Computer Methods and Programs in Biomedicine 261 (2025) 108605

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

VAE-Surv: A novel approach for genetic-based clustering and prognosis prediction in myelodysplastic syndromes

Cesare Rollo [a], Corrado Pancotti [a,1], Flavio Sartori [a], Isabella Caranzano [a], Saverio D'Amico [b,c], Luciana Carota [d], Francesco Casadei [e], Giovanni Birolo [a], Luca Lanino [b], Elisabetta Sauta [b], Gianluca Asti [b], Alessandro Buizza [b], Mattia Delleani [b,c], Elena Zazzetti [b], Marilena Bicchieri [b], Giulia Maggioni [b], Pierre Fenaux [f], Uwe Platzbecker [g], Maria Diez-Campelo [h], Torsten Haferlach [i], Gastone Castellani [d,j], Matteo Giovanni Della Porta [b,k], Piero Fariselli [a,*], Tiziana Sanavia [a]

[a] Computational Biomedicine Unit, Department of Medical Sciences, University of Torino, Via Santena 19, 10126, Torino, Italy
[b] IRCCS Humanitas Research Hospital, via Manzoni 56, 20089 Rozzano – Milan, Italy
[c] Train s.r.l., via Alessandro Manzoni 56, 20089 Rozzano – Milan, Italy
[d] Department of Medical and Surgical Sciences (DIMEC), University of Bologna, 40126 Bologna, Italy
[e] IRCCS Istituto delle Scienze Neurologiche di Bologna, 40138 Bologna, Italy
[f] Hematology and Bone Marrow Transplantation, Hôpital Saint-Louis/University Paris 7, Paris, France
[g] Medical Clinic and Policlinic 1, Hematology and Cellular Therapy, University Hospital Leipzig, Germany
[h] Hematology Department, Hospital Universitario de Salamanca, Salamanca, Spain
[i] MLL Munich Leukemia Laboratory, Max-Lebsche-Platz 31, 81377 Munich, Germany
[j] IRCCS Azienda Ospedaliero-Universitaria di Bologna S.Orsola, 40138 Bologna, Italy
[k] Department of Biomedical Sciences, Humanitas University, via Montalcini 4, 20072 Pieve Emanuele – Milan, Italy

ARTICLE INFO

Keywords:
Survival analysis
Deep Learning
Variational Autoencoder
Myelodysplastic syndrome
Genetic-based clustering

ABSTRACT

Background and Objectives Several computational pipelines for biomedical data have been proposed to stratify patients and to predict their prognosis through survival analysis. However, these analyses are usually performed independently, without integrating the information derived from each of them. Clustering of survival data is an underexplored problem, and current approaches are limited for biomedical applications, whose data are usually heterogeneous and multimodal, with poor scalability for high-dimensionality.

Methods We introduce VAE-Surv, a multimodal computational framework for patients' stratification and prognosis prediction. VAE-Surv integrates a Variational Autoencoder (VAE), which reduces the high-dimensional space characterizing the molecular data, with a deep survival model, which combines the embedded information with the clinical features. The VAE embedding step prioritizes local coherence within the feature space to detect potential nonlinear relationships among the molecular markers. The latent representation is then exploited to perform K-means clustering. To test the clinical robustness of the algorithm, VAE-Surv was applied to the Genomed4all cohort of Myelodysplastic Syndromes (MDS), comparing the identified subtypes with the World Health Organization (WHO) classification. The survival outcome was compared with the state-of-the-art Cox model and its penalized versions. Finally, to assess the generalizability of the results, the method was also validated on an external MDS cohort.

Results Tested on 2,043 patients in the GenomMed4All cohort, VAE-Surv achieved a median C-index of 0.78, outperforming classical approaches. In addition, the latent space enhanced the clustering performance compared to a traditional approach that applies the clustering directly to the input data. Compared to the WHO 2016 MDS subtypes, the analysis of the identified clusters showed that the proposed framework can capture existing clinical categorizations while also suggesting novel, data-driven patient groups. Even tested in an external MDS cohort of 2,384 patients, VAE-Surv achieved a good prediction performance (median C-index=0.74), preserving the interpretability of the main clinical and genetic features.

Conclusions VAE-Surv enables automatic identification of patients' clusters, while outperforming the traditional CoxPH model in survival prediction tasks at the same time. Applied to MDS use case, the obtained genetic-based clusters exhibit a clear survival stratification, and the application of the clinical information allowed high performance in prognosis prediction.

https://doi.org/10.1016/j.cmpb.2025.108605
Received 8 July 2024; Received in revised form 13 December 2024; Accepted 12 January 2025
Available online 20 January 2025

1. Introduction

Survival analysis is one of the most used approaches in clinical prognostic investigations, aiming to predict the time to an event, such as death or the progression of a disease. A key aspect of this analysis is the presence of censored data, indicating that the event of interest did not occur during the study, therefore requiring the use of specialized statistical methods. Traditionally, the Cox proportional hazards model (CoxPH) [1] has been the most widely used technique to analyze censored data, but it was designed for small data sets and does not scale well to high dimensions. In addition, most of the state-of-the-art methods assume only linear relationships among covariates. Recently, several machine learning (ML) algorithms have been adapted to work with censored data and can give more accurate results than traditional statistical methods since they are able to consider nonlinear relationships among the features [2].

Another main goal in clinical studies when censored data are analyzed is disease subtyping through risk stratification of patients. To address this task, cluster analysis of survival data can identify similar groups of patients in time-to-event distributions, in order to enhance a precision medicine approach for clinical decision-making. So far, only a few approaches have been suggested to cluster patients according to their survival function, yet this topic is beginning to gain increased attention. Liverani et al. proposed a Dirichlet process mixture model with cluster-specific parameters for the Weibull distribution, addressing cases where it is difficult to apply usual survival models due to multicollinearity [3]. Chapfuwa et al. proposed a Bayesian non-parametric time-to-event approach with structured latent representations that can be clustered through a prior for infinite mixture of distributions [4]. However, it is difficult to transfer these approaches to medical studies when multiple types of data are available (omics, clinical, imaging, etc.), which are often high-dimensional, heterogeneous and with missing information, representing challenges to current statistical methods.

