*Article*

# Artificial Intelligence Analysis of the Gene Expression of Follicular Lymphoma Predicted the Overall Survival and Correlated with the Immune Microenvironment Response Signatures

**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, UAE

[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 or r.hamoudi@ucl.ac.uk (R.H.); Tel.: +81-0463-93-1121 (ext. 3170) (J.C.); +00971650-57758 (R.H.); Fax: +81-0463-91-1370 (J.C.); +06-5585879 (R.H.)

**Abstract:** Follicular lymphoma (FL) is the second most common lymphoma in Western countries. FL is characterized by being incurable, usually having an indolent clinical course with frequent relapses, and an eventual patient's death or transformation to Diffuse Large B-cell Lymphoma. The immune response and the tumoral immune microenvironment, including FOXP3+Tregs, PD-1+TFH cells, TNFRSF14 (HVEM), and BTLA play a role in the pathogenesis. We aimed to analyze the gene expression of FL by Artificial Intelligence (machine learning, deep learning), to identify genes associated with the prognosis of the patients and with the microenvironment in terms of overall survival (OS). A series of 184 cases of the GSE16131 dataset was analyzed by multilayer perceptron (MLP) and radial basis function (RBF) neural networks. In the analysis, MLP and RBF had a synergistic effect. From an initial set of 22,215 genes probes, a final set of 43 genes was highlighted. These 43 genes predicted the OS and correlated with the immune microenvironment: in a multivariate Cox analysis, 18 genes were associated with a poor prognosis (namely, *MED8, KRT19, CDC40, SLC24A2, PRB1, KIAA0100, EVA1B, KLK10, TMEM70, BTN2A3P, TRPM4, MED6, FRYL, CBFA2T2, RANBP9, BNIP2, PTP4A2* and *ALDH1L1*) and 25 genes were associated with a good prognosis of the patients. Gene set enrichment analysis (GSEA) confirmed these findings and showed a typical sinusoidal-like shape. Some of the most relevant genes for poor OS were *EVA1B, KRT19, BTN2A3P, KLK10, TRPM4, TMEM70,* and *SLC24A2* (hazard risk = from 1.7 to 4.3, $p < 0.005$) and for good OS, these were *TDRD12* and *ZNF230* (HR = 0.34 and 0.28, $p < 0.001$). *EVA1B, KRT19, BTN2AP3, KLK10,* and *TRPM4* also associated with M2-like macrophage markers including *CD163, MRC1 (CD206),* and *IL10* in the core enrichment for dead OS outcome by GSEA and to poor OS by Kaplan–Meier with Log rank test. The scientific literature showed that some of these genes also play a role in other types of cancer. In conclusion, by Artificial Intelligence, we have identified new biomarkers with prognostic relevance in FL.

---

## 1. Introduction

Predictive neural networks are usually the chosen tool for many data-mining applications because they can handle complex processes, such as the analysis of gene expression data of human pathological conditions. The multilayer perceptron (MLP) and radial basis function (RBF) networks are supervised methods and their architecture is feedforward [1,2]. The choice between MLP or RBF depends on the data and the level of complexity that you seek to uncover. In general, the MLP procedure can find more complex relationships, while the RBF is generally faster [3–5].

Lymphomas are a type of hematologic neoplasia of the immune system. The 2016 revision of the World Health Organization classification of lymphoid neoplasms classifies the entity of follicular lymphoma into the Mature B-cell neoplasms (as a part of the non-Hodgkin lymphomas), which includes other relevant subtypes such as the chronic lymphocytic leukemia/small lymphocytic lymphoma, splenic marginal zone lymphoma, plasma cell myeloma, extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma), mantle cell lymphoma, and diffuse large b-cell lymphoma, among others [6]. All these tumors are originated from B-lymphocytes in a defined stage of B-cell maturation and differentiation. Lymphocytes are one of the five types of the white blood cells. B-lymphocytes are produced in the bone marrow; they have a B-cell receptor specific for a specific antigen, differentiate into plasma cells, and secrete antibodies or become memory cells. Mantle cell lymphoma and diffuse large b-cell lymphoma behave aggressively. On the other hand, chronic lymphocytic leukemia and marginal zone lymphomas are relatively benign. Follicular lymphoma (FL) is a heterogeneous clinicopathological entity derived from germinal center B-lymphocytes, and it is one of most common non-Hodgkin lymphomas in Western countries. Therefore, it is clinically relevant [6,7].

Follicular lymphoma is characterized by the translocation (14;18) that results in *BCL2* overexpression due to the *IGH/BCL2* fusion gene. *BCL2* is an oncogene with anti-apoptosis function. Molecular genetics using next-generation sequencing (NGS) has shown recurrent mutations of several genes including *KMT2D*, *EZH2*, *ARID1A*, *EP300*, *MEF2B*, and *FOXO1*. The prognosis of follicular lymphoma is variable, and to date, not many biological markers have been identified to predict the prognosis of the patients [7]. The Follicular Lymphoma International Prognostic Index (FLIPI) analyzes five adverse prognostic factors to identify risk groups with significantly different overall survival. The FLIPI includes the following variables: (1) Age >60 years; (2) Stage III or IV; (3) Hemoglobin level <12.0 g/dL; (4) Number of involved nodal areas >4; and (5) Serum lactate dehydrogenase level greater than the upper limit of normal. In addition to the FLIPI, gene expression analysis highlighted the role of the tumoral immune microenvironment. An immune response type 1, rich in T cells, was associated to good prognosis. In the context of the immune response type 1, FOXP3+regulatory T lymphocytes (Tregs) and PD-1+follicular T helper cells (TFH cells) are also included. Conversely, an immune response type 2, rich in macrophages, correlated to a poor prognosis; in this context, high TNFRSF14 and low BTLA are also included [8–13].

The purpose of this work was to re-analyze the gene expression data of a robust series of 184 cases of follicular lymphoma using an Artificial Intelligence approach in order to prove that our methodological approach was feasible and to identify new genes that were associated to the prognosis of the patients. We searched for genes that correlated with the survival outcome (dead/alive) as well as other clinicopathological variables included in the FLIPI and the immune microenvironment (immune response 2 versus 1). This work is significant because to our knowledge, the use of Artificial Intelligence

is a novel approach. We found a new set of 43 genes associated with the prognosis of the follicular lymphoma patients.

## 2. Materials and Methods

### 2.1. Statistical Analysis

All the statistical analysis was performed using both R programming language with R version 3.6.3 (2020-02-29) and RStudio (version 1.3.959), and IBM SPSS (version 25; IBM, Armonk, NY, USA). The analysis was performed following the manufacturers' instructions that can be found at the following webpages: (1) https://www.r-project.org/, (2) https://rstudio.com/, and (3) https://www.ibm.com/support/pages/node/618179 [3].

Comparisons between groups were performed with crosstabulations and Chi-square tests that included the Pearson Chi-square, Likelihood ratio, and the Fisher's exact test when required; non-parametric tests included the two-independent-samples test (the Mann–Whitney U test). Survival analysis was performed with the Kaplan–Meier with Log rank, Breslow, and Tarone–Ware tests; and the Cox regression in the univariate analysis (enter method) as well as in the multivariate analysis (backward conditional method). A *p* value of less than 0.05 was considered as statistically significant. The definition of overall survival was the standard.

### 2.2. Follicular Lymphoma Gene Expression Dataset

The gene expression omnibus (GEO) series GSE16131 of follicular lymphoma [14] was downloaded from the National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/geo/). The samples corresponded to follicular lymphoma cases from lymph nodes; fresh-frozen tumor-biopsy specimens were from 184 untreated patients. The series was last updated in 10 August 2018.

The RNA had been extracted from the biopsy specimens (Fast track 2.0 mRNA isolation kit, extraction method: oligo dt cellulose; Invitrogen, Carlsbad, CA, USA). Biotinylated cRNA had been prepared according to the standard Affymetrix protocol from 1 microg mRNA (Expression analysis technical manual, 2001, Affymetrix). The RNA had been examined for gene expression with the use of Affymetrix U133A and U133B microarrays. Following fragmentation, 20 micrograms of cRNA had been hybridized for 16 hours at 45C on U133A/B GeneChips. GeneChips had been washed and stained in the Affymetrix Fluidics Station 400. Scanning had been performed by the Affymetrix 3000 Scanner. Calculation method: The data had been analyzed with Microarray suite version 5.0 (MAS 5.0) using Affymetrix default analysis settings and global scaling as the normalization method. The trimmed mean target intensity of each array was arbitrarily set to 500. The data were normalized and log2 transformed.

Definition of the immune response type 1 and 2 signatures in follicular lymphoma as originally described by Dave SS et al. published in N Engl J Med 2004 [12]. From the total number of genes of the chip, a set of 3299 genes were identified as being associated with survival. The genes that were associated with good prognosis (1568 genes) and poor prognosis (1731 genes) were hierarchically clustered separately. Five gene signatures were found among the genes predicting good prognosis, and five signatures were found among the genes predicting poor prognosis. Within each signature, the expression levels of the component genes were averaged, resulting in 10 signature averages associated with each sample. By multivariate survival analysis, it was noted that a combination of two signatures provided the better prediction. These two signatures were later named immune-response 1 and immune-response 2 based on the known function of the constituent genes. The coefficients in the final model were derived from the Cox model applied to the training set. For each sample, the final model generated a survival-predictor score according to the formula: survival-predictor score = (2.71 × immune-response 2 signature average) − (2.36 × immune-response 1 signature average) [12].

The genes of the immune-response 1 signature were *ACTN1, ATP8B2, BIN2, C1RL, C6orf37, C9orf52, CD7, CD8B1, DDEF2, DKFZP566G1424, DKFZP761D1624, FLJ32274, FLNA, FLT3LG, GALNT12, GNAQ, HCST, HOXB2, IL7R, IMAGE:5289004, INPP1, ITK, JAM, KIAA1128, KIAA1450, LEF1, LGALS2, LOC340061, NFIC, PTRF, RAB27A, RALGDS, SEMA4C, SEPW1, STAT4, TBC1D4, TEAD1, TMEPAI, TNFRSF1B, TNFRSF25, TNFSF12*, and *TNFSF13B* [12].

The genes of the immune-response 2 signature were *BLVRA, C17orf31, C1QA, C1QB, C3AR1, C4A, C6orf145, CEB1, DHRS3, DUSP3, F8, FCGR1A, GPRC5B, HOXD8, LGMN, ME1, MITF, MRVI1, NDN, OASL, PELO, SCARB2, SEPT10*, and *TLR5* [12].

