AI

MDPI

Article

A Combination of Multilayer Perceptron, Radial Basis Function Artificial Neural Networks and Machine Learning Image Segmentation for the Dimension Reduction and the Prognosis Assessment of Diffuse Large B-Cell Lymphoma

Joaquim Carreras [1,*], Yara Yukie Kikuti [1], Masashi Miyaoka [1], Shinichiro Hiraiwa [1], Sakura Tomita [1], Haruka Ikoma [1], Yusuke Kondo [1], Atsushi Ito [1], Naoya Nakamura [1] and Rifat Hamoudi [2,3,*]

[1] Department of Pathology, School of Medicine, Tokai University, 143 Shimokasuya, Isehara 259-1193, Japan; ki285273@tsc.u-tokai.ac.jp (Y.Y.K.); mm946645@tsc.u-tokai.ac.jp (M.M.); hiraiwa19@tokai-u.jp (S.H.); hs800759@tsc.u-tokai.ac.jp (S.T.); oh298955@tsc.u-tokai.ac.jp (H.I.); kondou@tokai-u.jp (Y.K.); ito.atsushi.s@tokai.ac.jp (A.I.); naoya@is.icc.u-tokai.ac.jp (N.N.)

[2] Department of Clinical Sciences, College of Medicine, University of Sharjah, P.O. Box 27272 Sharjah, United Arab Emirates

[3] Division of Surgery and Interventional Science, University College London, Gower Street, London WC1E-6BT, UK

* Correspondence: joaquim.carreras@tokai-u.jp (J.C.); rhamoudi@sharjah.ac.ae (R.H.)

Citation: Carreras, J.; Kikuti, Y.Y.; Miyaoka, M.; Hiraiwa, S.; Tomita, S.; Ikoma, H.; Kondo, Y.; Ito, A.; Nakamura, N.; Hamoudi, R. A Combination of Multilayer Perceptron, Radial Basis Function Artificial Neural Networks and Machine Learning Image Segmentation for the Dimension Reduction and the Prognosis Assessment of Diffuse Large B-Cell Lymphoma. *AI* **2021**, *2*, 106–134. https://doi.org/10.3390/ai2010008

Academic Editor: Joaquín Torres-Sospedra

Received: 20 October 2020
Accepted: 22 February 2021
Published: 8 March 2021

Abstract: The prognosis of diffuse large B-cell lymphoma (DLBCL) is heterogeneous. Therefore, we aimed to highlight predictive biomarkers. First, artificial intelligence was applied into a discovery series of gene expression of 414 patients (GSE10846). A dimension reduction algorithm aimed to correlate with the overall survival and other clinicopathological variables; and included a combination of Multilayer Perceptron (MLP) and Radial Basis Function (RBF) artificial neural networks, gene-set enrichment analysis (GSEA), Cox regression and other machine learning and predictive analytics modeling [C5.0 algorithm, logistic regression, Bayesian Network, discriminant analysis, random trees, tree-AS, Chi-squared Automatic Interaction Detection CHAID tree, Quest, classification and regression (C&R) tree and neural net)]. From an initial 54,613 gene-probes, a set of 488 genes and a final set of 16 genes were defined. Secondly, two identified markers of the immune checkpoint, PD-L1 (*CD274*) and IKAROS (*IKZF4*), were validated in an independent series from Tokai University, and the immunohistochemical expression was quantified, using a machine-learning-based Weka segmentation. High PD-L1 associated with poor overall and progression-free survival, non-GCB phenotype, Epstein–Barr virus infection (EBER+), high RGS1 expression and several clinicopathological variables, such as high IPI and absence of clinical response. Conversely, high expression of IKAROS was associated with a good overall and progression-free survival, GCB phenotype and a positive clinical response to treatment. Finally, the set of 16 genes (*PAF1, USP28, SORT1, MAP7D3, FITM2, CENPO, PRCC, ALDH6A1, CSNK2A1, TOR1AIP1, NUP98, UBE2H, UBXN7, SLC44A2, NR2C2AP* and *LETM1*), in combination with *PD-L1*, *IKAROS*, *BCL2*, *MYC*, *CD163* and *TNFAIP8*, predicted the survival outcome of DLBCL with an overall accuracy of 82.1%. In conclusion, building predictive models of DLBCL is a feasible analytical strategy.

Keywords: overall survival; diffuse large B-cell lymphoma; artificial intelligence; Multilayer Perceptron; Radial Basis Function; PD-L1 (CD274); IKAROS (IKZF4)

1. Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most common histologic subtype of non-Hodgkin lymphoma (NHL). DLBCL accounts for approximately 25 percent of adult NHL cases. It is increasingly appreciated that the diagnostic category of "DLBCL" is quite heterogeneous in terms of morphology, genetics and biologic behavior. DLBCL is

curable in approximately half of cases with current therapy, particularly in those who achieve a complete remission with first-line treatment. The molecular pathogenesis of DLBCL is a complex, multistep process that ultimately results in the transformation and expansion of a malignant B-cell clone. This neoplastic B-cell is of germinal or post-germinal B cell-of-origin. Two molecular subtypes are identified according to the gene expression: the "germinal center B-cell-like" (GCB) and the "activated B-cell-like" (ABC), including a third subtype that is "unclassified". Based on the immunohistochemistry of CD10, BCL6 and MUM1 (*IRF4*), the Hans classifier also stratifies the samples into GCB and non-GCB (ABC). Despite that some of the molecular mechanisms have been elucidated, most of the pathogenesis remains unknown [1–5]. Therefore, new approaches in analysis may help to clarify the remaining unknown pathogenic factors.

Deep learning (also known as deep structured learning or differential programming) is part of a broader family of machine learning methods based on artificial neural networks with representation learning; the learning can be supervised, semi-supervised or unsupervised [6–9]. Artificial neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering of the raw input data/information. The patterns that artificial neural networks recognize are numerical, contained in vectors, into which all real-world data (be it images, sound, text or time series) must be translated [10]. In this project, we have used the Trainable Weka Segmentation method to quantify the raw colors of the immunohistochemical protein expression of markers that are present in the DLBCL tissue. The Trainable Weka Segmentation is a plugin and library that combines a collection of machine learning algorithms with a set of selected image features to produce pixel-based segmentations.

Artificial neural networks allow to cluster and to classify, they can group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on [10]. With classification, the predictive analytics can be performed by correlations between present and future events. For example, a correlation between the gene or protein expressions levels of DLBCL samples and future events such as the patients' outcome (alive or dead) or other clinicopathological characteristics including the International Prognostic Index (IPI).

Artificial neural networks are the preferred tool for many predictive data-mining applications, because of their power, flexibility and ease of use [11]. Artificial neural networks used in predictive applications, such as the Multilayer Perceptron (MLP) and the Radial Basis Function (RBF) networks, are supervised in the sense that the model-predicted results can be compared against known values of the target variables [11]. Both MLP and RBF have a structure known as a "feedforward architecture" because the connections in the network flow forward from the input layer to the output layer without any feedback loops. The architecture composition is the following: (1) an input layer that contains the predictors; (2) a hidden layer with unobservable nodes, or units; and (3) the output layer that contains the responses. The value of each hidden unit is some function of the predictors. The exact form of the function depends, in part, upon the network type and in part upon user-controllable specifications. The choice of procedure, MLP or RBF, is influenced by the type of data and the level of complexity to uncover. While the MLP procedure can find more complex relationships, the RBF procedure is generally faster [11].

We have recently described the use of MLP for the analysis of gene expression of DLBCL, using a series of 100 cases [12,13]. This continuation project was characterized by (1) the series of cases expanded up to 414 cases (four times larger), (2) the artificial neural network analysis included the comparison and integration of both MLP and RBF methods, and (3) the aim to predict several clinicopathological characteristics, in addition to the overall survival. Finally, after the integration of the results we validated two of the most relevant markers by immunohistochemistry in another series of 113 cases from Tokai University Hospital. The digital quantification of the validation marker was also performed by using machine learning and the Waikato Environment for Knowledge Analysis (Weka).

2. Materials and Methods

2.1. Study Subjects

The subjects of the study for the artificial neural network of gene expression data were obtained from a well-known and robust series of DLBCL from Caucasian subjects (Table 1). This series belongs to the Lymphoma/Leukemia Molecular Profiling Project (LLMPP) and the patients are from several institutions in Europe and USA. This series is publicly available from the NCBI GEO datasets as GSE10846 and comprises 414 cases [14]. The sample data are from an Affymetrix Human Genome U133 Plus 2.0 Array and the processing used the MAS 5.0 Data Processing Protocol. The data were analyzed with the Microarray Suite version 5.0 (MAS 5.0), using the Affymetrix default analysis settings and global scaling as normalization method. The trimmed mean target intensity of each array was arbitrarily set to 500. The post-normalized data were log-2 scaled. For the MLP and RBF analysis on this discovery set all the 414 cases were selected. The clinicopathological characteristics of the discovery series is shown in the Table 1. In summary, the age ranged from 14 to 92 years old, with a median of 62.5 years, and 226 were men (54.6%). According to the cell-of-origin (COO) molecular classification of DLBCL based on the gene expression [1–5], 44.2% of the cases were of germinal center B-cell subtype (GCB), 40.3% of activated B-cell subtype (ABC) and 15.5% of unclassified. Fifty-six percent of the cases had received rituximab, cyclophosphamide, doxorubicin hydrochloride, vincristine sulfate, and prednisone (RCHOP)-like therapy. As expected in a conventional series of DLBCL, several variables correlated with the overall survival of the patients. Patients that had an unfavorable prognosis associated with several clinical variables including age > 60 years, high LDH, Eastern Cooperative Oncology Group (ECOG) Performance Status $\geq$ 2, clinical stage III/IV, extranodal sites >1, higher National Comprehensive Cancer Network International Prognostic Index (NCCN IPI) score and an activated B-cell (ABC) molecular subtype. The alive/dead ratio of this series was 1.51. The clinicopathological characteristics of this series is in concordance with a standard series of DLBCL.

Table 1. Clinicopathological characteristics of the discovery series (GSE10846).

Variable	no.	%	*p*-Value	Hazard Risk	95.0% CI for HR	
					Lower	Upper
Sex Male	224/414	54.6	0.9	1.021	0.744	1.402
Age > 60	226/414	54.6	2×10^{-6}	2.209	1.59	3.069
LDH ratio > 1	182/351	51.9	5.1×10^{-8}	2.723	1.899	3.905
LDH ratio > 3	32/351	9.1	2.9×10^{-8}	3.673	2.319	5.818
ECOG Performance Status $\geq$ 2	93/389	23.9	3.1×10^{-10}	2.835	2.049	3.921
Clinical stage III or IV	218/406	53.7	2.5×10^{-4}	1.834	1.326	2.537
Extranodal disease site > 1	30/383	7.8	0.014	1.927	1.144	3.246
NCCN IPI						
Low risk	54/321	16.8	5.2×10^{-08}	-	-	-
Low-intermediate risk	152/321	47.4	3.8×10^{-4}	5.221	2.096	13.004
High-intermediate risk	98/321	30.5	4×10^{-6}	8.74	3.493	21.871
High risk	17/321	5.3	6.9×10^{-8}	17.761	6.244	50.521
Cell-of-origin molecular subtype						
Germinal center B-cell (GCB)	183/414	44.2	2.8×10^{-8}	-	-	-
Activated B-cell (ABC)	167/414	40.3	1.1×10^{-8}	2.75	1.944	3.891
Unclassified	64/414	15.5	0.2	1.389	0.84	2.298
Treatment						
RCHOP-like	233/414	56.3	7.8×10^{-5}	0.52	0.376	0.719
CHOP-like	181/414	43.7	-	-	-	-
Overall survival (outcome)						
Dead	165/414	39.9	-	-	-	-
Alive	249/414	60.1	-	-	-	-
Overall survival						
Dead < 1.5 years	115/414	27.8	3.3×10^{-9}	113.448	23.654	544.122
Alive $\geq$ 7 years	40/414	9.7	-	-	-	-

The first three columns of the table show the clinicopathological characteristics of patients of this series GSE10846 of diffuse large B-cell lymphoma (DLBCL), with the frequencies of cases per each variable. The variables include clinical variables such as sex, age, the International Prognostic Index, etc., as well as biological factors, such as the cell-of-origin molecular subtypes. Columns 4 to 7 show the prognostic relevance of the variables, using a univariate Cox regression analysis for overall survival. The data show the statistical *p*-value, the Hazard Risk and the 95% confidence interval (CI) for the Hazard Risk. ECOG, Eastern Cooperative Oncology Group. NCCN IPI, National Comprehensive Cancer Network International Prognostic Index. RCHOP, rituximab, cyclophosphamide, doxorubicin hydrochloride, vincristine sulfate, and prednisone.

The validation set consisted of 113 cases of DLBCL from Tokai University Hospital. This validation set was used for the immunohistochemical quantification of the protein expression of two of the most relevant genes, previously identified in the artificial neural network analysis and later confirmed by Cox analysis. One of the markers was associated to a poor and the other to a good prognosis of the patients. The clinicopathological characteristics of the validation set is shown in the Table 2. In summary, the age ranged from 14 to 97 years old, with a median of 67 years, and 64 were men (55.8%). According to the cell-of-origin classification, using the Hans classifier, 33.6% were GCB and 66.4% non-GCB. The Hans classifier uses three markers that can be tested by immunohistochemistry (CD10, BCL-6 and MUM-1), and the cases can be assigned to either of two groups, using a new nomenclature: GC-group and non-GC group. Ninety-five percent of the cases had received RCHOP or RCHOP-like therapy. Patients that had an unfavorable prognosis associated with age > 60 years, high LDH, high sIL2RA, ECOG Performance status $\geq$ 2, clinical stage III/IV, extranodal sites > 1, higher IPI score, a non-GCB cell-of-origin subtype and positivity for Epstein–Barr virus (EBER+). The alive/dead ratio was 0.82. The clinicopathological characteristics of this series from Tokai University are also in concordance with a standard series of DLBCL.

Table 2. Clinicopathological characteristics of the validation series (Tokai cases).

