Computer Methods and Programs in Biomedicine 244 (2024) 107966

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

SurvIAE: Survival prediction with Interpretable Autoencoders from Diffuse Large B-Cells Lymphoma gene expression data

Gian Maria Zaccaria [a,1], Nicola Altini [a,1,*], Giuseppe Mezzolla [a], Maria Carmela Vegliante [b], Marianna Stranieri [a], Susanna Anita Pappagallo [b], Sabino Ciavarella [b], Attilio Guarini [b], Vitoantonio Bevilacqua [a,c]

[a] Department of Electrical and Information Engineering (DEI), Polytechnic University of Bari, Via Edoardo Orabona, 4, Bari 70126, Italy
[b] Hematology and Cell Therapy Unit, IRCCS Istituto Tumori "Giovanni Paolo II", Via O. Flacco, 65, Bari 70124, Italy
[c] Apulian Bioengineering srl, Via delle Violette, 14, Modugno 70026, Italy

ARTICLE INFO

Keywords:
Gene expression data
Survival prediction
Autoencoder
Explainable Artificial Intelligence

ABSTRACT

Background: In Diffuse Large B-Cell Lymphoma (DLBCL), several methodologies are emerging to derive novel biomarkers to be incorporated in the risk assessment. We realized a pipeline that relies on autoencoders (AE) and Explainable Artificial Intelligence (XAI) to stratify prognosis and derive a gene-based signature.
Methods: AE was exploited to learn an unsupervised representation of the gene expression (GE) from three publicly available datasets, each with its own technology. Multi-layer perceptron (MLP) was used to classify prognosis from latent representation. GE data were preprocessed as normalized, scaled, and standardized. Four different AE architectures (Large, Medium, Small and Extra Small) were compared to find the most suitable for GE data. The joint AE-MLP classified patients on six different outcomes: overall survival at 12, 36, 60 months and progression-free survival (PFS) at 12, 36, 60 months. XAI techniques were used to derive a gene-based signature aimed at refining the Revised International Prognostic Index (R-IPI) risk, which was validated in a fourth independent publicly available dataset. We named our tool SurvIAE: Survival prediction with Interpretable AE.
Results: From the latent space of AEs, we observed that scaled and standardized data reduced the batch effect. SurvIAE models outperformed R-IPI with Matthews Correlation Coefficient up to 0.42 vs. 0.18 for the validation-set (PFS36) and to 0.30 vs. 0.19 for the test-set (PFS60). We selected the SurvIAE-Small-PFS36 as the best model and, from its gene signature, we stratified patients in three risk groups: R-IPI Poor patients with High levels of *GAB1*, R-IPI Poor patients with Low levels of *GAB1* or R-IPI Good/Very Good patients with Low levels of *GPR132*, and R-IPI Good/Very Good patients with High levels of *GPR132*.
Conclusions: SurvIAE showed the potential to derive a gene signature with translational purpose in DLBCL. The pipeline was made publicly available and can be reused for other pathologies.

1. Introduction

Diffuse Large B-Cell Lymphoma (DLBCL) is a heterogeneous disease because of genetic alterations, morphology, and clinical context. Some subtypes are aggressive and chemo-refractory; however, other subtypes have shown prolonged survival after tailored treatment [1,2]. The most used clinical prognostic tool is the International Prognostic Index (IPI) which has been refined and adapted over time as in the Revised IPI (R-IPI) [3,4]. Gene Expression Profiling (GEP) studies unveiled prognostic Cell-Of-Origin (COO) subtypes of DLBCL, named Germinal Center B-Cell like (GCB) and Activated B-Cell like (ABC), driven by peculiar oncogenic pathways [5–7]. To better characterize DLBCL at diagnosis, among emerging prognosticators, transcriptome determinants were shown to predict patients' risk, also related to the cellular/extracellular microenvironment [8–11].

The application of Artificial Intelligence (AI) and Machine Learning (ML) tools in biology and medicine is currently on the rise [12–15]. AI and ML have been applied to drug discovery and GEP as well as to the development of novel clinical prognosticators [16–18]. Among these tools, models based on autoencoders (AE) are still emerging. The first

* Corresponding author.
E-mail address: nicola.altini@poliba.it (N. Altini).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.cmpb.2023.107966
Received 23 October 2023; Received in revised form 24 November 2023; Accepted 1 December 2023
Available online 2 December 2023

examples in oncology are focused on multi-omics integration for Colorectal (CRC) and breast cancers [19,20] and on spatial transcriptomics [21,22]. Indeed, multi-omics data are high-dimensional, posing problems for the generalization capabilities of downstream classifiers. AE architectures offer a powerful methodology to learn unsupervised representations from the data itself, reducing the dimensionality and unveiling underlying feature patterns. Furthermore, combined models, composed of AEs and classifiers, can be exploited to perform end-to-end feature extraction and classification. Such models can be investigated with eXplainable AI (XAI), a term that refers to those methods used to provide an understandable explanation of ML models' predictions and decisions [23–26].

In this paper, we propose an explainable pipeline to reduce dimensionality and subsequently stratify the prognosis of DLBCL patients starting from transcriptomic data. Even though AEs have been used with success for diverse applications, there are no general guidelines for model architecture design and data preprocessing in the context of lymphoma GE data. In the first part of our work, we systematically compared four AE architectures and three preprocessing for six different prognostic outcomes. We performed our analysis on three different datasets, each with its own technology for GE data acquisition. Later, we focused on the explainable module of our pipeline. By exploiting the SHapley Additive exPlanation (SHAP) algorithm [27] on the joint model composed by AE and the neural network for classifying the severity of the prognosis, we were able to unveil the most important genes for prognosis. Since the joint model exploits both AEs and neural networks to stratify the prognosis, i.e., survival prediction, we named it SurvIAE (Survival prediction with Interpretable AE). It is worth noting that the interpretability part of our pipeline had an important role, since it posed the basis for performing the validation on a fourth independent dataset by exploiting a new signature with the extracted genes and clinical features which are traditionally considered for DLBCL patients.

In summary, we can affirm that our work brings the following four contributions:

1. A systematic comparison of AE architectures and data preprocessing strategies. The analysis involved three free-publicly available datasets of both microarrays from different platforms and RNA-seq, so that we also estimated the batch effect and devised which methodologies can be adopted to reduce it.
2. A thorough investigation of the most relevant genes for each architecture, preprocessing, and outcome. Even though several works used XAI methodologies before, this is an in-depth investigation available in the context of DLBCL.
3. The development of the SurvIAE tool, whose code was made publicly available to ensure reproducibility and allow other researchers to reuse it on other pathologies. SurvIAE effectively combines AE, neural networks, and interpretability to unveil prognostic biomarkers.
4. An examination of the clinical translation of SurvIAE. First, we considered the risk scores of SurvIAE to construct a prognostic model and compared it with the recognized clinical scores for DLBCL, including R-IPI and COO. Then, we plugged the most relevant genes found by the XAI module to create an enhanced R-IPI prognostic score.

2. Related works

Studies involving AE models in oncology are generally aimed at reducing the dimensionality of omics data to predict cancer subtyping and discover molecular patterns associated with new potential biomarkers. In Lupat et al., the authors used a semi-supervised AE to generalize the combination of GE, copy number, and somatic mutation data from a dataset of breast cancer. Model performances were independently validated on a cohort of The Cancer Genome Atlas (TCGA) samples achieving more than 85 % of accuracy in subtype identification [20]. Way and Greene trained variational AEs on TCGA pan-cancer RNA-seq data, unveiling specific patterns in the encoded features. Interestingly, they identified both primary and metastatic tumors of skin cutaneous melanoma as well as a lower dimensional manifold of high-grade serous ovarian cancer subtypes [28]. Dwivedi et al. reduced the GE dimensionality of a TCGA dataset of Non-Small Cell Lung Cancer (NSCLC) with AE involvement. Authors trained different classifiers such as Multi-Layer Perceptron (MLP), Logistic Regression (LR), eXtreme Gradient Boosting (XGB), and Support Vector Machine (SVM) for biomarker discovery purposes. They compared an XAI-based feature selection with classical ML algorithms such as Random Forest (RF), SVM-RF, Least Absolute Shrinkage and Selection Operator (LASSO), Mutual Information (MI), and ReliefF. Applying the Leave-One-Out-Validation-set, the MLP classifier achieved Area Under the Receiver Operating Characteristic curve (AUROC) of 98.89 % whereas the XAI-based feature selection outperformed other methods in terms of accuracy (95.74 % vs. 93.62 %, 93.80 %, 93.89 %, 91.76 %, 92.47 %, and 92.20 % of SVM-RFE, MI, ReliefF, LASSO, XGB, and RF, respectively) [29]. This research group used a similar pipeline to predict breast cancer subtypes either from Copy Number Variation (CNV) or DNA methylation data [30,31]. In the first work, the XAI-CNVMarker classifier, on the same TCGA breast cancer dataset, outperformed previous tools with an accuracy of 0.712 vs. 0.705 and 0.706 allowing to discover a signature of 44 genes, more clinically applicable than others with more than 200 genes. In the second work, the XAI-MethylMarker classifier comprises an AE for dimensionality reduction, a feed-forward neural network to classify breast cancer subtypes, and a biomarker discovery algorithm employing XAI. The classifier achieved an accuracy of 0.815 and led to the discovery of a predictive signature of 52 genes via XAI.

Focusing on prognosis stratification, Wang and Lee identified two prognostic subgroups of luminal-A breast cancer. Authors trained an AE using GEP of luminal-A breast cancer to predict subgroups through the latent features and unsupervised learning. Afterward, they validated latent features independently proving that only these features allowed to identify distinct statistically significant prognostic groups (p-value: 5.82E-05) with respect to all GEP (p-value: 0.566), the most variable 5000 genes (p-value = 0.426) and 64 or 2-dimensional feature sets generated by traditional Principal Component Analysis (PCA) from whole GEP or top 5000 genes (p-values: 0.608, 0.183, 0.136, and 0.385, respectively) [32]. This point was also demonstrated by Song et al. in a multi-omics integration project from publicly available CRC datasets. Their AE-based prognostic model outperformed other strategies of dimensionality reduction with a concordance (C)-index of 0.781 vs. 0.665, 0.766, 0.632, and 0.755 for PCA, t-distributed Stochastic Neighbor Embedding (t-SNE), Non-Negative Matrix Factorization (NMF), and Cox Proportional Hazard (Cox-PH) model, respectively [19].

AE involvement in onco-hematology is still rare. However, researchers understood the high potentiality of using XAI tools to optimize the findability of robust and validated prognostic determinants [33–36]. In Maiseles et al., the authors used XAI for chronic lymphocytic leukemia treatment prediction. Starting from a limited cohort of real-life patients, they used several ML models (GB Machine [GBM], Generalized Linear Model, RF, Adaboost, SVM, and Catboost) to find the most accurate for predicting the clinical outcome after 2 years of treatment. For the best model (GBM model with 0.880 of AUROC and 0.78 of Area Under the Precision-Recall Curve [AUPRC]), the SHAP algorithm helped in identifying Red Blood Cells, Beta2-microglobulins, lymphocytes, and platelets levels as most important clinical determinants. On the other hand, SHAP selected 11q deletion as the most important mutational feature [37].

Due to epidemiologic reasons and the aggressiveness of the disease, DLBCL is one of the most studied hematological cancers and several works proposed innovative pipelines to find novel prognostic markers. For instance, LASSO logistic analysis has been used to identify novel prognostic signatures from GE data. In Wang et al., the authors

discovered a metabolic signature showing an AUROC of 0.725, 0.716, and 0.752 at 3 years, for the training-set, first and second validation-sets, respectively [38]. Again, Xiong et al. designed a risk score based on 19 survival-related genes associated with ferroptosis. This tool achieved an AUROC of 0.801 for the training-set and 0.708 for the validation-set at 3 years of FU [39]. Jiang et al. developed a prognostic tool based on Differentially Expressed Genes (DEG) from mRNA according to high and low immune infiltration groups obtained via unsupervised hierarchical clustering. Although this contribution lacked an independent validation-set, authors proposed a 16-genes signature showing an AUROC of 0.775 at 3 years of FU for the training-set, outperforming IPI (0.714), and comparably with other relevant works [40–44].

3. Materials and methods

3.1. Pipeline

According to the Fig. 1, the pipeline includes four blocks: Data Preparation, Model Design, Models Evaluation, and Prognostic Translation. The pseudocode which describes the pipeline is reported in Supplementary Algorithm 1.

3.2. Datasets

We considered four publicly available datasets with expression matrices:

- The GSE117556 which comprises GEP data from formalin-fixed paraffin-embedded (FFPE) samples from n. 928 DLBCL patients

