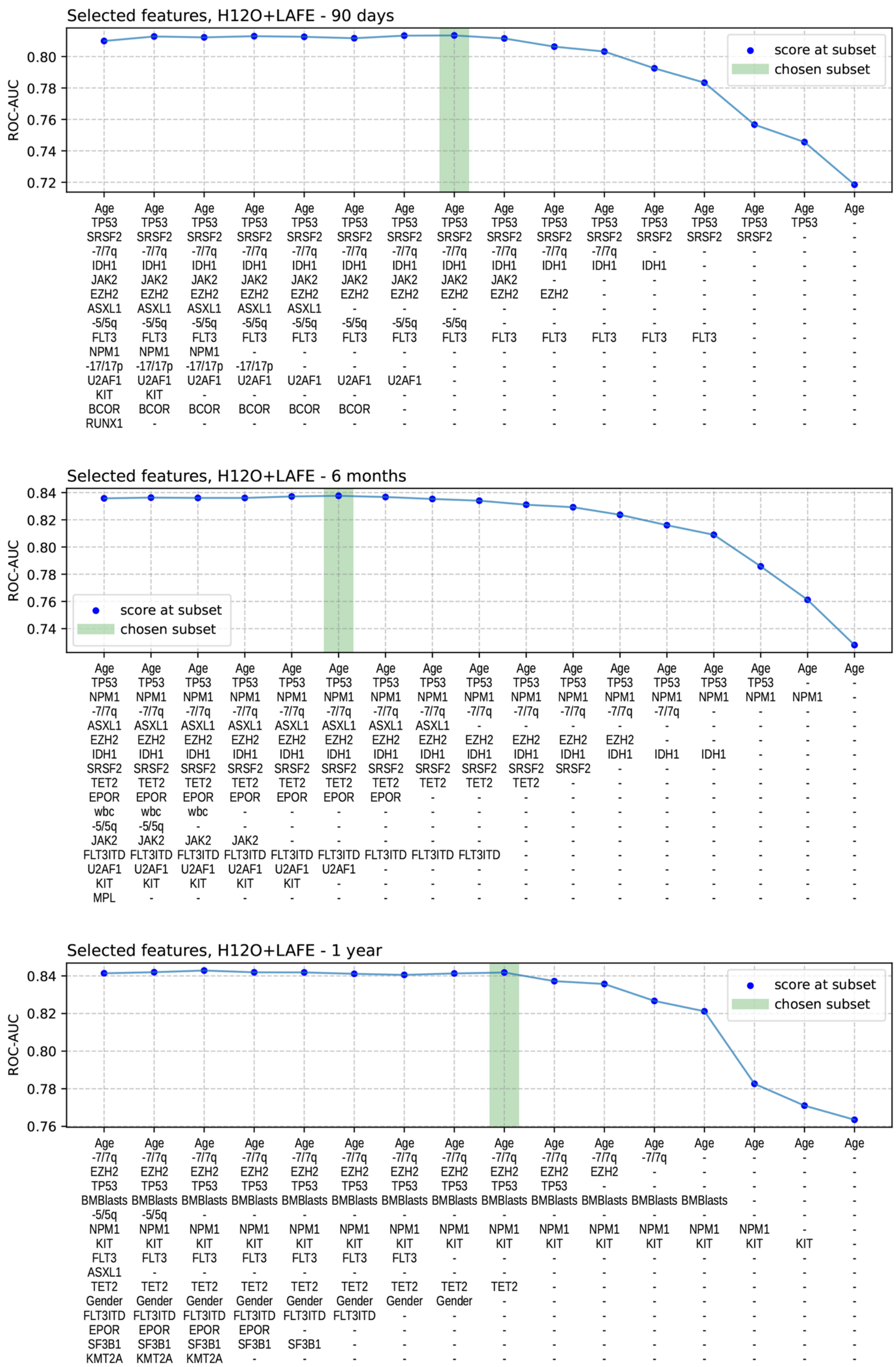

Fig. 7 Features selected by BSFS for the combined dataset. (Top) 3-month target; (Middle) 6-month target; (Down) 1-year target. Each dot represents the performance of a model fitted with the remaining feature subset in each step of the selection process. The optimal feature subset is highlighted in green

Among the relevant features identified, nine are consistently important across all three time periods: Age, TP53 (adverse), – 7/7q (adverse), EZH2 (adverse), KIT, NPM1 (favorable), ASXL1 (adverse), SRSF2 (adverse) and – 5/5q (adverse). The risk associations provided in parentheses correspond to those listed in the European LeukemiaNet 2022 (ELN2022) classification [3]. While KIT is the only gene that is not explicitly linked to a risk category in ELN2022, it is still listed as one of the additional recommended genes to be tested at diagnosis (Table 4 in [3]).