Recent research has shown that the integration of multiple sources of data (e.g. clinical, genomic data) leads to better prediction of the prognostic risk than the use of a single source [5]. In this context, some studies proved the effectiveness of Autoencoder architectures in integrating clinical and multiomics data, exploiting the learned latent representations to predict the outcomes of interest [6–8]. Kim et al. [9] applied a Variational Autoencoder (VAE) model to pan-cancer RNA-seq gene expression data from TCGA, combined through a transfer learning setting with a neural network for survival analysis. However, the authors focused the approach on the risk prediction, without exploiting the latent features obtained by the autoencoder and without integrating the clinical variables. In the work by Hira et al. [10] instead, multiple VAEs were used to integrate RNA-seq, CNVs and DNA-methylaytion data in the context of ovarian cancer, identifying disease subtypes through a classification schema applied to the latent features. This approach optimizes a combined loss that accounts for both the VAE reconstruction and the classification task. The latent features were also used to provide a stratification of the patients. However, the authors focused their approach on the classification task and performed linear CoxPH survival predictions to each classified subgroup only *a posteriori*.

In this study, we propose VAE-Surv, a computational framework based on a VAE with the aim to both stratify the patients and assess their prognostic risk by optimizing both data reconstruction and the survival prediction task. Specifically, the framework integrates genetic, cytogenetic and clinical features to uncover novel insights into genetic-based prognostic risk predictions. To show the robustness of our approach, we focused our application on the prognosis prediction of Myelodysplastic Syndromes (MDS), which represent a heterogeneous group of clonal hematopoietic stem cell disorders characterized by ineffective hematopoiesis. These syndromes predominantly affect elderly populations and manifest as bone marrow failure, leading to varying degrees of cytopenia (low blood cell counts) in one or more myeloid lineages. Additionally, MDS has a substantial risk of progression to Acute Myeloid Leukemia (AML). From a genetic standpoint, MDS showcases an intricate mutational landscape with numerous genetic abnormalities, including point mutations, chromosomal deletions, and translocations [11–13]. Although these mutations contribute to the onset and progression of the disease, their predictive utility for treatment response remains a topic of ongoing research. The current WHO 2016 classification of MDS patients [14], which is based on morphological, clinical, and genetic features, cannot fully encapsulate the complex genetic and cytogenetic landscape of this disorder, limiting the granularity of the patients' stratification [15,16] and the therapeutic options, which require personalized approaches [17].

Here, we will apply VAE-surv to two of the largest cohorts of MDS patients currently available (2,043 patients from Genomed4all MDS cohort [16] and 2,384 patients from the International Working Group for the study of Prognosis in MDS cohort [18]) to demonstrate that this computational model both provides clinically valid stratification of different MDS subtypes and achieves high accuracy in predicting patients' survival.

2. Materials and methods

2.1. Datasets

The proposed framework was built considering an international retrospective cohort of 2,043 MDS patients, available through the GenoMed4All consortium [16,19,20]. The study included patients diagnosed with MDS based on the 2016 WHO classification criteria. Laboratory and clinical data were collected at diagnosis or within six months of diagnosis for consistency. DNA sequencing was performed in bone marrow mononuclear cells or peripheral blood granulocytes, ensuring robust genomic profiling. Patients with therapy-related myeloid neoplasms, paroxysmal nocturnal hemoglobinuria, aplastic anemia, or MDS/myeloproliferative neoplasm with ring sideroblasts and thrombocytosis were excluded, as described in Supplementary File 2 of Bersanelli et al. study [15]. The dataset collects clinical, demographic and molecular features (a selected panel of 58 genetic and cytogenetic mutations deemed relevant for MDS). Demographic and clinical covariates include age of diagnosis (AOD), gender, neutrophils, hemoglobin, platelets, and bone marrow blasts (BMB). At diagnosis, cytogenetic analysis was performed using standard G-banding, and karyotypes were classified using the International System for Cytogenetic Nomenclature Criteria. Mutational screening of 46 genes related to myeloid neoplasms was performed on DNA from peripheral blood granulocytes or bone marrow mononuclear cells. Overall survival was considered as the outcome of interest for our analyses. Both survival time and censoring status were available for each patient (71% censoring rate). Further details are reported in Bersanelli et al. [15].

In order to assess the generalizability of the results, the model was also validated on an external cohort of 2,384 MDS patients, provided by the International Working Group for the study of Prognosis in MDS (IWG-PM) [18]. The inclusion criteria for this cohort were: age of diagnosis $\geq$ 18 years, a diagnosis of MDS according to WHO 2016 criteria and information available on demographics, clinical features, mutational screening/chromosomal abnormalities, treatment and overall survival. A comparison between IWG-PM and Genomed4all cohorts is provided in Supplementary Material, Section 1.

* Corresponding author.
E-mail address: piero.fariselli@unito.it (P. Fariselli).
[1] Co-first author.