The signatures in the survival model were named based on the biologic function of certain genes within each signature. The immune-response 1 signature includes genes encoding T-cell markers (e.g., *CD7, CD8B1, ITK, LEF1*, and *STAT4*) and genes that are highly expressed in macrophages (e.g., *ACTN1* and *TNFSF13B*). Notably, the immune-response 1 signature is not merely a surrogate for the number of T cells in the tumor-biopsy specimen, since many other standard T-cell genes (e.g., *CD2, CD4, LAT, TRIM*, and *SH2D1A*) were not associated with survival. The immune-response 2 signature includes genes known to be preferentially expressed in macrophages, dendritic cells, or both (e.g., *TLR5, FCGR1A, SEPT10, LGMN*, and *C3AR1*). It is noteworthy that the RNA of each sample is a mixture of the RNA of different cell populations, including the tumoral B-lymphocytes, T-lymphocytes, macrophages, and dendritic cells, among others [12].

It is noteworthy that in this research, we have used a variation to define the immune response variable. Knowing that the high immune response 2 signature would associate to poor prognosis due to the high macrophage component, we created a ratio between the gene expression values of the immune response 2 and 1 as present in the series matrix file, and the best cut-off for the overall survival was searched (which corresponded to 0.97 value).

## 2.3. Patients and Clinicopathological Characteristic

The series was comprised of 184 cases of follicular lymphoma with diagnostic biopsies (pre-treatment). Regarding the grade, the series included 152 grades 1/2 and 32 grade 3A. In 4 cases, the follow-up was missing. The follow-up time ranged from 0.01 to 19.3 years, with a mean of 7.0 years (±4.5 STD) and a median of 6.5 years. The 1, 3, 5, and 10-year overall survival was 92% (95% CI, 88–96%), 83% (77–88%), 71% (64–78%), and 50% (42–58%), respectively. At the end of the follow-up time, 92 of 180 (51.1%) had died. The immune response scores were available for all 184 cases. The immune response score 1 (representative of T-lymphocytes) ranged from 7.3 to 10.2, with a mean of 9.3 (±0.4 STD) and a median of 9.3. The immune response score 2 (representative of macrophages) ranged from 8.1 to 9.9, with a mean of 8.7 (±0.3) and median of 8.7. An immune response ratio 2:1 high (>0.97) was found in 26.1% of the cases. This cut-off of 0.97 had been found using a receiver operating characteristic (ROC) curve and the outcome (dead/alive) of the patients. The rest of the clinicopathological characteristics of the series is shown in Table 1. In summary, the clinicopathological variables that associated with a poor prognosis of the patients were age >60-years old (Hazard risk = 2.7), number of extranodal sites >1 (HR = 1.8), increased lactate dehydrogenase levels in serum (LDH; HR = 2.0), international prognostic index (IPI; HR = 3.1), and immune response ratio 2:1 high (HR = 3.3) (Table 2). In Figures 1–4, the distribution of the data of the different variables is shown, including the histograms as well as the overall survival plots. Follicular lymphoma is characterized by an indolent clinical course and frequent relapses, eventually leading to death or transformation to another lymphoma subtype with more aggressivity (Diffuse large b-cell lymphoma). We performed gene set enrichment analysis (GSEA) to identify biological pathways that could explain the pathological mechanism of survival outcome (dead/alive). First, the pan cancer immune oncology pathway was tested, and the result showed enrichment toward the dead phenotype, with the presence of markers related to macrophages. Of note, macrophages are associated to an immune response type 2 signature. Later, several individual pathways were tested, including *ATM, BCR, CASPASE, CD8, E2F, HDAC, MAPK TRK, MYC, NFKB canonical and atypical, NOTCH, RAS, RB, RET, TP53, VEGF*, and *WTN* canonical pathways. The most relevant,

which were found to be associated to a favorable outcome, were the BCR (B-cell receptor), the *MAPK TRK* and atypical *NFKB* pathways. The relevant pathways and genes will be of interest in the neural network analysis in association to the clinical variables. In summary, follicular lymphoma is an indolent hematological neoplasia with favorable prognosis. Nevertheless, some patients have a more aggressive clinical course. To date, the most important prognostic variables are the IPI and the immune response signature. Nevertheless, the pathological mechanism is still not completely understood. Therefore, there is a need to find prognostic biomarkers that can be correlated with these parameters.

### 2.4. Multilayer Perceptron and Radial Basis Function Neural Network Analysis

Multilayer perceptron and radial basis function neural networks were performed as previously described [4,5]. In summary, the analysis aimed to predict a single variable or various dependent variables based on several factors (categorial variable) or covariates (continuous variable).

In the multilayer perceptron analysis setup, the rescaling of covariates was standardized. The cases were randomly assigned based on the relative numbers of cases with a training partition of 70%, test partition of 30%, and holdout partition of 0%. The architecture was comprised of hidden layers and an output layer. The hidden layers included the setup of the number of hidden layers (one or two), the activation function (hyberbolic tangent, sigmoid), and the number of units (automatically computed or custom). The output layer was defined by the activation function (identify, softmax, hyperbolic tangent and sigmoid) and by the rescaling of scale-dependent variables (standardized, normalized, adjusted normalized, or none). Of note, the activation function chosen for the output layer determined which rescaling methods were available. The type of training could be batch, online, and mini-batch, and the optimization algorithm could use the scaled conjugate gradient or the gradient descent. The network structure included the description, the diagram, and the synaptic weights. The network performance included the model summary, the classification results, the ROC curve, cumulative gains chart, lift chart, predicted by observed chart, and residual by predicted chart. The output also included a case processing summary and the analysis of the independent variable importance. The predicted value or category and the predicted pseudo-probability for each dependent variable were saved. The synaptic weight estimates were also exported to an xml file. As stopping rules, the maximum training epochs were computed automatically, the minimum relative changes in the training error was 0.0001, and the training error ratio was 0.001.

**Table 1.** Clinicopathological characteristics of the follicular lymphoma series.

| Variable | Frequency | Percentage |
|---|---|---|
| Age > 60 years | 61 (182) | 22.5 |
| Number of extranodal sites > 1 | 24 (184) | 13 |
| LDH levels ratio > 1 | 46 (160) | 28.7 |
| Stage > 2 | 129 (180) | 71.7 |
| IPI score 2–3 | 74 (160) | 46.3 |
| Immune Response ratio 2:1 high (≥0.97) | 48 (184) | 26.1 |
| With translocation (14;18) positive | 147 (164) | 89.6 |
| Survival outcome dead | 92 (180) | 51.1 |
| Survival dead before 5 years | 35 (84) | 41.7 |
| Survival alive from 10 years | 49 (84) | 58.3 |

In the radial basis function setup, the parameters are similar to the multilayer perceptron but the activation function for the hidden layer can be a normalized radial basis function or an ordinary radial basis function, and the overlap among hidden units can be automatically computed or specified.

## Clinicopathological characteristics



**Figure 1.** Clinicopathological characteristics of the series. The histograms show the distribution of the data for each of the clinical variables including the immune response gene signature 2:1 ratio and the presence of translocation (14;18) involving the IGH/BCL2 genes.

The analysis follows a series of steps. (1) The multilayer perceptron and radial basis function neural networks analysis were performed using the gene expression data of 184 patients with follicular lymphoma. The input data were the 22,215 gene probes (the predictors, covariates) and the predicted variables (the dependent variables) were the survival output (dead/alive) as well as the other relevant clinicopathological variables. The genes were ranked according to their normalized importance for prediction. The genes with more than 70% of normalized importance were selected and pooled. The genes above 1% of averaged normalized importance were also selected and pooled. (2) The final list of genes was subjected to a univariate Cox regression analysis for the prediction of survival outcome (dead/alive, enter method). Each gene was used in the univariate Cox as a quantitative