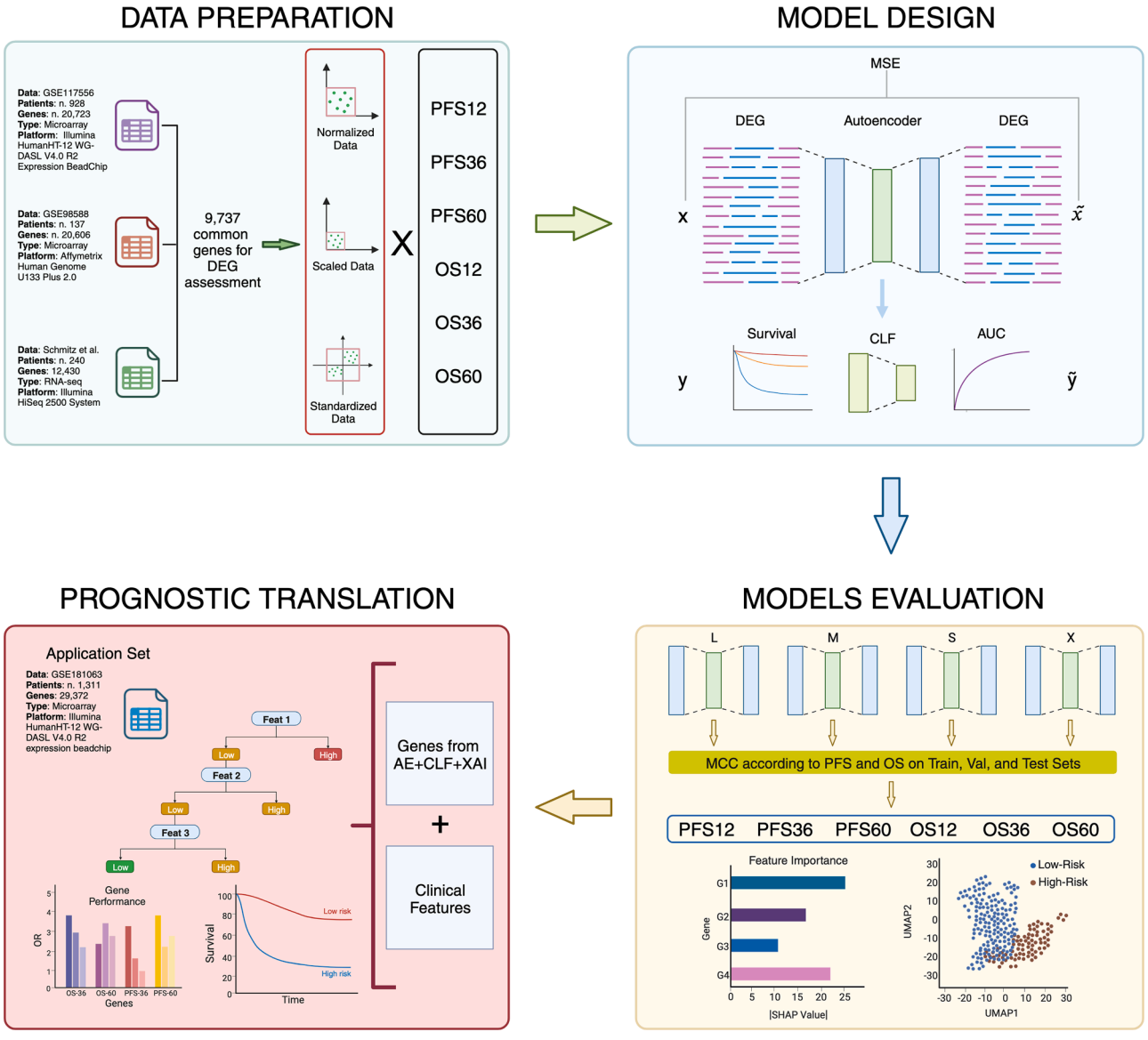

Fig. 1. Schematic description of the pipeline used for data processing. The Data Preparation block (light green) depicts the datasets' selection, the gene expression preprocessing strategy, and the clinical outcome definition. The Model Design block (light blue) describes the general AE architecture and the classifier modeling. The Models Evaluation block (light yellow) describes the strategy for selecting the best AE architecture according to clinical outcomes and the XAI involvement in selecting the most important biological features. The Prognostic Translation block (light red) shows how clinical and biological validation was performed. Abbreviations. DEG, Differential Expression Genes; PFS, Progression Free Survival; OS, Overall Survival; MSE, Mean Squared Error; CLF, classifier; AUC, Area Under Curve; L, Large; M, Medium; S, Small; X, extra-small; MCC, Matthews Correlation Coefficient; Train, Training; Val, Validation-set; SHAP, SHapley Additive exPlanation; UMAP, Uniform Manifold Approximation and Projection; Feat, feature; OR, Odds-Ratio.

(Microarray technology by Illumina® HumanHT-12 WG-DASL V4.0 R2 expression beadchip [11]).
- The GSE98588, including GEP data from frozen samples from n. 137 DLBCL patients (Microarray technology by Affymetrix® Human Genome U133 Plus 2.0 [9]).
- The Schmitz et al. dataset, which includes RNA-seq data from frozen samples from n. 240 DLBCL patients by Illumina® HiSeq 2500 System [7].
- The GSE181063 dataset, which includes GEP data from FFPE samples from n. 1,311 DLBCL patients (Microarray technology by Illumina® HumanHT-12 WG-DASL V4.0 R2 expression beadchip [45]).

We assumed the GSE117556 as the training-set, the GSE98588 as the validation-set, and the dataset from Schmitz et al. as the test-set. Those datasets were exploited for model training, evaluation, and extraction of the prognostic signature. Finally, the GSE181063 was considered as the application-set to check the significance of the devised prognostic signature. Patients with missing clinical data were excluded for further analysis. All patients were diagnosed with nodal, *de novo* DLBCL, not otherwise specified (NOS) and homogeneously treated with front-line "R-CHOP/R-CHOP-like" immunochemotherapy.

3.3. Preprocessing

After selecting common genes between datasets, three preprocessing procedures were compared:

1. **Normalized Data**. Expression data from the GSE117556 dataset (Illumina platform) was analyzed using the authors' normalization settings [11]. GSE98588 raw data (Affymetrix® technology) were summarized and normalized using the Robust Multi-array Averaging (RMA) method by means of affy (v. 1.70.0) package in R software (v. 4.2.1) [9]. RNA-seq data from Schmitz et al. were analyzed using the authors' normalization settings including counts per million, transcripts per million, and fragments per kilobase of transcript per million space, respectively [7]. Consistently, data from GSE181063 was analyzed using the authors' normalization settings.
2. **Scaled Data**. For each dataset, normalized expression data from point n. 1 underwent min-max scaling.
3. **Standardized Data**. For each dataset, normalized expression data from point n. 1 underwent a Z-score transformation.

Points 2 and 3 were carried on by using sklearn.preprocessing package (scikit-learn v. 1.2.2, Python v. 3.8.16).

Subsequently, the DEG analysis was conducted on the training-set, considering Progression Free Survival (PFS) and Overall Survival (OS) clinical outcomes at different times of FUs as 12 months (PFS12, OS12), 36 months (PFS36, OS36), and 60 months (PFS60, OS60). We selected DEGs based on adjusted p-value < 0.05 and log fold change > 0.2. DEG analysis was performed by limma R package (v. 3.48.0).

3.4. SurvIAE model

The autoencoder architecture exploited by SurvIAE consists of contracting and expanding blocks, in a symmetrical fashion, to create an intermediate latent representation of the data which may be useful for data compression or to perform other downstream tasks. The input to the autoencoder consisted of DEGs extracted for the various possible outcomes.

To design the architecture, we considered a symmetrical structure with two layers of encoding and two layers of decoding. Particularly, we defined four sizes starting from this base AE structure, resulting in the following topologies:

- **AE-L** (**L**arge). First layer of encoding and last layer of decoding with 512 neurons. Second layer of encoding and first layer of decoding with 256 neurons. Latent representation extracted from 128 neurons.
- **AE-M** (**M**edium). First layer of encoding and last layer of decoding with 256 neurons. Second layer of encoding and first layer of decoding with 128 neurons. Latent representation extracted from 64 neurons.
- **AE-S** (**S**mall). First layer of encoding and last layer of decoding with 128 neurons. Second layer of encoding and first layer of decoding with 64 neurons. Latent representation extracted from 32 neurons.
- **AE-X** (e**X**tra small). First layer of encoding and last layer of decoding with 64 neurons. Second layer of encoding and first layer of decoding with 32 neurons. Latent representation extracted from 16 neurons.

AEs were trained on the training-set defined in Section 3.2. For the training process, the ADAM [46] optimizer with a Mean Squared Error (MSE) loss was adopted. MSE and Mean Absolute Error (MAE) were monitored during the training process, which lasted 2,000 epochs, and used as evaluation metrics for the AE.

Latent representations extracted from the intermediate layer of the AE were qualitatively investigated, through the adoption of Uniform Manifold Approximation and Projection (UMAP) [47], between the different datasets, to check the occurrence of batch effects or other anomalies related to the adoption of different technologies of data acquisition.

Particularly, to quantify the batch effect among the different normalization techniques, the Silhouette score (Sil) was calculated as the average of the Silhouette Coefficient of all samples, using the implementation from sklearn.metrics (scikit-learn v. 1.2.2, Python v. 3.8.16), and portrayed on the embedding plots. As data points, the UMAP representations obtained on top of the latent representation of the AE were employed. As labels, the dataset from which the data point originates (training-set, validation-set, test-set). In this way, a higher value of Sil (close to 1) means that there is a strong batch effect, whereas a lower value of Sil (negative or close to 0) means that the batch effect is negligible.

Then, a classifier, consisting of an MLP with one hidden layer, with the size defined according to AE size, and one output layer, was trained for 200 epochs with the ADAM optimizer with the binary cross-entropy loss function. To mitigate the risk of overfitting, an early stopping criterion was implemented. In fact, after that the patience reached 100 times on the validation-set defined in Section 3.2, the training process was halted, and the best weights were restored. During the training of the classifier, the weights of the layers belonging to the AE were not updated.

We refer to the joint model of autoencoder and MLP as **SurvIAE**, and to its variants as follows:

- **SurvIAE-L**. Autoencoder architecture: AE-L; hidden layer of the MLP classifier with 64 neurons.
- **SurvIAE-M**. Autoencoder architecture: AE-M; hidden layer of the MLP classifier with 32 neurons.
- **SurvIAE-S**. Autoencoder architecture: AE-S; hidden layer of the MLP classifier with 16 neurons.
- **SurvIAE-X**. Autoencoder architecture: AE-X; hidden layer of the MLP classifier with 8 neurons.

The SurvIAE architecture was implemented with TensorFlow (tensorflow v. 2.10.1, Python v. 3.8.16).

Data representation obtained from the hidden layer of the MLP was investigated in conjunction with the latent representation of the AE. Again, the results of the classification stage were assessed in terms of AUROC and AUPRC. Also, for the hidden layer of the classification model, UMAP was used to qualitatively assess its internal representations.

3.5. Comparison with baseline models

The classification performances of the different joint AE-MLP architectures were compared with respect to reference ensemble methods that can be used for this task. Indeed, Gradient Boosting [48], AdaBoost [49], Random Forests [50], and Extra Trees [51] were borrowed from sklearn.ensemble (scikit-learn v. 1.2.2, Python v. 3.8.16), whereas XGB (eXtreme Gradient Boosting) from xgboost [52] (xgboost v. 1.7.5, Python v. 3.8.16).

3.6. Model interpretability

Interpreting a model can be thought as the task of finding an explanation model g which is an interpretable approximation (and hence a simpler model) of an original model f.

The authors of SHAP [27] introduced the concept of additive feature attribution methods, in which the explanation model g can be expressed as a linear combination of binary terms: $g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$, where $z' \in \{0,1\}^M$, M signifies the number of simplified input features, and $\phi_i \in \mathbb{R}$. Explainability methodologies which match this definition attribute an effect ϕ_i to every feature, and the sum of the effects of all feature attributions approximates the output $f(x)$ of the original model.

DeepLIFT was first proposed as an approach for recursively providing explanations of deep learning models' predictions [53]. It works by assigning a value $C_{\Delta x_i \Delta y}$ to every input x_i. The term $C_{\Delta x_i \Delta y}$ signifies the impact of adjusting the input to a reference value instead of its initial value. In DeepLIFT, a "summation-to-delta" property is enforced with $\sum_{i=1}^{n} C_{\Delta x_i \Delta o} = \Delta o$, where $o = f(x)$ represents the output of the model, $\Delta o = f(x) - f(r)$, $\Delta x_i = x_i - r_i$, and r signifies the reference output. Considering $\phi_i = C_{\Delta x_i \Delta o}$ and $\phi_o = f(r)$, we can see that DeepLIFT is an additive feature attribution method. Lundberg and Lee modified DeepLIFT by transforming it into a compositional approximation of SHAP values, resulting in the creation of Deep SHAP [27].

In SurvIAE, we exploited the Deep SHAP [27] algorithm to extract an interpretable approximation of the joint AE-MLP classification system. Starting from the obtained Shapley values, the most significant genes for prognosis were individuated and used for subsequent analysis to build a prognostic signature. This signature was then evaluated on the independent application-set, to demonstrate the prognostic translation of our model, as detailed in next section.

3.7. Clinical validation and prognostic translation

Clinical validation was fulfilled for each data preprocessing setup evaluated according to each categorical clinical outcome from Section 3.3 and exploiting each AE architecture from Section 3.4. Firstly, for the three datasets used for model training, evaluation, and extraction of the prognostic signature, i.e., training-set, validation-set, and test-set, we merged clinical and expression data excluding incomplete cases. Every SurvIAE-based prognostic model was then evaluated after the application of univariate LR, in terms of odds-ratio and significance level between Cluster 1 vs. Cluster 2. The p-values were derived from pairwise comparisons using z-statistic. To measure its prognostic value, for both validation-set and test-set, each SurvIAE-based model was compared to R-IPI categorized merging patients classified as Very Good and Good vs. Poor class which was assumed as reference. Again, for both validation-set and test-set, each model was thus evaluated in terms of Matthews Correlation Coefficient (MCC). MCC was measured by ROCR R package (v. 1.0.11). We only selected the SurvIAE-based models outperforming R-IPI models with higher MCC for both datasets. From each selected model, we extracted the top ten genes ordered by the sum of SHAP values from validation-set and test-set. We refer to these top ten genes as SHAP-derived signature for a SurvIAE model. The expression values of these genes were dichotomized from the training-set, according to a cutoff identified by maximally selected rank statistics, in 2 groups ("high" or "low") using the function surv_cutpoint as implemented in the survminer R package (v. 0.4.9). Hence, for all datasets, each gene was evaluated in terms of significance-level after the application of the univariate LR model. Indeed, only genes discriminating the clinical outcomes with a p-value < 0.1 were selected. Again, the p-values were derived from pairwise comparisons using z-statistic.

Then, the best SurvIAE model was identified with the following two conditions: (i) it possessed the highest difference of MCC with respect to the R-IPI, on both validation-set and test-set; (ii) its SHAP-derived signature includes at least one gene with significant impact on the clinical outcome. Finally, on the signature associated with the best model, we applied a multivariate LR analysis including also COO and R-IPI.

For prognostic translational purposes, the selected genes were included again with both COO and R-IPI features in a recursive decision-tree model using the partykit R package (v. 1.2.20) [54]. Decision-tree modeling was implemented on the training-set and the output groups of patients obtained were then identified in the application-set comprising newly diagnosed DLBCL patients treated with R-CHOP with complete data. PFS and OS analyses were performed on application-set according to novel classes obtained post decision-tree application, with Kaplan-Meier (K-M) method. The models were compared by assessing C-index and Brier score. The survival analysis was implemented with survival R package (v. 3.5.5).

4. Results

4.1. Autoencoder

Originally, the training-set included n. 928 patients and n. 20,723 transcripts, the validation-set n. 137 patients and n. 20,606 transcripts, and the test-set n. 240 patients and n. 12,430 transcripts. After merging the datasets, n. 9,737 common genes were retained. After filtering for patients with available outcomes, the training-set contained n. 928 cases for both OS and PFS, the validation-set n. 101 cases for OS and n. 98 cases for PFS, and the test-set n. 234 cases for both OS and PFS. After retaining only the common genes, the training-set was used to perform the DEG analysis, which revealed n. 415 DEGs for OS12, n. 827 for OS36, n. 860 for OS60, n. 288 for PFS12, n. 391 for PFS36, and n. 394 for PFS60 (details portrayed in Fig. 2).

The different AE models trained displayed the performance, in terms of MSE and MAE, reported in Tables 1 and 2, for the validation-set and test-set, respectively. Generally, it is possible to see that larger AEs tend to perform better on the reconstruction task, displaying lesser values of MSE and MAE. Also, because of the nature of the preprocessing, AEs trained on scaled data have very low values, since the scale is smaller.