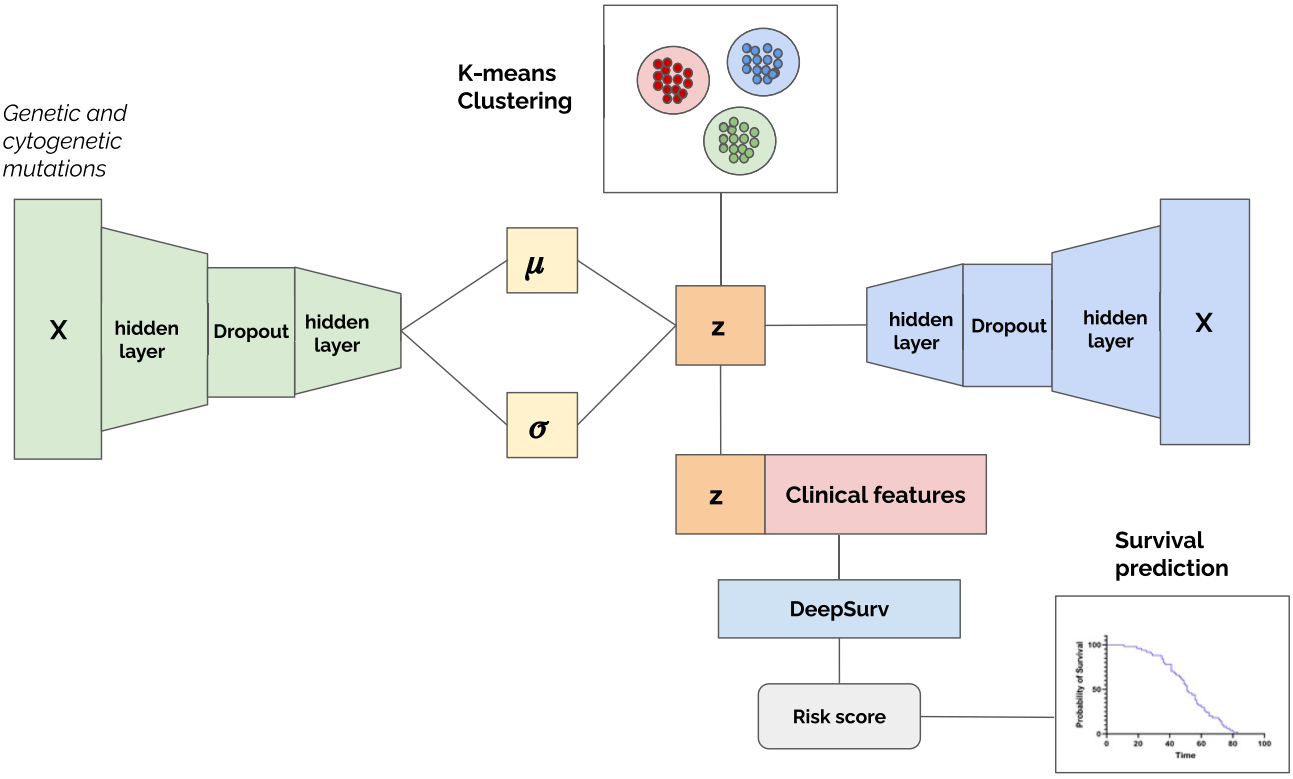

Fig. 1. Schematic representation of the VAE-Surv computational framework. Genetic and cytogenetic markers are treated as binary input vectors of a Variational Autoencoder (VAE) for dimensionality reduction. The latent feature space, obtained by the VAE, is concatenated with clinical variables to form an enriched feature set. This combined feature set is then exploited by a deep learning generalization of the Cox proportional hazards (DeepSurv) model for survival risk prediction. The dimensionality of the latent space z is optimized as a hyperparameter of the framework.

2.2. Model architecture

A schematic representation of the framework is presented in Fig. 1. The core of the computational approach is a Variational Autoencoder (VAE) [21], designed to intake genetic and cytogenetic markers as binary vectors (presence/absence of mutations and abnormalities). The goal of the VAE is to reduce the sparse and high-dimensional genetic-based input space into a latent feature space at lower dimensionality. This latent representation is then concatenated with the patients' clinical variables to generate an enriched feature set. A deep learning generalization of the Cox proportional hazards (DeepSurv [22]) model is applied to this enriched feature set for a survival analysis task.

In conventional VAE architectures, a loss function often includes both a reconstruction term $\mathcal{L}_{recon}$ and a Kullback–Leibler divergence term:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\theta,\phi;x) &= \mathcal{L}_{\text{recon}} + \text{KL-term}\\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - D_{KL}\left(q_\phi(z \mid x) \parallel p(z)\right),\end{aligned}$$

where $p_\theta(x \mid z)$ is the decoder parametrized by the weights of the neural network θ, $q_\phi(z \mid x)$ approximates the posterior distribution of the latent variables (encoder parametrized by ϕ) and $p(z)$ is the prior distribution over the latent variables assumed to be Gaussian ($z \sim p(z) = \mathcal{N}(0, I)$). Typically, the KL-term regularizes the latent space by encouraging $q_\phi(z \mid x)$ to adhere closely to the Gaussian prior. This allows the model to learn a smooth and continuous latent space, which is particularly beneficial for generative tasks. However, in our approach, we experiment a removal of the KL-term from the loss function. This decision is motivated by the fact that we are not primarily interested in the generative capabilities of the model or in ensuring that the latent space adheres to a predefined Gaussian prior. Instead, our focus is on learning compact representations of the genetic input, with a particular emphasis on enabling the latent space to exhibit clustered structures. Moreover, by removing the KL divergence, we prevent the phenomenon of posterior collapse [23], where the learned posterior distribution $q(z \mid x)$ becomes overly similar to the prior distribution $p(z)$, effectively ignoring the input x. Therefore, the risk is that the latent variables z become uninformative, and the decoder almost entirely relies on its learned capacity to reconstruct x without using the latent representation.

The VAE was initially pretrained separately from DeepSurv model, with the aim of minimizing the reconstruction error through a $logcosh$ loss function over a predefined number of epochs. Once a stable latent feature representation was obtained, the VAE's parameters were frozen. After the VAE pretraining, DeepSurv model was fine-tuned considering as input feature set the frozen latent representation from the VAE concatenated with the clinical variables, in order to predict the survival risk. Hyperparameters were optimized through a 10-fold cross-validation scheme stratified by censoring indicator, using the Concordance Index (C-Index), a widely used ranking metric that quantifies the model's ability to correctly order survival times [24]. C-Index ranges from zero to one, with a value of 0.5 corresponding to the performance of a random ranking and 1 to perfect discrimination. We compared the performance of our VAE-Surv framework against baseline linear Cox Proportional Hazards (CoxPH) regression models, which do not exploit latent feature extraction, and against the Survival Cluster Analysis (SCA) nonlinear method [4], using the median C-Index obtained across a 10-fold cross-validation as the principal performance metric. Specifically, CoxPH models were implemented with various regularization strategies, including no regularization, Lasso (L1), Ridge (L2), and ElasticNet penalties [25]. Regularization is a method for preventing overfitting by controlling the model complexity. It accomplishes this by penalizing the coefficients of predictors that do not provide meaningful information to the model. Lasso penalization allows for the generation of a sparse model (setting unimportant covariates to

0), while Ridge penalization avoids the feature selection set by Lasso by encouraging a grouping effect of the covariates used in the model. Finally, Cox regression combined with ElasticNet penalization combines both Lasso and Ridge penalty terms to identify the more representative covariates in each group that contribute most to modeling the outcome. On the other hand, SCA is a Bayesian non-parametric approach that represents the patients in a clustered latent space and encourages latent representations to behave as a mixture of distributions, following a Dirichlet Process structure via a distribution matching approach, to provide clusters (subpopulations of patients) with distinct risk profiles.