variable (predictor). Subsequently, a multivariate Cox regression analysis was performed only with the significant genes of the univariate analysis. In this multivariate analysis, the method was the backward conditional. (3) The final list of significant genes from the multivariate analysis was subjected to a second round of multilayer perceptron analysis to confirm their association with the most relevant variables, the survival outcome, the international prognostic index, and the immune microenvironment immune response.
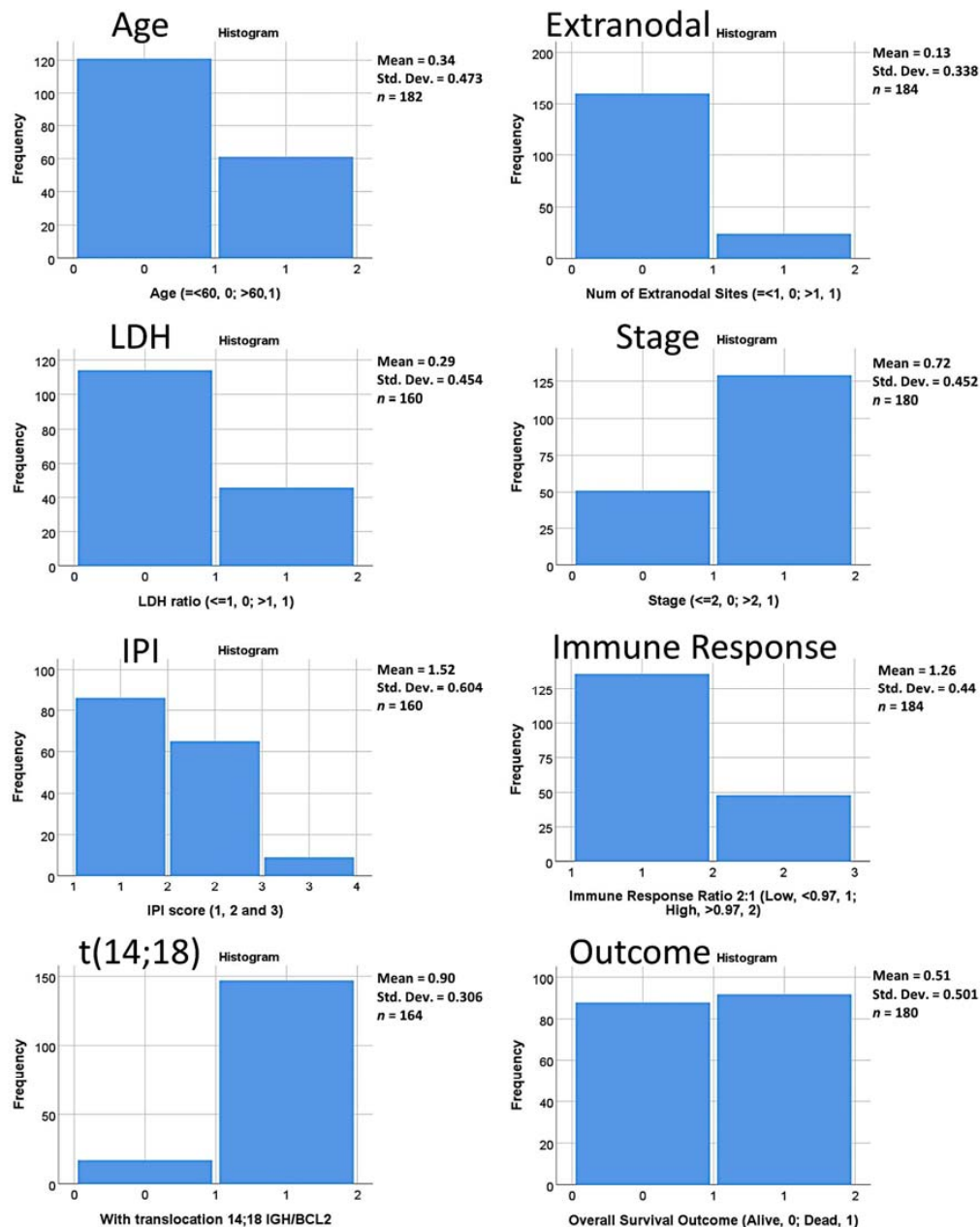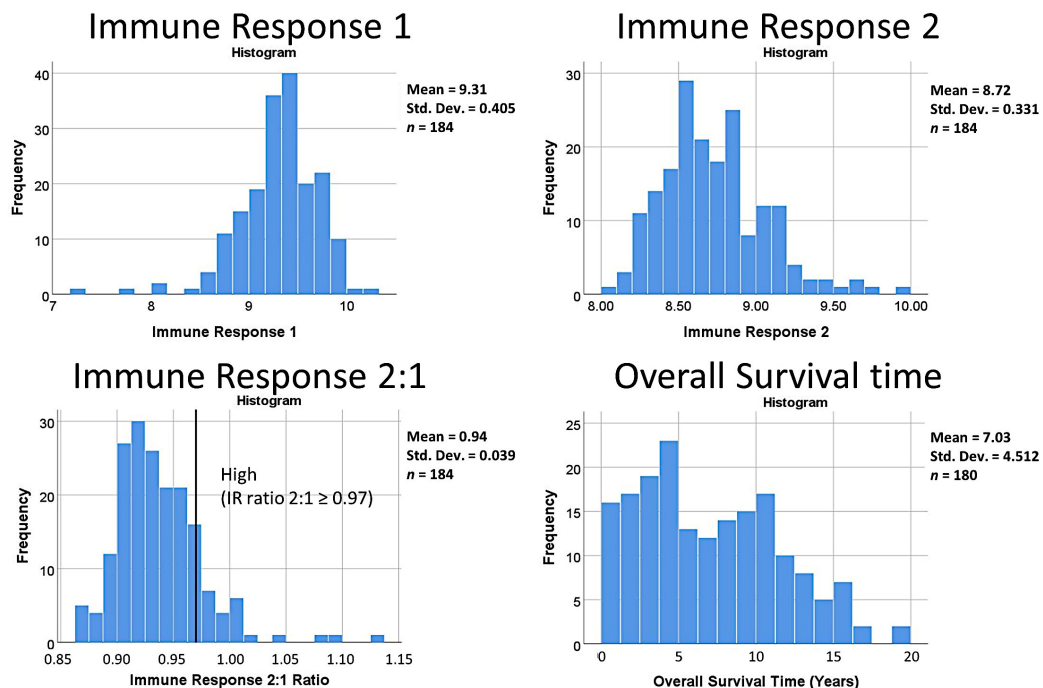
## Clinicopathological characteristics



**Figure 2.** Clinicopathological characteristics of the series. The histograms show the distribution of the data for the immune response variables (1, 2 and the 2:1 ratio). The immune response variables are based on the gene expression of markers that are characteristic of the tumoral immune microenvironment and host immune response (T cells, macrophages, etc.), as originally described by Dave SS et al. and published in *N Engl J Med* 2004 [12]. In this project, we created an immune response 2:1 ratio and the cut-off value of 0.97 was use in the overall survival analysis. The overall survival time of the series is also shown. The distribution and correlation of the different variables is also shown in the hierarchical clustering with dendrogram and heatmap.

# Clinicopathological characteristics
# Overall Survival
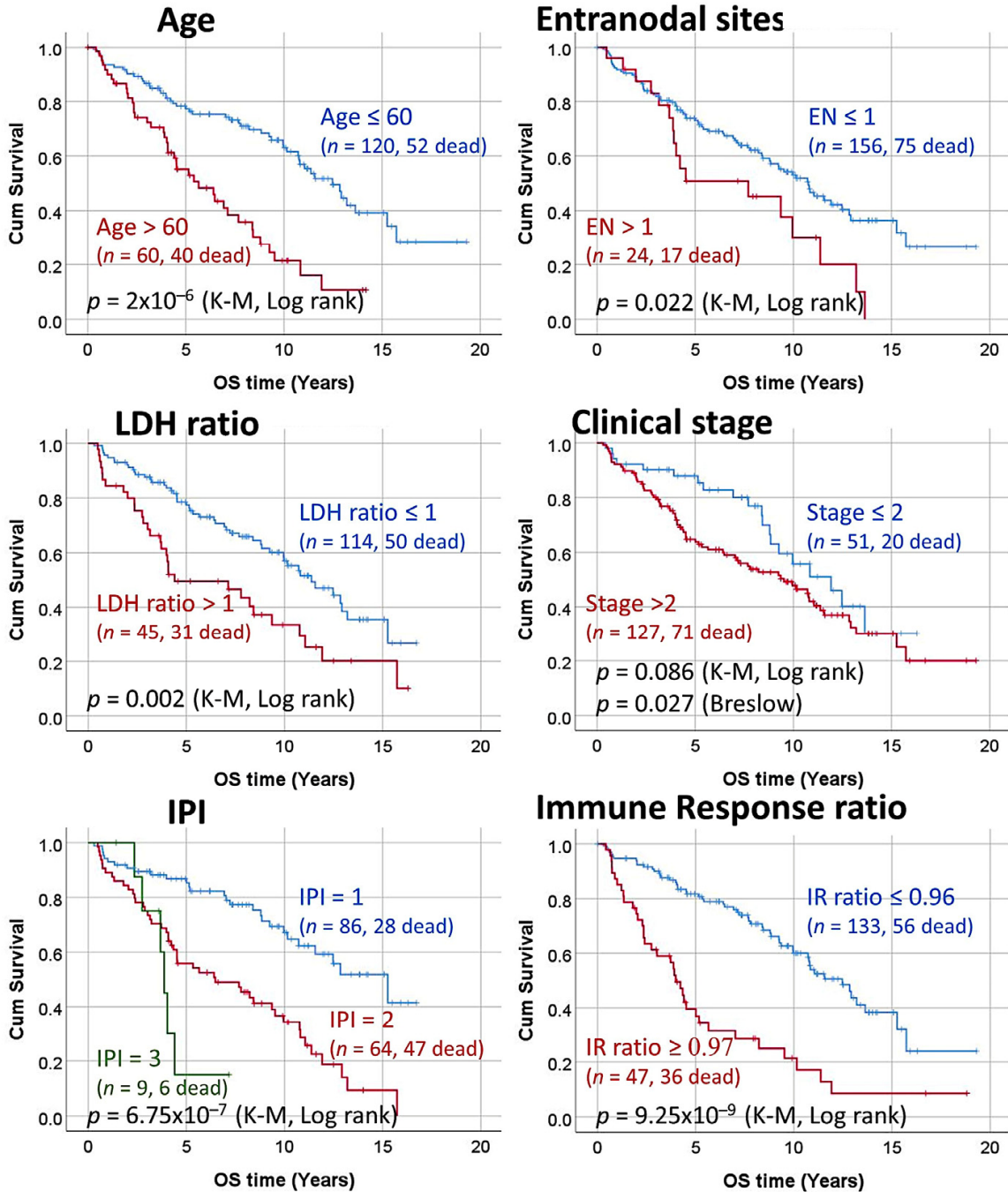


**Figure 3.** Clinicopathological characteristics of the series. Correlation between the clinicopathological characteristics and the overall survival of the patients. The clinical evolution of FL is usually indolent but heterogeneous. Some clinical variables such as the IPI (FLIPI) and the immune response allows stratifying the patients according to their risk.