Table 10 Selected variables for each scenario, considering varying time periods and hospitals. Variables that appear in both the combined scenario and either of the individual hospital scenarios within the same time period are highlighted in bold

Time	H12O+LAFE	H12O	LAFE
90 days	**Age**	**Age**	**Age**
	TP53_VAF	**TP53_VAF**	ASXL1_VAF
	SRSF2_VAF	**SRSF2_VAF**	**SRSF2_VAF**
	−7/7q	**−7/7q**	**−7/7q**
	IDH1_VAF	**IDH1_VAF**	−17/17p
	JAK2_VAF	**JAK2_VAF**	NPM1_VAF
	−5/5q	EPOR_VAF	**−5/5q**
	EZH2_VAF		
	FLT3_VAF		
180 days	**Age**	**Age**	**Age**
	TP53_VAF	**TP53_VAF**	**TP53_VAF**
	NPM1_VAF	FLT3_ITK	**NPM1_VAF**
	−7/7q	**−7/7q**	**−7/7q**
	ASXL1_VAF	JAK2_VAF	**ASXL1_VAF**
	EZH2_VAF	**EZH2_VAF**	−5/5q
	IDH1_VAF	KIT_VAF	**IDH1_VAF**
	SRSF2_VAF	**SRSF2_VAF**	wbc
	TET2_VAF		**TET2_VAF**
	EPOR_VAF	**EPOR_VAF**	
	U2AF1_VAF	**U2AF1_VAF**	
	FLT3_ITD		
1 year	**Age**	**Age**	**Age**
	−7/7q	**−7/7q**	**−7/7q**
	EZH2_VAF	**EZH2_VAF**	−17/17p
	TP53_VAF	FLT3_ITK	**TP53_VAF**
	BM_Blasts	Gender	**BM_Blasts**
	NPM1_VAF	KMT2A_VAF	**NPM1_VAF**
	KIT_VAF	**KIT_VAF**	ASXL1_VAF
	TET2_VAF	ETV6_VAF	**TET2_VAF**
			Gender

Table 11 Mean ROC-AUC with 95% confidence interval (5-Fold cross-validation) at different times after diagnosis using three distinct classifiers: XGBoost (XGB), multilayer perceptron (MLP), decision trees (DT), and logistic regression (LR)

Time	Variables	Model	ROC-AUC (95% CI)
90-day	Selected	XGB	0.82 (0.78 — 0.85)
		MLP	0.80 (0.76 — 0.85)
		DT	0.77 (0.70 — 0.84)
		LR	0.78 (0.72 — 0.85)
	All	XGB	0.80 (0.77 — 0.84)
		MLP	0.75 (0.69 — 0.81)
		DT	0.69 (0.65 — 0.74)
		LR	0.74 (0.65 — 0.82)
6-month	Selected	XGB	0.84 (0.79 — 0.89)
		MLP	0.81 (0.73 — 0.88)
		DT	0.78 (0.72 — 0.84)
		LR	0.82 (0.75 — 0.89)
	All	XGB	0.84 (0.75 — 0.92)
		MLP	0.79 (0.75 — 0.83)
		DT	0.74 (0.69 — 0.79)
		LR	0.79 (0.69 — 0.89)
1-year	Selected	XGB	0.82 (0.76 — 0.87)
		MLP	0.82 (0.75 — 0.89)
		DT	0.81 (0.77 — 0.85)
		LR	0.83 (0.79 — 0.87)
	All	XGB	0.84 (0.82 — 0.87)
		MLP	0.75 (0.63 — 0.86)
		DT	0.72 (0.70 — 0.75)
		LR	0.77 (0.71 — 0.83)

In the 3-month and 6-month target scenarios, common relevant features include IDH1, JAK2 (additional gene recommended to test at diagnosis by ELN2022), U2AF1 (adverse), and white-blood-cell count. The successful development of new therapeutic agents, including IDH1 inhibitors, is referenced in [43]. For predicting longer-term complications (6 and 12 months), the relevant features are TET2 (an additional recommended gene to be tested at diagnosis [3]), EPOR, FLT3-ITD (intermediate), and bone marrow blasts.

When analyzing the relevant variables (all exceeding random-target performance) ranked by their mean importance (Table 8), we observe that age consistently shows the strongest importance across all three time horizons, with its importance remaining stable over time. Variables exhibiting increasing importance over time are – 7/7q, EZH2, and bone-marrow blasts. Alternatively, variables with decreasing importance include SRSF2, IDH1, JAK2, EPOR, and U2AF1. These findings could help to identify the key variables at each time point after diagnosis when estimating the risk of complications. Variables that are particularly relevant at an early stage may be more closely linked to leukemia's refractoriness to treatment. Alternatively, variables that gain importance at 6 or 12 months may be more associated with leukemia relapse. These insights are of significant value for patient management, as they could guide the timing of treatment intensification according to the patient's risk at different stages. Further research is warranted to explore these observations in greater detail.

Furthermore, we have presented the selected variables found for each scenario and collaborating hospital (Table 10). Age, TP53 (VAF), – 7/7q, and EZH2 (VAF) were consistently found to be intersectional across all three time period targets. Additionally, we observed several variables appearing exclusively in the results from one of the hospitals. The two cohorts originate from two hospitals and exhibit variation in terms of size and the number of genes. While this circumstance is suboptimal, the differences may also more accurately reflect the intrinsic heterogeneity of the disease. The results suggest that the multicentric aspect of this work is key. By combining multiple datasets, each cohort offers complementary information. This heterogeneity of data should enhance