2.3. Post-training analysis

Once the best set of hyperparameters had been selected (see Section 2 of the Supplementary Materials for further details), the model was retrained on the whole dataset, and a K-means clustering on the latent representations generated by the VAE was performed using the Euclidean distance metric. Importantly, these latent variables were devoid of the clinical features and the survival outcome, preserving their nature as a specific representation of genetic and cytogenetic markers. To determine the optimal number of clusters k, we selected the one maximizing the quantity:

$$Q_k = Silh_{mean}(k) * \frac{N_{min}(k)}{N_{max}(k)}$$

where $Silh_{mean}$ is the mean Silhouette score [26], N_{max} and N_{min} are the sizes of the largest and the smallest clusters, respectively. The rationale behind this optimization was to maximize the cluster quality, measured by the Silhouette score, while simultaneously avoiding the presence of excessively small or large clusters.

Once the optimal clustering was determined, survival curves were then generated to evaluate the clinical validity of these clusters, and gene mutation frequencies were examined to understand the underlying genetic and cytogenetic landscape.

3. Results

VAE-Surv's performance assessment focused on both the accuracy of survival predictions and the quality of the genetic-based clusters.

3.1. VAE-surv prediction performance

The results of the survival predictions on the cross-validation sets are summarized in Table 1. The performance of the VAE-Surv model was systematically compared against widely adopted state-of-the-art survival analysis methods. Specifically, we benchmarked it against Cox Proportional Hazards regression models, implemented with various regularization strategies, including no regularization, Lasso (L1), Ridge (L2), and ElasticNet penalties [25]. Finally, the Survival Cluster Analysis (SCA) [4] framework was also employed as a baseline for comparative evaluation. It is important to highlight that none of these methods provides the capability to treat genetic and cytogenetic data independently of clinical and demographic information.

Table 1
Comparison of models' performance in terms of median C-Index and Confidence Intervals, evaluated in the 10-folds CV splits.

	Median C-Index	95% CI
CoxPH	0.754	(0.726, 0.784)
CoxPH - L2	0.754	(0.729, 0.785)
CoxPH - L1	0.775	(0.745, 0.796)
Elastic-Net	0.775	(0.745, 0.796)
SCA	0.744	(0.725, 0.758)
VAE-Surv w/ KL	0.751	(0.719, 0.767)
VAE-Surv w/o KL	**0.780**	**(0.747, 0.790)**

Among all comparisons, VAE-Surv consistently achieved either superior (p = 0.01 vs. SCA, Wilcoxon Rank Sum Test) or comparable ($p \geq 0.16$ vs. all Cox models, Wilcoxon Rank Sum Test) median C-Index scores, highlighting its robust predictive accuracy for survival outcomes. Notably, excluding the KL-term from the VAE loss function improves the performance. The effect of the KL-term on the learned latent space is clearly visible from Figure S4 in the Supplementary Materials. Furthermore, in order to assess the reliability of the survival predictions obtained from VAE-Surv, we divided the predicted risk scores into 3 risk groups based on percentiles (Low, Medium and High risk) and plotted the corresponding Kaplan–Meier curves (see Fig. 2). A pairwise log-rank test revealed that the survival distributions of the three groups are statistically different, with all the p-values < 0.005.

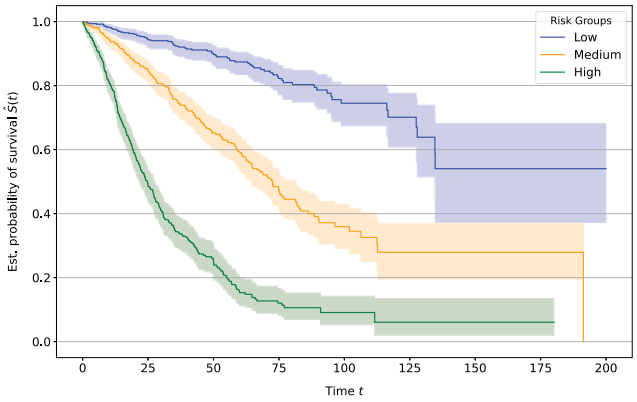

Fig. 2. Kaplan–Meier curves stratified by risk groups based on percentiles.

Additionally, we assessed whether the dimensionality reduction of genetic features applied by the VAE module provides a significant advantage in the prediction performance over other linear encoding representations of the input data. Therefore, we applied a Principal Component Analysis (PCA) to the raw input genetic data. We then concatenated the first two principal components (explained variance: 21% and 15%) with the clinical covariates. This concatenated vector was subsequently used to train the CoxPH models. VAE-Surv consistently exhibited a higher C-Index with respect to the combination of a PCA followed by a CoxPH model (Fig. 3).

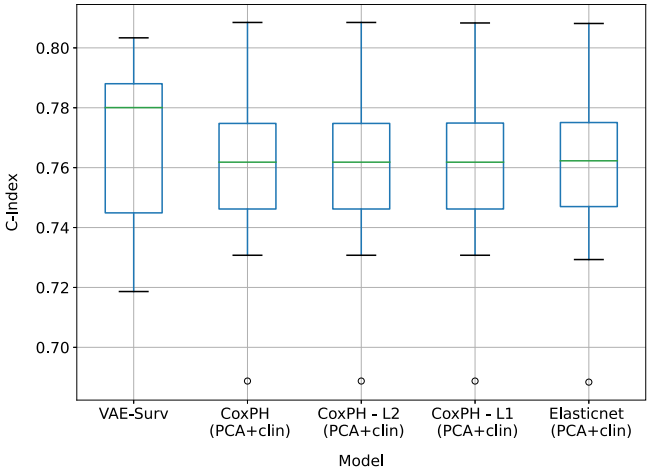

Fig. 3. Comparison of VAE-Surv and CoxPH models using PCA-reduced genetic features and clinical covariates.

This comparative analysis supports the ability of VAE-Surv framework in providing consistent survival risk predictions, outperforming a combination of traditional linear approaches, especially when high-dimensional genetic and cytogenetic data are considered.