# Gene set enrichment analysis (GSEA)
# for OS outcome
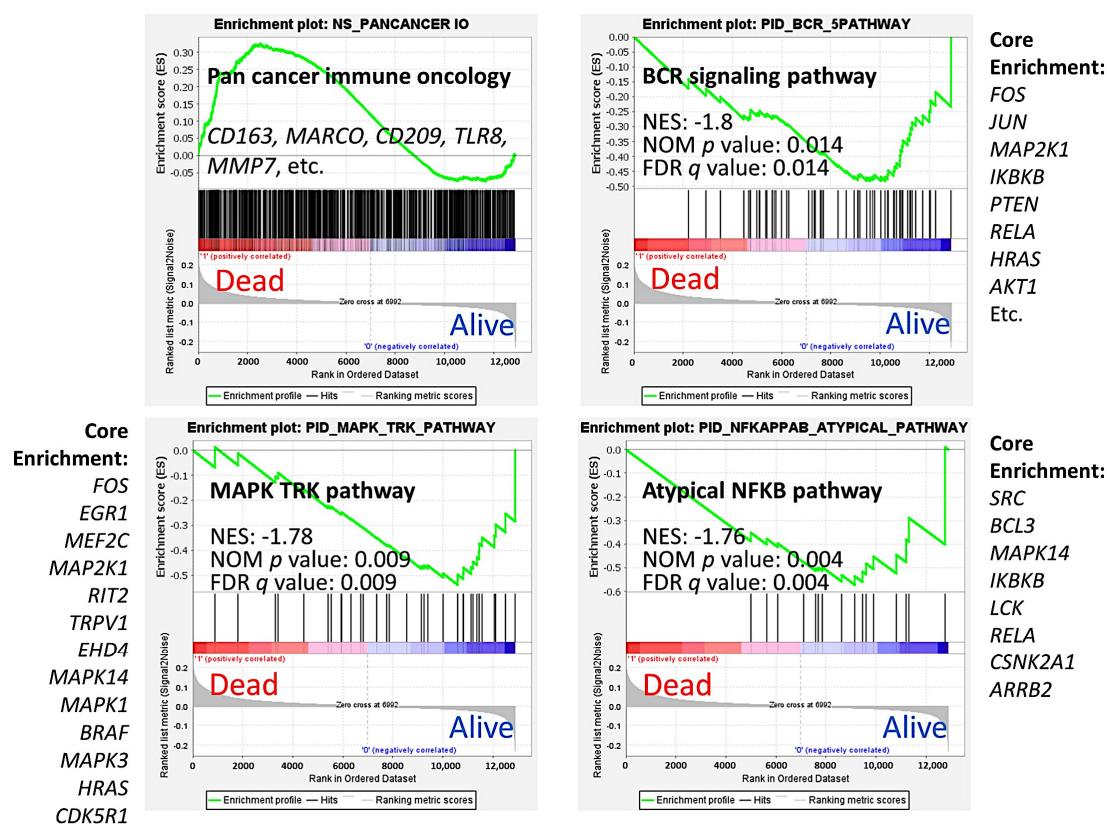


**Figure 4.** Clinicopathological characteristics of the series: Gene set enrichment analysis (GSEA). GSEA was performed to identify pathways related to the overall survival outcome (dead/alive). A pan cancer immune oncology pathway was used to identify which markers were potentially responsible for a poor outcome (dead phenotype). This pathway included genes involved in the complex interplay between the tumoral B-lymphocytes, the tumor immune microenvironment, and the host immune response. The plot showed an enrichment toward the dead phenotype with the presence of genes related to macrophages. Other pathways that were highlighted were the B-cell receptor (BCR) signaling, mitogen-activated protein kinase (MAPK)—tyrosine receptor kinase (TRK), and the atypical nuclear factor-kappa B (NFKB) pathways, which were associated to a favorable outcome.

## 2.5. Gene Set Enrichment Analysis

The gene set enrichment analysis (GSEA) was performed using the GSEA software from the UC San Diego, Broad Institute (version 4.1.0) following the manufacturer's instructions. In summary, three types of files were created: (1) the gene cluster text file (*.gct) that contains the gene expression data of the GSE16131 GPL96 follicular lymphoma series ($n$ = 180); (2) the phenotype data as a categorical class (dead/alive) file format (*.cls); and (3) the gene set database as a gene matrix file format (*.gmx). When running the GSEA, the following parameters were used: number of permutations (1000), collapse to gene symbols, permutation type (phenotype), chip platform (Affymetrix HG U133), enrichment statistic (weighted), metric for ranking genes (signal2noise), gene list sorting mode (real), gene list ordering mode (descending), max size (500), and min size (15). The Java web start option (1 GB for 64-bit) was launched using Java 8 (version 261, build 1.8.O_261-b12, Oracle Corporation, 500 Oracle Parkway, Redwood Shores CA, 94065, USA).

**Table 2.** Overall survival analysis of the clinicopathological variables.

| Variable | Log Rank *p* Value | Cox *p* Value | Hazard Risk | 95% CI for HR | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| Age > 60 years | $2 \times 10^{-6}$ | $5 \times 10^{-6}$ | 2.7 | 1.8 | 4.2 |
| Number of extranodal sites > 1 | 0.022 | 0.025 | 1.8 | 1.1 | 3.1 |
| LDH levels ratio > 1 | 0.002 | 0.002 | 2.0 | 1.3 | 3.2 |
| Stage > 2 [1] | 0.086 | 0.088 | 1.5 | 0.9 | 2.5 |
| IPI score 2–3 | $4 \times 10^{-7}$ | $2 \times 10^{-6}$ | 3.1 | 2.0 | 5.0 |
| With translocation (14;18) positive | 0.249 | 0.253 | 1.6 | 0.7 | 3.7 |
| Immune Response ratio 2:1 high (≥0.97) | $9 \times 10^{-9}$ | $5.3 \times 10^{-8}$ | 3.3 | 2.1 | 5.0 |
| Survival: Dead up to 5-y vs. Alive from 10-years | $5 \times 10^{-21}$ | $1.7 \times 10^{-5}$ | 209.2 | 18.3 | 2393.0 |

[1] In the Kaplan–Meier survival analysis of the variable stage, the *p* = 0.027 in the Breslow test and *p* = 0.04 in the Tarone–Ware test.

## 3. Results

### 3.1. Multilayer Perceptron and Radial Basis Function Neural Network Analysis

In Tables 3 and 4, all the data of the results of the MLP and RBF analyses are shown. The results include the case processing summary, the network information, the hidden layer, the output layer, the model summary, the model summary testing, the classification, and the area under the curve. The results are individualized for each of the dependent variables, as shown in the Figure 5, including survival outcome, age, extranodal sites, lactate dehydrogenase (LDH), clinical stage, international prognostic index (IPI), immune response ratio (2:1), the status of the presence of t(14;18), and the combined variable. The MLP analysis used the hyperbolic tangent activation function in the hidden layer, and the output layer used the softmax activation function and cross-entropy error function. Conversely, the RBF used the softmax activation function in the hidden layer and the identity activation function and sum of squares error function in the output layer. In both the training model and in the testing model, the error of the MLP was higher than in the RBF ($p < 0.05$). Conversely, the training time was higher in the RBF analysis: MLP vs. RBF, training time in seconds, 11.9 ± 20.9 vs. 384.3 ± 168.0% ($p = 4.871 \times 10^{-4}$), respectively. Importantly, the area under the curve of the ROC curve was higher in the MLP than in the RBF: 0.76 vs. 0.62 ($p = 4.871 \times 10^{-4}$). Therefore, the results indicate that the MLP is slightly more efficient than the RBF neural network. From an initial set of 22,215 gene probes, after several neural network analyses, the threshold for the 70% normalized importance and 1% averaged normalized importance, the pooling and deletion of duplicates, we finished with a set of 1005 genes. This approximately accounts for a 20-times reduction (Figure 5).

### 3.2. Univariate and Multivariate Cox Survival Analysis

A univariate overall survival analysis with the Cox regression analysis was performed in each of the 1005 genes. In each regression analysis, the gene expression values were analyzed as a continuous variable. As a result, the list was reduced to 89 genes. In order to highlight the most important, a survival analysis with a multivariate Cox regression analysis (backward conditional method) was performed, and in the last step, the list of the most relevant genes was reduced to 43 genes: 18 were associated to a poor prognosis (Table 5) and 25 were associated to a good prognosis of the patients (Table 6). In the tables, the genes are ranked according to their hazard ratio for risk of survival (dead/alive).

Each clinicopathological-dependent variable had a contribution of gene probes in the final set of 1005 genes, which was variable in case of MLP or RBF neural network as well as for Normalized Importance (NI) >70% or Averaged Normalized Importance. In the case of only gene probes with a normalized importance >70% and MLP analysis, the contribution was as follows: OS, 8 gene probes; Age, 47 gene probes; Extranodal, 3 gene probes; LDH, 32 gene probes; Stage, 41; IPI, 4; Immune Response, 8; OS 5-10y, 84; combined variables, 451. In case of RBF analysis, the contribution (also only gene probes with a normalized importance >70%) was as follows: OS, 56 gene probes; Age, 50 gene probes;

Extranodal, 52 gene probes; LDH, 184 gene probes; Stage, 24; IPI, 66; Immune Response, 64; OS 5-10y, 85; combined variables, 451. In summary, MLP contributed in 678 gene probes into the pooling, and RBF contributed in 1032 gene probes.

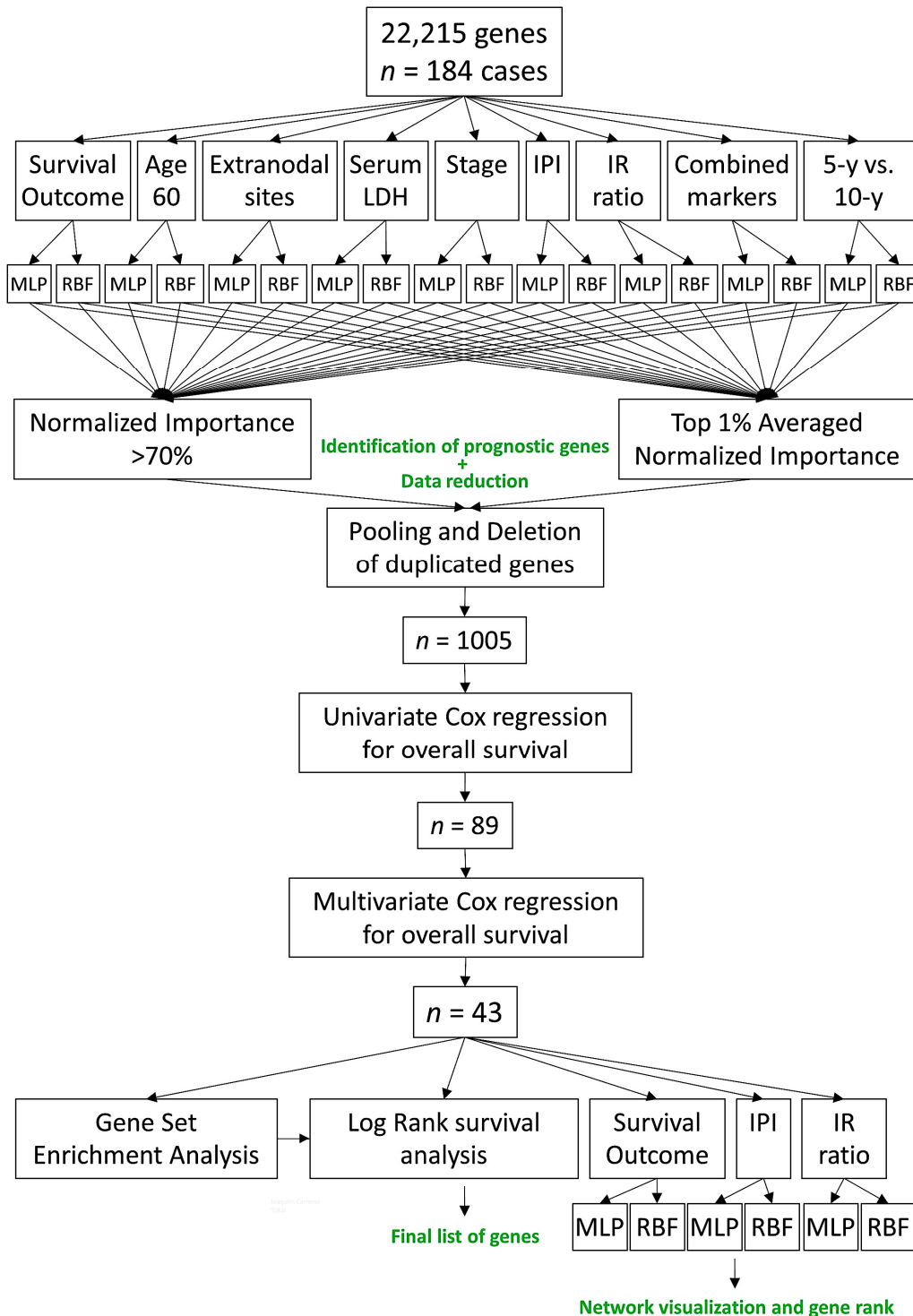**Figure 5.** Artificial Intelligence analysis of follicular lymphoma. This figure shows the method used to analyze the gene expression data, which included a multistep process. From an initial set of 22,215 gene probes, a final set of 43 genes were identified, of which 18 were associated to a poor overall survival of the patients and 25 were associated to a good overall survival of the patients.