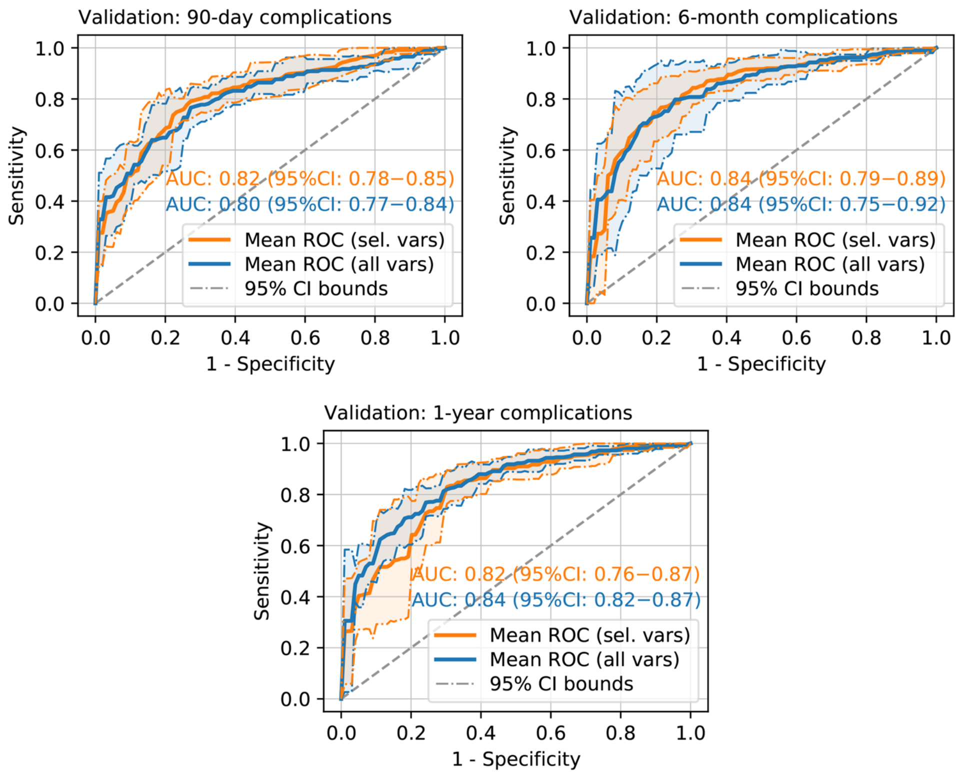

Fig. 8 ROC curves computed for the H12O+LAFE dataset using XGBoost through 5-fold cross-validation at different time horizons, using all variables (blue) and selected variables (orange). In each case, the mean ROC-AUC with 95% confidence interval is shown

Table 12 Mean validation metric values (5-Fold) at different percentile decision thresholds, 90 days after diagnosis

Thres.	TPR	TNR	PPV	NPV	F1-Score	J
p25	0.88±0.05	0.51±0.12	0.7±0.07	0.77±0.12	0.78±0.04	0.39±0.12
p50	0.73±0.05	0.79±0.02	0.81±0.02	0.69±0.08	0.77±0.02	0.52±0.05
p75	0.4±0.03	0.95±0.02	0.91±0.04	0.56±0.06	0.56±0.03	0.35±0.05
max(J)	0.75±0.07	0.81±0.05	0.83±0.04	0.71±0.08	0.78±0.03	0.55±0.04

Table 13 Mean validation metric values (5-Fold) at different percentile decision thresholds, 6 months after diagnosis

Thres.	TPR	TNR	PPV	NPV	F1-Score	J
p25	0.91±0.03	0.53±0.05	0.76±0.04	0.79±0.08	0.83±0.03	0.45±0.08
p50	0.7±0.04	0.85±0.06	0.88±0.05	0.63±0.06	0.78±0.05	0.55±0.1
p75	0.38±0.03	0.96±0.03	0.94±0.04	0.48±0.05	0.54±0.04	0.33±0.06
max(J)	0.77±0.07	0.82±0.09	0.88±0.05	0.69±0.07	0.82±0.04	0.59±0.07

Table 14 Mean validation metric values (5-Fold) at different percentile decision thresholds, 1 year after diagnosis

Thres.	TPR	TNR	PPV	NPV	F1-Score	J
p25	0.85±0.02	0.65±0.07	0.9±0.02	0.56±0.09	0.87±0.01	0.51±0.08
p50	0.6±0.02	0.83±0.06	0.92±0.03	0.38±0.06	0.72±0.02	0.42±0.09
p75	0.31±0.0	0.96±0.04	0.96±0.04	0.29±0.03	0.47±0.0	0.26±0.04
max(J)	0.81±0.06	0.77±0.08	0.92±0.03	0.55±0.1	0.86±0.03	0.58±0.06

the generalisability of predictive models and facilitate the identification of relevant variables.