Variable	no.	%	*p*-Value	Hazard Risk	95.0% CI for HR	
					Lower	Upper
Sex Male	63/113	55.8	0.9	1.045	0.6	1.821
Age > 60	78/112	69.6	0.01	2.553	1.24	5.253
Location						
Nodal (+spleen)	63/113	55.8	0.4	-	-	-
Extranodal						
Waldeyer's ring	11/113	9.7	0.2	0.486	0.147	1.613
Gastrointestinal	10/113	8.8	0.6	0.735	0.223	2.422
Other extranodal	29/113	25.7	0.4	1.326	0.725	2.427
LDH High (>219)	70/112	62.5	1.8×10^{-3}	3.03	1.51	6.083
Seric IL2RA High (>530)	83/106	78.3	1.4×10^{-2}	3.627	1.299	10.125
ECOG Performance Status $\geq$ 2	15/90	16.7	6.2×10^{-4}	3.466	1.701	7.062
Clinical stage III or IV	52/105	49.5	0.01	2.138	1.17	3.905
Extranodal disease site >1	20/86	23.3	5.1×10^{-5}	3.985	2.041	7.78
B symptoms	24/94	25.5	0.2	1.557	0.805	3.011
International Prognostic Index (IPI)						
Low risk (L)	34/96	35.4	$1.9-10^{-2}$	-	-	-
Low-intermediate risk (LI)	29/96	30.2	0.01	3.265	1.392	7.656
High-intermediate risk (HI)	20/96	20.8	0.02	2.99	1.193	7.495
High risk (H)	13/96	13.5	4.9×10^{-3}	4.326	1.558	12.016
Cell-of-origin subtype (Hans)						
GCB	37/110	33.6	-	-	-	-
Non-GCB	73/110	66.4	1.4×10^{-2}	2.318	1.186	4.529
Epstein–Barr virus, EBER+	16/111	14.4	2.5×10^{-2}	2.291	1.11	4.729
Treatment						
RCHOP	79/106	74.5	0.3	-	-	-
RCHOP-like	22/106	20.8	0.1	1.677	0.873	3.219
Others	5/106	4.7	0.5	1.701	0.406	7.134
Response to treatment						
CR	72/101	63.7	-			
PR+PD+SD+NC	29/101	28.7	2.9×10^{-13}	11.467	5.956	22.076
Overall survival (outcome)						
Dead	51/113	45.1	-	-	-	-
Alive	62/113	54.9	-	-	-	-

The first 3 columns show the frequencies of each clinicopathological variables. Columns 4 to 7 show the results of the univariate Cox regression analysis for overall survival.

No cases of primary mediastinal B-cell lymphoma (PMBL) were included in this research.

This study was conducted in compliance with the Helsinki Declaration, and the project was approved by the Institutional Review Board (IRB14R-080).

2.2. Statistical Analysis

Statistical analyses were performed by using R programming language version 3.6.3 (29 February 2020) and RStudio (version 1.3.959) [15], and with IBM SPSS (Statistics version 26 and Modeler version 18.0; IBM, New York, United States), following the manufacturers' instructions. Comparisons of means was performed with independent-samples T-Test or with non-parametric two-independent-samples test (Mann–Whitney U test) when required. The criteria for overall survival was based on the time calculated from the date of diagnosis to the date of death or last follow-up. The survival analysis was performed, using Kaplan–Meier (with Log rank, Breslow and Tarone–Ware tests) and Cox regression, method (Enter), contrast (Indicator) and reference category (First). Multivariate Cox regression was also performed with the backward conditional method. Hazard Ratios/Risks (HRs) were calculated with Cox regression. The odds ratios (ORs) were determined with binary logistic regression. The significance level was set up at a priori with a p-value < 0.05. R programming language software, instructions and methods can be found at http://cran.r-project.org (accessed on 4 March 2021). Instructions for RStudio are found at https://rstudio.com/ and for SPSS at https://www.ibm.com/jp-ja/analytics/spss-statistics-software (accessed on 4 March 2021).

2.3. Artificial Neural Network Analysis of Gene Expression Data

MLP and RBF analysis on the discovery series was performed, using similar strategy as previously described [12,13]. The desktop workstation had a Ryzen 7 3700X CPU and 16 GB of RAM.

For MLP analysis, the setup included a series of items. The dependent variable (i.e., the nominal variable that we want to predict) was the outcome of the overall survival as well as several other clinicopathological features including the cell-of-origin classification, the National Comprehensive Cancer Network International Prognostic Index (NCCN-IPI), stage, extranodal disease, etc. (Figure 1). The predictor variables were the genes that were specified as covariates (scale). The rescaling of the covariates was standardized. Partition dataset: The cases were randomly assigned based on the relative number of cases.

The partitions were the training (relative number = 7, 70%), the test (3, 30%), the hold-out (0, 0%). The architecture can be automatically selected, with a minimum number of units in the hidden layer from 1 to 50 or can be a custom architecture. A custom architecture includes a setup for the hidden layers and the output layer. In the hidden-layers section, the following options can be arranged: (1) number of hidden layers (one, two), (2) activation function (hyperbolic tangent, sigmoid) and (3) number of units (automatically compute or custom for hidden layer 1 and 2). In the output layer section, the options are: (1) activation function (identity, softmax, hyperbolic tangent and sigmoid), (2) rescaling of scale dependent variables (standardized, normalized (correction 0.02 or another value), adjusted normalized (0.002 or another value) or none). Of note, the activation function chosen for the output layer determines which rescaling methods are available. The training can be of batch, online or mini-batch type. The optimization algorithm can be scaled conjugant gradient or gradient descent. The training options were initial lambda (0.0000005), initial sigma (0.00005), interval center (0) and interval offset (+/-0.5).

The output includes a network structure (description, diagram and synaptic weights) and the network performance (model summary, classification results, ROC curve, cumulative gains chart, lift chart, predicted by observed chart and residual by predicted chart). In addition, the cases processing summary and the independent variable importance analysis is also performed. The synaptic weight estimates were exported to a XML file. As options, the user-missing values were excluded. The stopping rules had a 0.0001 min-

imum relative change in the training error and 0.001 minimum relative change in the training error ratio.

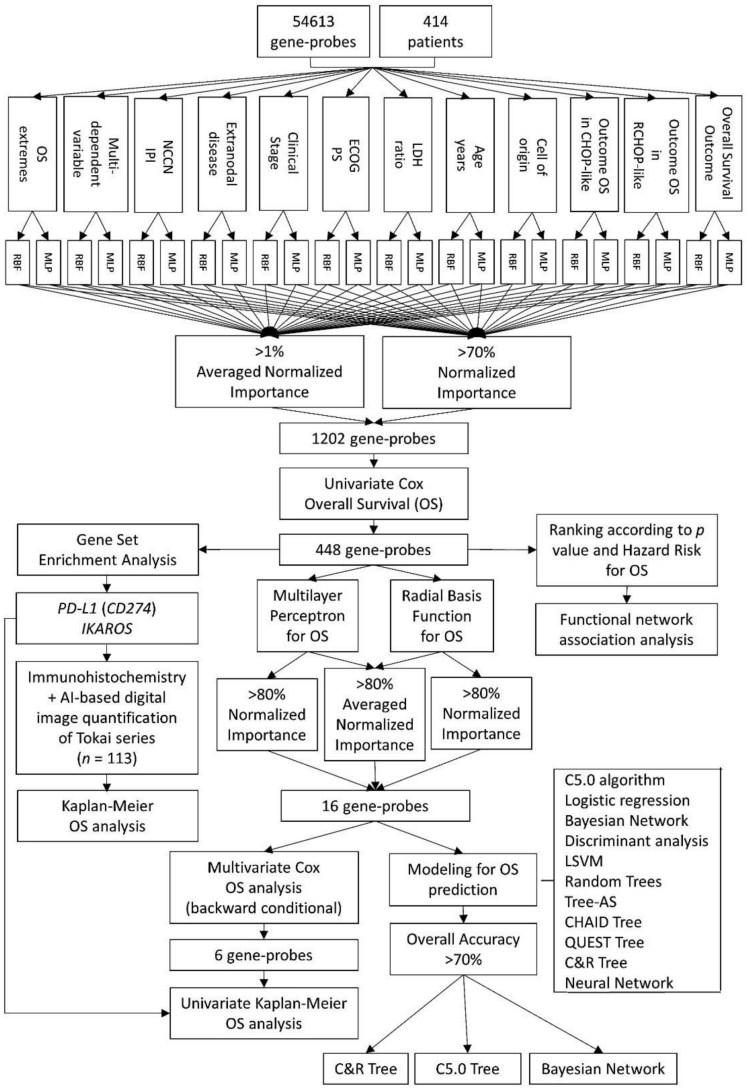

Figure 1. Analysis algorithm. In this project, two types of artificial neural network analyses were performed: Multilayer Perceptron (MLP) and Radial Basis Function (RBF). The input data were the gene expression of 54,613 gene-probes from 414 patients with diffuse large B-cell lymphoma (DLBCL). The target variables were the outcome of the overall survival (dead versus alive), as well as several relevant clinicopathological characteristics, including the cell-of-origin classification and the International Prognostic Index (IPI). The gen-probes were ranked according to their normalized importance (NI). A cutoff of >70% of normalized importance and >1% of averaged normalized importance was applied. Cox regression analysis (univariate and multivariate) reduced the final list to the most relevant genes (n = 448). The gene-set enrichment analysis (GSEA) technique confirmed the association toward bad or good prognosis and *PD-L1* and *IKAROS* were validated in an independent series of Tokai University. Additional data reduction was performed with Cox and Kaplan–Meier overall survival analyses, second round of artificial neural networks and predictive modeling in a multistep process up to a final 16 and 6 genes sets.

For the RBF setup, the architecture can have several units that can be automatically computed or specified. The activation function for the hidden layer can be normalized or ordinary Radial Basis Function and the overlap among hidden units can be automatically computed or specified. Receiver Operating Characteristic (ROC) curve displays a curve

for each categorical dependent variable. It also displays a table giving the area under each curve. For a given dependent variable, the ROC chart displays one curve for each category. Independent variable importance analysis performs a sensitivity analysis, which computes the importance of each predictor in determining the artificial neural network. The analysis is based on the combined training and testing samples or only on the training sample if there is no testing sample. This creates a table and a chart displaying importance and normalized importance for each predictor.

In this study, the analysis method included several individual MLP and RBF analysis (Figure 1). Each individual artificial neural network analysis included the gene-expression data of 414 cases (GSE10846, Affymetrix U133 Plus 2.0 Array, 54,613 gene-probes), which were correlated with a single target (dependent) variable. The target variables were the following: survival outcome (dead or alive), outcome only in the RCHOP-like subgroup, outcome only in the CHOP-like subgroup, cell-of-origin (GCB vs. ABC, the unclassified was excluded), age ($\leq$60 vs. >60), LDH ratio ($\leq$1 vs. >1), LDH ratio ($\leq$3 vs. >3), ECOG PS (<2 vs. $\geq$2), stage (I/II vs. III/IV), extranodal disease ($\leq$1 vs. >1), gender (male vs. female), NCCN-IPI (low + low-intermediate vs. high-intermediate + high), multi-dependent variable and survival extremes (dead < 1.5 years vs. alive > 7 years). In total, 28 AI individual analyses were performed. Each AI analysis provided an output in which the 54,613 gene probes were ranked according to their importance for prediction of the target variable. Then, the normalized importance values were processed as follows: (1) The gene probes with a normalized importance $\geq$70% in each target variable (with exclusion of gender) were selected and merged in a new database. (2) The normalized importance of each gene probe was averaged for all the predictive variables (with exclusion of gender) and the averaged values ranked from most to least important. Therefore, the results comprised 4 lists, top 1% of the averaged normalized importance for MLP and RBF analyses and the merged $\geq$70% normalized importance for MLP and RBF. Then, the gene-probes were merged and the duplicated were deleted.

The relevance of each identified gene-probe was tested in the series, using univariate and multivariate Cox regression analyses [12,13], with a second round of MLP and RBF, gene-set enrichment analysis (GSEA), and finally with an overall survival modeling and screening based on the overall accuracy for prediction of the overall survival outcome variable. This modeling included the following model types: C5.0 algorithm, logistic regression, Bayesian Network, discriminant analysis, linear support vector machine (LSVM), random trees, tree-AS, Chi-squared automatic interaction detection (CHAID) tree, Quick, unbiased, efficient statistical (QUEST) tree, classification and regression (C&R) tree and neural network. The GSEA was performed as described earlier by Carreras et al. [12], Hamoudi et al. [16] and Subramanian et al. [17]. In Figure 1, the details of the analysis's algorithm are shown.

2.4. Immunohistochemistry

Immunohistochemistry was performed in a Bond-Max Fully automated immunohistochemistry (IHC) and in situ hybridization (ISH) equipment following the manufacturer's instructions (Leica K.K., Tokyo, Japan) and using the DAB-based BOND Polymer Refine Detection kit (#DS9800). In summary, the immunohistochemical protocol follows the next steps: bake, dewax, rehydrate, antigen retrieval, block endogenous peroxidase, primary antibody, detection of bond antibody (post-primary antibody and polymer), color development with 3'-Diaminobenzidine (DAB), counterstain with hematoxylin and mounting. Each step is followed by washes (standard Bond wash). Mounting was performed, using a Leica CV5030 Fully Automated Glass Coverslipper.

For the cell-of-origin classification with the Hans classifier, the following antibodies were used: CD10 antigen (1:100, Clone 56C6, Novocastra, Leica K.K., Tokyo, Japan), BCL-6 oncoprotein (1:100, LN22, Novocastra) and Multiple Myeloma Oncogene 1 (MUM-1, also known as IRF4) (1:100, EAU32, Novocastra) can we put the dilutions used for each Ab. Epstein–Barr virus (EBV) infection status was assessed by in situ hybridization of

EBV-encoded mRNA (EBER, #BP0589, #AR0833, Novocastra). Validation of the prognostic markers from artificial neural network analysis was made targeting PD-L1 (Extracellular Domain Specific) (1:100, enhancer A Toyobo, retrieval in pressure cooker, E1J2J, Cell Signaling Technology K.K., Tokyo, Japan) and IKAROS (1:100, D6N9Y, CST). The antigen retrieval solution was EDTA-based (Leica BOND epitope retrieval solution 2 for 20 min for all antibodies, with exception of CD10 that was for 30 min).