**Table 3.** Multilayer perceptron analysis (MLP).

| Multilayer Perceptron | Dependent Variable | Survival Outcome | Age 60 | Extranodal Sites | LDH | Stage | IPI | IR 2:1 Ratio | Translocation | Combined | 5 vs. 10-y | Mean | STD | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case processing summary | Training | 129 | 116 | 137 | 108 | 130 | 109 | 132 | 109 | 44 | 59 | 107.1 | 33.3 | 116 |
| | Training Percentage | 71.7 | 63.7 | 74.5 | 67.5 | 72.2 | 68.1 | 71.7 | 66.5 | 71 | 69.4 | 70.0 | 3.2 | 71 |
| | Testing | 51 | 66 | 47 | 52 | 50 | 51 | 52 | 55 | 18 | 26 | 45.9 | 14.7 | 51 |
| | Testing Percentage | 28.3 | 36.3 | 25.5 | 32.5 | 27.8 | 31.9 | 28.3 | 33.5 | 29 | 30.6 | 30.0 | 3.2 | 29 |
| | Valid | 180 | 182 | 184 | 160 | 180 | 160 | 184 | 164 | 62 | 85 | 153 | 46.4 | 180 |
| | Excluded | 4 | 2 | 0 | 24 | 4 | 24 | 0 | 20 | 122 | 99 | 31 | 46.4 | 4 |
| | Total | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 0 | 184 |
| Network information | Num of Units | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | - | - | - |
| | Rescaling Method for covariates | | | | | Standarized | | | | | | - | - | - |
| Hidden layer | Num Hidden Layers | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| | Num Units in Hidden Layer | 9 | 8 | 5 | 15 | 10 | 8 | 11 | 6 | 8 | 8 | 9.1 | 2.8 | 8 |
| | Activation Function | | | | | Hyperbolic tangent | | | | | | - | - | - |
| Output layer | Dep Variable | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 9 | 1 | - | - | - |
| | Num of Units | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 18 | 2 | 3.8 | 5.3 | 2 |
| | Activation Function | | | | | Softmax | | | | | | - | - | - |
| | Error Function | | | | | Cross-entropy | | | | | | - | - | - |
| Model summary training | Cross Entropy Error | 87.8 | 70.7 | 50.9 | 51.4 | 70.8 | 70.5 | 59.3 | 30.2 | 173.5 | 30.0 | 73.9 | 40.8 | 70.5 |
| | Percent of Incorrect Predictions | 40.3 | 35.3 | 14.6 | 22.2 | 23.8 | 33.0 | 16.7 | 11.0 | 19.9 | 22.0 | 25.3 | 8.8 | 22.2 |
| | Stopping Rule Used * | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | - | - | - |
| | Time in Seconds | 84.8 | 83.5 | 101.2 | 71.7 | 99.2 | 79.1 | 97.9 | 76.4 | 41.9 | 44.3 | 78.2 | 22.2 | 83.5 |
| Model sum. testing | Cross Entropy Error | 28.1 | 32.7 | 11.3 | 27.8 | 24.0 | 30.2 | 23.9 | 9.2 | 84.2 | 13.9 | 30.7 | 21.3 | 27.8 |
| | Percent Incorrect Predictions | 23.5 | 22.7 | 8.5 | 23.1 | 18.0 | 33.0 | 23.1 | 7.3 | 25.9 | 26.9 | 22.7 | 6.7 | 23.1 |
| Classification | Training Overall Percent | 59.7 | 64.7 | 85.4 | 77.8 | 76.2 | 67.0 | 83.3 | 89.0 | 80.1 | 78.0 | 74.7 | 8.8 | 77.8 |
| | Testing Overall Percent | 76.5 | 77.3 | 91.5 | 76.9 | 82.0 | 66.7 | 76.9 | 92.7 | 74.1 | 73.1 | 77.2 | 6.8 | 76.9 |
| Area under the curve | Alive | 0.7 | 0.7 | 0.8 | 0.8 | 0.7 | 0.7 | 0.8 | 0.9 | 0.8 | 0.8 | 0.8 | 0.0 | 0.8 |
| | Dead | 0.7 | 0.7 | 0.8 | 0.8 | 0.7 | 0.7 | 0.8 | 0.9 | 0.8 | 0.8 | 0.8 | 0.0 | 0.8 |

LDH, lactate dehydrogenase; IPI, international prognostic index; IR, immune response (type 2 and type 1); STD, standard deviation. * Consecutive with no decrease in error.

**Table 4.** Radial basis function analysis (RBF).

| Radial Basis Function | Dependent Variable | Survival Outcome | Age 60 | Extranodal Sites | LDH | Stage | IPI | IR 2:1 Ratio | Translocation | Combined | 5 vs. 10-y | Mean | STD | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case processing summary | Training | 127 | 123 | 125 | 116 | 127 | 109 | 134 | 119 | 47 | 66 | 108.2 | 30.5 | 123 |
| | Training Percentage | 70.6 | 67.6 | 67.9 | 72.5 | 70.6 | 68.1 | 72.8 | 72.6 | 75.8 | 77.6 | 71.5 | 3.5 | 70.6 |
| | Testing | 53 | 59 | 59 | 44 | 53 | 51 | 50 | 45 | 15 | 19 | 44.8 | 16.4 | 51 |
| | Testing Percentage | 29.4 | 32.4 | 32.1 | 27.5 | 29.4 | 31.9 | 27.2 | 27.4 | 24.2 | 22.4 | 28.5 | 3.5 | 29.4 |
| | Valid | 180 | 182 | 184 | 160 | 180 | 160 | 184 | 164 | 62 | 85 | 153 | 46.4 | 180 |
| | Excluded | 4 | 2 | 0 | 24 | 4 | 24 | 0 | 20 | 122 | 99 | 31 | 46.4 | 4 |
| | Total | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 184 | 0 | 184 |
| Network information | Num of Units | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | 22,215 | - | - | - |
| | Rescaling Method for Covariates | | | | | | Standarized | | | | | - | - | - |
| Hidden layer | Num Hidden Layers | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| | Num Units in Hidden Layer | 2 | 2 | 3 | 7 | 6 | 2 | 10 | 2 | 3 | 2 | 4.1 | 2.9 | 3 |
| | Activation Function | | | | | | Softmax | | | | | - | - | - |
| Output layer | Dep Variable | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 9 | 1 | - | - | - |
| | Num of Units | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 18 | 2 | 3.8 | 5.3 | 2 |
| | Activation Function | | | | | | Identity | | | | | - | - | - |
| | Error Function | | | | | | Sum of Squares | | | | | - | - | - |
| Model summary training | Sum of Squares Error | 31.7 | 26.7 | 12.1 | 25.5 | 27.6 | 27.2 | 20.1 | 9.7 | 72.2 | 14.7 | 28.7 | 17.5 | 26.7 |
| | Percent of Incorrect Predictions | 47.2 | 32.5 | 11.2 | 36.2 | 33.1 | 47.7 | 21.6 | 9.2 | 26.5 | 36.4 | 32.5 | 11.6 | 33.1 |
| | Stopping Rule Used | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | Time in Seconds | 487.5 | 481.9 | 451.5 | 436.2 | 477.2 | 336.7 | 587.6 | 406.0 | 84.4 | 94.4 | 381.9 | 178.0 | 451.5 |
| Model sum. testing | Sum of Squares Error | 13.3 | 12.8 | 8.5 | 6.9 | 8.8 | 12.8 | 6.6 | 5.2 | 29.8 | 4.6 | 11.6 | 7.5 | 8.8 |
| | Percent Incorrect Predictions | 56.6 | 35.6 | 16.9 | 11.4 | 17.0 | 56.9 | 18.0 | 13.3 | 35.6 | 36.8 | 31.6 | 17.1 | 35.6 |
| Classification | Training Overall Percent | 52.8 | 67.5 | 88.8 | 63.8 | 66.9 | 52.3 | 78.4 | 90.8 | 73.5 | 63.6 | 67.5 | 11.6 | 66.9 |
| | Testing Overall Percent | 43.4 | 64.4 | 83.1 | 88.6 | 83.0 | 43.1 | 82.0 | 86.7 | 64.4 | 63.2 | 68.4 | 17.1 | 64.4 |
| Area under the curve | Alive | 0.5 | 0.6 | 0.6 | 0.6 | 0.6 | 0.5 | 0.8 | 0.6 | 0.7 | 0.7 | 0.6 | 0.1 | 0.6 |
| | Dead | 0.5 | 0.6 | 0.6 | 0.6 | 0.6 | 0.5 | 0.8 | 0.6 | 0.7 | 0.7 | 0.6 | 0.1 | 0.6 |

LDH, lactate dehydrogenase; IPI, international prognostic index; IR, immune response (type 2 and type 1); STD, standard deviation.