In the validation results using the combined dataset for both training and evaluation within a K-Fold setting, the XGBoost classifier consistently demonstrates the highest average ROC-AUC values across different time points and variable sets. Nonetheless, the multilayer perceptron model shows competitive performance, especially when using the selected variables. Decision trees, serving as a baseline model, yield the lowest ROC-AUC values overall, although the upper bounds of their confidence intervals overlap with the mean ROC-AUC of the top-performing

Table 15 Mean 5-Fold cross-validation ROC-AUC of models trained with all variables (left number) versus models trained with the selected subsets for each scenario (right number). The direction of arrows between both numbers indicates if the score has increased ($\nearrow$), maintained (the difference is less than 0.01) ($\rightarrow$) or decreased ($\searrow$)

Time	H12O+LAFE	H12O	LAFE
90 days	$0.804 \nearrow 0.816$	$0.833 \nearrow 0.860$	$0.713 \rightarrow 0.713$
6 months	$0.836 \rightarrow 0.840$	$0.854 \rightarrow 0.857$	$0.773 \nearrow 0.807$
1 year	$0.842 \searrow 0.817$	$0.834 \nearrow 0.848$	$0.705 \nearrow 0.827$

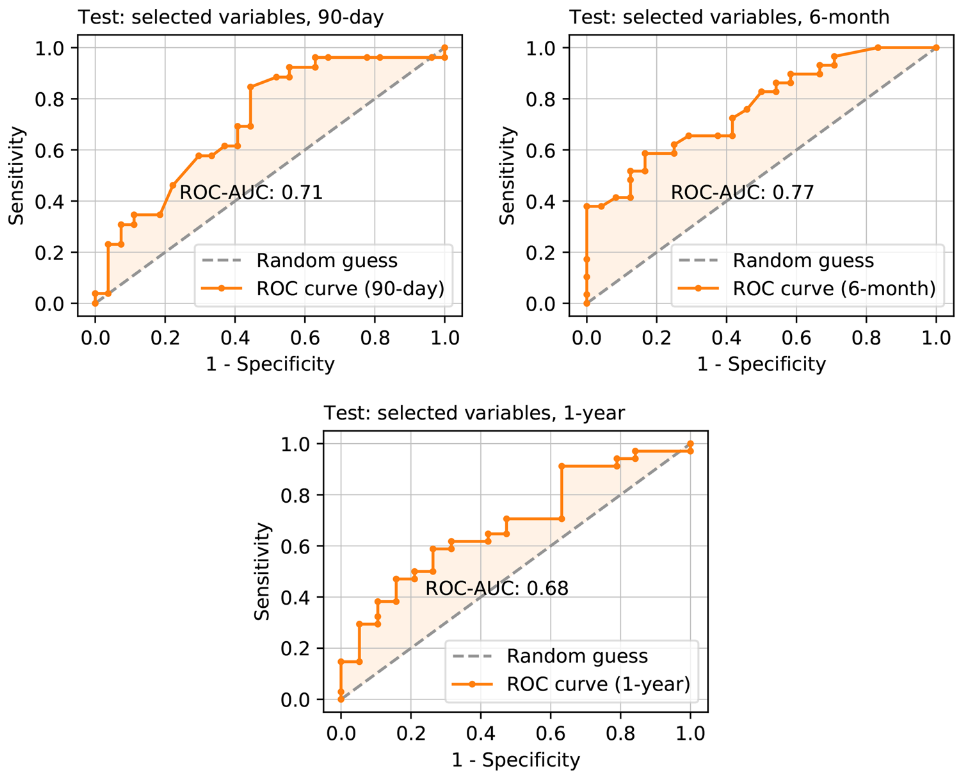

Fig. 9 ROC curves and AUC values computed for the external test set when predicting at different time horizons, using only selected variables

Table 16 External test performances of a model trained with H12O+LAFE dataset for predicting 90-day complications. 25, 50 and 75 percentiles are shown, as well as that which maximizes Youden's index (J)

Threshold	TPR	TNR	PPV	NPV	F1	J
p25	0.92	0.44	0.62	0.86	0.74	0.37
p50	0.62	0.63	0.62	0.63	0.62	0.25
p75	0.35	0.85	0.69	0.57	0.46	0.2
p34/max(J)	0.85	0.56	0.65	0.79	0.73	0.4

Table 17 External test performances of a model trained with H12O+LAFE dataset for predicting 6-month complications. 25, 50 and 75 percentiles are shown, as well as that which maximizes Youden's index (J)

Threshold	TPR	TNR	PPV	NPV	F1	J
p25	0.86	0.42	0.64	0.71	0.74	0.28
p50	0.66	0.71	0.73	0.63	0.69	0.36
p75	0.38	0.96	0.92	0.56	0.54	0.34
p66/max(J)	0.52	0.88	0.83	0.6	0.64	0.39

models. It is worth noting that logistic regression, despite being a simple classifier, achieved competitive results, comparable to both MLP and XGBoost in several scenarios, and even outperformed XGBoost when using selected variables in the 1-year scenario. The relatively strong performance of logistic regression could be indicative of the underlying linearity or independent relationships between the selected variables and the outcome,

Table 18 External test performances of a model trained with H12O+LAFE dataset for predicting 1-year complications. 25, 50 and 75 percentiles are shown, as well as that which maximizes Youden's index (J)

Threshold	TPR	TNR	PPV	NPV	F1	J
p25	0.79	0.37	0.69	0.5	0.74	0.16
p50	0.59	0.68	0.77	0.48	0.67	0.27
p75	0.32	0.89	0.85	0.42	0.47	0.22
p53/max(J)	0.59	0.74	0.8	0.5	0.68	0.33

suggesting that complex models may not always offer substantial improvements in predictive accuracy.

Given that XGBoost's performance led in most of the scenarios, demonstrating robust performance across time horizons and variable sets, our subsequent discussion focuses on its results. Nonetheless, the competitiveness of logistic regression highlights the potential of simpler, more interpretable models, particularly in scenarios where the relationship between features and outcomes is less complex or when the goal is to maintain transparency.