For **normalized data**, MSE ranged between 3.268–9.091 (5.475 $\pm$ 1.471) and 3.437–13.021 (7.268 $\pm$ 2.299) for the validation-set and test-set, respectively, whereas MAE ranged between 1.405–2.121 (1.650 $\pm$ 0.170) and 1.312–2.369 (1.772 $\pm$ 0.225) for the validation-set and test-set, respectively. For **scaled data**, MSE ranged between 0.039–0.054 (0.047 $\pm$ 0.004) and 0.026-0.046 (0.035 $\pm$ 0.005) for the validation-set and test-set, respectively, whereas MAE ranged between 0.155–0.182 (0.170 $\pm$ 0.007) and 0.127-0.166 (0.146 $\pm$ 0.010) for the validation-set and test-set, respectively. For **standardized data**, MSE ranged between 0.759–0.948 (0.866 $\pm$ 0.060) and 0.743-0.919 (0.834 $\pm$ 0.052) for the validation-set and test-set, respectively, whereas MAE ranged between 0.682–0.760 (0.726 $\pm$ 0.024) and 0.659–0.739 (0.701 $\pm$ 0.024) for the validation-set and test-set, respectively.

The quality of the internal representation of the AE intermediate layer can be seen by embedding plots obtained with UMAP in Figs. 3 and 4, for OS60 and PFS60, respectively. Interestingly, the batch effects among different datasets were considerably reduced by using standardized data, suggesting the possibility of multi-technology data integration. AE models that were trained on standardized data suffered less from batch effect than those trained on simply normalized data, as is

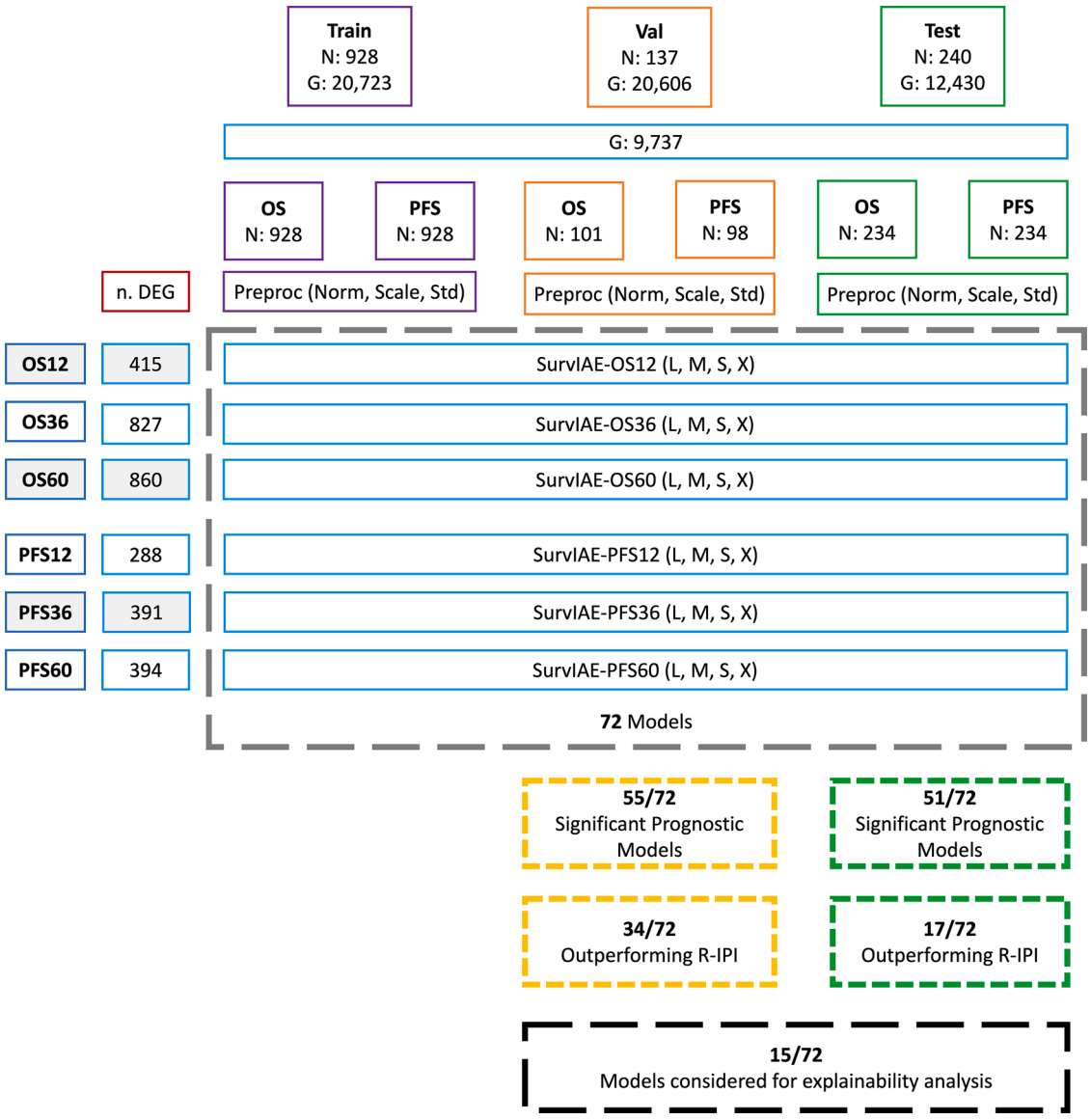

Fig. 2. Model selection workflow. Three normalization and four architectures were considered for each prognostic outcome, resulting in 12 models for each. Six different prognostic outcomes (OS and PFS, at FU times of 12, 36, 60 months) were considered, resulting in 72 models compared for this study. Models considered for XAI are only those which are statistically significant as prognosis predictors, and which outperform R-IPI, resulting in 15 models. Those models were then used to devise novel gene signatures to incorporate into clinically applicable prognostic models for DLBCL. Abbreviations. Train, Training-set; Val, Validation-set; Test, Test-set; G, Genes; OS, Overall Survival; PFS, Progression Free Survival; DEG, Differential Expression Genes; Preproc, Preprocessing; Norm, Normalized; Std, Standardized; L, Large; M, Medium; S, Small; X, extra-small; R-IPI, Revised International Prognostic Index.

quantified by Sil calculated as described in Section 3.4. Indeed, Sil values for standardized data are close to 0 or slightly negative.

The different SurvIAE models trained displayed the performance, in terms of AUROC and AUPRC, reported in Tables 3 and 4, for the validation-set and test-set, respectively. In this case, it is possible to see that some outcomes are difficult to predict, such as OS12. On the other hand, other outcomes are easier to predict, such as OS60 and PFS60, resulting in generally higher values of AUROC and AUPRC. In those tables, five reference ensemble methods are also considered for the comparison: GradientBoosting, AdaBoost, RandomForest, ExtraTrees, and XGB.

On the validation-set, the only case in which ensemble models perform better than SurvIAE models on both AUROC and AUPRC is PFS60. For OS12, OS60, and PFS36, SurvIAE improves over ensemble methods on both AUROC and AUPRC. For OS36, SurvIAE presents a better AUROC but a slightly lesser AUPRC. For PFS12, SurvIAE possesses a better AUPRC but a worse AUROC.

On the test-set, the only case in which ensemble models perform better than SurvIAE models on both AUROC and AUPRC is PFS36. In all other cases, SurvIAE displayed better AUROC and AUPRC than ensemble models.

For SurvIAE, the ROC curves for OS60 and PFS60 are portrayed in Figs. 5 and 6, respectively. It is possible to see that, for the PFS60 outcome, simply normalized data poses more generalization problems, with SurvIAE-S displaying unsatisfactory performance on validation-set and test-set, and SurvIAE-M on test-set. Indeed, the adoption of standardized data led to more stable performance among the datasets.

For **normalized data**, AUROC ranged between 0.387–0.754 (0.635 $\pm$ 0.084) and 0.388–0.654 (0.571 $\pm$ 0.063) for the validation-set and test-set, respectively, whereas AUPRC ranged between 0.146–0.608 (0.415 $\pm$ 0.138) and 0.183–0.542 (0.379 $\pm$ 0.094) for the validation-set and test-set, respectively. For **scaled data**, AUROC ranged between 0.537–0.824 (0.671 $\pm$ 0.057) and 0.548–0.679 (0.614 $\pm$ 0.036) for the validation-set and test-set, respectively, whereas AUPRC ranged

Table 1
Performance in terms of MSE and MAE of the different AE models on the validation-set. Abbreviations. AE, autoencoder; L, large; M, medium; S, small; X; extra-small; OS, Overall Survival; PFS, Progression Free Survival; MSE, Mean Squared Error; MAE, Mean Absolute Error.

Validation-set													
Model	Preprocessing	OS12		OS36		OS60		PFS12		PFS36		PFS60	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AE-L	**Normalized**	5.638	1.614	3.873	1.497	5.151	1.569	3.268	1.405	4.023	1.490	3.499	1.433
	Scaled	0.050	0.176	0.049	0.176	0.050	0.177	0.041	0.159	0.042	0.163	0.043	0.164
	Standardized	0.802	0.694	0.838	0.714	0.846	0.717	0.764	0.683	0.759	0.682	0.771	0.691
AE-M	**Normalized**	7.160	1.921	5.143	1.607	4.953	1.607	4.790	1.569	4.085	1.519	3.552	1.416
	Scaled	0.045	0.167	0.049	0.175	0.047	0.171	0.039	0.155	0.045	0.169	0.042	0.163
	Standardized	0.854	0.718	0.876	0.731	0.873	0.728	0.796	0.699	0.802	0.704	0.814	0.710
AE-S	**Normalized**	5.630	1.672	5.533	1.750	6.127	1.686	6.308	1.696	5.100	1.633	5.040	1.650
	Scaled	0.043	0.162	0.046	0.169	0.049	0.175	0.045	0.165	0.045	0.167	0.044	0.166
	Standardized	0.948	0.755	0.899	0.739	0.904	0.740	0.924	0.753	0.896	0.741	0.906	0.745
AE-X	**Normalized**	9.091	2.121	5.517	1.603	6.359	1.730	7.653	1.846	8.058	1.891	5.849	1.684
	Scaled	0.050	0.175	0.054	0.182	0.051	0.177	0.052	0.177	0.050	0.174	0.047	0.169
	Standardized	0.944	0.756	0.897	0.739	0.911	0.745	0.948	0.760	0.893	0.737	0.922	0.747

Table 2
Performance in terms of MSE and MAE of the different AE models on the test-set. Abbreviations. AE, autoencoder; L, large; M, medium; S, small; X; extra-small; OS, Overall Survival; PFS, Progression Free Survival; MSE, Mean Squared Error; MAE, Mean Absolute Error.

Test-set													
Model	Preprocessing	OS12		OS36		OS60		PFS12		PFS36		PFS60	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AE-L	**Normalized**	7.200	1.711	4.697	1.529	6.412	1.672	3.437	1.312	5.471	1.611	4.725	1.584
	Scaled	0.040	0.158	0.035	0.148	0.037	0.153	0.027	0.131	0.027	0.130	0.029	0.136
	Standardized	0.786	0.679	0.806	0.685	0.812	0.690	0.752	0.659	0.743	0.659	0.743	0.660
AE-M	**Normalized**	9.126	2.016	6.714	1.757	6.436	1.720	5.574	1.561	4.836	1.583	4.851	1.551
	Scaled	0.038	0.153	0.036	0.149	0.035	0.147	0.026	0.127	0.030	0.136	0.029	0.135
	Standardized	0.830	0.702	0.846	0.705	0.847	0.707	0.781	0.676	0.784	0.681	0.780	0.680
AE-S	**Normalized**	7.898	1.807	7.279	1.843	8.695	1.878	7.396	1.705	7.017	1.834	6.595	1.745
	Scaled	0.039	0.154	0.035	0.147	0.037	0.151	0.032	0.138	0.033	0.142	0.032	0.141
	Standardized	0.919	0.739	0.873	0.716	0.874	0.717	0.894	0.728	0.859	0.714	0.861	0.714
AE-X	**Normalized**	13.021	2.369	7.477	1.815	8.886	1.903	10.595	1.980	11.716	2.184	8.375	1.860
	Scaled	0.046	0.166	0.043	0.162	0.041	0.158	0.038	0.151	0.036	0.147	0.037	0.149
	Standardized	0.915	0.737	0.863	0.713	0.864	0.715	0.883	0.722	0.841	0.705	0.857	0.713

between 0.179–0.688 (0.468 ± 0.127) and 0.235–0.585 (0.426 ± 0.092) for the validation-set and test-set, respectively. For **standardized data**, AUROC ranged between 0.512–0.742 (0.643 ± 0.056) and 0.488–0.680 (0.600 ± 0.042) for the validation-set and test-set, respectively, whereas AUPRC ranged between 0.131–0.636 (0.451 ± 0.139) and 0.226–0.572 (0.411 ± 0.096) for the validation-set and test-set, respectively.

4.2. Prognostic evaluation of the AE-based strategy and comparison with R-IPI

Overall, we tested 72 models considering the six categorical clinical outcomes, the four AE architectures and the three data preprocessing techniques. Table 5 shows odds-ratios and p-values for each univariate LR model comparing, on one hand, two subgroups of patients clustered by the SurvIAE-based strategy (Cluster 1 vs. Cluster 2), and, on the other hand, two subgroups of patients according to R-IPI clinical prognostic tool (Good/Very Good vs. Poor) for the validation-set and test-set. Firstly, R-IPI demonstrated a prognostic value with significant levels for OS12, OS36, OS60, and PFS12 outcomes on the validation-set, and for all clinical outcomes on the test-set.

Considering all preprocessing strategies (normalized, scaled, and standardized expression data), 55/72 and 51/72 SurvIAE-based models demonstrated statistically significant levels of survival between Cluster 1 and Cluster 2 for the validation-set and the test-set, respectively (details shown in Fig. 2).

As shown in Table 6, the SurvIAE-based strategy outperformed the R-IPI tool for 15/72 models for both validation-set and test-set. For the PFS60 outcome, MCC from SurvIAE-L models starting from normalized, scaled, and standardized expression datasets were 0.30, 0.28, and 0.22 for the validation-set as well as 0.22, 0.23, and 0.23 for the test-set which were higher than respective R-IPI models. For the same outcome, SurvIAE-M models from scaled and standardized expression datasets outperformed R-IPI, obtaining an MCC of 0.31 and 0.23 for validation-set and 0.34 and 0.21 for the test-set. Interestingly, for this outcome, models with SurvIAE-S outperforming R-IPI resulted only for normalized and transformed expression datasets, whereas SurvIAE-X models outperforming R-IPI were only for the normalized and scaled expression datasets.

For the PFS36 outcome, MCC from SurvIAE-L models from standardized expression datasets was 0.23 for both the validation-set and the test-set outperforming respective R-IPI model with 0.18 for the validation-set and 0.21 for the test-set. For the same outcome, MCC from the SurvIAE-S vs. R-IPI models from scaled and standardized expression datasets were 0.26 and 0.41 vs. 0.18 for the validation-set, and 0.22 and 0.24 vs. 0.21 for the test-set. Finally, MCC from SurvIAE-X models for normalized, scaled, and standardized datasets outperformed R-IPI for both validation and test-sets. We excluded OS-based models for further steps, as no SurvIAE model outperformed R-IPI. For brevity, in the following, we will refer to model size and outcome as SurvIAE-{size}-{outcome}.