**Table 5.** Genes associated to poor prognosis in the multivariate Cox survival analysis.

| Gene | B | *p* Value | HR | HR Lower | HR Higher | Cytoband |
|---|---|---|---|---|---|---|
| *FRYL* | 1.84 | 0.00963 | 6.27 | 1.56 | 25.16 | 4p11 |
| *KIAA0100* | 1.82 | 0.00005 | 6.16 | 2.55 | 14.90 | 17q11.2 |
| *CDC40* | 1.64 | 0.00004 | 5.13 | 2.34 | 11.24 | 6q21 |
| *MED8* | 1.57 | $1.6 \times 10^{-8}$ | 4.79 | 2.78 | 8.25 | 1p34.2 |
| *PTP4A2* | 1.51 | 0.05252 | 4.55 | 0.98 | 21.00 | 1p35 |
| *BNIP2* | 1.48 | 0.04624 | 4.38 | 1.02 | 18.73 | 15q22.2 |
| *TMEM70* | 1.46 | 0.00308 | 4.30 | 1.64 | 11.31 | 8q21.11 |
| *MED6* | 1.41 | 0.00627 | 4.08 | 1.49 | 11.20 | 14q24.2 |
| *SLC24A2* | 1.39 | 0.00005 | 4.03 | 2.06 | 7.89 | 9p22.1 |
| *KLK10* | 1.34 | 0.00265 | 3.81 | 1.59 | 9.13 | 19q13 |
| *RANBP9* | 1.29 | 0.01825 | 3.63 | 1.24 | 10.60 | 6p23 |
| *PRB1* | 1.08 | 0.00005 | 2.94 | 1.74 | 4.95 | 12p13.2 |
| *EVA1B* | 1.00 | 0.00041 | 2.71 | 1.56 | 4.72 | 1p34.3 |
| *CBFA2T2* | 0.99 | 0.01269 | 2.69 | 1.24 | 5.86 | 20q11 |
| *ALDH1L1* | 0.74 | 0.08266 | 2.09 | 0.91 | 4.80 | 3q21.3 |
| *KRT19* | 0.71 | 0.00002 | 2.04 | 1.47 | 2.81 | 17q21.2 |
| *BTN2A3P* | 0.71 | 0.00320 | 2.03 | 1.27 | 3.25 | 6p22.1 |
| *TRPM4* | 0.56 | 0.00449 | 1.76 | 1.19 | 2.60 | 19q13.33 |

This contribution may be of interest if we want to know which markers are associated to each of the dependent variables as well as to know which variables are the most relevant for predicting the overall survival of the patients in terms of gene-probes contribution.

For example, if we focus on the last step of multivariate Cox survival analysis in which a final set of 43 genes was identified, the MLP and RBF contributions are variable depending on the type of neural network and the variable tested. In general, RBF neural network contributed more frequently to the identification of the final set of genes than the MLP (35 vs. 28, respectively). Nevertheless, in the averaged normalized importance variable, the MLP identified more genes than the RBF (17 vs. 14, respectively). In summary, the combination of both MLP and RBF produces a synergic effect, the combined variable is the most efficient of the individual variables, and the averaged normalized importance is the most efficient of all. The contribution of each variable is shown in Figure 6.

### 3.3. Gene Set Enrichment Analysis (GSEA)

GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g., phenotypes). In our case, the two biological states were the outcome status of dead vs. alive. GSEA was performed on the set of 43 genes, and the result was a sinusoid-like plot, with some genes associated to good and other to poor prognosis. This is the same result as in the multivariate Cox regression analysis for overall survival. The GSEA association was also confirmed for the set of 18 genes (good prognosis) and the set of 25 (good). In addition, we included in the GSEA immune genes, oncogenes, and tumor-suppressor genes known to play a role in the pathogenesis of follicular lymphoma. All the results, including the genes of the core enrichment, are shown in Figure 7.

### 3.4. Univariate Survival Analysis with Kaplan–Meier and Log Rank Test

Based on the results of the GSEA analysis, a small group of genes was selected, and a cut-off of gene expression was searched (Figure 8). We confirmed that a high expression of *EVA1B*, *KRT19*, *BTN2A3P*, *KLL10*, and *TRPM4* was associated to a poor OS of the patients, while high *TDRD12* and *ZNF230* was associated to a favorable OS ($p < 0.05$).

**Table 6.** Genes associated to good prognosis in the multivariate Cox survival analysis.

| Gene | B | *p* Value | HR | HR Lower | HR Higher | Cytoband |
|------|------|-----------|------|----------|-----------|----------|
| HSF2 | −0.48 | 0.04873 | 0.62 | 0.38 | 1.00 | 6q22.31 |
| ATPAF2 | −0.48 | 0.04152 | 0.62 | 0.39 | 0.98 | 17p11.2 |
| SLC7A11 | −0.51 | 0.00208 | 0.60 | 0.43 | 0.83 | 4q28.3 |
| PTAFR | −0.64 | 0.00060 | 0.53 | 0.37 | 0.76 | 1p35-p34.3 |
| TTLL3 | −0.74 | 0.01720 | 0.48 | 0.26 | 0.88 | 3p25.3 |
| TCP10L | −0.75 | 0.03536 | 0.47 | 0.24 | 0.95 | 21q22.11 |
| DNAAF1 | −0.8 | 0.00807 | 0.45 | 0.25 | 0.81 | 16q24.1 |
| PRH1 | −0.85 | $1 \times 10^{-5}$ | 0.43 | 0.29 | 0.62 | 12p13.2 |
| NSDHL | −0.89 | 0.04102 | 0.41 | 0.17 | 0.96 | Xq28 |
| TAF12 | −0.99 | 0.01139 | 0.37 | 0.17 | 0.80 | 1p35.3 |
| TSPAN3 | −1 | 0.00040 | 0.37 | 0.21 | 0.64 | 15q24.3 |
| AKIRIN1 | −1.03 | 0.00195 | 0.36 | 0.19 | 0.69 | 1p34.3 |
| ITK | −1.04 | 0.00102 | 0.35 | 0.19 | 0.66 | 5q31-q32 |
| TDRD12 | −1.09 | 0.00392 | 0.34 | 0.16 | 0.70 | 19q13.11 |
| LPP | −1.12 | 0.00097 | 0.33 | 0.17 | 0.63 | 3q28 |
| BTD | −1.13 | $9.5 \times 10^{-6}$ | 0.32 | 0.20 | 0.53 | 3p25 |
| SIRT5 | −1.22 | 0.04956 | 0.30 | 0.09 | 1.00 | 6p23 |
| ZNF230 | −1.29 | 0.00002 | 0.27 | 0.15 | 0.50 | 19q13.31 |
| ABHD6 | −1.38 | $7.2 \times 10^{-5}$ | 0.25 | 0.13 | 0.50 | 3p14.3 |
| TOP2B | −1.49 | 0.01673 | 0.23 | 0.07 | 0.76 | 3p24 |
| ARPC2 | −1.7 | 0.00804 | 0.18 | 0.05 | 0.64 | 2q36.1 |
| ASAP2 | −1.96 | 0.00003 | 0.14 | 0.06 | 0.36 | 2p25|2p24 |
| IDH3A | −2.03 | 0.00009 | 0.13 | 0.05 | 0.36 | 15q25.1-q25.2 |
| PSMF1 | −2.44 | 0.00415 | 0.09 | 0.02 | 0.46 | 20p13 |
| ARFGEF1 | −2.69 | 0.00000 | 0.07 | 0.02 | 0.20 | 8q13 |

### 3.5. Final MLP and RBF Neural Network Analysis

Based only on the set of 43 genes and the 184 patients, a final MLP and RBF neural network analysis was formed. The aim was to confirm the quality control parameters of our model using the ROC curve, the cumulative gains chart, and the lift chart. In addition, the 43 were ranked according to their normalized importance to predict the OS of the patients, the immune response, and the IPI. The results are shown in Figures 9–11.

### 3.6. Correlation between the 7 Highlighted Genes and the Clinicopathological Characteristics of the Patients

In this project, we identified seven genes that were associated with the overall survival of the patients. In five genes (*EVA1B*, *KRT19*, *BTN2A3P*, *KLK10*, and *TRPM4*), high expression correlated to a poor overall survival; and in two genes (*TDRD12* and *ZNF230*), high expression correlated to a favorable overall survival. In order to understand the reason for this association, we correlated using simple crosstabulations with the clinicopathological characteristics of the patients. The associations can be seen in Table 7.

**Table 7.** Clinicopathological correlations with the final set of seven prognostic genes.

| Crosstabulations | OS Prognosis | Outcome Dead/Alive | Age | EN | LDH | Stage | IPI | IR |
|------------------|--------------|--------------------|-----|-----|-----|-------|-----|-----|
| EVA1B | Bad | <u>0.0005</u> | 0.7840 | 0.7045 | 0.1580 | 0.1046 | 0.0668 | 0.1657 |
| KRT19 | Bad | <u>0.0148</u> | <u>0.0017</u> | <u>0.0384</u> | 0.1236 | 0.1815 | 0.4020 | 0.9878 |
| BTN2A3P | Bad | <u>0.0094</u> | 0.0820 | 0.7045 | 0.5451 | 0.5724 | <u>0.0440</u> | 0.0751 |
| KLK10 | Bad | 0.4673 | <u>0.0481</u> | <u>0.0184</u> | 0.6009 | 0.8798 | <u>0.0258</u> | 0.6764 |
| TRPM4 | Bad | 0.1646 | 0.3882 | 0.1212 | 0.1698 | 0.1710 | 0.9865 | 0.0974 |
| TDRD12 | Good | <u>0.0094</u> | 0.5468 | 0.8321 | 0.1934 | 0.1616 | 0.4593 | 0.8869 |
| ZNF230 | Good | 0.0698 | <u>0.0310</u> | 0.9695 | 0.6875 | 0.2887 | 0.5414 | <u>0.0009</u> |

OS, overall survival; EN, extranodal; IPI, international prognostic index, IR, immune response. *p* values < 0.05 are underlined. *p* values < 0.01 are double underlined.