We observed that the mean model performance remained stable across the different time periods. Although this may appear counterintuitive, as extending the time interval typically introduces greater uncertainty, the stable performance could be explained by shifts in the distribution of the outcome variable over time. As patients advance through their treatment, they transition from a state of resistant disease to other outcomes such as remission, relapse, or death (Fig. 4). Resistant disease could be considered an intermediate condition, which may be more difficult to distinguish from complete remission (a non-complication). However, as time passes, the dataset is increasingly populated by patients who either experience complications or approach exitus. Consequently, in the longer term, the predictive task is simplified to essentially differentiating between exitus and non-exitus, wich may be less complex and could explain the stable model performance over time despite the increasing uncertainty.

In regard to the results obtained with the external test set, a decrease in model performance is observed when compared to the cross-validation results (approximately from 0.8 ROC-AUC to 0.7). This is to be expected but still represents a promising result in that it maintains a ROC-AUC of approximately 0.7 when applied to unseen observations during the training. This ROC-AUC decrease can be attributed to several factors. The external test set is relatively small, comprising 54 patients. At the time of this study's publication, further test data are not yet available for analysis. Indeed, acute myeloid leukemia is a rare disease, with approximately 3.7 cases diagnosed per 100,000 inhabitants per year in Spain. Consequently, the collection of data from new patients is a medium to long-term ongoing process. Nevertheless, the outcomes yielded by this preliminary cohort are promising in terms of predictive power, and the model is anticipated to demonstrate robust generalisation while augmenting the sample size. Secondly, the external cohort comprises solely LAFE patients. As demonstrated in Table 1, models trained and evaluated on the LAFE data consistently exhibited lower scores, indicating that predictions on this dataset may be more challenging. This may have contributed to the lower performance observed in the external test. Besides, the external cohort data was collected at a later stage and under potentially different conditions, raising the possibility of data drift (i.e., the distribution of the data may have shifted over time).

The results presented in Table 15 show primarily consistent model performance when comparing models trained using all variables with those trained on only selected subsets of variables. Interestingly, there is an overall tendency for improved performance in the latter scenario, where only selected variables are used. As observed in Table 11, this trend holds across the three classification models evaluated in this study. Specifically, performance improvements with selected variables were most pronounced for the multilayer perceptron and decision tree models. It is worth noting that neither of these two model types served as inductors within SEQENS. This outcome underscores SEQENS' capability to effectively identify the relevant features that have the highest predictive power for estimating the risk of complications. Although alternative feature selection algorithms could be tested (e.g., exploring more variable combinations or genetic approaches), BSFS has demonstrated its effectiveness in isolating the smallest subset of features that retain the most critical information.

Conclusions

In this paper, we present machine learning models for predicting the risk of complications at 3, 6, and 12 months in AML patients, utilizing demographic, clinico-biological, genetic, and cytogenetic data available at diagnosis. We propose the use of an enhanced version of SEQENS as a relevant feature identification methodology that preserves valuable potential interactions, which is essential for addressing a complex, multifactorial disease. This way, we identified a set of relevant features for each distinct time period, with common variables across all time periods including Age, TP53, −7/7q, EZH2, KIT, NPM1, ASXL1, SRSF2, and −5/5q. Among these, backward feature selection consistently selected Age, TP53, −7/7q, and EZH2 in all three time intervals. Notably, most of these variables align with the risk factors outlined by the 2022 European LeukemiaNet [3].

Using the features selected by the SEQENS-based methodology, the XGBoost predictive models achieved mean ROC-AUCs of 0.82 (95% CI: 0.78−0.85) at 90 days, 0.84 (95% CI: 0.79−0.89) at 6 months, and 0.82 (95% CI: 0.76−0.87) at 12 months post-diagnosis. When evaluated on a separate cohort of 54 unseen patients, we obtained ROC-AUCs of 0.71, 0.77, and 0.68, respectively. In most cases, the predictive power of models using the selected variables was equal to or greater than those using the full feature set, validating the effectiveness of the proposed methodology. This highlight the practical benefits of reducing the number of features, which can decrease the costs of data collection while maintaining predictive performance.

The model described in this article is available for trial at https://lmacre.iti.es, where users can input patient data via a form to receive a risk percentile relative to the training population. Those interested in testing the tool can request access by contacting aml-team@iti.es.

The findings of this study, which are currently undergoing further validation in collaborating hospitals at the time of publication, suggest that the proposed model could serve as a valuable decision support tool for managing AML patients. The results of this ongoing validation, which involve new, unseen patient data, will be crucial in determining the long-term applicability and usefulness of our model. By offering risk-based insights derived from a reduced set of highly predictive variables, the model may assist in therapeutic decision-making and guide the frequency of follow-up visits for each patient.

As future work, ongoing efforts include collecting additional data to expand the external test cohort, thereby increasing the confidence in the results. This process also allows for the gradual expansion of the training dataset, enabling the evaluation of its impact on model predictive power and generalisation. We also recommend exploring predictive models based on survival analysis. Such techniques would enhance the handling of censored patients and provide temporal cohesion in predictions across different time intervals, thereby improving the robustness and clinical applicability of the models. Additionally, further research could focus on AI explainability, addressing a key barrier to the clinical adoption of AI tools by making the models' predictions more transparent and interpretable.