4.3. Gene signatures from XAI and clinical translation from the best model

From each of the fifteen models, as described in Section 4.2, we verified, from a predictive point of view, each signature including top-ten genes after the XAI application as shown in Table 7. Since

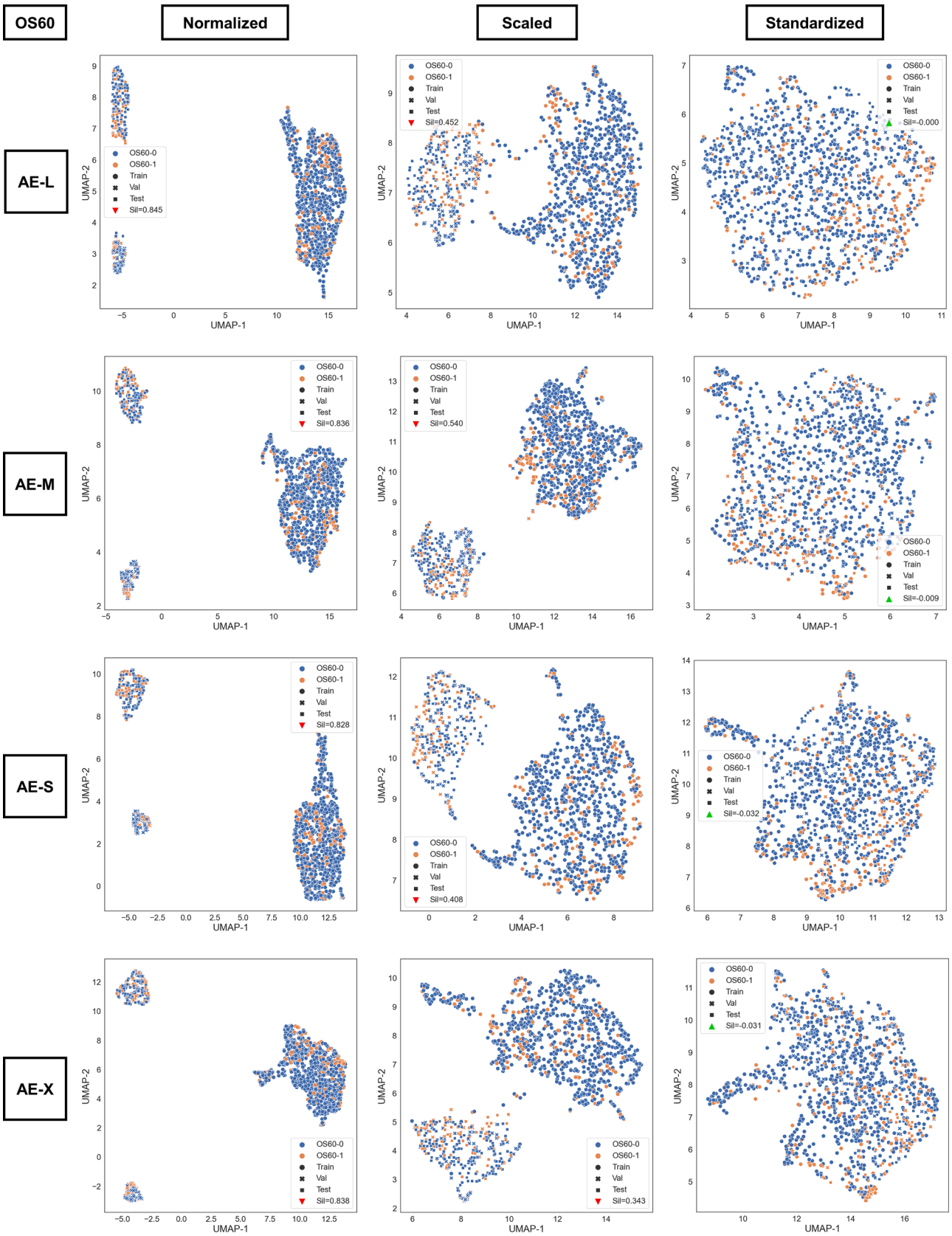

Fig. 3. Latent representation of different AE models with 2D UMAP scatter plots for the OS60 outcome. The embedding representations are portrayed for the three different preprocessing. Batch effect among the different datasets is particularly observable in Normalized and Scaled data. On Standardized data, the batch effect is negligible. Abbreviations. OS, Overall Survival; AE, autoencoder; L, Large; M, Medium; S, Small; X, extra-small; UMAP, Uniform Manifold Approximation and Projection; Train, training-set; Val, validation-set; Sil, Silhouette.

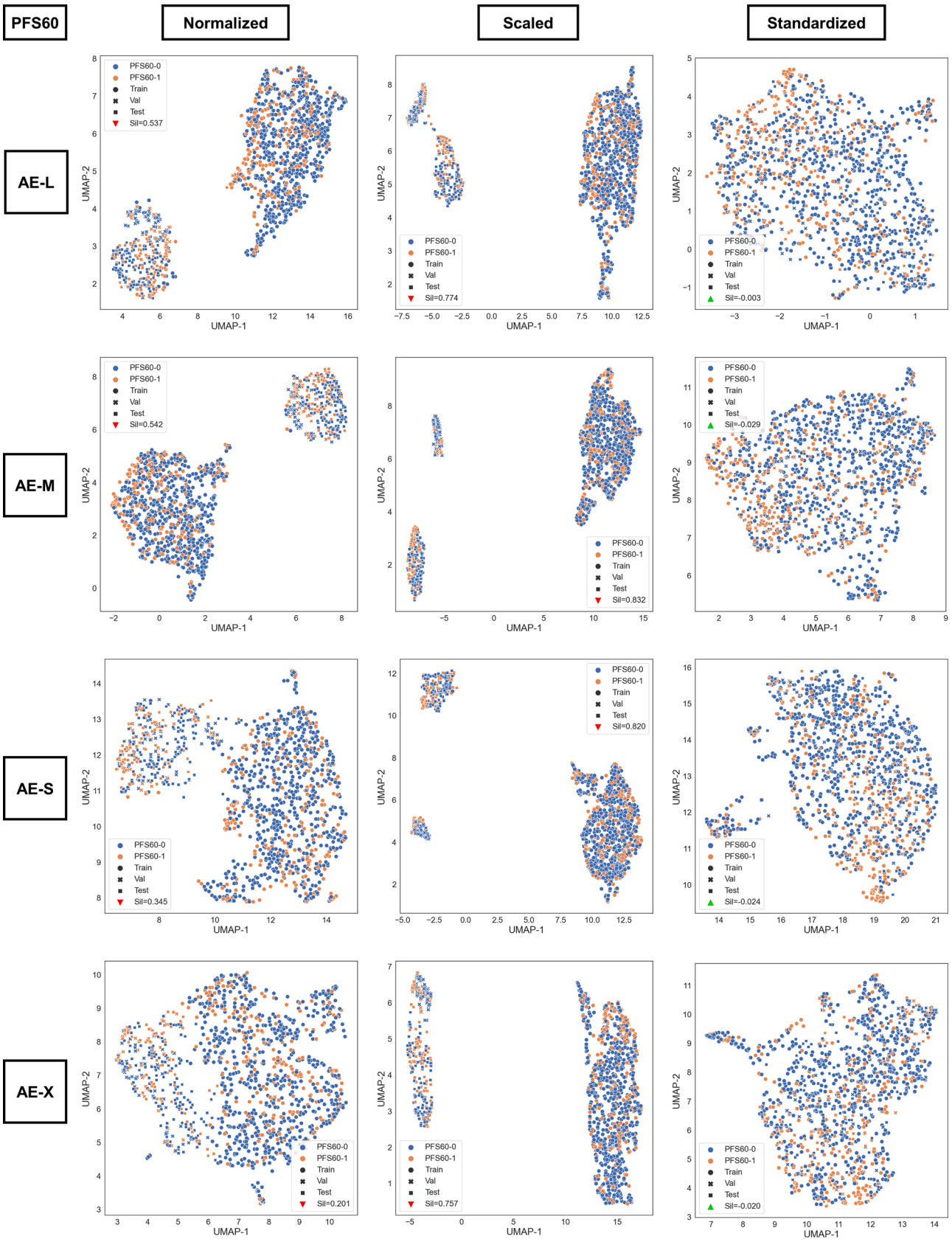

Fig. 4. Latent representation of different AE models with 2D UMAP scatter plots for the PFS60 outcome. The embedding representations are portrayed for the three different preprocessing. Batch effect among the different datasets is particularly observable in Normalized and Scaled data. On Standardized data, the batch effect is negligible. Abbreviations. OS, Overall Survival; AE, autoencoder; L, Large; M, Medium; S, Small; X, extra-small; UMAP, Uniform Manifold Approximation and Projection; Train, training-set; Val, validation-set; Sil, Silhouette.

Table 3
Performance in terms of AUROC and AUPRC of the different SurvIAE models on the validation-set. Five reference ensemble methods are also considered for the comparison: GradientBoosting, AdaBoost, RandomForest, ExtraTrees, XGB. The highest value of AUROC or AUPRC for each column is in bold font. Abbreviations. L, large; M, medium; S, small; X; extra-small; OS, Overall Survival; PFS, Progression Free Survival.

Validation-set													
Model	Preprocessing	OS12		OS36		OS60		PFS12		PFS36		PFS60	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
SurvIAE-L	**Normalized**	0.698	0.252	0.650	0.502	0.660	0.509	0.711	0.491	0.577	0.367	0.674	0.608
	Scaled	0.671	0.299	0.740	0.571	**0.824**	**0.688**	0.569	0.446	0.685	0.617	0.668	0.540
	Standardized	0.588	0.149	0.664	0.490	0.724	0.538	0.675	0.477	0.639	0.507	0.634	0.516
SurvIAE-M	**Normalized**	0.490	0.146	0.700	0.421	0.716	0.488	0.556	0.394	0.692	0.482	0.642	0.532
	Scaled	**0.719**	0.215	0.640	0.446	0.657	0.482	0.712	**0.546**	0.667	0.588	0.635	0.513
	Standardized	0.541	0.131	0.742	0.511	0.687	0.592	0.640	0.410	0.702	**0.636**	0.636	0.535
SurvIAE-S	**Normalized**	0.625	0.215	0.577	0.318	0.665	0.506	0.627	0.313	0.541	0.393	0.387	0.312
	Scaled	0.645	0.243	0.682	0.476	0.714	0.501	0.647	0.432	0.637	0.445	0.537	0.473
	Standardized	0.614	**0.322**	0.582	0.498	0.678	0.542	0.638	0.366	0.685	0.587	0.623	0.489
SurvIAE-X	**Normalized**	0.657	0.182	**0.754**	0.560	0.731	0.559	0.568	0.249	0.683	0.570	0.653	0.585
	Scaled	0.613	0.179	0.680	0.467	0.711	0.558	0.702	0.399	0.703	0.585	0.640	0.532
	Standardized	0.512	0.188	0.610	0.373	0.666	0.490	0.582	0.358	**0.717**	0.523	0.646	0.600
GradientBoosting	**Normalized**	0.696	0.195	0.548	0.271	0.507	0.290	0.500	0.204	0.421	0.281	0.641	0.481
	Scaled	0.480	0.111	0.632	0.319	0.704	0.477	0.730	0.348	0.648	0.483	**0.732**	**0.614**
	Standardized	0.627	0.205	0.506	0.246	0.598	0.337	0.576	0.253	0.616	0.410	0.552	0.442
AdaBoost	**Normalized**	0.512	0.122	0.488	0.248	0.528	0.306	0.673	0.324	0.564	0.372	0.497	0.347
	Scaled	0.475	0.112	0.612	0.471	0.733	0.550	0.722	0.428	0.613	0.488	0.694	0.593
	Standardized	0.582	0.198	0.506	0.347	0.647	0.448	0.610	0.295	0.605	0.406	0.683	0.483
RandomForest	**Normalized**	0.486	0.108	0.578	0.281	0.545	0.315	0.514	0.213	0.485	0.308	0.498	0.351
	Scaled	0.546	0.142	0.621	0.309	0.511	0.319	0.521	0.208	0.458	0.306	0.556	0.381
	Standardized	0.664	0.182	0.607	0.364	0.606	0.338	0.608	0.333	0.603	0.371	0.614	0.462
ExtraTrees	**Normalized**	0.451	0.100	0.458	0.235	0.486	0.287	0.425	0.198	0.465	0.294	0.432	0.319
	Scaled	0.594	0.208	0.643	0.357	0.591	0.361	0.538	0.264	0.614	0.383	0.633	0.449
	Standardized	0.608	0.179	0.615	0.306	0.621	0.392	0.524	0.239	0.556	0.329	0.563	0.399
XGB	**Normalized**	0.545	0.132	0.682	**0.574**	0.500	0.300	0.292	0.153	0.473	0.418	0.573	0.486
	Scaled	0.540	0.126	0.746	0.519	0.751	0.603	**0.751**	0.450	0.563	0.395	0.591	0.481
	Standardized	0.694	0.226	0.683	0.464	0.678	0.534	0.651	0.420	0.676	0.531	0.619	0.477

Table 4
Performance in terms of AUROC and AUPRC of the different SurvIAE models on the test-set. Five reference ensemble methods are also considered for the comparison: GradientBoosting, AdaBoost, RandomForest, ExtraTrees, XGB. The highest value of AUROC or AUPRC for each column is in bold font. Abbreviations. L, large; M, medium; S, small; X; extra-small; OS, Overall Survival; PFS, Progression Free Survival.