2.5. Conventional and Machine-Learning-Based Digital Image Analysis

Slides were visualized in an optical microscope (Olympus BX63, Olympus K.K., Tokyo, Japan) and later digitalized, using a digital slide scanner (NanoZoomer S360, Hamamatsu Photonics, Hamamatsu City, Japan). Both conventional and machine-learning-based digital-image analysis were performed, using Fiji software. Fiji is an image processing package based on ImageJ, that contains scientific image analysis functions. Fiji is an open-source project hosted on GitHub (https://github.com/fiji) (accessed on 4 March 2021) maintained by the Eliceiri/LOCI group at the University of Wisconsin-Madison and the Jug and Tomancak labs at the MPI-CBG in Dresden (https://fiji.sc/) (accessed on 4 March 2021). For conventional analysis, the image processing was carried out on the RGB stack and the positive/negative pixel identification made use of the threshold function. This RGB method is the gold standard method and depends on the pathologist direct supervision to define the positive and negative areas (pixels). Percentage quantification was calculated in excel, using the following formula:

Percentage of positive cells = Area of positive pixels ÷ by the total area (i.e., positive + negative pixels) × 100.

The machine-learning-based image analysis quantified the marker based on the Waikato Environment for Knowledge Analysis (Weka), developed at the University of Waikato, New Zealand, version 3.2.24. The Weka can be downloaded from GitHub (https://github.com/fiji/Trainable_Segmentation) (accessed on 4 March 2021). The raw immunohistochemical image was loaded into the analysis software and directly analyzed without type change. For the training input, three types of pixels were selected: Class 1 (positive staining, DAB), Class 2 (negative areas) and Class 3 (absence of cellularity). Around 30 different areas for each color class in 6 characteristic cases were trained to set up the classifier properties, which was later used to create a result. Then, the same classifier was automatically applied to the rest of the cases. Of note, in 13 cases of PD-L1 staining the classifier made a result that was discordant with the conventional RGB-based method and the ordinal evaluation by the pathologist (Joaquim Carreras). These 13 cases were re-evaluated, using a new trained classifier that was more sensitive for a lower and diffuse expression of the PD-L1 marker. The segmentation settings included as training features the Gaussian blur, Hessian, membrane projections, Sobel filter and difference of Gaussians. The membrane thickness was set at value 1, membrane patch size at 19, minimum sigma at 1.0 and maximum sigma at 16.0. The training of the classifier included fast random forest. The classifying of the whole image used all available CPU threads. Classifying a characteristic whole image datum took from 80,113 to 89,663 milliseconds. Finally, the segmentation of the whole image was performed, and each class area was inked with a different color and quantified.

3. Results

3.1. Artificial Neural Network Analysis of Gene Expression Data

The core of the analysis comprises 28 artificial-neural-network-based analyses (14 MLP and 14 RBF) that were run independently (Figure 1). The aim was to identify which gene-probes among the 54,613 input probes had a higher importance for prediction of the 14 target variables. The target variables included the survival outcome (dead or alive) but also other relevant clinicopathological variables such as the IPI and cell-of-origin as germinal centre B-cell (GCB) vs. activated B-cell–like (ABC) that also are relevant for the prognosis of the DLBCL patients. The genes above 70% of normalized importance for

each variable were selected and pooled with the top 1% averaged normalized important as shown in the Figure 1. After deleting the duplicates, the resulting set comprised 1202 gene-probes, which equals to a 45 times reduction.

The results are shown in the Tables 3–5 and Figure 2. The two artificial neural network methodologies of MLP and RBF used distinct activation and error functions. Both methods had comparable overall performances, with similar training set percentages of cases (≈70% of the total series), testing set cases of percentages (≈30%), percentages of incorrect predictions in the training and testing sets (≈30%), and overall percentages of correct classifications in the training and testing sets (≈70%). Nevertheless, they differed in the number of units in the hidden layers (nine in MLP and six in RBF), in the training time (≈7 min for MLP and ≈114 min for RBF) and in the ROC area under the curve (0.7 in MLP and 0.6 in RBF). Of note, the classification model's performance differed according the target variable. For instance, in MLP, the artificial neural network ability to predict the binary target variable was higher in the variables "Extranodal disease" (Area under the curve of 0.88), "Alive < 1.5 years vs. Alive => 7 years" (AUC of 0.84), "cell of origin" (0.80), "Outcome Dead CHOP-like only" (0.76), "LDH ratio >3" (0.75), "ECOG $\geq$ 2" (0.73) and "Outcome Dead" (0.70).

Table 3. Multilayer Perceptron (MLP) artificial neural network analysis for the prediction of DLBCL prognosis.

Multilayer Perception	Dependent Variable	Outcome Dead	Outcome Dead RCHOP-Like Only	Outcome Dead CHOP-Like Only	Cell-of-Origin Activated B-Cell-Like	Age > 60	LDH Ratio ≥ 1	LDH Ratio > 3	ECOG ≥ 2	Stage III/IV	Extranodal Sites > 1	Sex Male	NCCN IPI–like HI+H	Dead < 1.5 vs. Alive => 7 y.	Multivariate
Case processing summary	Training	283	161	131	252	279	253	239	264	295	260	268	222	104	187
	Training Percentage	68.40	69.10	72.40	72.00	67.40	72.10	68.10	67.90	72.70	67.90	67.70	69.20	67.10	69.00
	Testing	131	72	50	98	135	98	112	125	111	123	128	99	51	84
	Testing Percentage	31.60	30.90	27.60	28.00	32.60	27.90	31.90	32.10	27.30	32.10	32.30	30.80	32.90	31.00
	Valid	414	233	181	350	414	351	351	389	406	383	396	321	155	271
	Excluded	6	0	0	0	6	69	69	31	14	37	24	99	265	149
	Total	420	233	181	350	420	420	420	420	420	420	420	420	420	420
Network information	Number of Units	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613
	Rescaling Method of Covariates	Standardized													
Hidden layer	Number of Hidden Layers	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Number of Units in Hidden Layer	6	6	9	15	5	9	15	9	8	8	8	8	10	7
	Activation Function	Hyperbolic tangent													
Output layer	Dependent variable	1	1	1	1	1	1	1	1	1	1	1	1	1	8
	Number of Units	2	2	2	2	2	2	2	2	2	2	2	2	2	16
	Activation Function	Softmax													
	Error Function	Cross-entropy													
Model summary training	Cross-Entropy Error	174.16	85.80	79.90	140.50	178.40	163.90	74.50	132.80	193.90	47.50	186.20	136.50	46.49	840.80
	Percent of Incorrect Predictions	33.90	24.80	33.60	27.80	38.40	36.80	10.50	24.60	39.30	6.50	45.40	33.80	19.20	30.20
	Stopping Rule Used *	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Time in Minutes	9.12	4.92	4.60	7.10	8.20	8.02	6.80	7.92	9.35	7.63	9.33	7.32	2.98	7.83
Model summary testing	Cross-Entropy Error	77.39	36.10	25.30	49.80	87.30	63.00	22.70	61.50	69.60	27.60	80.60	59.10	18.80	397.90
	Percent of Incorrect Predictions	32.10	26.40	26.00	25.50	39.30	35.70	6.30	23.20	34.20	8.90	35.20	30.30	19.60	32.90
Classification	Training Overall Percent	66.10	75.20	66.40	72.20	61.60	63.20	89.50	75.40	60.70	93.50	54.50	66.20	80.80	69.80
	Testing Overall Percent	67.90	73.60	74.00	74.50	60.70	64.30	93.80	76.80	65.80	91.10	64.80	69.70	80.40	67.10
Area under the curve	Alive	0.70	0.69	0.76	0.80	0.68	0.67	0.75	0.73	0.66	0.88	0.60	0.68	0.84	0.66
	Dead	0.70	0.69	0.76	0.80	0.68	0.67	0.75	0.73	0.66	0.88	0.60	0.68	0.84	0.66

RCHOP, rituximab, cyclophosphamide, doxorubicin hydrochloride, vincristine sulfate and prednisone; LDH, lactate dehydrogenase; IPI, International Prognostic Index; HI, high intermediate; H, high; * consecutive with no decrease in error.

Table 4. Radial Basis Function (RBF) artificial neural network analysis for the prediction of DLBCL prognosis.

Radial Basis Function	Dependent Variable	Outcome Dead	Outcome Dead RCHOP-Like Only	Outcome Dead CHOP-Like Only	Cell-of-Origin Activated B-Cell-Like	Age > 60	LDH Ratio ≥ 1	LDH Ratio > 3	ECOG ≥ 2	Stage III/IV	Extranodal Sites > 1	Sex Male	NCCN IPI–Like HI+H	Dead < 1.5 vs. Alive => 7 years	Multivariate
Case processing summary	Training	283	161	131	239	301	239	240	264	295	261	269	222	104	187
	Training Percentage	68.30	69.10	72.40	68.30	72.70	68.10	68.40	67.90	72.70	68.10	67.90	69.20	67.10	69.00
	Testing	131	72	50	111	113	112	111	125	111	122	127	99	51	84
	Testing Percentage	31.60	30.90	27.60	31.70	27.30	31.90	31.60	32.10	27.30	31.90	32.10	30.80	32.90	31.00
	Valid	414	233	181	350	414	351	351	389	406	383	396	321	155	271
	Excluded	6	0	0	70	6	69	69	31	14	37	24	99	265	149
	Total	420	233	181	420	420	420	420	420	420	420	420	420	420	420
Network information	Number of Units	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613	54,613
	Rescaling Method of Covariates	Standardized													
Hidden layer	Number of Hidden Layers	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Number of Units in Hidden Layer	10	2	3	10	2	10	9	3	2	10	2	8	8	8
	Activation Function	Softmax													
Output layer	Dependent variable	1	1	1	1	1	1	1	1	1	1	1	1	1	8
	Number of Units	2	2	2	2	2	2	2	2	2	2	2	2	2	16
	Activation Function	Identity													
	Error Function	Sum of Squares													
Model summary training	Cross Entropy Error	57.50	31.20	31.30	47.60	74.20	55.90	23.70	48.40	73.30	15.50	66.60	49.40	16.60	293.60
	Percent of Incorrect Predictions	33.60	26.70	41.20	27.60	46.20	39.70	11.30	24.20	46.10	7.70	46.10	34.70	23.10	31.10
	Stopping Rule Used *	-													
	Time in Minutes	171.08	32.70	21.07	127.07	221.92	127.37	117.73	145.20	145.75	141.83	171.65	107.20	13.33	57.70
Model summary testing	Cross Entropy Error	27.80	12.90	12.50	24.90	28.00	27.00	5.30	22.30	27.70	8.60	31.20	22.60	9.30	138.30
	Percent of Incorrect Predictions	32.10	23.60	44.00	36.00	43.30	41.10	4.50	23.20	46.80	8.20	48.80	35.40	29.40	33.60
Classification	Training Overall Percent	66.40	73.30	58.80	72.40	53.80	60.30	88.80	75.80	53.90	92.30	53.90	65.30	76.90	68.90
	Testing Overall Percent	67.90	76.40	56.00	64.00	56.60	58.90	95.00	76.80	53.20	91.80	51.20	64.60	70.60	66.40
Area under the curve	Alive	0.71	0.58	0.54	0.75	0.54	0.62	0.59	0.52	0.49	0.83	0.52	0.58	0.75	0.64
	Dead	0.71	0.58	0.54	0.75	0.54	0.62	0.59	0.52	0.49	0.83	0.52	0.58	0.75	0.64

RCHOP, rituximab, cyclophosphamide, doxorubicin hydrochloride, vincristine sulfate and prednisone; LDH, lactate dehydrogenase; IPI, International Prognostic Index; HI, high-intermediate; H, high.
* Consecutive with no decrease in error.

Table 5. Comparison of Performance between MLP and RBF artificial neural networks for prediction of DLBCL prognosis.

Artificial Neural Network Parameters	MLP (n = 14)	RBF (n = 14)	p-Value
Training set number	228.4 ± 59.7	228.3 ± 60.9	0.995
Training set percentage	69.4% ± 2.0	69.2% ± 1.9	0.864
Testing set number	101.2 ± 28.2	101.4 ± 26.9	0.989
Testing set percentage	30.6% ± 2.0	30.7% ± 1.9	0.872
Valid number of cases	329.6 ± 86.7	329.6 ± 86.7	1
Number of gene-probes (units)	54,613	54,613	1
Rescaling method for covariates	Standardized	Standardized	N/A
Hidden layer(s)			
Number of hidden layers	1	1	N/A
Number of units in hidden layer	8.8 ± 2.9	6.2 ± 3.6	0.048
Activation function	Hyperbolic tangent	Softmax	N/A
Output layer			
Dependent (target) variable	1 (8 for multivariate)	1 (8 for multivariate)	1
Number of units	2 (16 for multivariate)	2 (16 for multivariate)	1
Activation function	Softmax	Identity	N/A
Error function	Cross-entropy	Sum of Squares	N/A
Model summary			
Training			
Cross-entropy error	177.2 ± 197.7	63.2 ± 69.1	0.058
Percent of incorrect predictions	28.9% ± 11.0	31.4% ± 12.3	0.559
Training time (min)	7.2 ± 1.8	114.4 ± 61.9	7.5×10^{-7}
Testing			
Cross-entropy error	76.9 ± 95.2	28.5 ± 32.8	0.091
Percent of incorrect predictions	26.8% ± 9.8	32.1% ± 13.5	0.245
Classification			
Training sample, overall percent correct	71.1% ± 11.0	68.6% ± 12.3	0.583
Testing sample, overall percent correct	73.2% ± 9.8	67.8% ± 13.4	0.239
Area under the curve (ROC)	0.7 ± 0.1	0.6 ± 0.1	0.007

After the +/- symbol, the standard deviation is shown.

The number of genes with a normalized importance >70% for each individual artificial neural network analysis of MLP ranged from 1 to 132, with an average of 34 and a median of 12. In case of RBF ranged from 3 to 84 with an average of 34 and a median of 24. After integration with the top 1% of normalized importance set and once the duplicates were deleted, the final set comprised 1202 gene-sets. In order to identify the most relevant ones we performed a univariate Cox regression analysis and only the probes with significant correlation for overall survival were selected (n = 448). The most relevant genes based on their p-value and Hazard Risk are shown in Tables 6–8.