Acknowledgements
We acknowledge clinicians and researchers at Instituto de Investigación Sanitaria La Fe, Universitat Politècnica de València, Hospital Universitario 12 de Octubre and Carlos III University for their valuable contribution to this work.

Author contributions
P.P., F.S., J.D., L.A, R.L., J.P., R.A. and E.B. participated in the conception and design of the work. P.P., F.S., J.D., N.A., C.S., S.D., J.G., M.L., R.A. and E.B made contributions on the acquisition, analysis and interpretation of the data. P.P., F.S., L.A. and J.D. worked on the creation of the software used in the work. P.P., F.S., R.A. and E.B. drafted and substantively revised the published work.

Funding
The publication of this article has been funded by the Generalitat Valenciana through IVACE (Valencian Institute of Business Competitiveness, https://www.ivace.es/index.php/es/) under project number IMAMCA/2025/11. This work has been supported by the Instituto de Salud Carlos III (ISCIII) through projects PI19/00730, PI19/01518 and PI22/1088, by the Instituto de Investigación Hospital 12 de Octubre (imas12), by the Generalitat Valenciana through IVACE and by the European Union through FEDER funding under project IMDEEA/2023/92.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The Institutional Ethics Committee for Clinical Research of Instituto de Investigación Sanitaria La Fe (IISLaFe, Valencia, Spain) approved this study with registration number 2019/0117. The Comité de Ética de la Investigación Clínica con Medicamentos (CEIm) of the Hospital 12 de Octubre (Madrid, Spain) approved this work with registration number 19/434. Written informed consent in accordance with the recommendations of the Declaration of Human Rights, the Conference of Helsinki, and institutional regulations were obtained from all patients.

Consent for publication
Not applicable—no individual details, images or videos were used in this work.

Competing interests
The authors declare no competing interests.

Author details
[1]ITI, Universitat Politècnica de València, Valencia, Spain
[2]Hospital Universitario 12 de Octubre, Imas12, Departament of Medicine, Complutense University, Madrid, Spain
[3]Instituto de Investigación Sanitaria La Fe, Valencia, Spain
[4]Altum Sequencing, s.l., Computer Science and Engineering Department, Carlos III University, Madrid, Spain

Received: 20 March 2024 / Accepted: 10 April 2025
Published online: 01 May 2025

References

1. Sargas C, Ayala R, Chillón MC, Larráyoz MJ, Carrillo-Cruz E, Bilbao C, Yébenes-Ramrez M, Llop M, Rapado I, Garca-Sanz R, et al. Networking for advanced molecular diagnosis in acute myeloid leukemia patients is possible: the pethema ngs-aml project. haematologica. 2021;106(12):3079.
2. PETHEMA F: Fundación PETHEMA - Programa Español de Tratamientos en Hematología. Accessed: 2025-03-05 (2025). https://www.fundacionpethema.es
3. Döhner H, Wei AH, Appelbaum FR, Craddock C, DiNardo CD, Dombret H, Ebert BL, Fenaux P, Godley LA, Hasserjian RP, Larson RA, Levine RL, Miyazaki Y, Niederwieser D, Ossenkoppele G, Röllig C, Sierra J, Stein EM, Tallman MS, Tien H-F, Wang J, Wierzbowska A, Löwenberg B. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. Blood. 2022;140(12):1345–77.
4. Radakovich N, Cortese M, Nazha A. Acute myeloid leukemia and artificial intelligence, algorithms and new scores. Best Pract Res Clin Haematol. 2020;33(3):101192.
5. Radakovich N, Nagy M, Nazha A. Machine learning in haematological malignancies. Lancet Haematol. 2020;7(7):541–50.