Test-set													
Model	Preprocessing	OS12		OS36		OS60		PFS12		PFS36		PFS60	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
SurvIAE-L	**Normalized**	0.567	0.284	0.603	0.340	0.652	**0.518**	0.470	0.273	0.584	0.448	0.620	0.542
	Scaled	0.585	0.235	0.568	0.367	0.616	0.484	0.556	0.333	0.575	0.455	0.611	0.497
	Standardized	0.588	0.274	0.576	0.366	0.647	0.473	0.559	0.344	0.625	0.473	0.631	0.534
SurvIAE-M	**Normalized**	0.587	0.266	0.560	0.365	0.653	0.492	0.454	0.260	0.587	0.453	0.532	0.410
	Scaled	0.665	0.303	0.626	0.458	0.624	0.439	0.572	0.357	0.635	0.497	0.647	0.518
	Standardized	**0.680**	0.306	0.612	0.397	0.621	0.446	0.621	**0.396**	0.629	0.510	0.619	0.522
SurvIAE-S	**Normalized**	0.541	0.183	0.537	0.326	0.548	0.354	0.567	0.356	0.589	0.465	0.388	0.347
	Scaled	0.579	0.281	0.647	0.431	0.627	0.452	0.585	0.363	0.614	0.497	0.608	0.526
	Standardized	0.601	0.226	0.488	0.311	0.618	0.458	0.573	0.366	0.625	0.480	0.649	0.572
SurvIAE-X	**Normalized**	0.597	**0.309**	0.606	0.393	0.601	0.411	0.583	0.329	0.623	0.468	0.654	0.513
	Scaled	0.548	0.280	**0.661**	**0.475**	**0.655**	0.516	**0.627**	0.381	0.634	0.490	**0.679**	**0.585**
	Standardized	0.543	0.255	0.540	0.365	0.570	0.447	0.576	0.339	0.601	0.471	0.611	0.522
GradientBoosting	**Normalized**	0.434	0.143	0.543	0.318	0.571	0.390	0.547	0.304	0.654	0.496	0.577	0.470
	Scaled	0.507	0.165	0.544	0.335	0.556	0.376	0.619	0.355	0.622	0.470	0.601	0.513
	Standardized	0.481	0.158	0.571	0.428	0.517	0.340	0.531	0.293	0.599	0.460	0.575	0.464
AdaBoost	**Normalized**	0.460	0.206	0.589	0.409	0.592	0.455	0.609	0.372	**0.674**	**0.583**	0.618	0.512
	Scaled	0.512	0.222	0.608	0.371	0.624	0.468	0.608	**0.396**	0.566	0.431	0.594	0.470
	Standardized	0.599	0.252	0.569	0.386	0.605	0.417	0.562	0.352	0.620	0.496	0.630	0.553
RandomForest	**Normalized**	0.521	0.177	0.542	0.315	0.469	0.321	0.544	0.292	0.425	0.346	0.567	0.472
	Scaled	0.663	0.249	0.507	0.289	0.544	0.386	0.510	0.279	0.519	0.413	0.640	0.507
	Standardized	0.580	0.260	0.508	0.321	0.543	0.372	0.609	0.363	0.597	0.443	0.564	0.466
ExtraTrees	**Normalized**	0.516	0.172	0.457	0.306	0.452	0.318	0.593	0.337	0.547	0.413	0.436	0.371
	Scaled	0.551	0.186	0.597	0.347	0.560	0.370	0.572	0.339	0.597	0.457	0.502	0.413
	Standardized	0.598	0.237	0.602	0.373	0.602	0.403	0.584	0.364	0.590	0.473	0.600	0.494
XGB	**Normalized**	0.586	0.227	0.594	0.371	0.602	0.476	0.509	0.273	0.562	0.465	0.606	0.515
	Scaled	0.545	0.237	0.600	0.403	0.621	0.457	0.597	0.362	0.648	0.510	0.664	0.562
	Standardized	0.593	0.281	0.577	0.378	0.573	0.408	0.602	0.389	0.625	0.487	0.659	0.562

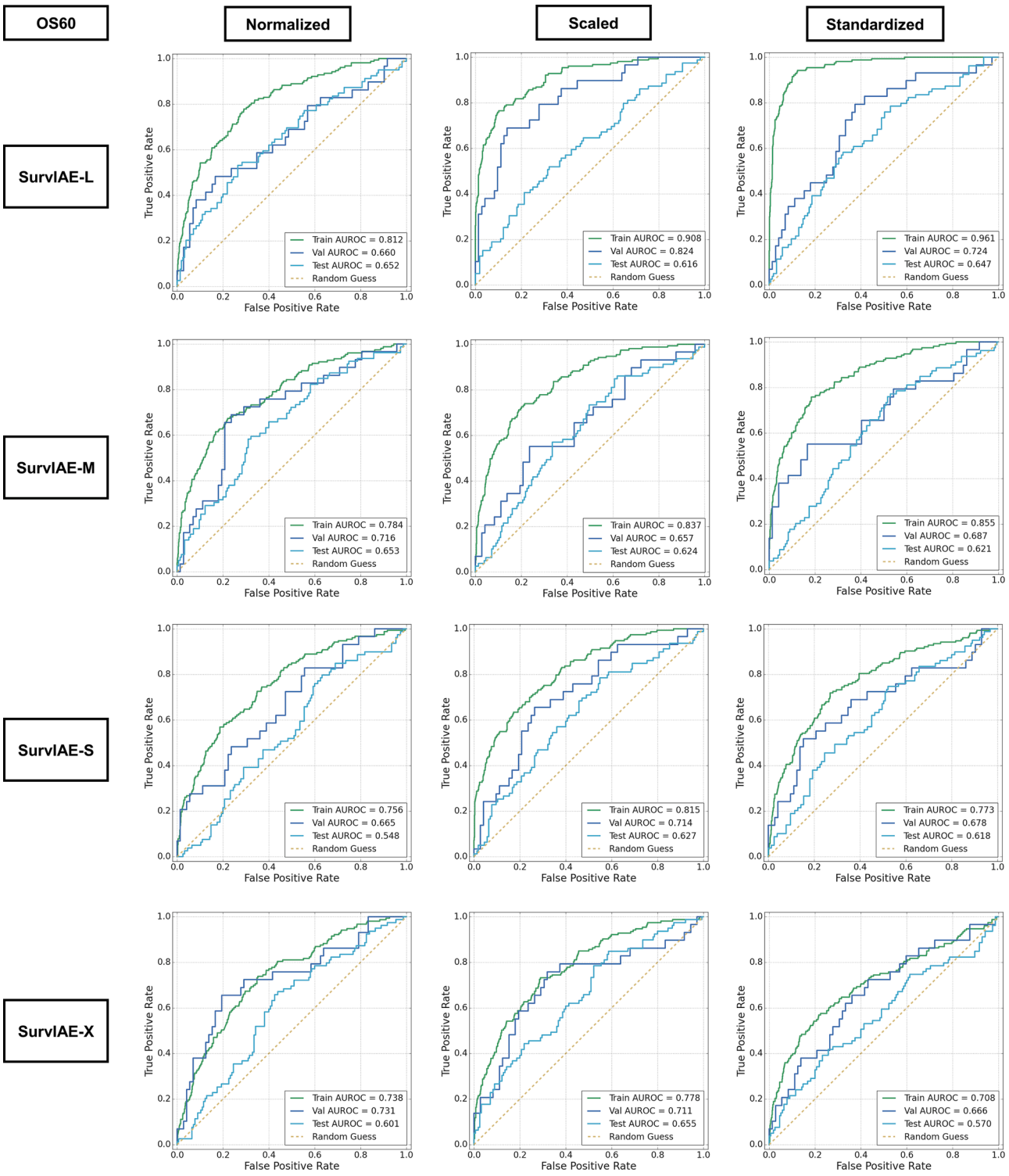

Fig. 5. ROC curves for the OS60 outcome. The curves are depicted for the three different preprocessing, the four SurvIAE architectures considered, and the three datasets. Abbreviations. OS, Overall Survival; L, Large; M, Medium; S, Small; X, extra-small; UMAP, Uniform Manifold Approximation and Projection; Train, training-set; Val, validation-set.

thresholding of GE levels failed in at least one among validation-set and test-set, we excluded gene signatures obtained from models that were trained from the normalized GE dataset. Among models trained from the scaled GE dataset, signatures including at least one gene with prognostic capability were those obtained from models SurvIAE-L-PFS60 (gene *SLC1A1*) and SurvIAE-M-PFS60 (*TMEM163*). Among models trained from the standardized GE dataset, those deriving genes with prognostic capability were those obtained from the SurvIAE-L-PFS36 (*GAB1*), the SurvIAE-S-PFS36 (signature including *CDC42EP4, GAB1*, and *GPR132*), the SurvIAE-S-PFS60 (signature including *DENDD3, A4GALT*, and *SERPINE1*), and the SurvIAE-X-PFS36 (*HES4*).

The SurvIAE-S-PFS36 was then identified as the best model, considering the metric measured from every LR SurvIAE model vs. R-IPI. Firstly, we applied a multivariate LR analysis across training, validation-

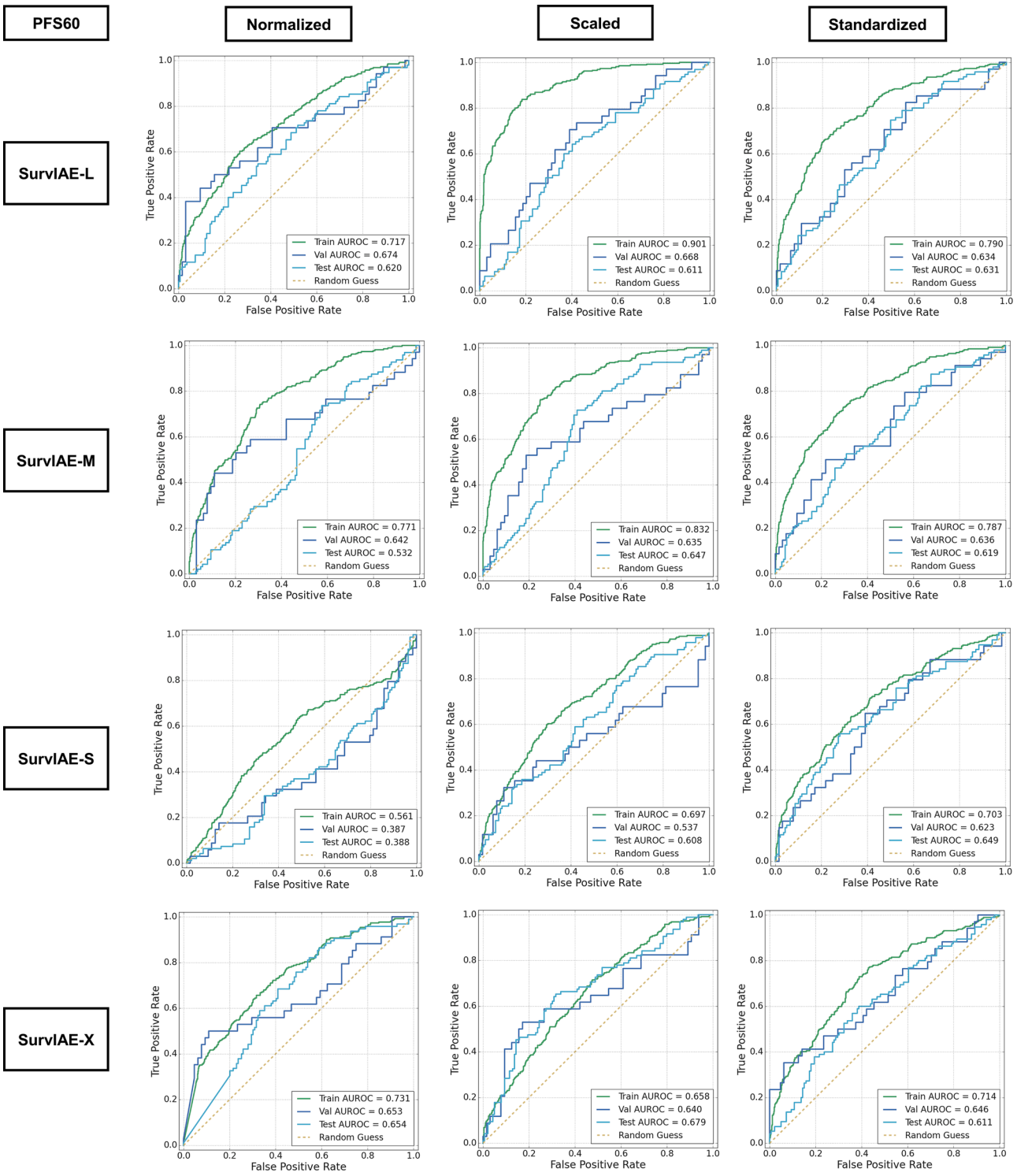

Fig. 6. ROC curves for the PFS60 outcome. The curves are depicted for the three different preprocessing, the four SurvIAE architectures considered, and the three datasets. Abbreviations. PFS, Progression Free Survival; L, Large; M, Medium; S, Small; X, extra-small; UMAP, Uniform Manifold Approximation and Projection; Train, training-set; Val, validation-set.

set, and test-set, including as predictors *GPR132, GAB1, CDC42EP4*, COO, and R-IPI determinants. Thus, the R-IPI retained a significant impact in predicting PFS36 (Fig. 7A) across datasets whether all genes were significant for the training-set. In fact, patients with "high" levels of *CDC42EP4*, patients with "low" levels of *GAB1*, and patients with "high" levels of *GPR132* had 0.64 (p-value < 0.01), 0.67 (p-value < 0.05), and 0.63 (p-value < 0.01) probability to occur a PFS36 event with respect to other patients. Patients with "high" levels of *CDC42EP4* had odds-ratio of 0.26 (p-value < 0.01) and 0.51 (p-value < 0.05) also for validation and test sets, respectively. Interestingly, patients with "low" levels of GAB1, retained their prognostic impact also for the application-set with an odds-ratio of 0.34 (p-value < 0.001). According to COO, we observed significance levels only for the test-set. In fact, for this dataset, patients classified as GCB vs. ABC+UNC reported a significant (p-value < 0.001) odds-ratio of 0.26, whereas odds-ratios were not significant for training-set, validation-set, and application-set.

Table 5
Odds-ratios and p-values for each univariate logistic regression model for validation-set and test-set. The p-values are derived from pairwise comparisons using z-statistic. $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$. Abbreviations. R-IPI, Revised-International Prognostic Index; Ref, reference; GE, gene expression; OS, Overall Survival; PFS, Progression Free Survival; Val, validation-set; OR, odds-ratio; p, p-value.