| N. | Gene-probe | Gene Symbol | Multilayer Perceptron (MLP) Neural Network | | | | | | | | | | Radial Basis Function (RBF) Neural Network | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | OS | Age | EN | LDH | Stage | IPI | IR | OS 1/5y | Comb | AVG NI top | OS | Age | EN | LDH | Stage | IPI | IR | OS 1/5y | Comb | AVG NI top |
| 1 | 213126_at | MED8 | | | | | | | | + | | | + | + | | + | | | | + | | + |
| 2 | 202955_s_at | ARFGEF1 | | | | | | | | | | | | | | + | | | | | | |
| 3 | 214116_at | BTD | | | | | | | | | | + | | | | | | | | | | |
| 4 | 205272_s_at | PRH1 | | + | | + | | | + | | | + | | | + | | | | | | | + |
| 5 | 201650_at | KRT19 | | | | | | | | | | + | | | | | | | | | | |
| 6 | 205791_x_at | ZNF230 | | | | | + | | | | | + | | + | | | | | | | | + |
| 7 | 206414_s_at | ASAP2 | | | | | | | | + | | + | | | | | | | | + | | |
| 8 | 203377_s_at | CDC40 | | | | | | | | + | | | + | | + | + | | | + | + | + | | + |
| 9 | 220867_s_at | SLC24A2 | | | | | | | | | + | | | | | | | | | | | + |
| 10 | 210597_x_at | PRB1 | | | | + | | | | | | + | | | | | | | | + | | |
| 11 | 201729_s_at | KIAA0100 | | | | | | + | + | | | + | | | | | + | | + | | | |
| 12 | 221552_at | ABHD6 | | | | | | | | | | + | | | | | | | | | | |
| 13 | 202069_s_at | IDH3A | | | | | | | | | | | | + | | + | | | | | | + |
| 14 | 200973_s_at | TSPAN3 | | | | | | | | + | | | | | + | | | | | + | | + |
| 15 | 221710_x_at | EVA1B | | | | | | | | | | + | | | | | | | | | | |
| 16 | 206278_at | PTAFR | | | | | | | | | + | | | | | | | | | | + | |
| 17 | 202822_at | LPP | | | | | | | | | + | | | | | | | | | | + | |
| 18 | 211339_s_at | ITK | | | | | | | | + | + | | | | | | | | | + | + | |
| 19 | 217893_s_at | AKIRIN1 | | | | | | | | | | + | | | + | + | | | + | | | + |
| 20 | 217678_at | SLC7A11 | | + | | | | | | | | + | | | | | | | | | | |
| 21 | 209792_s_at | KLK10 | | | | | | | | | | + | | | | | | | | | | |
| 22 | 219449_s_at | TMEM70 | | | | | | | | | | | | | + | | | | | | + | |
| 23 | 208585_at | BTN2A3P | | | | | | | | | | | | | | | | | | | + | |
| 24 | 215356_at | TDRD12 | | | | | | | | + | | | + | | + | | | | | | | + |
| 25 | 201052_s_at | PSMF1 | | | | | | | | | | | | | + | | | | | | | |
| 26 | 219360_s_at | TRPM4 | | | | + | | | | | | + | | | | | | | + | | | |
| 27 | 207079_s_at | MED6 | | | | | | | | + | | + | + | | + | | | | + | | + | + |
| 28 | 208679_s_at | ARPC2 | | | | | | | | | | | | | | | | | | | + | |
| 29 | 222068_s_at | DNAAF1 | | | | | | | | | | | | | | | | | + | | | |
| 30 | 212546_s_at | FRYL | | | | | | | | | | | | | | | | | | | | + |
| 31 | 209463_s_at | TAF12 | | | | | | | | | | | | | | | | | | | | + |
| 32 | 207625_s_at | CBFA2T2 | | | | | | | | | | + | | | | | | | | | | |
| 33 | 211987_at | TOP2B | | | | + | | | | | | | | | | | | | | | + | |
| 34 | 210129_s_at | TTLL3 | | | | | | | | | | | | | | | | | + | | | |
| 35 | 202583_s_at | RANBP9 | | | | | | | | | | | | | + | | | | | | | + |
| 36 | 220826_at | TCP10L | | | | | | | | + | | | | | | | | | | | + | |
| 37 | 209279_s_at | NSDHL | | | | | | | | + | | | | | | | | | | | + | |
| 38 | 214330_at | ATPAF2 | | | | | | | | | | + | | | | | | | | | | |
| 39 | 209308_s_at | BNIP2 | | | | | | | | | | | | | + | | | | | | | |
| 40 | 211220_s_at | HSF2 | | | | | | | | | | | + | | | | | | | | | + |
| 41 | 219185_at | SIRT5 | | | | | | | | + | | | | | | | | | | | + | |
| 42 | 216988_s_at | PTP4A2 | | | | | | | | | | | | | | + | | | | | | + |
| 43 | 205208_at | ALDH1L1 | | | | | | | | | | | | | | + | | | | | | |