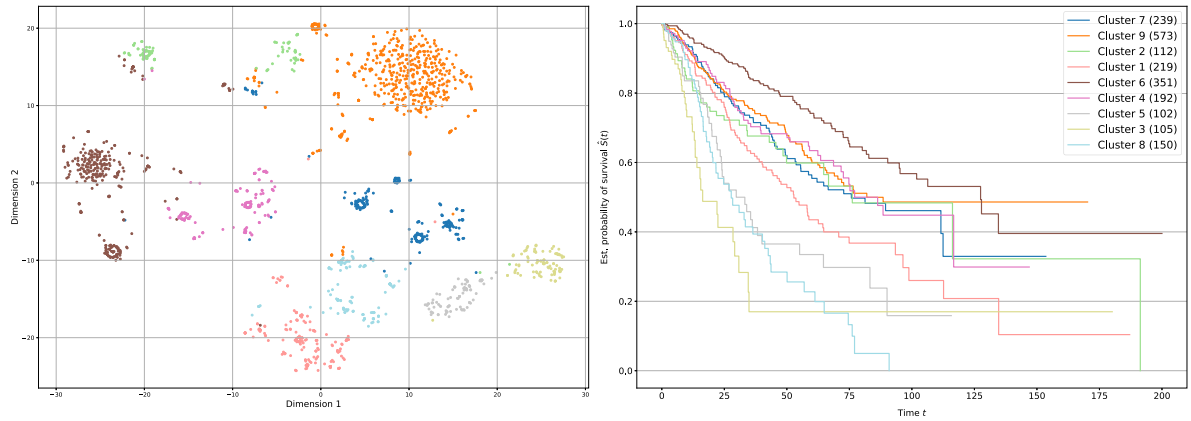

Fig. 4. Left panel: t-SNE representation of the patients' genetic and cytogenetic latent space learned from VAE, colored according to the cluster assignment. Right panel: Kaplan–Meier survival curves stratified by the identified cluster. The size of each group is specified in the legend.

3.2. VAE-surv latent representation and clustering performance

The dimensionality of z was optimized as a hyperparameter of the model and fixed to $z_{dim} = 7$. We employed K-means clustering on the z space to explore the utility of the latent representation created by the VAE. The optimal number of clusters was determined according to the quality metric Q_k. As shown in Supplementary Figure S1, the optimal configuration that ensures the maximum efficiency is when $k = 9$.

A t-SNE visualization of the latent representation points, colored by the assigned cluster, is displayed in Fig. 4, left panel. The Kaplan–Meier survival curves (Fig. 4, right panel) reveal distinct survival profiles, thereby validating that the clusters obtained from the latent space (whose creation was not survival-informed) provide meaningful stratification of patients' survival.

Table 2
Performance comparison including K-modes clustering on the input space, Survival Cluster Analysis, and K-means clustering on the latent representations learned by VAE-Surv.

	k	Q_k	Silhouette
K-modes on input space	3	0.041	0.15
Survival Cluster Analysis	3	0.035	0.26
K-means on VAE latent space	9	0.046	0.32

To evaluate the effectiveness of the clustering on the VAE-Surv latent space, we compared the K-means with two alternative methods: (1) K-modes clustering, applied directly to the raw genetic and cytogenetic data, and (2) the Survival Cluster Analysis (SCA) framework [4], which integrates survival data into its clustering process. For both approaches, the optimal number of clusters was $k = 3$ according to the predefined quality metrics Q_k (Supplementary Figures S2, S3). As summarized in Table 2, K-means clustering on the VAE-Surv latent space outperformed both K-modes and SCA in terms of both Silhouette score and Q_k metric. Despite SCA integrates survival data into the clustering process, its lower performance compared to VAE-surv indicates that it does not leverage genetic and cytogenetic data as effectively as the VAE-Surv latent representation.

3.3. Biological consistency of the MDS clusters

To investigate the biological consistency of the clusters defined by the genetic and cytogenetic features, we focused both on the distribution of MDS subtypes within the clusters and on the mutational profiling of each group. Supplementary Figures S6 and S7 show the number of mutations observed and the median values of the most relevant clinical attributes per cluster, respectively.

The clustermap reported in Fig. 5 and the Sankey plot displayed in Supplementary Figure S8 compare the identified clusters with the state-of-the-art MDS subtypes from WHO 2016 classification, based on both genetic and clinical features (including morphological criteria and haematologic parameters). The color intensity in each cell of the clustermap correlates with the percentage of patients in a given cluster belonging to a specific MDS subtype. While the clustermap explains the composition of each cluster according to the WHO MDS classification, the Sankey plot highlights how the MDS subtypes are distributed across all the new groups.

Focusing on each cluster, we found that cluster 6 collects patients with both subtypes of MDS with Ring Sideroblasts (RS-MLD and RS-SLD), which are not significantly present in the other clusters, as can also be noticed in the Sankey plot. Cluster 2 has a composite characterization, including Excess Blasts-1 (EB1), Excess Blasts-2 (EB2) and MDS with deletions in the long arm of chromosome 5 (5Q-) subtypes. It is worth highlighting that the MDS subtypes EB2, EB1 and the multilineage dysplasia (MLD) are spread across the clusters, with clusters 8, 3 and 9 mainly represented by these classes, respectively (Fig. 5). Therefore, it seems that some subtypes from WHO-based classification can be explained by potential subgroups identified by VAE-Surv. Indeed, focusing on the mutational landscape, the VAE-Surv clusters are represented by different underlying genetic features. Clusters 1, 6, 4, 8, 5 and 3 are strongly characterized by ASXL1, SF3B1, DNMT3 A, RUNX1, Gainofchr8 and Lossofchr7ordel7q alterations, respectively, as displayed in Fig. 6(a). This clustermap associates each identified cluster with the frequency of mutated genes or cytogenetic alterations in that cluster. Cluster 7 shows higher frequencies of SRSF2 and TET2 gene mutation with respect to the other groups, while

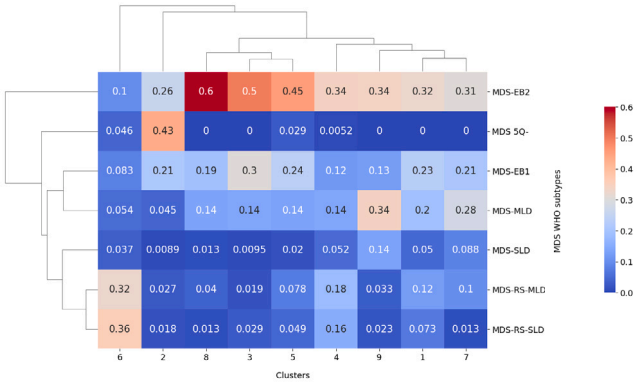

Fig. 5. Clustermap showing the distribution of WHO 2016-defined MDS subtypes (y-axis) within each of the nine clusters identified (x-axis). Each column sums to 1. The color intensity correlates with the percentage of patients in a given cluster belonging to a specific MDS subtype.