Model	GE Data Preprocessing	OS12				OS36				OS60				PFS12				PFS36				PFS60			
		Val		Test		Val		Test		Val		Test		Val		Test		Val		Test		Val		Test	
		OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*	OR	*p*
R-IPI (ref)	-	0.07	*	0.18	***	0.22	**	0.21	***	0.23	**	0.22	***	0.27	*	0.20	***	0.45	*0.072*	0.27	**	0.43	*0.055*	0.31	*
SurvIAE-L	Normalized	0.14	**	0.49	*0.140*	0.27	*	0.43	**	0.28	**	0.30	***	0.12	***	0.74	*0.709*	0.55	*0.182*	0.60	*0.086*	0.27	**	0.40	**
	Scaled	0.11	**	0.57	*0.241*	0.23	**	0.61	*0.120*	0.09	***	0.54	*	0.27	*	0.62	*0.153*	0.14	***	0.57	*0.058*	0.30	**	0.39	***
	Standardized	1.37	*0.776*	0.31	**	0.35	*	0.64	*0.152*	0.18	***	0.38	**	0.37	*0.076*	0.37	**	0.35	*	0.38	***	0.39	*	0.36	***
SurvIAE-M	Normalized	0.25	*0.079*	0.39	*	0.13	***	0.46	*	0.16	***	0.37	***	0.37	*0.076*	0.87	*0.710*	0.18	***	0.56	*0.052*	0.25	**	0.52	*
	Scaled	0.19	*	0.35	*	0.27	*	0.39	**	0.27	**	0.36	***	0.11	***	0.38	**	0.17	***	0.43	**	0.26	**	0.23	***
	Standardized	0.46	*0.353*	0.28	**	0.26	**	0.43	*	0.24	**	0.41	**	0.27	*	0.36	**	0.14	***	0.43	**	0.36	*	0.43	**
SurvIAE-S	Normalized	0.34	*0.163*	0.67	*0.381*	0.39	*0.063*	0.58	*0.096*	0.46	*0.101*	0.45	*	0.32	*	0.52	*0.054*	0.67	*0.380*	0.53	*	0.21	***	0.44	**
	Scaled	0.20	*	0.47	*0.087*	0.21	**	0.37	**	0.19	***	0.37	***	0.33	*	0.50	*	0.31	*	0.39	**	0.46	*0.080*	0.59	*0.074*
	Standardized	0.39	*0.287*	0.39	*	0.19	**	0.90	*0.747*	0.28	**	0.46	*	0.24	**	0.41	**	0.15	***	0.36	***	0.37	*	0.30	***
SurvIAE-X	Normalized	0.20	*	0.29	**	0.21	**	0.61	*0.128*	0.16	***	0.41	**	0.45	*0.132*	0.40	**	0.23	**	0.39	**	0.32	*	0.32	***
	Scaled	0.30	*0.091*	0.57	*0.241*	0.24	**	0.34	***	0.19	***	0.43	**	0.24	**	0.40	**	0.23	**	0.29	***	0.25	**	0.28	***
	Standardized	0.79	*0.834*	0.79	*0.642*	0.35	*	0.77	*0.431*	0.31	*	0.68	*0.220*	0.44	*0.138*	0.39	**	0.20	***	0.29	***	0.44	*0.063*	0.43	**

Thus, for the training-set, after including those five features (R-IPI, COO, *CDC42EP4, GAB1,* and *GPR132*) in a decision-tree model (Fig. 7B), the Rev-IPI was selected as the first feature splitting the entire training cohort into Good/Very Good ($N = 482$) and Poor ($N = 446$) subgroups of patients (p-value < 0.001). The more favorable one was further subdivided according to the level of *GPR132* into "high" ($N = 360$) and "low" ($N = 122$) subsets differing significantly in terms of PFS36 (p-value $= 0.029$). Conversely, the less favorable one was further subdivided according to the level of *GAB1* into "high" ($N = 206$) and "low" ($N = 240$) subsets differing significantly in terms of PFS36 (p-value $=$ 0.002). For clinical applicability, we identified as Group 1 patients classified as Poor R-IPI with "high" levels of *GAB1,* as Group 2 Poor patients with "low" levels of *GAB1* or those classified as Good/Very Good R-IPI expressing "low" levels of *GPR132,* and we identified as Group 3 patients classified as Good/Very Good R-IPI expressing "high" levels of *GPR132*. Fig. 7C shows nine univariate K-M curves (PFS) for training, validation, and test datasets according to the expression of each gene included in the final signature. Generally, genes retained their prognostic impact across datasets. Fig. 7D **left** shows that, for the application-set, patients included into Group 1 had a PFS at 36 months of 46 % (95 % confidence interval [CI]: 35 %–60 %) which significantly differed from patients from Group 2, with PFS at 36 months of 73 % (CI: 65 %–82 %) and from patients from Group 3, with PFS at 36 months of 80 % (CI: 73 %–88 %). We also verified the tool assuming OS as clinical outcome and for the application-set, patients included into Group 1 had OS at 36 months of 47 % (CI: 36 %–62 %) which significantly differed from patients from Group 2, with OS at 36 months of 78 % (CI: 71 %–86 %) and from patients from Group 3, with OS at 36 months of 90 % (CI: 85 %–96 %) (Fig. 7D**, center**). Fig. 7D**, right** displays that the proposed score slightly outperformed R-IPI with a C-index of 0.73 vs. 0.70, and 0.70 vs. 0.69, and with a Brier Score of 0.19 vs. 0.20 and 0.16 vs. 0.17 for PFS36 and OS36, respectively. As expected, in all tested cases, COO resulted less accurate in discriminating outcomes than both new score and R-IPI.

5. Discussion

We compared four different AE architectures, three preprocessing methodologies, and six different outcomes to build a deeper understanding on how to use such tools in the analysis of GE data. To corroborate the validity of our findings, we made our analysis exploiting four different datasets, each with its own technology for GE data acquisition. We trained both the AE models and the classifiers on only one of these datasets and showed that these results can be generalized on unseen data acquired in different modalities.

The comparison of SurvIAE with ensemble models showed that the classification performances obtained with the proposed approach are promising. Indeed, on the test-set, SurvIAE outperformed the ensemble models for the outcomes OS12, OS36, OS60, PFS12, and PFS60, leaving only PFS36 to ensembles.

From our analysis, it emerged that some preprocessing techniques are more suitable to tackle GE data belonging to different datasets. Indeed, some AE models suffered from batch effect, as visible from 2D embeddings of their latent representation reported in Figs. 3 and 4, when data were normalized only with the platform-specific normalization procedure. On the other hand, the adoption of standardized data solved this issue. The batch effect hardens the models' training, as can be seen in Fig. 6, where SurvIAE-S displayed particularly bad performance on the validation-set and test-set with simply normalized data. Models trained with standardized data tend to have a better generalization capability.

We proposed a novel generalizable approach to train different AE architectures from omics data to find out the best model for prognostic purposes [29–31]. To make other researchers interested in the analysis of GE data for survival prediction benefit from our efforts, we made our tool, SurvIAE, publicly available (https://github.com/Nicolik/SurvIAE).

Table 6
Performance according to MCC for each univariate LR model for validation-set and test-set. Cells marked with * represent SurvIAE models with MCC higher than the reference in the validation-set, whereas cells labeled with ** indicate the same result in the test-set. Abbreviations. MCC, Matthews Correlation Coefficient; R-IPI, Revised-International Prognostic Index; Ref, reference; GE, gene expression; OS, Overall Survival; PFS, Progression Free Survival; Val, validation-set.

Model	GE Data Preprocessing	OS12		OS36		OS60		PFS12		PFS36		PFS60	
		Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
R-IPI (ref)	-	0.32	0.28	0.31	0.26	0.31	0.25	0.26	0.26	0.18	0.21	0.20	0.19
SurvIAE-L	Normalized	0.29	0.11	0.26	0.19	0.27	0.28**	0.41*	0.03	0.14	0.12	0.30*	0.22**
	Scaled	0.33*	0.08	0.30	0.11	0.51*	0.14	0.25	0.10	0.42*	0.13	0.28*	0.23**
	Standardized	-0.03	0.21	0.22	0.10	0.35*	0.22	0.18	0.21	0.23*	0.23**	0.22*	0.23**
SurvIAE-M	Normalized	0.19	0.15	0.41*	0.17	0.39*	0.23	0.18	0.03	0.38*	0.14	0.32*	0.15
	Scaled	0.25	0.17	0.26	0.21	0.29	0.23	0.43*	0.21	0.38*	0.20	0.31*	0.34**
	Standardized	0.10	0.23	0.28	0.18	0.30	0.20	0.25	0.21	0.42*	0.20	0.23*	0.21**
SurvIAE-S	Normalized	0.15	0.06	0.19	0.12	0.17	0.17	0.23	0.14	0.09	0.15	0.35*	0.20**
	Scaled	0.22	0.12	0.31	0.22	0.36*	0.23	0.21	0.15	0.26*	0.22**	0.18	0.13
	Standardized	0.11	0.17	0.34*	0.02	0.27	0.18	0.27*	0.19	0.41*	0.24**	0.23*	0.28**
SurvIAE-X	Normalized	0.24	0.21	0.32*	0.11	0.39*	0.21	0.15	0.20	0.33*	0.22**	0.26*	0.27**
	Scaled	0.18	0.08	0.29	0.24	0.35*	0.19	0.28*	0.20	0.33*	0.28**	0.32*	0.30**
	Standardized	0.02	0.03	0.22	0.06	0.25	0.09	0.15	0.20	0.36*	0.25**	0.19	0.20**

Table 7
List of the top 10 genes for SurvIAE models overperforming R-IPI. Gene labels with * are those with dichotomized levels of expressions not significant according to the outcome for at least one between validation-set and test-set. Gene labels with ** are those with dichotomized levels of expressions that are significant according to the outcome for both validation-set and test-set. Abbreviations. PFS, Progression Free Survival; L, large; M, medium; S, small; X, extra-small.

Model	GENE 1	GENE 2	GENE 3	GENE 4	GENE 5	GENE 6	GENE 7	GENE 8	GENE 9	GENE 10
Normalized Data SurvIAE-L PFS60	*ACY3*	*KCNIP2*	*CD1D*	*COBLL1*	*CPT1C*	*EFNB1*	*ALDH1A3*	*FOS**	*UPP1*	*NSBP1**
Normalized Data SurvIAE-S PFS60	*VASH1*	*FOSB*	*KCNMB1*	*TMEM119*	*HTRA1*	*FOS**	*KATNAL2*	*A4GALT*	*ACY3*	*PDPN*
Normalized Data SurvIAE-X PFS36	*ACY3*	*PEG10*	*CRABP2*	*KATNAL2*	*PDPN*	*FOSB*	*KIAA1377*	*WASF1*	*CPNE5*	*RBP7*
Normalized Data SurvIAE-X PFS60	*PEG10*	*ACY3*	*CRABP2*	*COL9A2*	*CAND2*	*UPP1*	*PDPN*	*RENBP*	*FGF11*	*KATNAL2*
Scaled Data SurvIAE-L PFS60	*CAND2**	*TRIP10**	*ANKRD13B**	*SLC1A1***	*MED12L**	*LMO2**	*ACY3**	*ALDH1A3**	*PTK7**	*LRRC32**
Scaled Data SurvIAE-M PFS60	*KCNIP2**	*SYTL4**	*TRIP10**	*ACY3**	*ZNF639**	*IL18R1**	*PDE8B**	*TMEM163***	*LOXL2**	*NSBP1**
Scaled Data SurvIAE-S PFS36	*DHRS9**	*RASL11A**	*ACY3**	*GAB1**	*TBC1D2**	*PLEKHG3**	*PARVA**	*ROR2**	*DPYSL3**	*CD24**
Scaled Data SurvIAE-X PFS36	*LMO2**	*CAPN5**	*DHRS9**	*PEG10**	*DPYSL3**	*GAB1**	*KATNAL2**	*RGS12**	*AVPI1**	*ZMYND15**
Scaled Data SurvIAE-X PFS60	*FSCN1**	*ITGA5**	*ACY3**	*AVPI1**	*PEG10**	*NOTCH3**	*CD300LF**	*PLXDC2**	*CLN5**	*EPAS1**
Standardized Data SurvIAE-L PFS36	*GAB1***	*PRDM1**	*KIAA1199**	*ADRA2A**	*TRPM4*	*ZC3H12A**	*ASF1A**	*FBXO6**	*PDK3**	*ARHGAP28**
Standardized Data SurvIAE-L PFS60	*SPHK1**	*KCNIP2**	*CKB**	*MXRA5**	*WASF1**	*SOCS3**	*COBLL1**	*MMAB**	*PBX4**	*HAPLN3**
Standardized Data SurvIAE-M PFS60	*MED12L**	*COBLL1**	*ELL2**	*CD4**	*ADAM28**	*EXTL2**	*SLC41A2**	*TNFRSF9**	*PBX1**	*IDE**
Standardized Data SurvIAE-S PFS36	*RARRES2**	*CDC42EP4***	*RENBP**	*FNDC1**	*SULF1**	*GAB1***	*SELM**	*LMO2**	*DUSP10**	*GPR132***
Standardized Data SurvIAE-S PFS60	*DENND3***	*CTSK**	*LOX**	*COL1A1**	*PAPLN**	*A4GALT***	*C1QTNF6**	*SCARA3**	*SERPINE1***	*ADRA2A**
Standardized Data SurvIAE-X PFS36	*VASN**	*CSPP1**	*AKAP1**	*HES4***	*SLAMF8**	*HSPB8**	*TNFRSF4**	*C5AR1**	*ELL2**	*SSPN**

Furthermore, even if we focused our analysis on the DLBCL, the pipeline can be used for any kind of GE data, and it is not limited to lymphomas.

With respect to previous works, we concentrated on realizing a pipeline that can be useful for prognostic purposes and can unveil relevant gene signatures. Indeed, other studies used AE and MLP to classify breast cancer subtypes [20,30–32] or lung cancer subtypes [29], but few works concentrated on prognosis (e.g., for CRC [19]). With respect to works related to DLBCL, we note that there is a lack of works investigating the role of AE models and their latent representations for prognosis stratification, while most works rely on more traditional techniques such as LASSO and hierarchical clustering [38–40].

Previous research considered all GE data [32] or genes with the highest variability by median absolute deviation [28] as input to the AE. Herein, we concentrated on a different approach, by only considering a subset of the genes, obtained via DEG analysis, to train and validate each AE. Furthermore, if the information carried out by those genes can be likely relevant from a prognostic point of view, conversely, AE-based models including a classifier whose input data are budded from a latent representation cannot be directly applicable in the clinical practice [32]. In our analysis, we systematically compared SurvIAE models with the recognized R-IPI for DLBCL [4]. Hence, with the aim of proposing a translational approach, our pipeline comprises the application of Deep SHAP to interpret the latent information by figuring out the most important genes according to the clinical outcomes. In more detail, after the XAI application, each gene was evaluated by adding SHAP contributions from the validation-set (from microarrays) and the test-set (obtained from RNA-seq) and thereafter reordered on its importance level (Part 3 from Algorithm 1).