6. Eckardt J-N, Bornhäuser M, Wendt K, Middeke JM. Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. Blood Adv. 2020;4(23):6077–85.
7. Karami K, Akbari M, Moradi M-T, Soleymani B, Fallahi H. Survival prognostic factors in patients with acute myeloid leukemia using machine learning techniques. PLoS One. 2021;16(7):0254976.
8. Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: introduction and review. J Biomed Informat. 2018;85:189–203. https://doi.org/10.1016/j.jbi.2018.07.014.
9. Shanbehzadeh M, Afrash MR, Mirani N, Kazemi-Arpanahi H. Comparing machine learning algorithms to predict 5-year survival in patients with chronic myeloid leukemia. BMC Med Inf Decis Making. 2022;22(1):236. https://doi.org/10.1186/s12911-022-01980-w.
10. Long F, Peng H, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach. 2005;27(08):1226–38. https://doi.org/10.1109/TPAMI.2005.159.
11. Gerstung M, Papaemmanuil E, Martincorena I, Bullinger L, Gaidzik VI, Paschka P, Heuser M, Thol F, Bolli N, Ganly P, Ganser A, McDermott U, Döhner K, Schlenk RF, Döhner H, Campbell PJ. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet. 2017;49(3):332–40.
12. Tazi Y, Arango-Ossa JE, Zhou Y, Bernard E, Thomas I, Gilkes A, Freeman S, Pradat Y, Johnson SJ, Hills R, Dillon R, Levine MF, Leongamornlert D, Butler A, Ganser A, Bullinger L, Döhner K, Ottmann O, Adams R, Döhner H, Campbell PJ, Burnett AK, Dennis M, Russell NH, Devlin SM, Huntly BJP, Papaemmanuil E. Unified classification and risk-stratification in acute myeloid leukemia. Nat Commun. 2022;13(1):4622. https://doi.org/10.1038/s41467-022-32103-8.
13. Siddiqui NS, Klein A, Godara A, Buchsbaum RJ, Hughes MC. Predicting in-hospital mortality after acute myeloid leukemia therapy: through supervised machine learning algorithms. JCO Clin Cancer Inform. 2022;6(6):2200044.
14. Eckardt J-N, Röllig C, Metzeler K, Kramer M, Stasik S, Georgi J-A, Heisig P, Spiekermann K, Krug U, Braess J, Görlich D, Sauerland CM, Woermann B, Herold T, Berdel WE, Hiddemann W, Kroschinsky F, Schetelig J, Platzbecker U, Müller-Tidow C, Sauer T, Serve H, Baldus C, Schäfer-Eckart K, Kaufmann M, Krause S, Hänel M, Schliemann C, Hanoun M, Thiede C, Bornhäuser M, Wendt K, Middeke JM. Prediction of complete remission and survival in acute myeloid leukemia using supervised machine learning. Haematologica. 2023;108(3):690–704.
15. Wei Y, Xia T, Ma J, Gao X, Ge Z. Machine Learning Applications for Prediction of Blood Transfusion and Survival in Acute Myeloid Leukemia. Blood. 2022;140(Supplement 1):2831–32. https://doi.org/10.1182/blood-2022-164949.
16. Mehrbakhsh Z, Hassanzadeh R, Behnampour N, Tapak L, Zarrin Z, Khazaei S, Dinu I. Machine learning-based evaluation of prognostic factors for mortality and relapse in patients with acute lymphoblastic leukemia: a comparative simulation study. BMC Med Inf Decis Making. 2024;24(1):261. https://doi.org/10.1186/s12911-024-02645-6.
17. Sanchis JAD, Pons-Suñer P, Alvarez N, Sargas C, Dorado S, Ort JVG, Signol F, Llop M, Arnal L, Llobet R, Perez-Cortes J-C, Ayala R, Barragán E. A supervised machine learning model to predict therapy response and mortality at 90 days after acute myeloid leukemia diagnosis. medRxiv. 2023. https://doi.org/10.1101/2023.06.26.23291731.
18. Signol F, Arnal L, Navarro-Cerdán JR, Llobet R, Arlandis J, Perez-Cortes J-C. Seqens: an ensemble method for relevant gene identification in microarray data. Comput. Biol. Med. 2023;152:106413. https://doi.org/10.1016/j.compbiomed.2022.106413.
19. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognit Lett. 1994;15(11):1119–25. https://doi.org/10.1016/0167-8655(94)90127-9.
20. Khoury JD, Solary E, Abla O, Akkari Y, Alaggio R, Apperley JF, Bejar R, Berti E, Busque L, Chan JK, et al. The 5th edition of the world health organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. leukemia. 2022;36(7):1703–19.
21. Association WM, et al. World medical association declaration of helsinki: ethical principles for medical research involving human subjects. Jama 2013;310:2191–94.
22. Bibbins-Domingo K, Brubaker L, Curfman G. The 2024 revision to the declaration of helsinki: modern ethics for medical research. JAMA. 2025;333(1):30–31. https://doi.org/10.1001/jama.2024.22530.
23. Sargas C, Ayala R, Chillón MC, Larráyoz MJ, Carrillo-Cruz E, Bilbao C, Yébenes-Ramrez M, Llop M, Rapado I, Garca-Sanz R, et al. Networking for advanced molecular diagnosis in acute myeloid leukemia patients is possible: the pethema ngs-aml project. Haematologica 2020;106:3079.
24. Biotechnology Information NC. Database of Single Nucleotide Polymorphisms (Dbsnp). National Center for Biotechnology Information, National Library of Medicine â€¦ 2015.
25. Project NGES: Exome Variant Server. 20 December 2019 (2017). https://evs.gs.washington.edu/EVS/
26. Ayala R, Rapado I, Onecha E, Martínez-Cuadrón D, Carreño-Tarragona G, Bergua JM, Vives S, Algarra JL, Tormo M, Martinez P, Serrano J, Herrera P, Ramos F, Salamero O, Lavilla E, Gil C, López Lorenzo JL, Vidriales MB, Labrador J, Falantes JF, Sayas MJ, Paiva B, Barragán E, Prosper F, Sanz MÁ, Martínez-López J, Montesinos P. Hemopatias Malignas (PETHEMA) Cooperative Study Group: the mutational landscape of acute myeloid leukaemia predicts responses and outcomes in elderly patients from the pethema-flugaza phase 3 clinical trial. Cancers. 2021;13(10). https://doi.org/10.3390/cancers13102458.