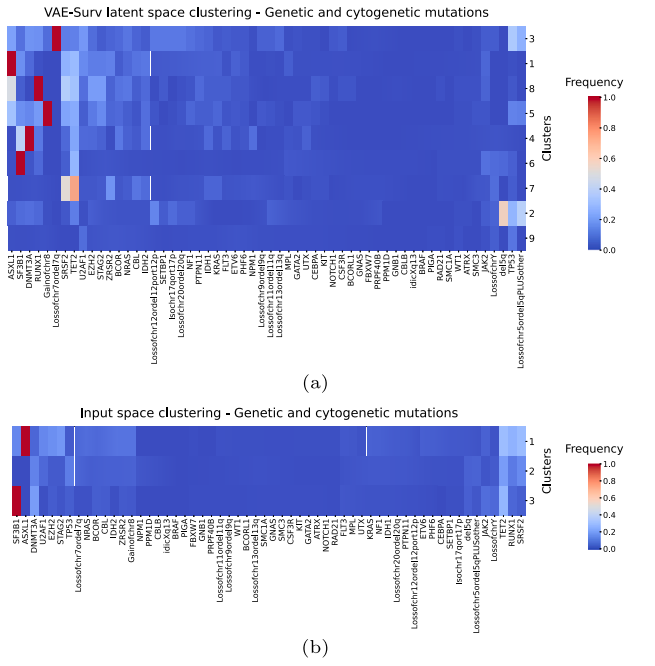

Fig. 6. Clustermaps displaying the frequency of gene mutations and cytogenetic alterations within each identified cluster. The color intensity corresponds to the prevalence of each genetic feature in the clusters, providing insights into their mutational landscape. Panels (a) and (b) show the results from clustering on the VAE latent space and on the input space, respectively.

cluster 2 is enriched with del5q, TP53 and Lossofchr5ordel5qPLUSother alterations. Notably, TP53 mutation is also frequent in cluster 3 (which is strongly defined by the Lossofchr7ordel7q alteration). In particular, it is worth highlighting that patients belonging to cluster 6, characterized by augmented SF3B1 gene mutations, exhibit improved survival outcomes, which is in agreement with the literature [27,28]; while patients from clusters 3, 5 and 8, which are associated respectively with Lossofchr7ordel7q, Gainofchr8 and RUNX1 mutations, exhibit poor survival outcomes. An intriguing insight emerging from our clustering approach lies in the distribution of patients within the '5Q-' subtype according to WHO 2016 MDS classification. These patients are mostly shared between clusters 2 (71%) and 6 (24%) and, despite sharing the same '5Q-' classification, the survival outcomes differ markedly. This divergence underscores the clinical utility of our model, offering a more nuanced stratification of the patients. Cluster 9 exhibits a flat mutational pattern, thus representing patients with a limited number of mutations. Indeed, as shown in Supplementary Figure S6, all patients with no mutations belong to this cluster. As expected, since cluster 9 includes patients with low mutation frequencies, they exhibit higher survival probabilities.

Finally, to further validate the biological relevance of the clustering methods, we analyzed the genetic mutational landscape within each cluster obtained by directly applying the clustering to the genetic input data, as done previously (Fig. 6(b)). In this case, since the optimal number of clusters k is 3, there is less differentiation in the prevalence of specific mutations compared to the VAE-Surv clusters. Indeed, cluster 2, which contains the vast majority of samples (1,230 patients, compared to 363 and 450 in cluster 1 and 3, respectively) fails to capture the intricate genetic and cytogenetic heterogeneity of the disease, which could be observed in a more granular way through the VAE-Surv approach. This suggests that the traditional clustering method is less effective at capturing the genetic diversity of MDS patients.

Table 3
Logistic Regression average performance for the cluster assignment task on the 100-times repeated internal test sets.

	Mean	Standard Deviation
Accuracy	0.995	0.004
MCC	0.994	0.005

4. Validation in the external cohort

To validate our framework, the VAE-Surv model was retrained on the entire GenoMed4all dataset, using the best hyperparameters optimized with the 10-fold cross-validation strategy. We then applied the model to the IWG-PW cohort, characterized by a censoring rate of 52% and a mean patients' overall survival time of 32.1 months. In order to perform a fair comparison of the results, the final model was tested using a bootstrap with resample technique, repeated 1,000 times and ensuring that, for each repetition, the censoring rate matched exactly that one of the training cohort (71.1%). The resulting median C-Index was equal to 0.74 (Standard Deviation = 0.01). Further details of the performance on the validation cohort are reported in Supplementary Materials (Sections S7, S8).

Then, a Logistic Regression classifier was trained on the latent space representation of the GenoMed4all patients and applied to the IWG-PW cohort to assign its patients to the previously defined genetic clusters. The performance of the classifier was assessed by creating random internal test sets within the training cohort and repeating the random split 100 times for statistical consistency. The results in terms of accuracy and Matthew's Correlation Coefficients are reported in Table 3. The latent representation of the training and validation cohorts are shown in Supplementary Figure S10.

Finally, we compared the interpretability of the results obtained from the survival models of VAE-Surv applied to both Genomed4all and IWG-PM cohorts. To this aim, we estimated the SHAP values, which quantify the contribution of each feature to the model's prediction for an individual patient's overall survival outcome, i.e. the relative risk. Originating from cooperative game theory, SHAP values are based on the Shapley values, which assign a fair distribution of the 'payout' (prediction effect) among the 'players' (features) [29]. By integrating SHAP with our VAE-Surv model, we can dissect the nonlinear, complex interactions captured by the deep learning architecture, providing insights into how specific genetic and clinical variables influence the survival predictions.

The Shapley values were calculated for the model retrained on the entire GenoMed4all cohort. Summary plots for the top-15 ranked features according to SHAP are shown in Supplementary Figures S11 and S12 for training and validation cohorts, respectively. It is possible to observe a consistency of the main clinical/genetic features involved in both models, confirming the reproducibility and generalizability of the results in different patients' cohorts.