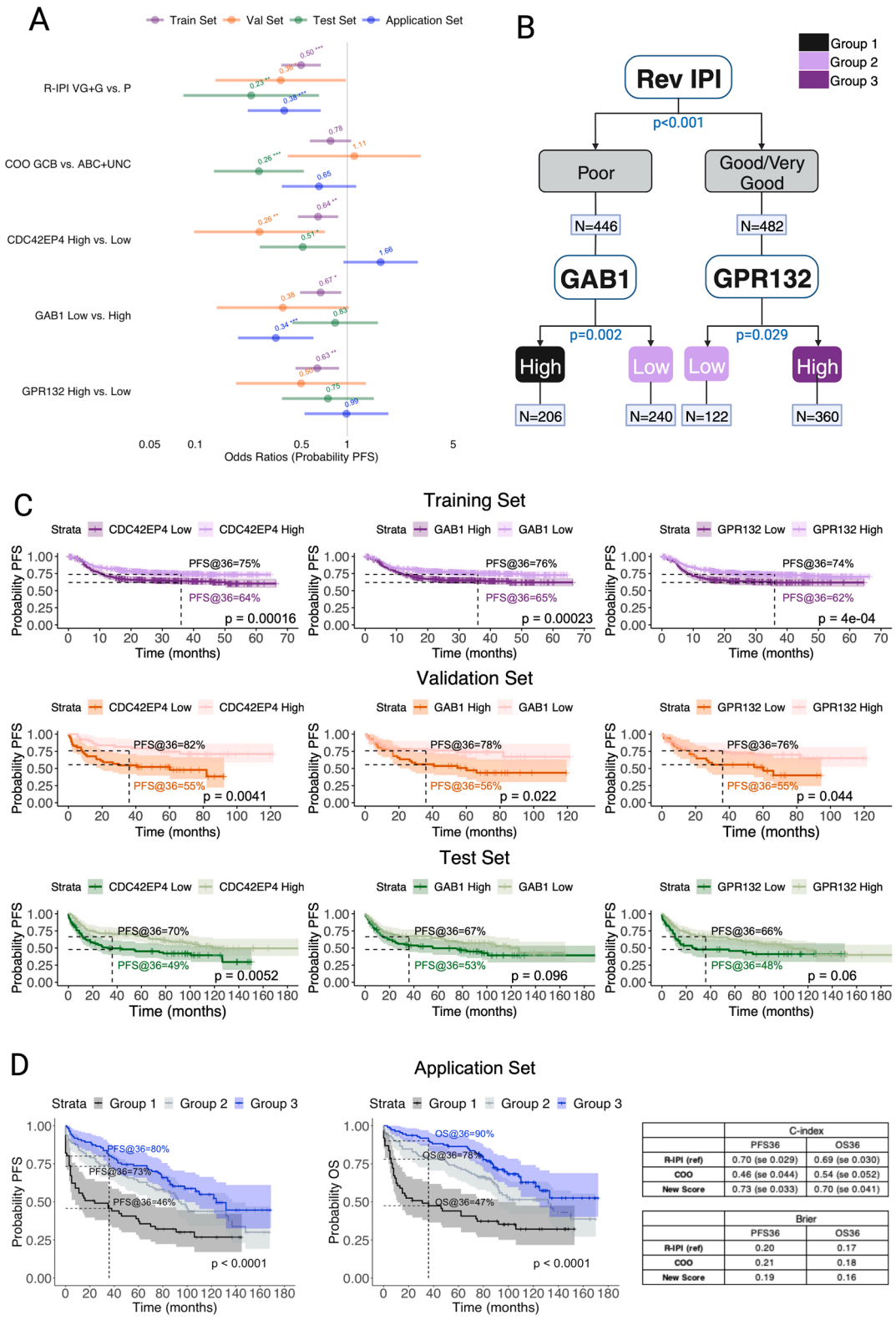

Fig. 7. Translational validation combining SurvIAE with R-IPI. (A) Multivariate logistic regression analysis including clinical determinants and the gene signature retrieved from the SurvIAE-S for the PFS36 outcome. (B) Decision tree depicting results of recursive models applied on clinical and biological features built on PFS36 in the training dataset. The most relevant groups were Group 1 (R-IPI Poor patients with "High" levels of *GAB1*), Group 2 (R-IPI Poor patients with "Low" levels of *GAB1* or R-IPI Good/Very Good patients with "Low" levels of *GPR132*), and Group 3 (R-IPI Good/Very Good patients with "High" levels of *GPR132*). (C) Univariate Kaplan-Meier curves (PFS) for training, validation, and test datasets according to the expression of each gene included in the final signature. (D) Kaplan-Meier survival plots for PFS (left) and OS (center) of Group 1 vs. Group 2 vs. Group 3 for the application-set with survival rates measured at 36 months and performance metrics measured across prognostic models (right). Abbreviations. Train, Training-set; Val, Validation-set; Test, test-set; COO, Cell-of-Origin; ABC, Activated B-Cell like; GCB, Germinal B-cell like; R-IPI, Revised-International Prognostic Index; VG: Very Good; G: Good; P, Poor; PFS, Progression Free Survival; OS, Overall Survival, C, Concordance.

Since the R-IPI was unable to significantly discriminate for the validation set Good/Very Good patients from Poor patients according to PFS at 36 and 60 months (Table 5), this affected the performance assessed by MCC from Table 6. In fact, SurvIAE models were, in most cases, superior independently of AE architectures as well as GE data preprocessing. However, identification of the SurvIAE-S-PFS36 model was also driven by the number of significant genes after the dichotomization of expression levels to "high" and "low" risk classes.

DLBCL is a heterogeneous disease with a very complex biology. Recent works proposed novel prognostic signatures [41,42,44,40,55]. Among these, Wang et al. performed an expression cluster analysis founding two epigenetic-related clusters [56]. DEG analysis allowed to find a subset of prognostic epigenetic-related genes overexpressed in both clusters. In our study, we implemented a pipeline to find out the best deep-learning architecture on GE data. Application of our pipeline on those data might suggest a potential signature to evaluate in terms of immune and therapeutic response. The same research group proposed an lnc-RNA-regulating epigenetic event signature (ELncSig) for predicting prognosis in DLBCL [57]. Interestingly, among DEGs between the high and low-risk ELncSig, we found the *LMO2* gene, which is a marker of longer survival of DLBCL patients following immunotherapy, in 3/15 signatures from Table 7.

SurvIAE-S-PFS36 model identified a signature including *GPR132, GAB1*, and *CDC42EP4* genes. The *GPR132* gene has been demonstrated to have a tumor-suppressive role since it is activated by ONC212, an anti-tumor molecule connected to leukemias [58]. Furthermore, a GE analysis from TCGA revealed that *GPR132* is expressed in a range of tumors with the highest expression also in lymphoma [59]. *GAB1,* which has been identified as a driver gene in DLBCL, can favor cancer progression when highly expressed [60,61]. In our analyses, we first evaluated the prognostic impact of each gene in a multivariate fashion adjusting for R-IPI and COO. Thus, after including those three genes with clinical determinants in a recursive decision tree model from the training-set, not surprisingly, R-IPI was selected as the first feature discriminating patients with Good/Very Good and Poor risks. Thus, for the Good/Very Good patients, our decision-tree model showed the capacity of *GPR132* to recognize "high" risk patients with lower expression, while identifying more favorable cases at higher expression. On the other hand, for Poor patients, our decision-tree model showed the capacity of *GAB1* to recognize "very high" risk patients with higher expression.

5.1. Limitations

We considered MLP as the downstream classifier after the latent representation of the AE was learned, but, of course, other classifiers could have been adopted. We limited our analysis to bulk GE data, but it would be interesting to generalize our workflow to single-cell or spatial GE data.

For the survival analysis, only dichotomized genes were considered, which on the one hand simplifies the clinical translation, but on the other hand, may reduce the accuracy of the prognosis.

6. Conclusions

In this work, we propose an interpretable, end-to-end pipeline, to perform prognosis and derive gene signatures from three datasets of DLBCL patients by means of autoencoders and classifiers. The devised prognostic signature was validated in an independent fourth dataset. We made our tool publicly available to enhance the reproducibility of our efforts. We focused on the analysis of DLBCL, but the pipeline can be easily adapted for other oncologic diseases. Future works will involve considering spatial profiling data, to perform an even more accurate prognosis and further increase the understanding of the biology of the tumor and its microenvironment.

Funding

The study was funded under the research projects:

- National Recovery and Resilience Plan (NRRP), project "BRIEF—Biorobotics Research and Innovation Engineering Facilities", Mission 4: "Istruzione e Ricerca", Component 2: "Dalla ricerca all'impresa", Investment 3.1: "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione", CUP: J13C22000400007, funded by European Union—NextGenerationEU.
- National Recovery and Resilience Plan (NRRP), project "National Centre for HPC, Big Data and Quantum Computing – CN HPC", Mission 4: "Istruzione e Ricerca", Component 2: "Dalla ricerca all'impresa", Investment 1.4: "Potenziamento strutture di ricerca e creazione di 'campioni nazionali di R&S' su alcune Key Enabling Technologies", CUP: D93C22000430001, funded by European Union—NextGenerationEU.
- Italian Ministry of Health - "Ricerca Corrente 2023" – deliberation n.187/2023.

CRediT authorship contribution statement

Gian Maria Zaccaria: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Nicola Altini:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Giuseppe Mezzolla:** Data curation, Formal analysis, Methodology, Writing – review & editing. **Maria Carmela Vegliante:** Data curation, Investigation, Writing – review & editing. **Marianna Stranieri:** Data curation, Writing – review & editing. **Susanna Anita Pappagallo:** Visualization, Writing – review & editing. **Sabino Ciavarella:** Methodology, Supervision, Writing – review & editing. **Attilio Guarini:** Supervision, Writing – review & editing. **Vitoantonio Bevilacqua:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107966.

References

[1] S.H. Swerdlow, E. Campo, S.A. Pileri, N.L. Harris, H. Stein, R. Siebert, R. Advani, M. Ghielmini, G.A. Salles, A.D. Zelenetz, E.S. Jaffe, The 2016 revision of the World Health Organization classification of lymphoid neoplasms, Blood 127 (2016) 2375–2390, https://doi.org/10.1182/blood-2016-01-643569.

[2] B. Coiffier, C. Thieblemont, E. Van Den Neste, G. Lepeu, I. Plantier, S. Castaigne, S. Lefort, G. Marit, M. Macro, C. Sebban, K. Belhadj, D. Bordessoule, C. Fermé, H. Tilly, Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: a study by the Groupe d'Etudes des Lymphomes de l'Adulte, Blood 116 (2010) 2040–2045, https://doi.org/10.1182/blood-2010-03-276246.

[3] A predictive model for aggressive non-Hodgkin's lymphoma, N. Engl. J. Med. 329 (1993) 987–994, https://doi.org/10.1056/NEJM199309303291402.

[4] L.H. Sehn, B. Berry, M. Chhanabhai, C. Fitzgerald, K. Gill, P. Hoskins, R. Klasa, K. J. Savage, T. Shenkier, J. Sutherland, R.D. Gascoyne, J.M. Connors, The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with Diffuse Large B-Cell Lymphoma treated with R-CHOP, Blood 109 (2007) 1857–1861, https://doi.org/10.1182/blood-2006-08-038257.

[5] R.A. Roberts, C.M. Sabalos, M.L. LeBlanc, R.R. Martel, Y.M. Frutiger, J.M. Unger, I. W. Botros, M.P. Rounseville, B.E. Seligmann, T.P. Miller, T.M. Grogan, L. M. Rimsza, Quantitative nuclease protection assay in paraffin-embedded tissue replicates prognostic microarray gene expression in diffuse large-B-cell lymphoma, Lab. Investig. 87 (2007) 979–997, https://doi.org/10.1038/labinvest.3700665.

[6] A. Rosenwald, G. Wright, W.C. Chan, J.M. Connors, E. Campo, R.I. Fisher, R. D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, J.M. Giltnane, E.M. Hurt, H. Zhao, L. Averett, L. Yang, W.H. Wilson, E.S. Jaffe, R. Simon, R.D. Klausner, J. Powell, P.L. Duffey, D.L. Longo, T.C. Greiner, D.D. Weisenburger, W.G. Sanger, B.J. Dave, J.C. Lynch, J. Vose, J.O. Armitage, E. Montserrat, A. López-Guillermo, T. M. Grogan, T.P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, L.M. Staudt, The use of molecular profiling to predict survival after chemotherapy for Diffuse Large-B-Cell Lymphoma, N. Engl. J. Med. 346 (2002) 1937–1947, https://doi.org/10.1056/NEJMoa012914.

[7] R. Schmitz, G.W. Wright, D.W. Huang, C.A. Johnson, J.D. Phelan, J.Q. Wang, S. Roulland, M. Kasbekar, R.M. Young, A.L. Shaffer, D.J. Hodson, W. Xiao, X. Yu, Y. Yang, H. Zhao, W. Xu, X. Liu, B. Zhou, W. Du, W.C. Chan, E.S. Jaffe, R. D. Gascoyne, J.M. Connors, E. Campo, A. Lopez-Guillermo, A. Rosenwald, G. Ott, J. Delabie, L.M. Rimsza, K. Tay Kuang Wei, A.D. Zelenetz, J.P. Leonard, N. L. Bartlett, B. Tran, J. Shetty, Y. Zhao, D.R. Soppet, S. Pittaluga, W.H. Wilson, L. M. Staudt, Genetics and pathogenesis of Diffuse Large B-Cell Lymphoma, N. Engl. J. Med 378 (2018) 1396–1407, https://doi.org/10.1056/NEJMoa1801445.

[8] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D. D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J. C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, Distinct types of Diffuse Large B-Cell Lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511, https://doi.org/10.1038/35000501.

[9] B. Chapuy, C. Stewart, A.J. Dunford, J. Kim, A. Kamburov, R.A. Redd, M. S. Lawrence, M.G.M. Roemer, A.J. Li, M. Ziepert, A.M. Staiger, J.A. Wala, M. D. Ducar, I. Leshchiner, E. Rheinbay, A. Taylor-Weiner, C.A. Coughlin, J.M. Hess, C.S. Pedamallu, D. Livitz, D. Rosebrock, M. Rosenberg, A.A. Tracy, H. Horn, P. Van Hummelen, A.L. Feldman, B.K. Link, A.J. Novak, J.R. Cerhan, T.M. Habermann, R. Siebert, A. Rosenwald, A.R. Thorner, M.L. Meyerson, T.R. Golub, R. Beroukhim, G.G. Wulf, G. Ott, S.J. Rodig, S. Monti, D.S. Neuberg, M. Loeffler, M. Pfreundschuh, L. Trümper, G. Getz, M.A. Shipp, Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes, Nat. Med. 24 (2018) 679–690, https://doi.org/10.1038/s41591-018-0016-8.

[10] S. Ciavarella, M.C. Vegliante, M. Fabbri, A. Enjuanes, S. Rega, A. Gulino, C. Agostinelli, A. Scattone, S. Tommasi, A. Mangia, F. Mele, G. Simone, A.F. Zito, G. Ingravallo, U. Vitolo, A. Chiappella, C. Tarella, A.M. Gianni, A. Rambaldi, P. L. Zinzani, B. Casadei, E. Derenzini, G. Loseto, A. Pileri, V. Tabanelli, S. Fiori, A. Rivas-Delgado, E. Campo, C. Tripodo, A. Guarini, S.A. Pileri, Dissection of DLBCL microenvironment provides a gene expression-based predictor of survival applicable to formalin-fixed paraffin-embedded tissue, Ann. Oncol. 29 (2018) 2363–2370, https://doi.org/10.1093/annonc/mdy450.

[11] C. Sha, S. Barrans, F. Cucco, M.A. Bentley, M.A. Care, T. Cummin, H. Kennedy, J. S. Thompson, R. Uddin, L. Worrillow, R. Chalkley, M. Van Hoppe, S. Ahmed, T. Maishman, J. Caddy, A. Schuh, C. Mamot, C. Burton, R. Tooze, A. Davies, M. Q. Du, P.W.M. Johnson, D.R. Westhead, Molecular high-grade B-cell lymphoma: Defining a poor-risk group that requires different approaches to therapy, J. Clin. Oncol. 37 (2019) 202–212, https://doi.org/10.1200/JCO.18.01314.

[12] I. Walsh, D. Fishman, D. Garcia-Gasulla, T. Titma, G. Pollastri, , ELIXIR Machine Learning Focus Group, E. Capriotti, R. Casadio, S. Capella-Gutierrez, D. Cirillo, A. Del Conte, A.C. Dimopoulos, V.D. Del Angel, J. Dopazo, P. Fariselli, J. M. Fernández, F. Huber, A. Kreshuk, T. Lenaerts, P.L. Martelli, A. Navarro, P.Ó. Broin, J. Piñero, D. Piovesan, M. Reczko, F. Ronzano, V. Satagopam, C. Savojardo, V. Spiwok, M.A. Tangaro, G. Tartari, D. Salgado, A. Valencia, F. Zambelli, J. Harrow, F.E. Psomopoulos, S.C.E. Tosatto, DOME: recommendations for supervised machine learning validation in biology, Nat. Methods 18 (2021) 1122–1127, https://doi.org/10.1038/s41592-021-01205-4.