27. Pons-Suñer P, Arnal L, Navarro-Cerdán JR, Signol F. ITI-IQA: a Toolbox for Heterogeneous Univariate and Multivariate Missing Data Imputation Quality Assessment. 2024. https://doi.org/10.48550/arXiv.2407.11767.
28. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern recognition letters. 2010;31(14):2225–36.
29. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature 1986;323(6088):533–36.
30. Chen T, Guestrin C: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD 16, pp. 785–94. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785
31. Karthikeyan A, Garg A, Vinod PK, Priyakumar UD. Machine learning based clinical decision support system for early COVID-19 mortality prediction. Front Public Health. 2021;9:626697.
32. Montomoli J, Romeo L, Moccia S, Bernardini M, Migliorelli L, Berardini D, Donati A, Carsetti A, Bocci MG, Wendel Garcia PD, Fumeaux T, Guerci P, Schüpbach RA, Ince C, Frontoni E, Hilty MP, Alfaro-Farias M, Vizmanos-Lamotte G, Tschoellitsch T, Meier J, Aguirre-Bermeo H, Apolo J, Martínez A, Jurkolow G, Delahaye G, Novy E, Losser M-R, Wengenmayer T, Rilinger J, Staudacher DL, David S, Welte T, Stahl K, Pavlos A, Aslanidis T, Korsos A, Babik B, Nikandish R, Rezoagli E, Giacomini M, Nova A, Fogagnolo A, Spadaro S, Ceriani R, Murrone M, Wu MA, Cogliati C, Colombo R, Catena E, Turrini F, Simonini MS, Fabbri S, Potalivo A, Facondini F, Gangitano G, Perin T, Grazia Bocci M, Antonelli M, Gommers D, Rodríguez-García R, Gámez-Zapata J, Taboada-Fraga X, Castro P, Tellez A, Lander-Azcona A, Escós-Orta J, Martín-Delgado MC, Algaba-Calderon A, Franch-Llasat D, Roche-Campo F, Lozano-Gómez H, Zalba-Etayo B, Michot MP, Klarer A, Ensner R, Schott P, Urech S, Zellweger N, Merki L, Lambert A, Laube M, Jeitziner MM, Jenni-Moser B, Wiegand J, Yuen B, Lienhardt-Nobbe B, Westphalen A, Salomon P, Drvaric I, Hillgaertner F, Sieber M, Dullenkopf A, Petersen L, Chau I, Ksouri H, Sridharan GO, Cereghetti S, Boroli F, Pugin J, Grazioli S, Rimensberger PC, Bürkle C, Marrel J, Brenni M, Fleisch I, Lavanchy J, Perez M-H, Ramelet A-S, Weber AB, Gerecke P, Christ A, Ceruti S, Glotta A, Marquardt K, Shaikh K, Hübner T, Neff T, Redecker H, Moret-Bochatay M, Bentrup F, Studhalter M, Stephan M, Brem J, Gehring N, Selz D, Naon D, Kleger G-R, Pietsch U, Filipovic M, Ristic A, Sepulcri M, Heise A, Franchitti Laurent M, Laurent J-C, Wendel Garcia PD, Schuepbach R, Heuberger D, Bühler P, Brugger S, Fodor P, Locher P, Camen G, Gaspert T, Jovic M, Haberthuer C, Lussman RF, Colak E. Machine learning using the extreme gradient boosting (xgboost) algorithm predicts 5-day delta of sofa score at icu admission in covid-19 patients. J Intensive Care Med. 2021;1(2):110–16. https://doi.org/10.1016/j.jointm.2021.09.002.
33. Budholiya K, Shrivastava SK, Sharma V. An optimized xgboost based diagnostic system for effective prediction of heart disease. J King Saud Univ Comput Inf Sci. 2022;34(7):4514–23. https://doi.org/10.1016/j.jksuci.2020.10.013.
34. Jiang J, Pan H, Li M, Qian B, Lin X, Fan S. Predictive model for the 5-year survival status of osteosarcoma patients based on the seer database and xgboost algorithm. Sci Rep. 2021;11(1):5542. https://doi.org/10.1038/s41598-021-85223-4.
35. Zhang C, Liu C, Zhang X, Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms. Expert Syst Appl. 2017;82:128–50. https://doi.org/10.1016/j.eswa.2017.04.003.
36. Florek P, Zagdański A. Benchmarking state-of-the-art gradient boosting algorithms for classification. arXiv preprint arXiv:2305.17094. 2023. https://doi.org/10.1109/tnet.2022.3155708.
37. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. 2 nd edn. New York, NY: Springer; 2009.

38. Chen T, Guestrin C: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD 16, pp. 785–94. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785
39. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion. 2022;81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011.
40. Didi I, Alliot J-M, Dumas P-Y, Vergez F, Tavitian S, Largeaud L, Bidet A, Rieu J-B, Luquet I, Lechevalier N, Delabesse E, Sarry A, De Grande A-C, Bérard E, Pigneux A, Récher C, Simoncini D, Bertoli S. Artificial intelligence-based prediction models for acute myeloid leukemia using real-life data: a dataml registry study. Leukemia Res. 2024;136:107437. https://doi.org/10.1016/j.leukres.2024.107437.
41. Watanabe S. Tree-structured parzen estimator: understanding its algorithm components and their roles for better empirical performance. 2023. https://doi.org/arXivpreprintarXiv:2304.111277. https://doi.org/10.48550/arXiv.2304.11127.
42. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.
43. Zarei M, Hue JJ, Hajihassani O, Graor HJ, Katayama ES, Loftus AW, Bajor D, Rothermel LD, Vaziri-Gohar A, Winter JM. Clinical development of IDH1 inhibitors for cancer therapy. Cancer Treat Rev. 2022;103(102334):102334.

Publisher's Note