5. Discussion

This study provides a robust computational framework for patient stratification in Myelodysplastic Syndromes (MDS) by integrating a Variational Autoencoder (VAE) with a DeepSurv model. The framework outperformed traditional CoxPH models in survival prediction tasks, showing that deep learning can add value to clinically characterize complex diseases like MDS.

The proposed model effectively exploits the rich genetic and cytogenetic landscape of MDS for feature extraction and subsequent risk stratification. The latent representations obtained by the VAE served as an informative summary of the genetic landscape, increasing the granularity of our survival model. The removal of the KL-term in the VAE loss function allowed a local interpretation of the feature space, helping to capture the inherent heterogeneity of the disease.

The application of K-means clustering on the latent space identified distinct patient groups, each with unique genetic and clinical characteristics. The visualization of these clusters in relation to WHO 2016 MDS subtypes indicates that our framework is capable of capturing existing clinical categorizations while also suggesting novel, data-driven patient groups. Indeed, according to the modern oncological view, WHO 2016 MDS classification is considered too coarse as it groups genetically different cancer subtypes into the same class [15,30,31]. VAE-Surv was able to refine WHO classification, and the non-uniform distribution of MDS subtypes across the new clusters implies that the model recognizes both well-defined and less-recognized patients' subgroups, potentially facilitating individualized treatment strategies. In contrast, we found that the direct clustering of the raw input genetic data reduces the number of cancer type classes to only 3, thus highlighting its inability to capture subtle distinctions in the input space. Furthermore, we also characterized the mutational landscape within each cluster, providing a more specific understanding of genetic factors driving different MDS subtypes.

Regarding the limitations of our study, its retrospective nature represents the primary limitation, including the potential for selection bias inherent in a single cohort and the high heterogeneity of the disease under investigation. To address these limitations, the model was validated in a second large cohort of patients. When comparing the two cohorts, the presence of significant differences, particularly in the WHO categories and risk classes, further highlights the heterogeneity of the disease under investigation. At the same time, these differences suggest the potential to generalize the findings to patients with characteristics distinct from those of the training cohort.

On the external validation cohort, consisting of 2,384 patients [18], VAE-Surv model performance resulted in an average C-Index equal to 0.74 and a cluster assignment accuracy of 0.995. These scores display the robustness and consistency of the model. Moreover, as shown by the SHAP analysis, the features that contribute most to the model's prediction are almost the same on both training and validation cohorts.

Finally, our findings align well with the previous results [15]. In particular, our analysis confirmed the pivotal role of gene SF3B1 mutation in improving the outcome and markedly characterizing a specific group of patients. In both works, patients with a flat mutational pattern fall into a well-defined cluster whose survival rate is higher than the average. The association of TP53 mutation with complex karyotypes, as highlighted by Bersanelli et al., is confirmed. However, our analysis delineates that this gene is related to extremely poor survival outcomes only when combined with Lossofchr7ordel7q (cluster 3), and not in cluster 2, where it co-occurs with del5q and Lossofchr5ordel5qPLUSother alterations.

6. Conclusion and future work

In summary, VAE-Surv framework demonstrates the power of deep learning in handling the intricate genetic landscape of MDS, offering a novel, robust methodology for patient stratification and survival prediction. It stands to contribute significantly to personalized medicine in the context of haematologic disorders, facilitating more accurate diagnosis and tailored treatment plans for MDS patients.

In future works, the integration with additional omics data like transcriptomics or epigenetics could be easily embodied in the model, hopefully providing a more comprehensive view of the disease. Moreover, the VAE-Surv architecture is not limited to the MDS case study. The tool can be extended to the study of other haematologic or even solid tumour malignancies where both clinical and genetic mutations data are available. By retraining the model on different disease-specific datasets, clinicians and researchers could gain similarly actionable insights for patient stratification and survival prediction across a wider array of conditions.

Data accessibility

The data that support the findings of this study are available from GenoMed4All consortium, but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with the permission of GenoMed4All consortium.

CRediT authorship contribution statement

Cesare Rollo: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Corrado Pancotti:** Writing – review & editing, Software, Methodology. **Flavio Sartori:** Writing – review & editing, Software, Methodology. **Isabella Caranzano:** Writing – review & editing, Software, Methodology. **Saverio D'Amico:** Validation, Data curation. **Luciana Carota:** Validation. **Francesco Casadei:** Validation. **Giovanni Birolo:** Writing – review & editing, Software, Methodology. **Luca Lanino:** Validation, Data curation. **Elisabetta Sauta:** Validation, Data curation. **Gianluca Asti:** Validation, Data curation. **Mattia Delleani:** Validation, Data curation. **Elena Zazzetti:** Validation, Data curation. **Marilena Bicchieri:** Validation, Data curation. **Giulia Maggioni:** Validation, Data curation. **Pierre Fenaux:** Validation, Data curation. **Uwe Platzbecker:** Validation, Data curation. **Maria Diez-Campelo:** Validation, Data curation. **Torsten Haferlach:** Validation, Data curation. **Gastone Castellani:** Validation. **Matteo Giovanni Della Porta:** Validation, Data curation. **Piero Fariselli:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Tiziana Sanavia:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by GenoMed4All consortium (Grant Agreement ID: 101017549) and PNRR M4C2 HPC-1.4 "CENTRI NAZIONALI"-Spoke 8. The funders played no role in study design, data collection, analysis, and interpretation of data, or the writing of this manuscript.

Corrado Pancotti was supported by a AIRC fellowship for Italy.

Code availability

The underlying code for this study is available at the link https://github.com/compbiomed-unito/VAE-Surv.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2025.108605.

References

[1] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. 34 (1972) 187–220.
[2] P. Wang, Y. Li, C.K. Reddy, Machine learning for survival analysis: A survey, ACM Comput. Surv. 51 (6) (2019) 1–36.
[3] S. Liverani, L. Leigh, I.L. Hudson, J.E. Byles, Clustering method for censored and collinear survival data, Comput. Statist. 36 (2021) 35–60.
[4] P. Chapfuwa, C. Li, N. Mehta, L. Carin, R. Henao, Survival cluster analysis, in: Proceedings of the ACM Conference on Health, Inference, and Learning, CHIL '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 60–68, http://dx.doi.org/10.1145/3368555.3384465.
[5] S. Boehm, P. Khosravi, R. Vanguri, J. Gao, S.P. Shah, Harnessing multimodal data integration to advance precision oncology, Nat. Rev. Cancer. 22 (2) (2022) 114–126.