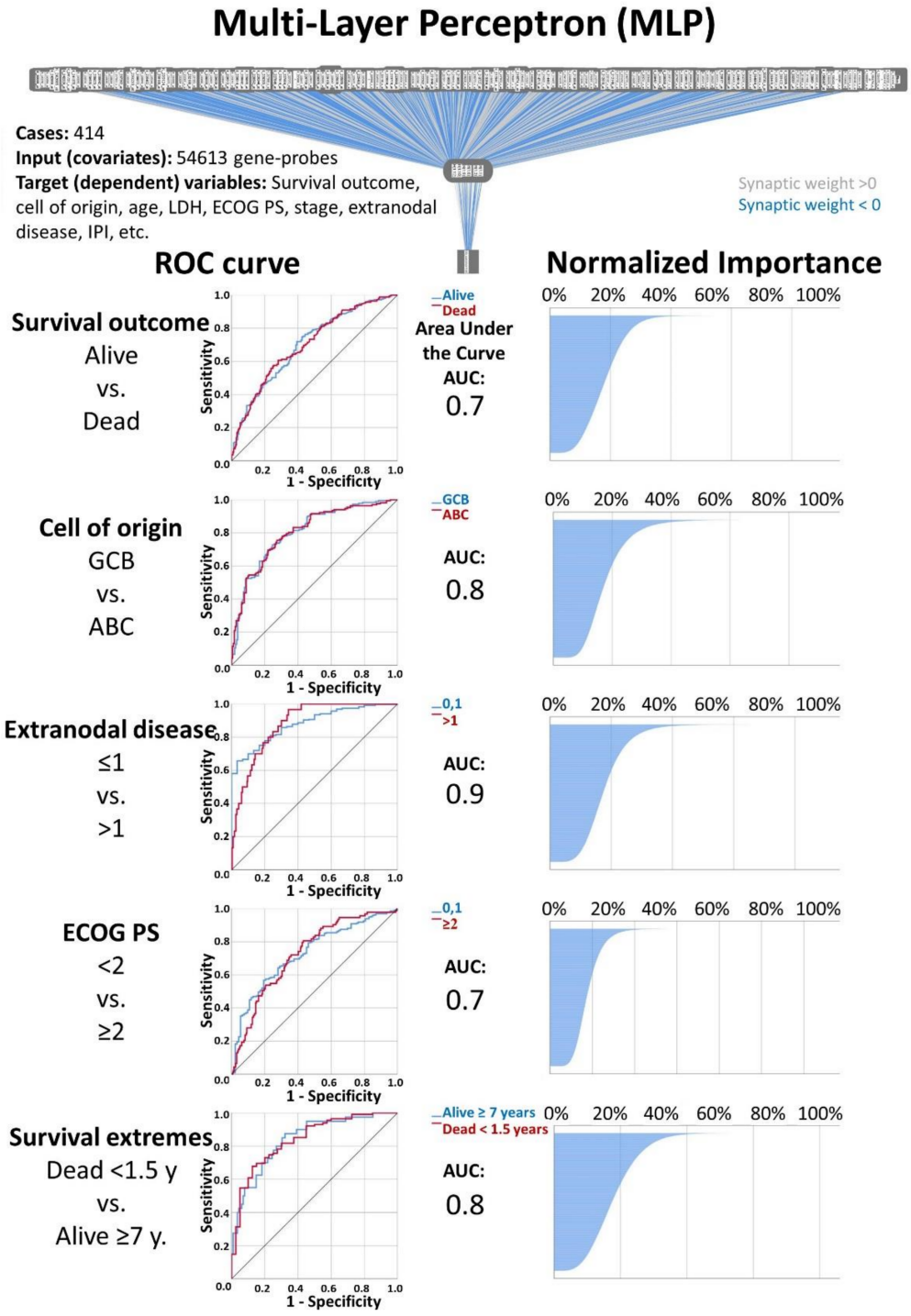

Figure 2. Multilayer Perceptron (MLP) analysis: The artificial neural network analysis consisted of applying of Multilayer Perceptron (MLP) and Radial Basis Function (RBF) artificial neural networks on publicly available gene-expression data from DLBCL patients. For both MLP and RBF, the inputs (covariates) were the 54,613 gene-probes, and the target variables (dependent variables) were the overall survival outcome (dead vs. alive) and a series of clinicopathological variables, including the cell of origin molecular classification, age, LDH, ECOG Performance Status, clinical stage, extranodal disease and IPI. A total of 26 individual AI analyses were performed. The most relevant genes were selected according to their normalized importance, following a strategy as described in Material and Methods and in Table 3. This figure shows part of the results of the MLP analysis. AUC, area under the curve.

Table 6. Top 10 genes according to the *p*-value (set of 448 gene-probes).

Gene	Probe	B	SE	Wald	df	p-Value	Hazard Risk	Lower	Upper
ALG3	207396_s_at	0.74	0.134	30.257	1	3.8×10^{-8}	2.095	1.61	2.727
UCK2	209825_s_at	0.589	0.11	28.454	1	9.6×10^{-8}	1.803	1.452	2.238
ZMYND19	227477_at	0.687	0.13	27.96	1	1.2×10^{-7}	1.988	1.541	2.564
ELFN1-AS1	231443_at	0.551	0.107	26.369	1	2.8×10^{-7}	1.735	1.406	2.141
PHTF2	1554780_a_at	-0.55	0.108	25.968	1	3.5×10^{-7}	0.577	0.467	0.713
EXOSC7	212627_s_at	0.546	0.11	24.453	1	7.6×10^{-7}	1.726	1.39	2.144
BCAT2	203576_at	0.594	0.121	23.946	1	9.9×10^{-7}	1.81	1.427	2.296
TBRG4	220789_s_at	0.645	0.133	23.646	1	1×10^{-6}	1.906	1.47	2.471
THOC1	204064_at	0.906	0.186	23.676	1	1×10^{-6}	2.476	1.718	3.566
KIF13B	202962_at	0.413	0.086	23.247	1	1×10^{-6}	1.511	1.278	1.787

Table 7. Top 10 bad prognostic genes according to the Hazard Risk (set of 448 gene-probes).

Gene	Probe	*B*	SE	Wald	df	*p*-Value	Hazard Risk	Lower	Upper
THOC1	204064_at	0.906	0.186	23.676	1	1×10^{-6}	2.476	1.718	3.566
TMX2	201175_at	0.878	0.184	22.698	1	2×10^{-6}	2.407	1.677	3.455
HNRNPC	214737_x_at	0.748	0.202	13.762	1	2×10^{-4}	2.113	1.423	3.138
ALG3	207396_s_at	0.74	0.134	30.257	1	3.8×10^{-8}	2.095	1.61	2.727
NELFA	203112_s_at	0.719	0.153	22.04	1	3×10^{-6}	2.052	1.52	2.77
PPP6R2	202791_s_at	0.695	0.158	19.292	1	1.1×10^{-6}	2.003	1.469	2.73
ZMYND19	227477_at	0.687	0.13	27.96	1	1.2×10^{-7}	1.988	1.541	2.564
TBRG4	220789_s_at	0.645	0.133	23.646	1	1×10^{-6}	1.906	1.47	2.471
GLO1	200681_at	0.643	0.157	16.795	1	4×10^{-5}	1.903	1.399	2.588
BORCS8	1553978_at	0.62	0.163	14.513	1	1×10^{-5}	1.859	1.351	2.558

Table 8. Top 10 good prognostic genes according to the Hazard Risk (set of 448 gene-probes).

Gene	Probe	*B*	SE	Wald	df	*p*-Value	Hazard Risk	Lower	Upper
TTC3	208663_s_at	−0.124	0.034	13.115	1	0.0002	0.884	0.826	0.945
YTHDC1	214814_at	−0.134	0.043	9.514	1	0.002	0.875	0.803	0.952
B3GALNT1	223374_s_at	−0.146	0.058	6.326	1	0.012	0.864	0.771	0.968
ZNF277	1555193_a_at	−0.152	0.074	4.246	1	0.039	0.859	0.744	0.993
RAB39B	238695_s_at	−0.154	0.063	5.87	1	0.015	0.857	0.757	0.971
ITPR1	211323_s_at	−0.156	0.065	5.671	1	0.017	0.856	0.753	0.973
CLIC5	213317_at	−0.157	0.066	5.745	1	0.017	0.854	0.751	0.972
SEL1L	202062_s_at	−0.16	0.077	4.335	1	0.037	0.852	0.732	0.991
N/A	242693_at	−0.167	0.069	5.962	1	0.015	0.846	0.74	0.968
MFSD6	219858_s_at	−0.171	0.064	7.251	1	0.0071	0.843	0.744	0.955

In general, MLP was more "efficient" than RBF. In comparison to the RBF artificial neural network, the MLP characterized by a significantly lower training time and better areas under the curve. In addition, although not statistically significant, MLP also had a lower percentage of incorrect predictions and higher overall percent correct.

The set of 448 genes was subjected to GSEA analysis in order to confirm the association to the prognosis outcome (dead vs. alive phenotype), using the same series of cases (Figure 3). Within the core enrichment 233 genes were found. The top five genes were *DMBT1, OR14J1, OCRL, DEFA1* and *ELFN1-AS1*. Within the core enrichment an important marker of the tumoral immune response with known potential relevance for the pathogenesis of DLBCL was highlighted, the Programmed Cell Death 1 Ligand 1 (*PD-L1*, *CD274*). Of note, PD-L1 can be targeted by immune checkpoint inhibitors. In the previous univariate Cox regression analysis, *PD-L1* was associated to a bad overall survival of the patients, with a Hazard Risk of 1.178 (95%CI 1.023–1.356, $p = 0.023$). Outside the core

enrichment, and towards the good prognosis phenotype we identified *IKAROS*. *IKAROS* also belongs to the immune checkpoint pathway, with a Hazard Risk of 0.488 (95%CI: 0.376–0.633, $p = 7.3 \times 10^{-8}$). Due to the biological importance of both *PD-L1* and *IKAROS* for the pathogenesis of DLBCL, these two markers were selected for validation in an independent series of DLBCL patients from Tokai University Hospital, the protein expression was evaluated by immunohistochemistry and AI-based image segmentation and digital quantification was performed.

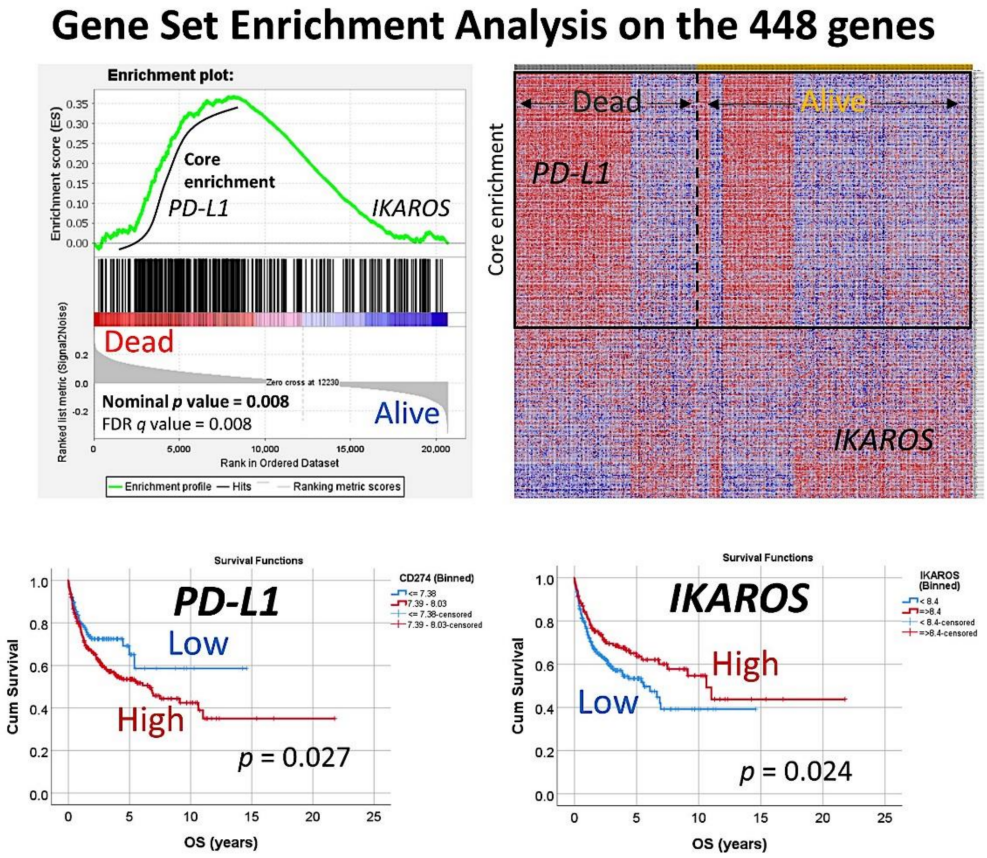

Figure 3. Gene-set enrichment analysis on the set of 448 genes. The set of 448 genes was used in a GSEA analysis, to confirm the association of this set to the overall survival outcome of the patients (dead vs. alive phenotype). In the core enrichment associated to poor prognosis (dead), the *PD-L1* (*CD274*) gene was identified. In the side of good prognosis (alive), the gene *IKAROS* was identified. Both markers that belong to the immune checkpoint pathway were further validated by immunohistochemistry in an independent series of DLBCL from Tokai University Hospital.

The set of 448 genes was also analyzed by a functional network association analysis; the results are shown in the Figure 4.

In order to highlight the most relevant markers and to reduce the number of genes within the set of 448, a second round of artificial neural network analysis was performed, including MLP and RBF, as shown in Figure 1. As a result, the set was reduced to 16 genes: *PAF1*, *USP28*, *SORT1*, *MAP7D3*, *FITM2*, *CENPO*, *PRCC*, *ALDH6A1*, *CSNK2A1*, *TOR1AIP1*, *NUP98*, *UBE2H*, *UBXN7*, *SLC44A2*, *NR2C2AP* and *LETM1* (Table 9).

A multivariate Cox regression for overall survival analysis, using the backward conditional, was applied to the set of 16 genes and the final step included only six genes (Table 10). In Figure 5, the relevance of these six genes for the overall survival of the patients is shown. A cutoff that stratified the patients according to the gene expression of each marker (70% vs. 30% approximately). In Figure 6, the correlation with known genes that are relevant for the pathogenesis of DLBCL is shown. The genes *BCL2* and *MYC* are relevant for the pathogenesis of the tumoral B-lymphocytes of DLBCL (anti-apoptosis

and cell cycle), *CD163* is a marker of M2-like tumor associates macrophages (TAMs) and *TNFAIP8* is an apoptosis inhibitor expressed by the B-lymphocytes of DLBCL as well as for the TAMs. The biological functions of the set of 16 genes is shown in the Table 11.