[13] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (2019) 44–56, https://doi.org/10.1038/s41591-018-0300-7.

[14] N. Altini, M. Rossini, S. Turkevi-Nagy, F. Pesce, P. Pontrelli, B. Prencipe, F. Berloco, S. Seshan, J.B. Gibier, A. Pedraza Dorado, G. Bueno, L. Peruzzi, M. Rossi, A. Eccher, F. Li, A. Koumpis, O. Beyan, J. Barratt, H.Q. Vo, C. Mohan, H.V. Nguyen, P. A. Cicalese, A. Ernst, L. Gesualdo, V. Bevilacqua, J.U. Becker, Performance and limitations of a supervised deep learning approach for the histopathological Oxford Classification of glomeruli with IgA nephropathy, Comput. Methods Progr. Biomed. 242 (2023), 107814, https://doi.org/10.1016/j.cmpb.2023.107814.

[15] N. Altini, T.M. Marvulli, F.A. Zito, M. Caputo, S. Tommasi, A. Azzariti, A. Brunetti, B. Prencipe, E. Mattioli, S. De Summa, V. Bevilacqua, The role of unpaired image-to-image translation for stain color normalization in colorectal cancer histology classification, Comput. Methods Progr. Biomed. 234 (2023), 107511, https://doi.org/10.1016/j.cmpb.2023.107511.

[16] S. Warnat-Herresthal, K. Perrakis, B. Taschler, M. Becker, K. Baßler, M. Beyer, P. Günther, J. Schulte-Schrepping, L. Seep, K. Klee, T. Ulas, T. Haferlach, S. Mukherjee, J.L. Schultze, Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics, iScience 23 (2020), 100780, https://doi.org/10.1016/j.isci.2019.100780.

[17] G.M. Zaccaria, M.C. Vegliante, G. Mezzolla, M. Stranieri, G. Volpe, N. Altini, G. Gargano, S.A. Pappagallo, A. Bucci, F. Esposito, G. Opinto, F. Clemente, A. Negri, P. Mondelli, M.S. De Candia, V. Bevilacqua, A. Guarini, S. Ciavarella, A decision-tree approach to stratify DLBCL risk based on stromal and immune microenvironment determinants, HemaSphere 7 (2023), https://doi.org/10.1097/HS9.0000000000000862 e862.

[18] G.M. Zaccaria, S. Ferrero, E. Hoster, R. Passera, A. Evangelista, E. Genuardi, D. Drandi, M. Ghislieri, D. Barbero, I. Del Giudice, M. Tani, R. Moia, S. Volpetti, M. G. Cabras, N. Di Renzo, F. Merli, D. Vallisa, M. Spina, A. Pascarella, G. Latte, C. Patti, A. Fabbri, A. Guarini, U. Vitolo, O. Hermine, H.C. Kluin-Nelemans, S. Cortelazzo, M. Dreyling, M. Ladetto, A clinical prognostic model Based on machine learning from the fondazione Italiana Linfomi (FIL) MCL0208 Phase III Trial, Cancers 14 (2021) 188, https://doi.org/10.3390/cancers14010188.

[19] H. Song, C. Ruan, Y. Xu, T. Xu, R. Fan, T. Jiang, M. Cao, J. Song, Survival stratification for colorectal cancer via multi-omics integration using an autoencoder-based model, Exp. Biol. Med. 247 (2022) 898–909, https://doi.org/10.1177/15353702211065010.

[20] R. Lupat, R. Perera, S. Loi, J. Li, Moanna: multi-omics autoencoder-based neural network algorithm for predicting breast cancer subtypes, IEEE Access 11 (2023) 10912–10924, https://doi.org/10.1109/ACCESS.2023.3240515.

[21] Y. Hu, Y. Zhao, C.T. Schunk, Y. Ma, T. Derr, X.M. Zhou, ADEPT: autoencoder with differentially expressed genes and imputation for robust spatial transcriptomics clustering, iScience 26 (2023), 106792, https://doi.org/10.1016/j.isci.2023.106792.

[22] K. Dong, S. Zhang, Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder, Nat. Commun. 13 (2022) 1739, https://doi.org/10.1038/s41467-022-29439-6.

[23] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.

[24] S.M. Lauritsen, M. Kristensen, M.V. Olsen, M.S. Larsen, K.M. Lauritsen, M. J. Jørgensen, J. Lange, B. Thiesson, Explainable Artificial Intelligence model to predict acute critical illness from electronic health records, Nat. Commun. 11 (2020) 3852, https://doi.org/10.1038/s41467-020-17431-x.

[25] N. Altini, A. Brunetti, E. Puro, M.G. Taccogna, C. Saponaro, F.A. Zito, S. De Summa, V. Bevilacqua, NDG-CAM: nuclei detection in histopathology images with semantic segmentation networks and Grad-CAM, Bioengineering 9 (2022) 475, https://doi.org/10.3390/bioengineering9090475.

[26] S.M. Hussain, D. Buongiorno, N. Altini, F. Berloco, B. Prencipe, M. Moschetta, V. Bevilacqua, A. Brunetti, Shape-based breast lesion classification using digital tomosynthesis images: the role of Explainable Artificial Intelligence, Appl. Sci. 12 (2022) 6230, https://doi.org/10.3390/app12126230.

[27] S.M. Lundberg, S.I. Lee. A unified approach to interpreting model predictions, Curran Associates Inc., Red, Hook, NY, USA, 2017, pp. 4768–4777, https://doi.org/10.5555/3295222.3295230.

[28] G.P. Way, C.S. Greene, Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. 23rd Pacific Symposium on Biocomputing, PSB 2018, Kohala Coast (2018), https://doi.org/10.1142/9789813235533_0008.

[29] K. Dwivedi, A. Rajpal, S. Rajpal, M. Agarwal, V. Kumar, N. Kumar, An explainable AI-driven biomarker discovery framework for non-small cell lung cancer classification, Comput. Biol. Med. 153 (2023), 106544, https://doi.org/10.1016/j.compbiomed.2023.106544.

[30] S. Rajpal, A. Rajpal, M. Agarwal, V. Kumar, A. Abraham, D. Khanna, N. Kumar, XAI-CNVMarker: explainable AI-based copy number variant biomarker discovery for breast cancer subtypes, Biomed. Signal Process. Control 84 (2023), 104979, https://doi.org/10.1016/j.bspc.2023.104979.

[31] S. Rajpal, A. Rajpal, A. Saggar, A.K. Vaid, V. Kumar, M. Agarwal, N. Kumar, XAI-MethylMarker: explainable AI approach for biomarker discovery for breast cancer subtype classification using methylation data, Expert Syst. Appl. 225 (2023), 120130, https://doi.org/10.1016/j.eswa.2023.120130.

[32] S. Wang, D. Lee, Identifying prognostic subgroups of luminal-A breast cancer using deep autoencoders and gene expressions, PLOS Comput. Biol. 19 (2023), https://doi.org/10.1371/journal.pcbi.1011197 e1011197.

[33] C. Ladbury, R. Zarinshenas, H. Semwal, A. Tam, N. Vaidehi, A.S. Rodin, A. Liu, , R. Salgia, A. Amini, Utilization of model-agnostic Explainable Artificial Intelligence frameworks in oncology: a narrative review, Transl. Cancer Res. 11 (2022), https://doi.org/10.21037/tcr-22-1626.

[34] W.H. Abir, Md.F. Uddin, F.R. Khanam, T. Tazin, M.M. Khan, M. Masud, S. Aljahdali, Explainable AI in diagnosing and anticipating leukemia using transfer learning method, Comput. Intell. Neurosci. 2022 (2022) 1–14, https://doi.org/10.1155/2022/5140148.

[35] M. Gimeno, E. San José-Enériz, S. Villar, X. Agirre, F. Prosper, A. Rubio, F. Carazo, Explainable Artificial Intelligence for precision medicine in acute myeloid leukemia, Front. Immunol. 13 (2022). https://www.frontiersin.org/articles/10.3389/fimmu.2022.977358. accessed April 12, 2023.

[36] J. Hoffmann, M.C. Thrun, M.A. Röhnert, M. von Bonin, U. Oelschlägel, A. Neubauer, A. Ultsch, C. Brendel, Identification of critical hemodilution by artificial intelligence in bone marrow assessed for minimal residual disease analysis in acute myeloid leukemia: the Cinderella method, Cytom. A 103 (2023) 304–312, https://doi.org/10.1002/cyto.a.24686.

[37] A. Meiseles, D. Paley, M. Ziv, Y. Hadid, L. Rokach, T. Tadmor, Explainable machine learning for chronic lymphocytic leukemia treatment prediction using only inexpensive tests, Comput. Biol. Med. 145 (2022), 105490, https://doi.org/10.1016/j.compbiomed.2022.105490.

[38] H. Wang, R. Shao, W. Liu, H. Tang, Y. Lu, Identification of a prognostic metabolic gene signature in Diffuse Large B-Cell Lymphoma, J. Cell. Mol. Med. 25 (2021) 7066–7077, https://doi.org/10.1111/jcmm.16720.

[39] D. Xiong, M. Li, C. Zeng, Construction and validation of a risk scoring model for Diffuse Large B-Cell Lymphoma based on ferroptosis-related genes and its association with immune infiltration, Transl. Oncol. 16 (2022), 101314, https://doi.org/10.1016/j.tranon.2021.101314.

[40] Y. Jiang, H. Sun, H. Xu, X. Hu, W. Wu, Y. Lv, J. Wang, S. Liu, Y. Zhai, L. Tian, Y. Wang, Z. Zhao, Immunophenotypic landscape and prognosis-related mRNA

signature in Diffuse Large B Cell Lymphoma, Front. Genet. 13 (2022), 872001, https://doi.org/10.3389/fgene.2022.872001.
[41] M. Li, F. Huang, Z. Xie, H. Hong, Q. Xu, Z. Peng, Identification of three small nucleolar RNAs (snoRNAs) as potential prognostic markers in Diffuse Large B-Cell Lymphoma, Cancer Med. 12 (2023) 3812–3829, https://doi.org/10.1002/cam4.5115.
[42] Y. Xie, X. Luo, H. He, T. Pan, Y. He, Identification of an individualized RNA binding protein-based prognostic signature for Diffuse Large B-Cell Lymphoma, Cancer Med. 10 (2021) 2703–2713, https://doi.org/10.1002/cam4.3859.
[43] R. Zhang, P. Lin, X. Yang, R.Q. He, H.Y. Wu, Y.W. Dang, Y.Y. Gu, Z.G. Peng, Z. B. Feng, G. Chen, Survival associated alternative splicing events in Diffuse Large B-Cell Lymphoma, Am. J. Transl. Res. 10 (2018) 2636–2647. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6129525/.
[44] H. Zhou, C. Zheng, D.S. Huang, A prognostic gene model of immune cell infiltration in Diffuse Large B-Cell Lymphoma, PeerJ 8 (2020), https://doi.org/10.7717/peerj.9658 e9658.
[45] S.E. Lacy, S.L. Barrans, P.A. Beer, D. Painter, A.G. Smith, E. Roman, S.L. Cooke, C. Ruiz, P. Glover, S.J.L. Van Hoppe, N. Webster, P.J. Campbell, R.M. Tooze, R. Patmore, C. Burton, S. Crouch, D.J. Hodson, Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a haematological malignancy research network report, Blood 135 (2020) 1759–1771, https://doi.org/10.1182/blood.2019003535.
[46] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, ICLR, 2015. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083951076&partnerID=40&md5=1512fe7d6538ffc6686cf01cba3c3460.
[47] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, ArXiv Prepr. arXiv:1802.03426. (2018).
[48] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. 29 (2001), https://doi.org/10.1214/aos/1013203451.
[49] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139, https://doi.org/10.1006/jcss.1997.1504.
[50] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[51] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.
[52] T. Chen, T. He, xgboost: eXtreme Gradient Boosting. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1–4.
[53] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, Proceedings of Machine Learning Research 70 (2019) 3145–3153.
[54] T. Hothorn, A. Zeileis, partykit: a modular toolkit for recursive partytioning in R, J. Mach. Learn. Res. 16 (1) (2015) 3905–3909.
[55] X. Ye, W. Ren, D. Liu, X. Li, W. Li, X. Wang, F.L. Meng, L.S. Yeap, Y. Hou, S. Zhu, R. Casellas, H. Zhang, K. Wu, Q. Pan-Hammarström, Genome-wide mutational signatures revealed distinct developmental paths for human B cell lymphomas, J. Exp. Med. 218 (2021), https://doi.org/10.1084/jem.20200573 e20200573.
[56] X. Wang, Y. Hong, S. Meng, W. Gong, T. Ren, T. Zhang, X. Liu, L. Li, L. Qiu, Z. Qian, S. Zhou, M. Zhao, Q. Zhai, B. Meng, X. Ren, H. Zhang, X. Wang, A novel immune-related epigenetic signature based on the transcriptome for predicting the prognosis and therapeutic response of patients with Diffuse Large B-Cell Lymphoma, Clin. Immunol. 243 (2022), 109105, https://doi.org/10.1016/j.clim.2022.109105.
[57] X. Wang, Y. Lu, Z. Liu, Y. Zhang, Y. He, C. Sun, L. Li, Q. Zhai, B. Meng, X. Ren, X. Wu, H. Zhang, X. Wang, A 9-LncRNA signature for predicting prognosis and immune response in Diffuse Large B-Cell Lymphoma, Front. Immunol. 13 (2022). https://www.frontiersin.org/articles/10.3389/fimmu.2022.813031.
[58] T. Nii, V.V. Prabhu, V. Ruvolo, N. Madhukar, R. Zhao, H. Mu, L. Heese, Y. Nishida, K. Kojima, M.J. Garnett, U. McDermott, C.H. Benes, N. Charter, S. Deacon, O. Elemento, J.E. Allen, W. Oster, M. Stogniew, J. Ishizawa, M. Andreeff, Imipridone ONC212 activates orphan G protein-coupled receptor GPR132 and integrated stress response in acute myeloid leukemia, Leukemia 33 (2019) 2805–2816, https://doi.org/10.1038/s41375-019-0491-z.
[59] V.V. Prabhu, N. Madhukar, R. Tarapore, M. Garnett, U. McDermott, C. Benes, N. Charter, S. Deacon, W. Oster, M. Andreeff, O. Elemento, M. Stogniew, J. Allen, Potent anti-cancer effects of selective GPR132/G2A agonist imipridone ONC212 in leukemia and lymphoma, Cancer Res. (2017) 1155, https://doi.org/10.1158/1538-7445.AM2017-1155. -1155.
[60] Z. Fan, R. Pei, K. Sha, L. Chen, T. Wang, Y. Lu, Comprehensive characterization of driver genes in diffuse large B cell lymphoma, Oncol. Lett. (2020), https://doi.org/10.3892/ol.2020.11552.
[61] J. Yan, W. Yuan, J. Zhang, L. Li, L. Zhang, X. Zhang, M. Zhang, Identification and validation of a prognostic prediction model in Diffuse Large B-Cell Lymphoma, Front. Endocrinol. 13 (2022), 846357, https://doi.org/10.3389/fendo.2022.846357.