[6] R. Wei, A. Mahmood, Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey, IEEE Access 9 (2020) 4939–4956.
[7] D. Wissel, D. Rowson, V. Boeva, Hierarchical autoencoder-based integration improves performance in multi-omics cancer survival models through soft modality selection, 2021, BioRxiv 2021-2009, Cold Spring Harbor Laboratory.
[8] L. Jiang, C. Xu, Y. Bai, A. Liu, Y. Gong, Y.P. Wang, H.W. Deng, Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data, NPJ Precis. Oncol. 8 (1) (2024) 4.
[9] S. Kim, K. Kim, J. Choe, I. Lee, J. Kang, Improved survival analysis by learning shared genomic information from pan-cancer data, Bioinformatics 36 (Supplement_1) (2020) i389–i398.
[10] M.T. Hira, M.A. Razzaque, C. Angione, J. Scrivens, S. Sawan, M. Sarker, Integrated multi-omics analysis of ovarian cancer using variational autoencoders, Sci. Rep. 11 (1) (2021) 6265.
[11] S. Ogawa, Genetics of MDS, Blood, J. Am. Soc. Hematol. 133 (10) (2019) 1049–1059.
[12] G. Garcia-Manero, Myelodysplastic syndromes: 2023 update on diagnosis, risk-stratification, and management, Am. J. Hematol. 98 (8) (2023) 1307–1325.
[13] R.M. Shallis, R. Ahmad, A.M. Zeidan, The genetic and molecular pathogenesis of myelodysplastic syndromes, Eur. J. Haematol. 101 (3) (2018) 260–271.
[14] D.A. Arber, A. Orazi, R. Hasserjian, J. Thiele, M.J. Borowitz, M.M. Le Beau, C.D. Bloomfield, M. Cazzola, J.W. Vardiman, The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia, Blood, J. Am. Soc. Hematol. 127 (20) (2016) 2391–2405.
[15] M. Bersanelli, E. Travaglino, M. Meggendorfer, T. Matteuzzi, C. Sala, E. Mosca, C. Chiereghin, N. Di Nanni, M. Gnocchi, M. Zampini, et al., Classification and personalized prognostic assessment on the basis of clinical and genomic features in myelodysplastic syndromes, J. Clin. Oncol. 39 (11) (2021) 1223.
[16] S. D'Amico, L. Dall'Olio, C. Rollo, P. Alonso, I. Prada-Luengo, D. Dall'Olio, C. Sala, M. Bersanelli, E. Sauta, M. Bicchieri, et al., Multi-modal analysis and federated learning approach for classification and personalized prognostic assessment in myeloid neoplasms, Blood 140 (Supplement 1) (2022) 9828–9830.
[17] M.A. Sekeres, J. Taylor, Diagnosis and treatment of myelodysplastic syndromes: a review, Jama 328 (9) (2022) 872–880.
[18] E. Bernard, H. Tuechler, P.L. Greenberg, R.P. Hasserjian, J.E. Arango Ossa, Y. Nannya, S.M. Devlin, M. Creignou, P. Pinel, L. Monnier, et al., Molecular international prognostic scoring system for myelodysplastic syndromes, NEJM Evid. 1 (7) (2022) EVIDoa2200008.
[19] F. Cremonesi, V. Planat, V. Kalokyri, H. Kondylakis, T. Sanavia, V. Miguel Mateos Resinas, B. Singh, S. Uribe, The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform, J. Biomed. Inform. 141 (2023) 104338.
[20] Saverio D'Amico, Lorenzo Dall'Olio, Cesare Rollo, Patricia Alonso, Iñigo Prada-Luengo, Daniele Dall'Olio, Claudia Sala, Elisabetta Sauta, Gianluca Asti, Luca Lanino, et al., MOSAIC: An artificial intelligence–based framework for multimodal analysis, classification, and personalized prognostic assessment in rare cancers, JCO Clin. Cancer Inform. 8 (2024) e2400008.
[21] D.P. Kingma, M. Welling, Auto-encoding variational bayes, 2013, arXiv preprint arXiv:1312.6114.
[22] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Med. Res. Methodol. 18 (1) (2018) 1–12.
[23] James Lucas, George Tucker, Roger Grosse, Mohammad Norouzi, Understanding posterior collapse in generative latent variable models, 2019.
[24] F.E. Harrell Jr., K.L. Lee, D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Stat. Med. 15 (4) (1996) 361–387.
[25] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for Cox's proportional hazards model via coordinate descent, J. Stat. Softw. 39 (2011) 1–13.
[26] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[27] L. Malcovati, K. Stevenson, E. Papaemmanuil, D. Neuberg, R. Bejar, J. Boultwood, D.T. Bowen, P.J. Campbell, B.L. Ebert, P. Fenaux, et al., SF3B1-mutant MDS as a distinct disease subtype: a proposal from the International Working Group for the Prognosis of MDS, Blood J. Am. Soc. Hematol. 136 (2) (2020) 157–170.
[28] L. Malcovati, M. Karimi, E. Papaemmanuil, I. Ambaglio, M. Jädersten, M. Jansson, C. Elena, A. Gallì, G. Walldin, M.G. Della Porta, et al., SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring sideroblasts, Blood, J. Am. Soc. Hematol. 126 (2) (2015) 233–241.
[29] A. Messalas, Y. Kanellopoulos, C. Makris, Model-agnostic interpretability with shapley values, in: 2019 10th International Conference on Information, Intelligence, Systems and Applications, IISA, IEEE, 2019, pp. 1–7.
[30] Y. Zhang, J. Wu, T. Qin, Z. Xu, S. Qu, L. Pan, B. Li, H. Wang, P. Zhang, X. Yan, et al., Comparison of the revised 4th (2016) and 5th (2022) editions of the World Health Organization classification of myelodysplastic neoplasms, Leukemia 36 (12) (2022) 2875–2882.
[31] A. Kündgen, M. Nomdedeu, H. Tuechler, G. Garcia-Manero, R.S. Komrokji, M.A. Sekeres, M.G. Della Porta, M. Cazzola, A.E. DeZern, G.J. Roboz, et al., Therapy-related myelodysplastic syndromes deserve specific diagnostic sub-classification and risk-stratification—an approach to classification of patients with t-MDS, Leukemia 35 (3) (2021) 835–849.