Prognostic modeling for overall survival outcome (dead vs. alive) was also applied to the set of 16 genes, and the tests with an overall accuracy above 70% were selected (Figures 7–9): classification and regression tree (C&R tree) (overall accuracy of 74.39%), C5 decision tree (72.46) and Bayesian Network (72.38).

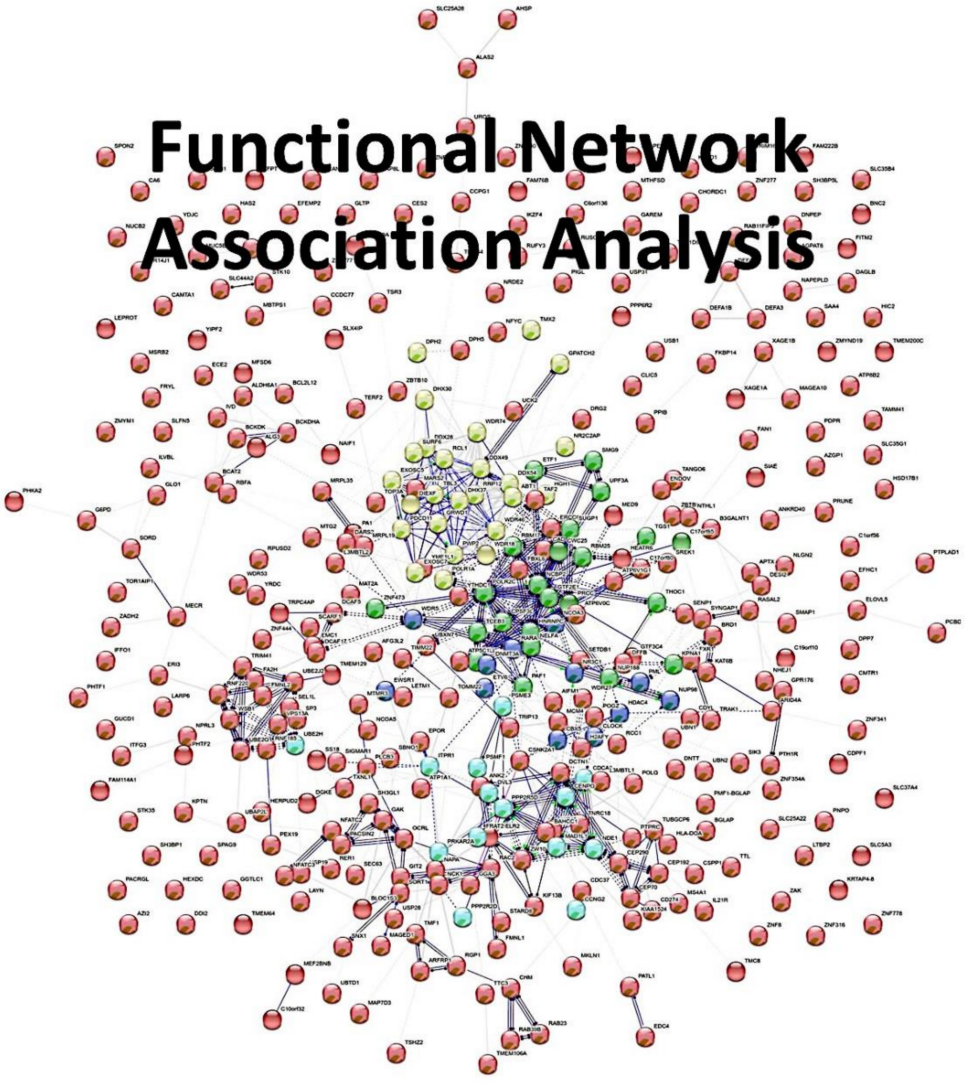

Figure 4. Functional network association analysis on the set of 448 genes. In order to analyze the set of 448 genes according to the biological processes, molecular function, cellular component and pathways, a network analysis was made. The network was characterized by 390 nodes, 791 edges, an average node degree of 4.06, average local clustering coefficient of 0.376 and a protein-protein interaction (PPI) enrichment p-value of 1×10^{-16}. In general, the set belonged to the Gene Ontology (GO) nuclei acid metabolic process (GO: 0090304, False Discovery Rate (FDR) = 0.00012). Of note, within the general network, five clusters could be identified.

Table 9. Set of 16 prognostic genes.

Gene	Probe	*B*	SE	Wald	df	*p*-Value	Hazard Risk	Lower	Upper
PAF1	202093_s_at	0.237	0.118	4.01	1	0.045	1.267	1.005	1.597
USP28	1552678_a_at	0.422	0.133	10.083	1	0.0015	1.526	1.175	1.98
SORT1	212807_s_at	0.177	0.081	4.758	1	0.03	1.194	1.018	1.401
MAP7D3	219626_at	0.376	0.135	7.79	1	0.005	1.456	1.118	1.896
FITM2	226805_at	0.302	0.148	4.162	1	0.04	1.352	1.012	1.807
CENPO	226118_at	0.324	0.108	8.946	1	0.003	1.383	1.118	1.71
PRCC	208938_at	0.229	0.117	3.856	1	0.05	1.258	1	1.581
ALDH6A1	221588_x_at	0.515	0.158	10.57	1	0.0012	1.673	1.227	2.282
CSNK2A1	212075_s_at	0.418	0.134	9.715	1	0.0018	1.52	1.168	1.977
TOR1AIP1	212409_s_at	0.384	0.162	5.607	1	0.018	1.468	1.068	2.017
NUP98	203194_s_at	0.339	0.131	6.718	1	0.009	1.404	1.086	1.814
UBE2H	221962_s_at	-0.415	0.121	11.699	1	0.0006	0.66	0.521	0.838
UBXN7	217100_s_at	0.269	0.108	6.187	1	0.013	1.309	1.059	1.618
SLC44A2	224609_at	0.251	0.107	5.503	1	0.019	1.286	1.042	1.586
NR2C2AP	226839_at	0.355	0.149	5.687	1	0.017	1.427	1.065	1.911
LETM1	222006_at	0.282	0.137	4.275	1	0.038	1.326	1.015	1.733

Table 10. Set of six prognostic genes.

Gene	Probe	*B*	SE	Wald	df	*p*-Value	Hazard Risk	Lower	Upper
USP28	1552678_a_at	0.44309	0.155462	8.123335	1	0.004	1.558	1.148	2.112
SORT1	212807_s_at	0.196301	0.082305	5.688408	1	0.017	1.217	1.036	1.430
ALDH6A1	221588_x_at	0.403888	0.167332	5.825904	1	0.016	1.498	1.079	2.079
CSNK2A1	212075_s_at	0.304248	0.152804	3.964474	1	0.047	1.356	1.005	1.829
TOR1AIP1	212409_s_at	0.313668	0.169528	3.423401	1	0.06	1.368	0.982	1.908
UBE2H	221962_s_at	−0.63261	0.113921	30.83679	1	2.8×10^{-8}	0.531	0.425	0.664

Multivariate Cox regression for overall survival analysis (backward conditional).

3.2. Machine-Learning-Based Quantification of the Immunohistochemical Expression and Correlation with the Clinicopathological Characteristics of the Patients

The results of this section are shown in Figures 10 and 11 and Tables 12 and 13.

Two markers were selected from the gene expression analysis for validation in an independent lymphoma series of DLBCL from Tokai University Hospital. PD-L1 and IKAROS were immuno-stained. After image digitization, the protein expression was quantified by using a machine-learning-trainable segmentation method.

The protein expression of *PD-L1* ranged from 0.01% to 92.5%, with a median of 16.9% and a mean of 25.0% ± 24.0 STD. The PD-L1 staining was also quantified by using a conventional RBG approach. Both quantifications had a good correlation (Pearson Correlation 0.853, $p = 4.6 \times 10^{-33}$) (Figure 11). The PD-L1 values from the AI Weka segmentation were ranked and the most significant cutoff point for overall survival was calculated (31%). The patients with a high PD-L1 expression had an 86% more risk of dying than the patients with low expression (Hazard Risk = 1.86, 95%CI 1.05–3.31). The five-years overall survival of the patients, high vs. low PD-L1, was 40% (95%CI 58–23%) vs. 67% (95%CI 77–56%) ($p = 0.031$), respectively. PD-L1 expression was also correlated with several clinicopathological characteristics. High PD-L1 expression correlated with a non-GCB phenotype, Epstein–Barr virus infection (EBER+), high RGS1 expression, high sIL2RA, clinical stage III/IV, presence of B symptoms and high to high-intermediate IPI. High PD-L1 also associated to a worse progression free survival ($p = 0.054$, "trend of association") (Table 12; Figures 10 and 11)

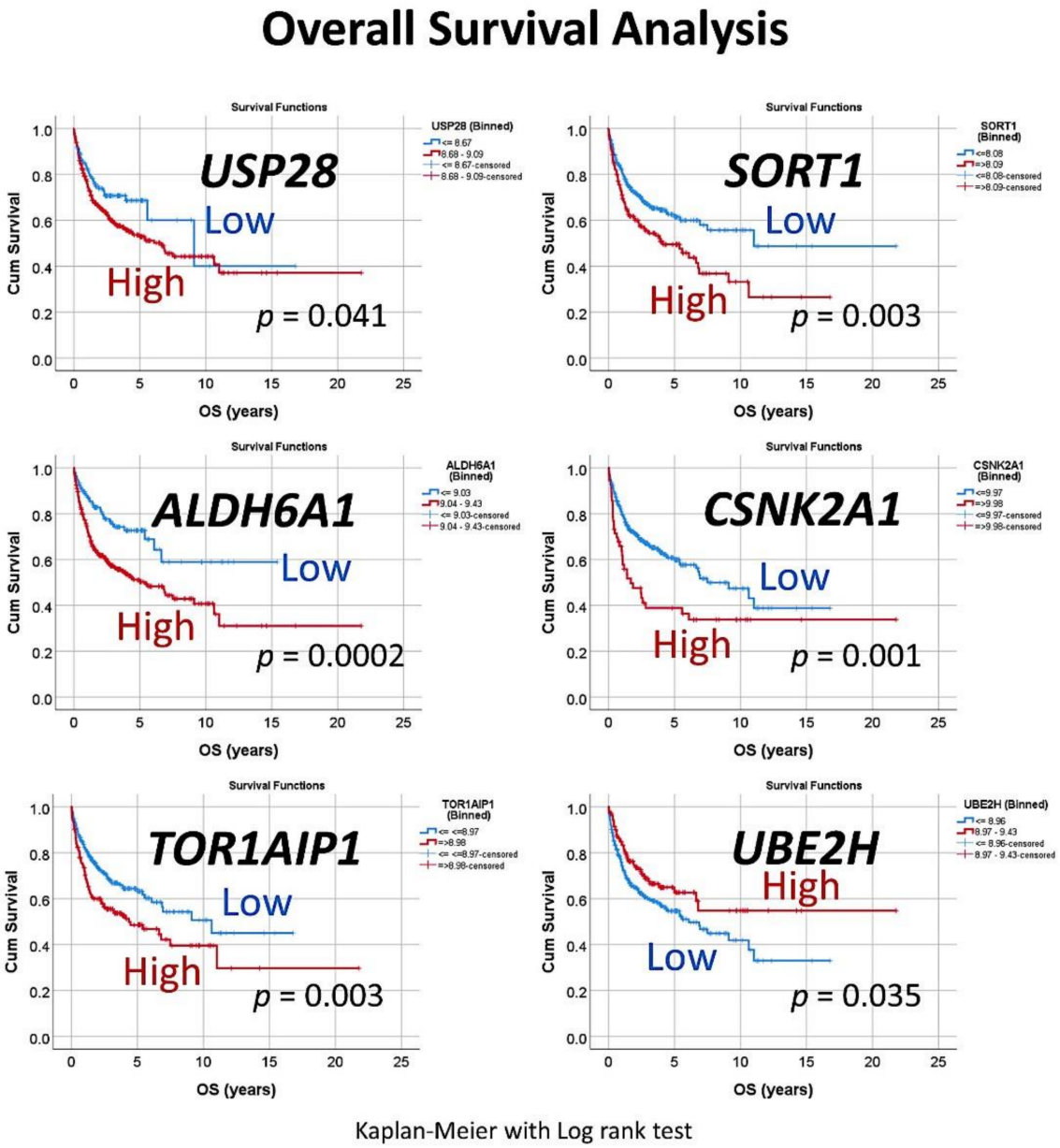

Figure 5. Univariate overall survival analysis of the set of six genes. For each of the six genes, a cutoff was searched to stratify the patients into high and low expression. Then, the overall survival for each marker was analyzed, using the Kaplan–Meier with Log rank test.

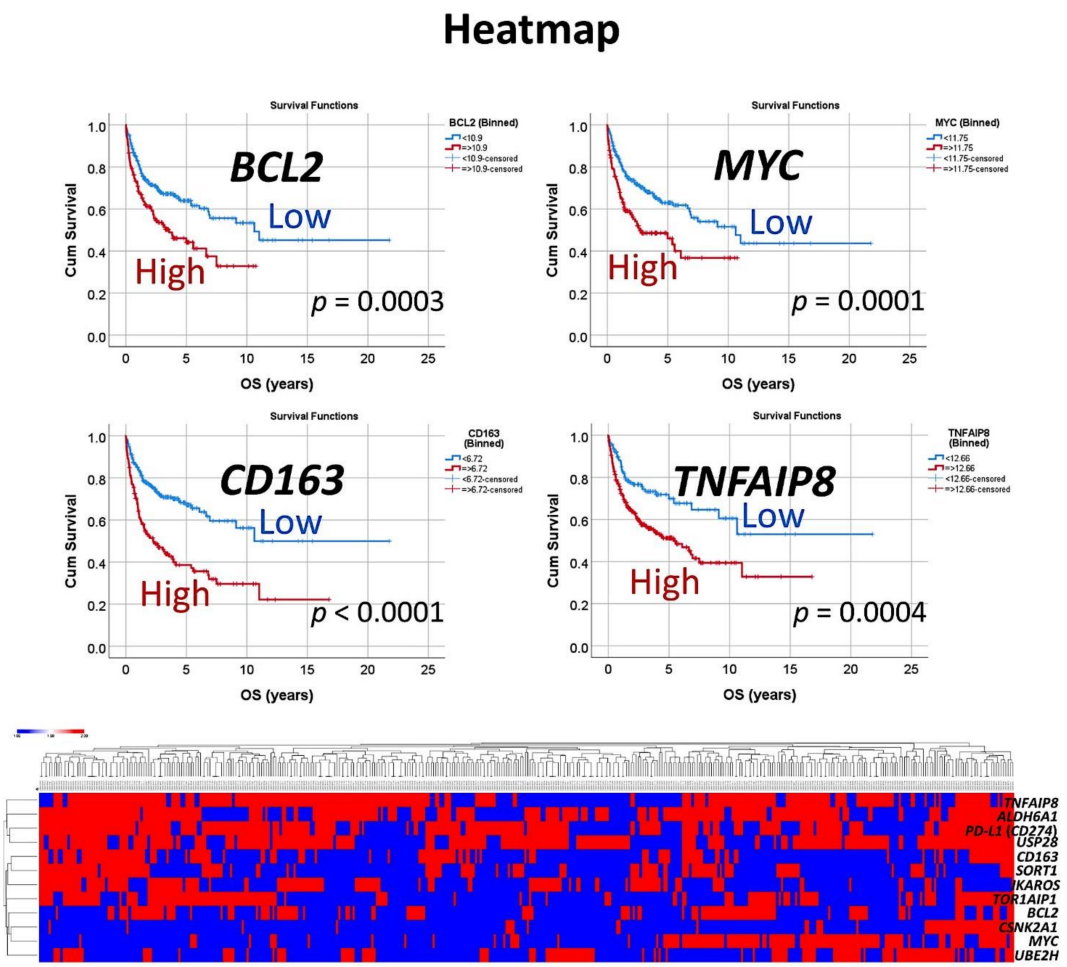

Figure 6. Correlation with known pathogenic biomarkers of DLBCL. The clinical relevance for overall survival of known pathogenic biomarkers, including *BCL2*, *MYC*, *CD163* and *TNFAIP8*, was tested in this series. After that, an unsupervised hierarchical clustering was performed with the set of six genes (*USP28*, *SORT1*, *ALDH6A1*, *CSNK2A1*, *TOR1AIP1* and *UBE2H*), *PD-L1 (CD204)*, *IKAROS*, *BCL2*, *MYC*, *CD163* and *TNFAIP8*. The dendrogram for the rows showed how *TNFAIP8*, *ALDH6A1*, *PD-L1* and *USP2B* clustered in the same group. In addition, *CSNK2A1* and *MYC* were also close.

Table 11. Biological function of the set of 16 prognostic genes.

Gene	Function
PAF1	Positive regulation of cell cycle G1/S phase transition
USP28	DNA damage response checkpoint and MYC proto-oncogene stability
SORT1	Endocytosis
MAP7D3	Microtubule cytoskeleton organization
FITM2	Cytoskeleton organization and lipid and energy homeostasis
CENPO	Mitotic progression and chromosome segregation
PRCC	Regulation of cell cycle progression
ALDH6A1	Pyrimidine metabolism, RNA binding
CSNK2A1	Cell cycle, apoptosis process
TOR1AIP1	Regulation of nuclear membrane integrity, protein localization to nucleus
NUP98	Role in the nuclear pore complex (NPC) assembly and/or maintenance
UBE2H	ATP binding, ubiquitin-protein transferase activity
UBXN7	Ubiquitin binding
SLC44A2	Positive regulation of I-kappaB kinase/NF-kappaB signaling
NR2C2AP	Transcription initiation from RNA polymerase II promoter
LETM1	Regulation of concentration of calcium ion.

Based on data provided by UniProt database (https://www.uniprot.org/) (accessed on 4 March 2021).

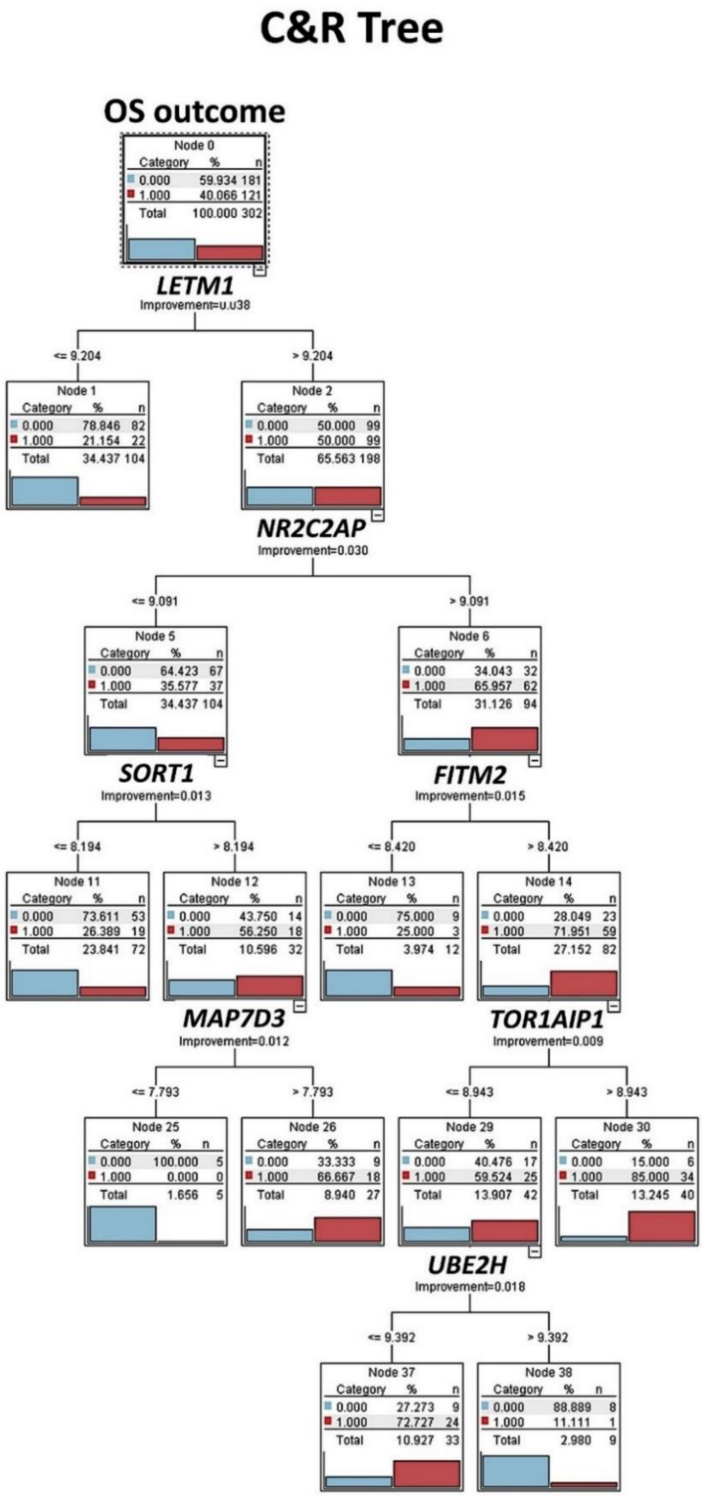

Figure 7. Classification and regression tree (C&R tree): Prognostic modeling was performed, using the set of 16 genes, as shown in Figure 1, and 12 different types of machine-learning analyses were tested. Then, the ones with >70% of overall accuracy were selected. This figure shows the result of the C&R tree. Decision list models identify subgroups or segments that show a higher or lower likelihood of a binary (yes or no) outcome relative to the overall sample. C&R tree node generates a decision tree that allows you to predict or classify future observations. The method uses recursive partitioning to split the training records into segments by minimizing the impurity at each step. Dead outcome, red color (number 1). Alive outcome, blue color (number 0).

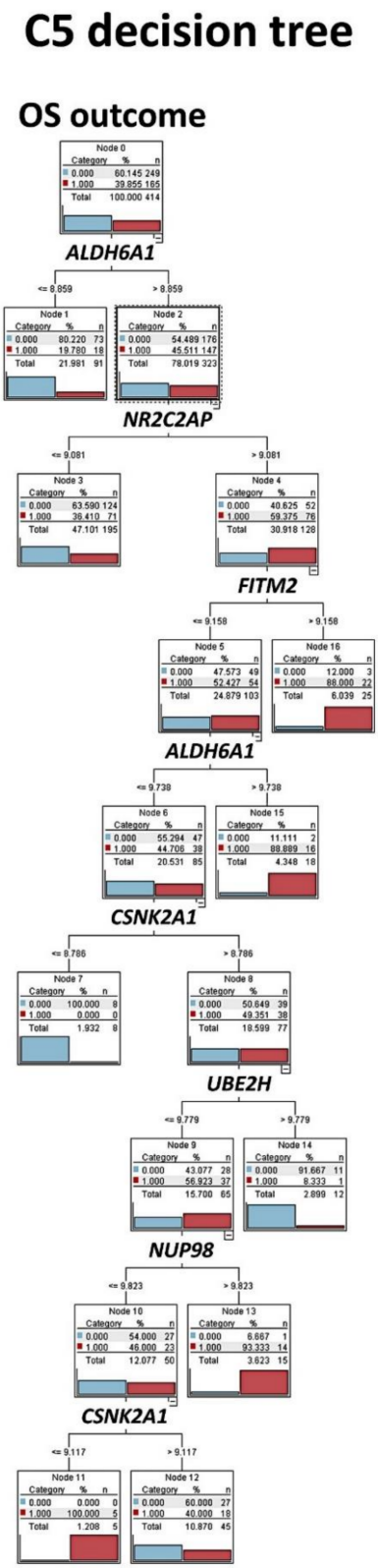

Figure 8. C5 decision tree: The set of 16 genes was also tested, using C5 decision tree. This result had an overall accuracy above 70%. The C5.0 node builds either a decision tree or a rule set. The model works by splitting the sample based on the field that provides the maximum information gain at each level. The target field must be categorical (in our case, the overall survival outcome as dead vs. alive). Multiple splits into more than two subgroups are allowed. Dead outcome, red color (number 1). Alive outcome, blue color (number 0).

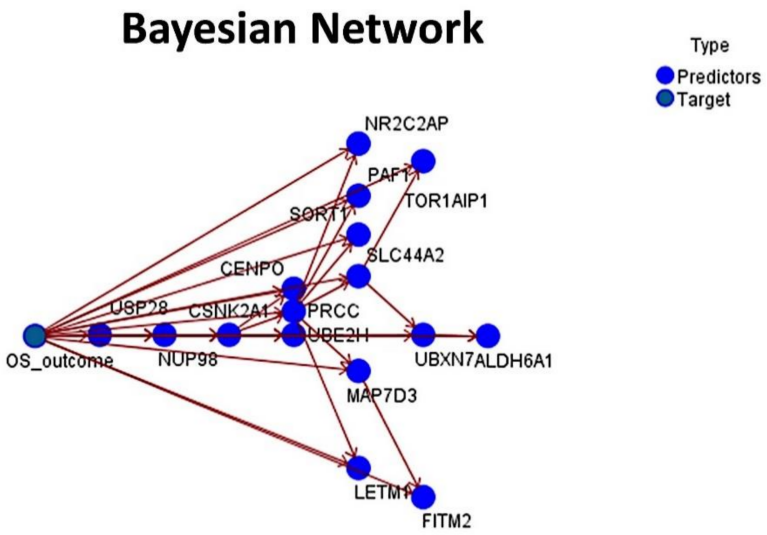

Figure 9. Bayesian Network: The set of 16 genes was also tested, using a Bayesian Network. The result had an overall accuracy above 70%. A Bayesian Network is a graphical model that displays variables (nodes) in a dataset and the probabilistic, or conditional, independencies between them. Causal relationships between the several variables may be represented by a Bayesian Network; however, the links (arcs) between the nodes do not necessarily represent a direct cause and effect.

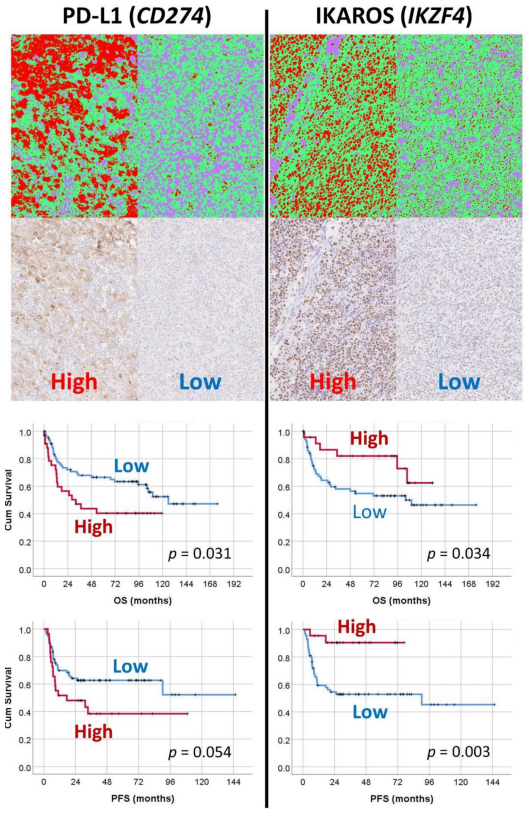

Figure 10. Machine-learning-based digital-image analysis of immunohistochemical expression of PD-L1 and IKAROS and their correlation with the survival of the patients. The markers of *PD-L1* and *IKAROS* were identified in the artificial neural network analysis of gene-expression data as bad prognosis and good prognosis markers, respectively. The immunohistochemical expression was tested in an independent DLBCL series from Tokai University. For the digital-image quantification, an AI-based segmentation method was used. Correlation with the survival of the patients confirmed the AI results.

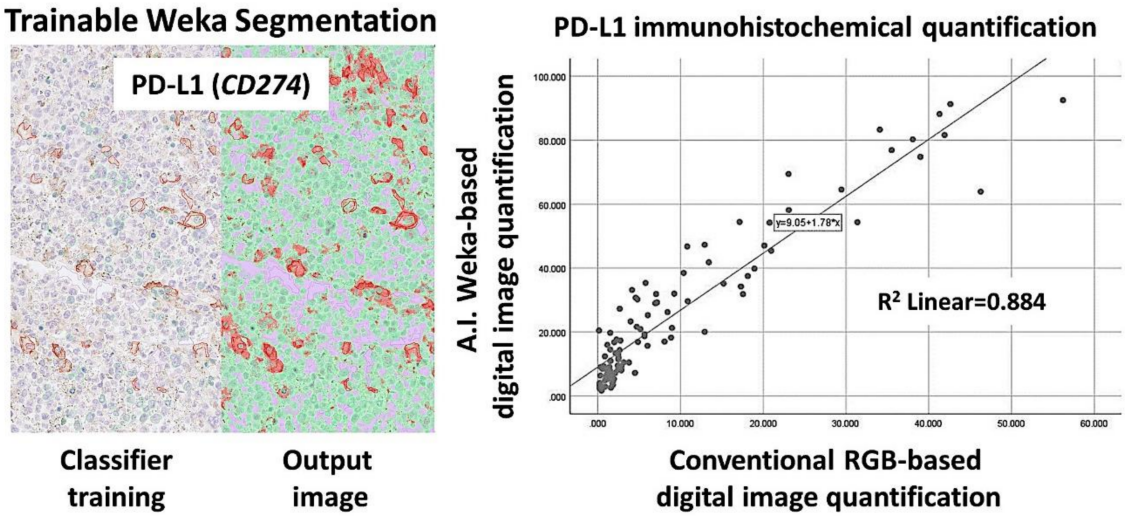

Figure 11. PD-L1 (CD274) marker validation, using digital image. Digital image quantification of PD-L1. PD-L1 was analyzed, using a conventional RGB-based analysis, as well as with a machine-learning trainable segmentation method. Good correlation was found between both methods.

Table 12. Correlation between PD-L1 and the clinicopathological features of the patients in the validation series (Tokai cases).

Predictors for High PD-L1	*p*-Value	Odds Ratio	95% CI for OR	
			Lower	Upper
Sex Male	0.211	1.699	0.741	3.898
Age > 60	0.994	1.004	0.415	2.429
Location				
Nodal (+spleen) (Reference)	-	-	-	-
Extranodal				
Waldeyer's ring	0.999	0	0	-
Gastrointestinal	0.756	1.242	0.317	4.875
Other extranodal	0.487	0.71	0.27	1.864
LDH High (>219)	0.115	2.037	0.841	4.933
Seric IL2RA high (>530)	0.016	12.453	1.598	97.065
ECOG Performance Status ≥ 2	0.207	2.111	0.661	6.741
Clinical stage III or IV	0.006	3.585	1.452	8.851
Extranodal disease site > 1	0.741	0.825	0.263	2.588
B symptoms	0.004	4.333	1.618	11.606
IPI HI+H	0.041	2.579	1.037	6.411
Non-GCB Subtype (Hans's algorithm)	0.014	3.757	1.307	10.794
Epstein–Barr virus, EBER+	0.005	4.931	1.620	15.005
High RGS1 protein expression	0.015	3.003	1.241	7.264
Absence of clinical response to treatment	0.078	2.284	0.912	5.717

Binary logistic regression setup: dependent variable (PD-L1) and predictors (the clinicopathological features).

The protein expression of *IKAROS* ranged from 0.53% to 44.1%, with a median of 18.9% and a mean of 18.0% ± 12.5. The cutoff for overall survival was 28.85%. High IKAROS expression associated with favorable prognosis. The five-years overall survival, high vs. low, 82% (95%CI 98–66%) vs. 55% (95%CI 67–43%), respectively (p = 0.034, Breslow). The correlation with the clinicopathological characteristics of the patients showed that high IKAROS associated with a GCB phenotype and with a positive clinical response to treatment. High IKAROS also correlated with a favorable progression free survival (p = 0.003) (Table 13, Figure 10).

Finally, no correlation was found between the expression of PD-L1 and IKAROS.

Table 13. Correlation between IKAROS and the clinicopathological features of the patients in the validation series (Tokai cases).

Predictors for High IKAROS	*p*-Value	Odds Ratio	95% CI for OR	
			Lower	Upper
Sex Male	0.216	0.549	0.213	1.418
Age > 60	0.768	1.173	0.405	3.4
Location				
Nodal (+spleen) (Reference)	-	-	-	-
Extranodal				
Waldeyer's ring	0.317	0.33	0.038	2.887
Gastrointestinal	0.869	1.133	0.256	5.004
Other extranodal	0.483	0.661	0.208	2.101
LDH high (>219)	0.407	0.669	0.259	1.728
Seric IL2RA high (>530)	0.189	0.481	0.162	1.433
ECOG Performance Status ≥ 2	0.632	0.711	0.176	2.873
Clinical stage III or IV	0.955	0.972	0.368	2.566
Extranodal disease site >1	0.802	0.86	0.264	2.796
B symptoms	0.635	0.739	0.213	2.566
IPI HI+H	0.731	1.206	0.414	3.512
GCB subtype (Hans's algorithm)	0.008	3.756	1.405	10.04
Epstein–Barr virus, EBER+	0.276	0.418	0.087	2.008
High RGS1 protein expression	0.112	0.459	0.176	1.199
Clinical response to treatment	0.031	9.767	1.226	77.796

Binary logistic regression setup: dependent variable (PD-L1) and predictors (the clinicopathological features).

3.3. Integration of the Data with Known Prognostic Biomarkers for the Assessment of the Overall Survival of the Patients with Diffuse Large B-Cell Lymphoma, Using Machine-Learning Analysis

Finally, the set of 16 genes, *PD-L1* (*CD274*) and *IKAROS* were merged with biomarkers known to play a role in the pathogenesis and prognosis of the patients with DLBCL, including *BCL2*, *MYC*, *CD163* and *TNFAIP8* [13]. Machine learning analysis was applied testing 11 models including C5, logistic regression, Bayesian Network, discriminant, LSVM, random trees, tree-AS, CHAID, Quest, C&R tree and neural net. After analysis, the models were ranked according to their overall accuracy (%) for prediction of the overall survival. The best models were C5 (82.126%), CHAID (81.401%) and Bayesian Network (79.286%). The result of the Bayesian Network and C5 decision tree are shown in Figures 12 and 13.

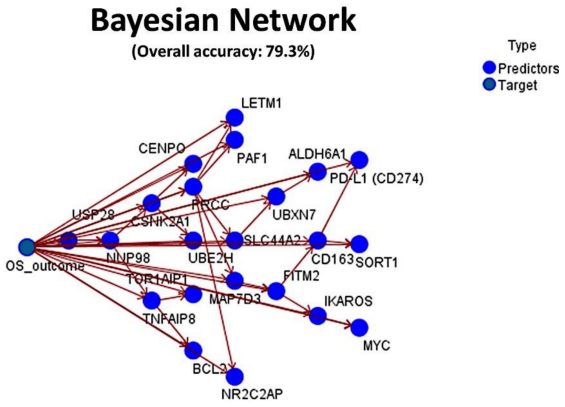

Figure 12. Final integrated Bayesian Network. The set of 16 genes, *PD-L1* (*CD274*) and *IKAROS* were merged with known biomarkers with prognostic relevance in diffuse large B-cell lymphoma (DLBCL), including *BCL2*, *MYC*, *CD163* and *TNFAIP8*. The resulting machine-learning analysis had an overall accuracy for prediction of the overall survival of 79.3%.

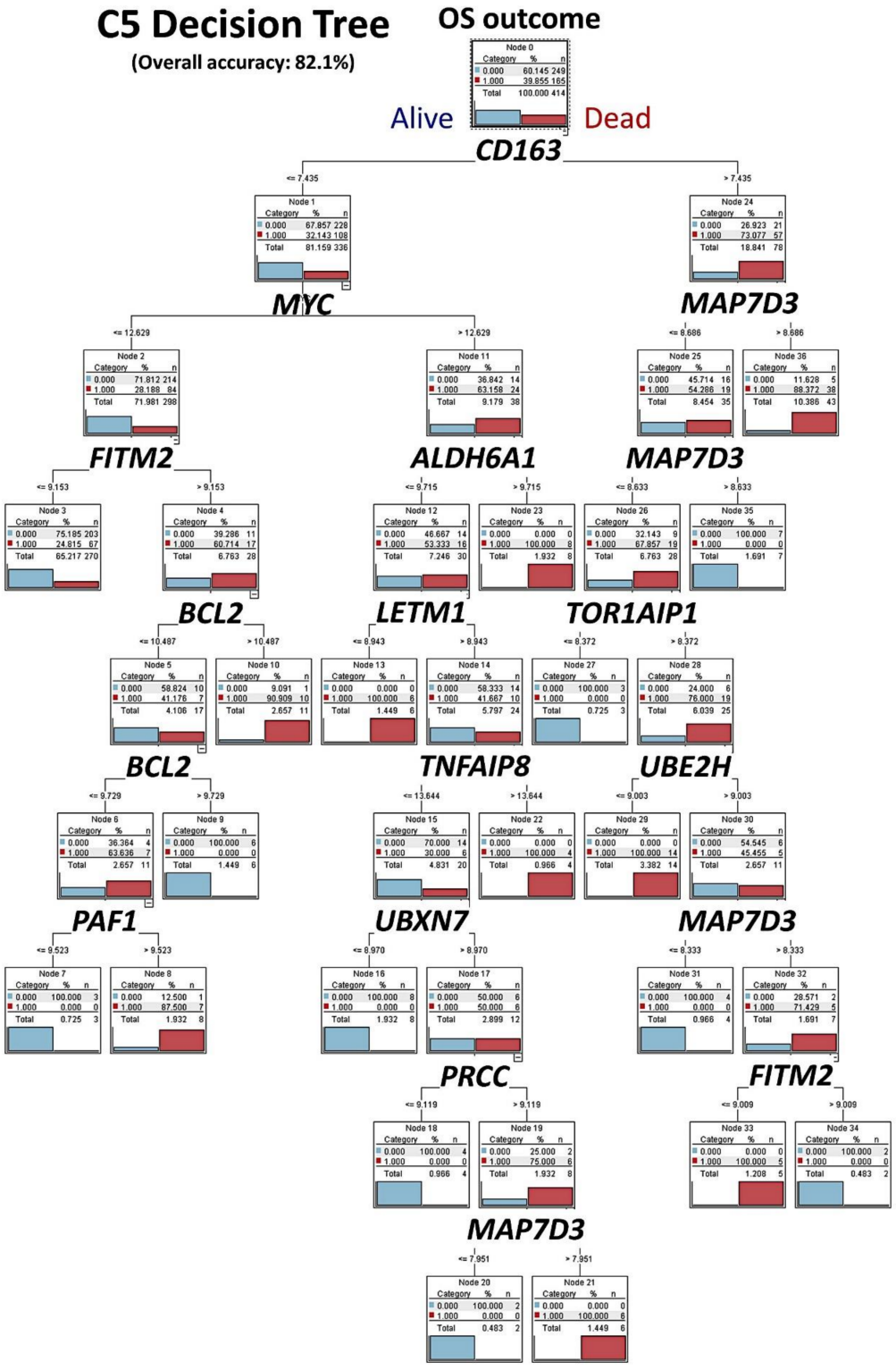

Figure 13. Final integrated C5.0 decision tree. The set of 16 genes, *PD-L1* (*CD274*) and *IKAROS* were merged with known biomarkers with prognostic relevance in diffuse large B-cell lymphoma (DLBCL), including *BCL2*, *MYC*, *CD163* and *TNFAIP8*. The resulting machine-learning analysis had an overall accuracy for prediction of the overall survival of 82.1%. Dead outcome, red color (number 1). Alive outcome, blue color (number 0).

4. Discussion

Diffuse large B-cell lymphoma is the most common subtype of non-Hodgkin lymphoma (NHL), accounting for approximately 25 percent of NHL cases. The diagnostic category of DLBCL is morphologically, genetically and biologically heterogeneous [1–5].

The molecular genetics of DLBCL have focused on the study of the cell-of-origin, which is based on the gene expression profiling (GEP). GEP is the gold standard for determining the cell-of-origin but this technique requires the use of RNA and frozen tissue. Therefore, alternative methods based on immunohistochemistry have been developed, such as the Hans classifier. The Hans classifier has a good correlation with the GEP. This classifier is an algorithm based on three markers: CD10, BCL-6 and MUM-1 (IRF4) [1–5]. Nowadays it is possible to perform GEP from paraffin-embedded formalin-fixed using the Lymph2Cx platform, which provides comparable results to the gold standard technique that is based on fresh frozen tissue [18]. In addition to the cell-of-origin analysis, the GEP has also identified different DLBCL subgroups that have distinct genetic profiles. These subtypes have been shown to influence the tumor biology and improve the prediction value of the gene-expression-based survival analysis [19]. A correlation between copy-number changes and GEP was performed and putative target genes were identified, such as *REL* and *XPO1* (2p14-p16); *PDCD10* and *TNFSF10* (3q); *PPHLN1*, *SENP1* and *MCRS1*; (12q) and *MADH2*, *MALT1* and *BCL2* (18q) [19]. The gene expression has also managed to characterize the tumoral immune microenvironment and has also enabled the prediction of patients' survival [20]. Recently, the gene expression analysis has also focused on specific subtypes such as the *IRF4*-rearranged DLBCL [21].

In this study, we have used as a discovery set the well-recognized series of GEP of DLBCL, the GSE10846 that is comprised of 414 cases. This series is relevant not only because it is has a lot of cases but also because it has served to develop the current cell-of-origin classification. In this research we have also used a validation set of 113 cases from Tokai University Hospital and for the cell-of-origin classification we have used the Hans algorithm. This algorithm is still valid in the modern rituximab-based therapy era [22]. In this research, we had the following aims: (1) to reanalyze the gene-expression data of GSE10846, using artificial intelligence (AI), based on artificial neural networks, in order to identify biomarkers; (2) to compare the efficiency between two techniques, the Multilayer Perceptron (MLP) and Radial Basis Function (RBF) networks, and to integrate the results; and (3) to validate the AI results in another series, using immunohistochemistry, by quantifying the protein expression by also using the AI-based Weka segmentation.

Artificial neural networks are the preferred tool for many predictive data-mining applications because of their power, flexibility and ease of use. Predictive artificial neural networks are particularly useful in applications where the underlying process is complex. We used both the MLP and RBF procedures. Both are supervised learning techniques as they map relationships implied by the data. Both use feedforward architectures, as the data move in only one direction, from the input nodes through the hidden layer of nodes to the output nodes. While MLP procedure can find more complex relationships, the RBF procedure is generally faster [23]. In this research we found that the performance comparison between MLP and RBF was similar in most of the parameters. Both methods managed to reduce a list of 54,613 gene-probes to a final set of 24 and 33, respectively, which accounts for more than a 99.9% reduction. Nevertheless, they differed in the activation and error functions, the number of units in the hidden layer, the training time and in the areas under the curve of the ROC analysis. In summary, we found that MLP has an overall better performance with shorter training time and a better predictive power, that is better areas under the curves. Therefore, MLP may be more appropriate for the analysis of this type of data. Both techniques managed to identify prognostically relevant markers, most of them not previously highlighted in the literature. Interestingly, a 30% of the identified genes were common between both techniques. On the other hand, a 70% of the final sets had different genes. All of them are potentially relevant and should be explored with more detail in future research. In this research we used the gene-probes without collapsing as

the start input for the artificial neural network analyses. Nevertheless, a more robust and reproducible approach would include the dataset collapsing, which could include the max, median, mean or sum of probes values.

From the final set of genes, we selected two biomarkers for validation by immunohistochemistry in another cohort from Tokai University. The two biomarkers are *CD274* (*PD-L1*) for the set of bad prognosis; and *IKZF4* (*IKAROS*) for the good prognosis set.

Programmed cell death 1 ligand 1 (PD-L1, *CD274*) plays a critical role in induction and maintenance of the immune tolerance to self [24]. As a ligand for the inhibitory receptor PDCD1/PD-1, modulates the activation threshold of T-cells and limits T-cell effector response [24]. Through a yet unknown activating receptor, PD-L1 may costimulate T-cell subsets that predominantly produce interleukin-10 (IL10) [25–27]. The PDCD1 (PD-1)-mediated inhibitory pathway is exploited by tumors to attenuate antitumor immunity and escape destruction by the immune system, thereby facilitating tumor survival [28,29]. The interaction with PD-1 inhibits cytotoxic T lymphocytes (CTLs) effector function [28]. The blockage of the PD-1-mediated pathway results in the reversal of the exhausted T-cell phenotype and the normalization of the antitumor response, providing a rationale for cancer immunotherapy [27,28]. Our data showed that a high expression of PD-L1 in DLBCL is associated to an unfavorable overall survival and progression-free survival of the patients. In addition, high PD-L1 levels also correlated with several unfavorable clinicopathological features such as a non-GCB cell-of-origin subtype (Hans's classifier), Epstein–Barr virus positivity, high RGS1 expression and IPI high/high-intermediate. Of note, our findings are in concordance with previous literature [29,30].

DNA-binding protein Ikaros (IKAROS, *IKZF1*) is a transcription regulator of hematopoietic cell differentiation. IKAROS binds gamma-satellite DNA and plays a role in the development of lymphocytes, B-cells and T-cells. IKAROS regulates transcription through association with both HDAC-dependent and HDAC-independent complexes and in adult erythroid cells increases normal apoptosis [27,31–33]. IKAROS has multiple functions in hematological malignancies (leukemia), solid tumors (lung, ovarian and colorectal cancer) and autoimmune diseases (systemic lupus erythematosus and Sjogren's syndrome) [34]. In solid cancers, high IKAROS has been associated with poor prognosis [34]. In our series of DLBCL, we have found that high IKAROS protein expression associated to a good prognosis of the patients, with a favorable overall survival and progression-free survival. In addition, high IKAROS also associated to a GCB cell of origin subtype and good clinical response to treatment.

In conclusion, artificial neural network analysis can be a useful computational tool to identify prognostic markers from gene-expression data and to quantify immunohistochemical biomarkers in the tumor samples; thus, it provides a complete tool to identify and validate diagnostic and prognostic disease-specific biomarkers. This study found that MLP is slightly more "efficient" than RBF artificial neural network, and the AI methodology identified two DLBCL prognostic biomarkers (*PD-L1* and *IKAROS*) that were validated.

Author Contributions: J.C., principal investigator, designed the project, data acquisition, performed analysis and wrote the manuscript. R.H. supervised the project and revised the paper. N.N. supervised the project and approved final submission. Y.Y.K., M.M., S.H., S.T., H.I., Y.K. and A.I. contributed to data acquisition and diagnosis of the cases. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grant KAKEN 18K15100 to Dr. Joaquim Carreras, grant-in-Aid for Early-Career Scientists from the Japanese Society for the Promotion of Science (JSPS) of the Ministry of Education, Culture, Sports, Science and Technology-Japan (MEXT). R. H is funded by Al-Jalila Foundation (AJF201741), the Sharjah Research Academy (Grant code: MED001) and University of Sharjah (Grant code: 1901090258).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board and the Ethics Committee of Tokai University, School of Medicine (protocol code IRB14R-080 and IRB-156).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The gene expression data of DLBCL (GEO dataset GSE10846) was obtained from the publicly available database of the NCBI resources webpage, located at https://www.ncbi.nlm.nih.gov/gds (accessed on 4 March 2021).

Acknowledgments: We would like to thank all the members of the Lymphoma/Leukemia Molecular Profiling Project (LLMPP) including Louis M. Staudt, Elias Campo, WC Chan and ES Jaffe (among others) for creating and sharing publicly the GSE10846 dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brown, J.R.; Freedman, A.S.; Aster, J.C. *Pathobiology of Diffuse Large B Cell Lymphoma and Primary Mediastinal Large B Cell Lymphoma; Lister, A.; Rosmarin, A.G., Eds.; Up-to-Date, Wolters Kluwer Health division of Wolters Kluwer (Philadelphia, Pennsylvania, USA) 2020*. Available online: http://www.uptodate.com (accessed on 3 April 2020).
2. Brown, J.R.; Freedman, A.S.; Aster, J.C. *Epidemiology, Clinical Manifestations, Pathologic Features, and Diagnosis of Diffuse Large B Cell Lymphoma; Lister, A., Rosmarin, A.G., Ed.; Up-to-Date, Wolters Kluwer Health division of Wolters Kluwer (Philadelphia, Pennsylvania, USA) 2020*. Available online: http://www.uptodate.com (accessed on 3 April 2020).
3. Brown, J.R.; Freedman, A.S.; Aster, J.C. *Prognosis of Diffuse Large B Cell Lymphoma*. Lister, A., Rosmarin, A.G., Eds.; Up-to-Date, Wolters Kluwer Health division of Wolters Kluwer (Philadelphia, Pennsylvania, USA) 2020. Available online: http://www.uptodate.com (accessed on 3 April 2020).
4. Morton, L.M.; Wang, S.S.; Devesa, S.S.; Hartge, P.; Weisenburger, D.D.; Linet, M.S. Lymphoma incidence patterns by WHO subtype in the United States, 1992–2001. *Blood* **2006**, *107*, 265–276. [CrossRef]
5. Swerdlow, S.H.; Campo, E.; Harris, N.L.; Jaffe, E.S.; Pileri, S.A.; Stein, H.; Thiele, J. (Eds.) *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*, revised 4th ed; International Agency for Research on Cancer (IARC): Lyon, France, 2017.
6. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
7. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
8. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
9. Deep Learning. (n.d.). Wikipedia. Available online: https://en.wikipedia.org/wiki/Deep_learning#cite_note-BENGIO2012-1 (accessed on 29 March 2020).
10. A.I. Wiki. (Chris Nicholson). Pathmind. Available online: https://pathmind.com/wiki/neural-network (accessed on 29 March 2020).
11. IBM SPSS Neural Networks 25. In *IBM SPSS Statistics 25 Documentation*; Document Number: 618179; IBM K.K.: Tokyo, Japan, 17 June 2018.
12. Carreras, J.; Hamoudi, R.; Nakamura, N. Artificial Intelligence Analysis of Gene Expression Data Predicted the Prognosis of Pa-tients with Diffuse Large B-Cell Lymphoma. *Tokai J. Exp. Clin. Med.* **2020**, *45*, 37–48.
13. Carreras, J.; Kikuti, Y.Y.; Miyaoka, M.; Hiraiwa, S.; Tomita, S.; Ikoma, H.; Kondo, Y.; Ito, A.; Shiraiwa, S.; Hamoudi, R.; et al. A Single Gene Expression Set Derived from Artificial Intelligence Predicted the Prognosis of Several Lymphoma Subtypes; and High Immunohistochemical Expression of TNFAIP8 Associated with Poor Prognosis in Diffuse Large B-Cell Lymphoma. *AI* **2020**, *1*, 23. [CrossRef]
14. Cardesa-Salzmann, T.M.; Colomo, L.; Gutierrez, G.; Chan, W.C.; Weisenburger, D.; Climent, F.; González-Barca, E.; Mercadal, S.; Arenillas, L.; Serrano, S.; et al. High microvessel density determines a poor outcome in patients with diffuse large B-cell lymphoma treated with rituximab plus chemotherapy. *Haematologica* **2011**, *96*, 996–1001. [CrossRef]
15. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018; Available online: https://www.R-project.org (accessed on 4 March 2021).
16. Hamoudi, R.A.; Appert, A.; Ye, H.; Ruskone-Fourmestraux, A.; Streubel, B.; Chott, A.; Raderer, M.; Gong, L.; Wlodarska, I.; De Wolf-Peeters, C.; et al. Differential expression of NF-kappaB target genes in MALT lymphoma with and without chromosome translocation: Insights into molecular mechanism. *Leukemia* **2010**, *24*, 1487–1497.
17. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef]
18. Scott, D.W.; Wright, G.W.; Williams, P.M.; Lih, C.-J.; Walsh, W.; Jaffe, E.S.; Rosenwald, A.; Campo, E.; Chan, W.C.; Connors, J.M.; et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood* **2014**, *123*, 1214–1217. [CrossRef]
19. Beà, S.; Zettl, A.; Wright, G.; Salaverría, I.; Jehn, P.; Moreno, V.; Burek, C.; Ott, G.; Puig, X.; Yang, L.; et al. Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood* **2005**, *106*, 3183–3190. [CrossRef]

20. Ciavarella, S.; Vegliante, M.C.; Fabbri, M.; De Summa, S.; Melle, F.; Motta, G.; De Iuliis, V.; Opinto, G.; Enjuanes, A.; Rega, S.; et al. Dissection of DLBCL microenvironment provides a gene expression-based predictor of survival applicable to formalin-fixed paraffin-embedded tissue. *Ann. Oncol.* **2018**, *29*, 2363–2370.
21. Ramis-Zaldivar, J.E.; Gonzalez-Farré, B.; Balagué, O.; Celis, V.; Nadeu, F.; Salmerón-Villalobos, J.; Andrés, M.; Martin-Guerrero, I.; Garrido-Pontnou, M.; Gaafar, A.; et al. Distinct molecular profile of IRF4-rearranged large B-cell lymphoma. *Blood* **2020**, *135*, 274–286. [CrossRef]
22. Ichiki, A.; Carreras, J.; Miyaoka, M.; Kikuti, Y.Y.; Jibiki, T.; Tazume, K.; Watanabe, S.; Sasao, T.; Obayashi, Y.; Onizuka, M.; et al. Clinicopathological Analysis of 320 Cases of Diffuse Large B-cell Lymphoma Using the Hans Classifier. *J. Clin. Exp. Hematop.* **2017**, *57*, 54–63. [CrossRef]
23. IBM SPSS Neural Networks. *New Tools for Building Predictive Models. IBM Software Business Analytics*; IBM Corporation Software Group: Somers, NY, USA, 2017.
24. Tamura, H.; Dong, H.; Zhu, G.; Sica, G.L.; Flies, D.B.; Tamada, K.; Chen, L. B7-H1 costimulation preferentially enhances CD28-independent T-helper cell function. *Blood* **2001**, *97*, 1809–1816. [CrossRef]
25. Freeman, G.J.; Long, A.J.; Iwai, Y.; Bourque, K.; Chernova, T.; Nishimura, H.; Fitz, L.J.; Malenkovich, N.; Okazaki, T.; Byrne, M.C.; et al. Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **2000**, *192*, 1027–1034.
26. Wang, S.; Bajorath, J.; Flies, D.B.; Dong, H.; Honjo, T.; Chen, L. Molecular Modeling and Functional Mapping of B7-H1 and B7-DC Uncouple Costimulatory Function from PD-1 Interaction. *J. Exp. Med.* **2003**, *197*, 1083–1091. [CrossRef]
27. The UniProt Consortium UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef]
28. Iwai, Y.; Ishida, M.; Tanaka, Y.; Okazaki, T.; Honjo, T.; Minato, N. Involvement of PD-L1 on tumor cells in the escape from host immune system and tumor immunotherapy by PD-L1 blockade. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12293–12297. [CrossRef]
29. Kataoka, K.; Shiraishi, Y.; Takeda, Y.; Sakata, S.; Matsumoto, M.; Nagano, S.; Maeda, T.; Nagata, Y.; Kitanaka, A.; Mizuno, S.; et al. Aberrant PD-L1 expression through 3′-UTR disruption in multiple cancers. *Nature* **2016**, *534*, 402–406. [CrossRef]
30. Cheng, Z.; Dai, Y.; Wang, J.; Shi, J.; Ke, X.; Fu, L. High PD-L1 expression predicts poor prognosis in diffuse large B-cell lymphoma. *Ann. Hematol.* **2018**, *97*, 1085–1088. [CrossRef]
31. Dijon, M.; Bardin, F.; Murati, A.; Batoz, M.; Chabannon, C.; Tonnelle, C. The role of Ikaros in human erythroid differentiation. *Blood* **2008**, *111*, 1138–1146. [CrossRef]
32. Ronni, T.; Payne, K.J.; Ho, S.; Bradley, M.N.; Dorsam, G.; Dovat, S. Human Ikaros function in activated T cells is regulated by coor-dinated expression of its largest isoforms. *J. Biol. Chem.* **2007**, *282*, 2538–2547.
33. Kim, J.-H.; Ebersole, T.; Kouprina, N.; Noskov, V.N.; Ohzeki, J.-I.; Masumoto, H.; Mravinac, B.; Sullivan, B.A.; Pavlicek, A.; Dovat, S.; et al. Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Res.* **2009**, *19*, 533–544. [CrossRef]
34. Chen, Q.; Shi, Y.; Chen, Y.; Ji, T.; Li, Y.; Yu, L. Multiple functions of Ikaros in hematological malignancies, solid tumor and autoimmune diseases. *Gene* **2019**, *684*, 47–52. [CrossRef]