**Figure 6.** Contribution of each dependent variable and neural network method in the final set of 43 predictive genes. Both multilayer perceptron (MLP) and radial basis function (RBF) contribute to the final set with a synergic effect. In general, the most efficient variables are the combined and the averaged normalized importance.
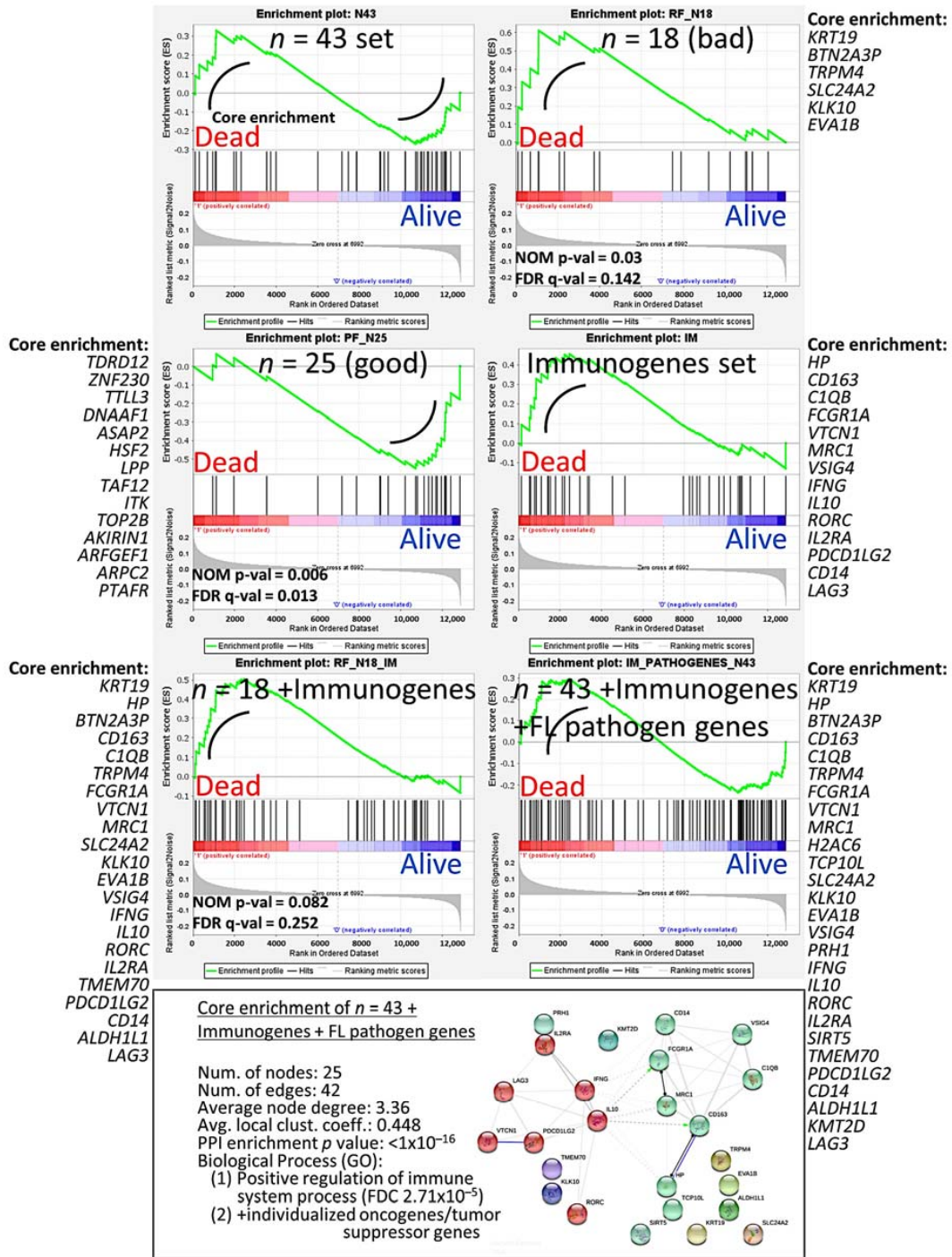
**Figure 7.** Gene set enrichment analysis (GSEA) was performed to confirm the results of the multivariate Cox regression for the overall survival analysis. The set of 43 were used in addition to genes of the immune response as well as oncogenes and tumor suppressor genes related to the pathogenesis of follicular lymphoma. The genes of the core enrichment are shown as well as a functional network association analysis.
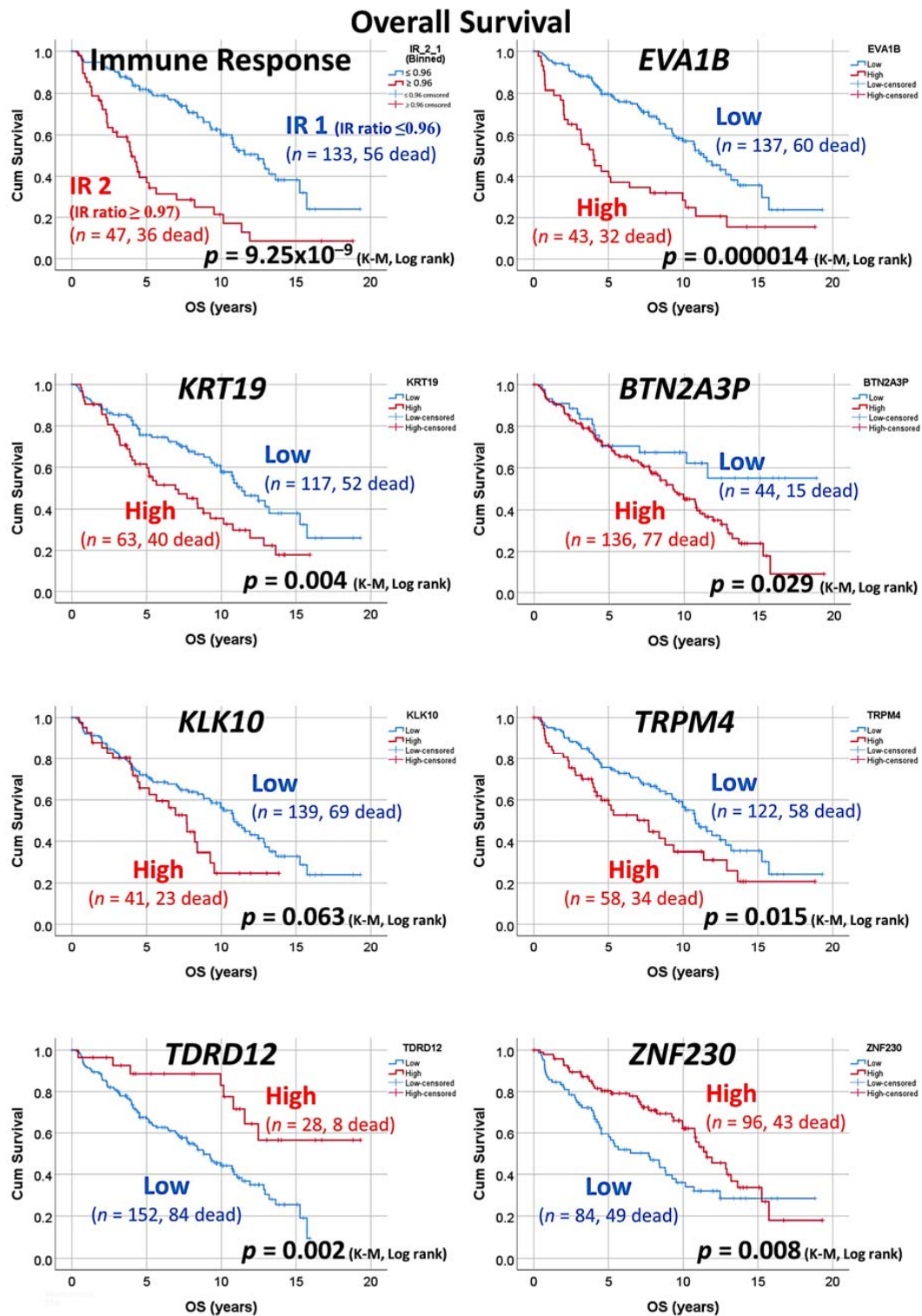
**Figure 8.** Overall survival with Kaplan–Meier analysis. Based on the GSEA results, a group of genes was selected. After finding the more adequate cut-off, a group of patients with high and low expression was found with correlation to the overall survival. Bad OS genes were *EVA1B*, *KRT19*, *BTN2A3P*, *KLL10*, and *TRPM4*. Good OS genes were *TDRD12* and *ZNF230*.
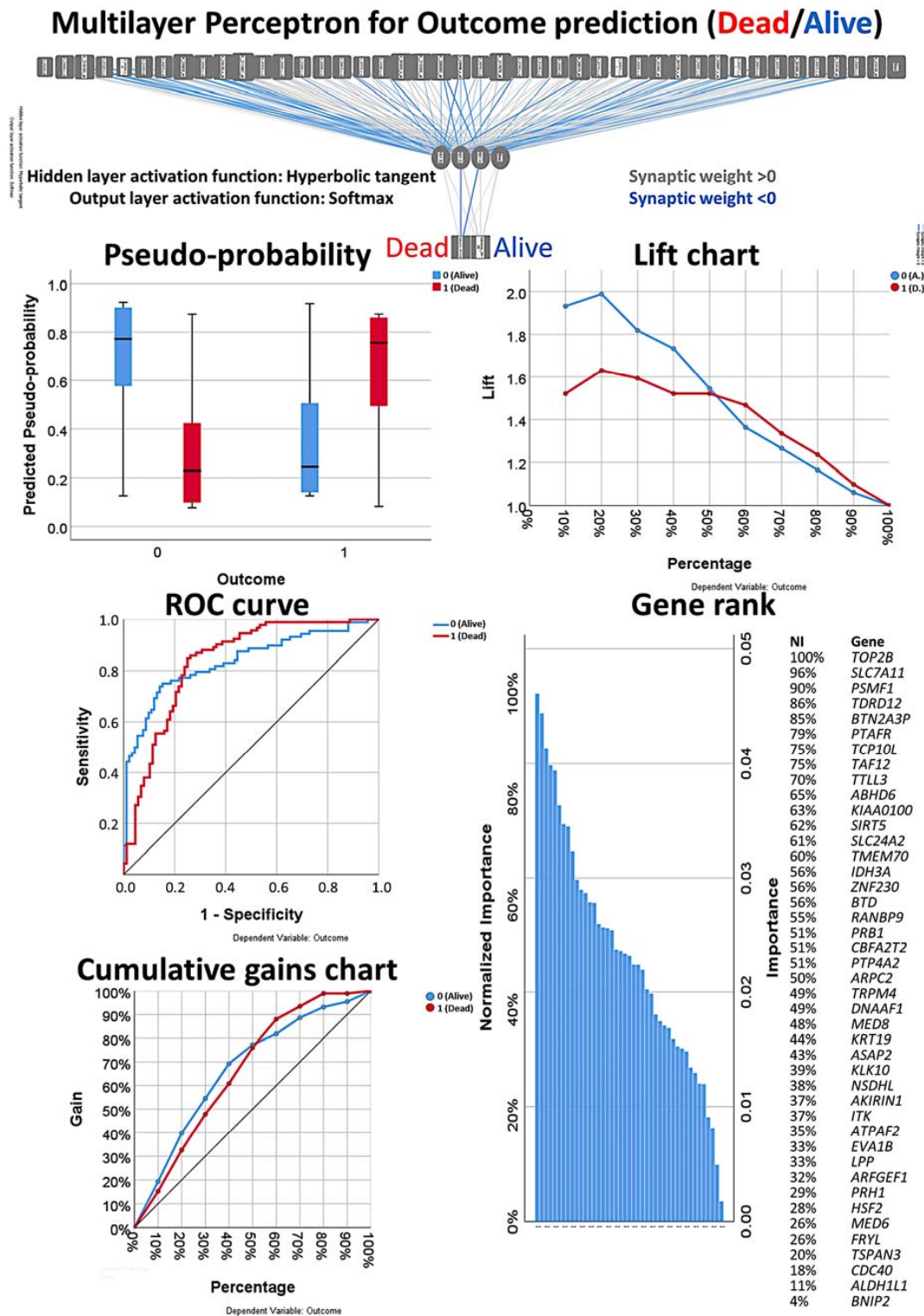
**Figure 9.** Multilayer perceptron analysis based on the set of 43 genes. This MLP analysis includes only the final set of 43 genes as input variables and predicts the survival of the patients based on the survival outcome (dead/alive). The quality of the neural network analysis can be confirmed with the receiver operating characteristic (ROC) curve, and the rank of the genes is shown according to their normalized importance for prediction.
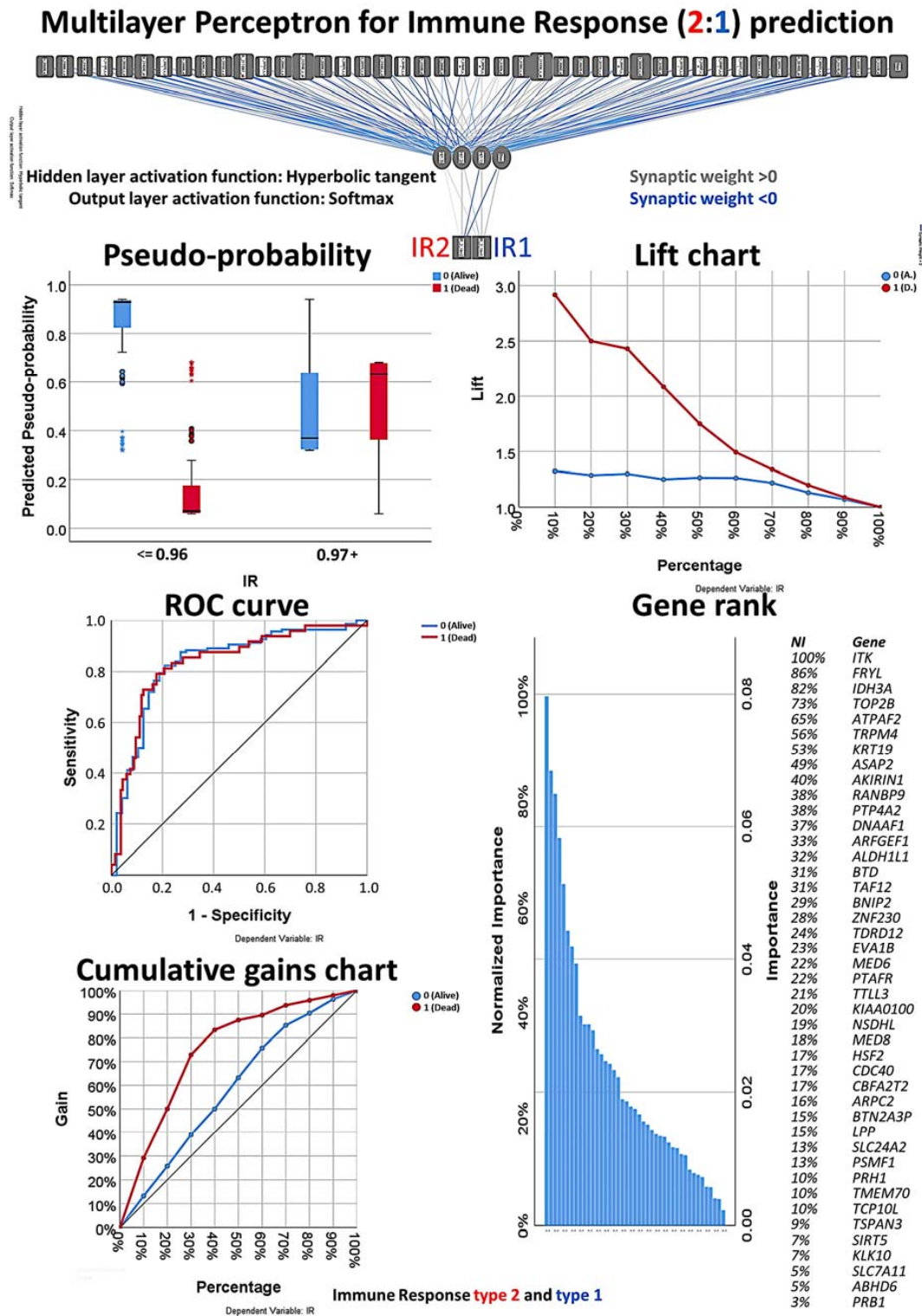
**Figure 10.** Multilayer perceptron analysis based on the set of 43 genes. This MLP analysis includes only the final set of 43 genes as input variables and predicts the immune response of the patients based on the immune response 2:1 ratio. The quality of the neural network analysis can be confirmed with the ROC curve, and the rank of the genes is shown according to their normalized importance for prediction.
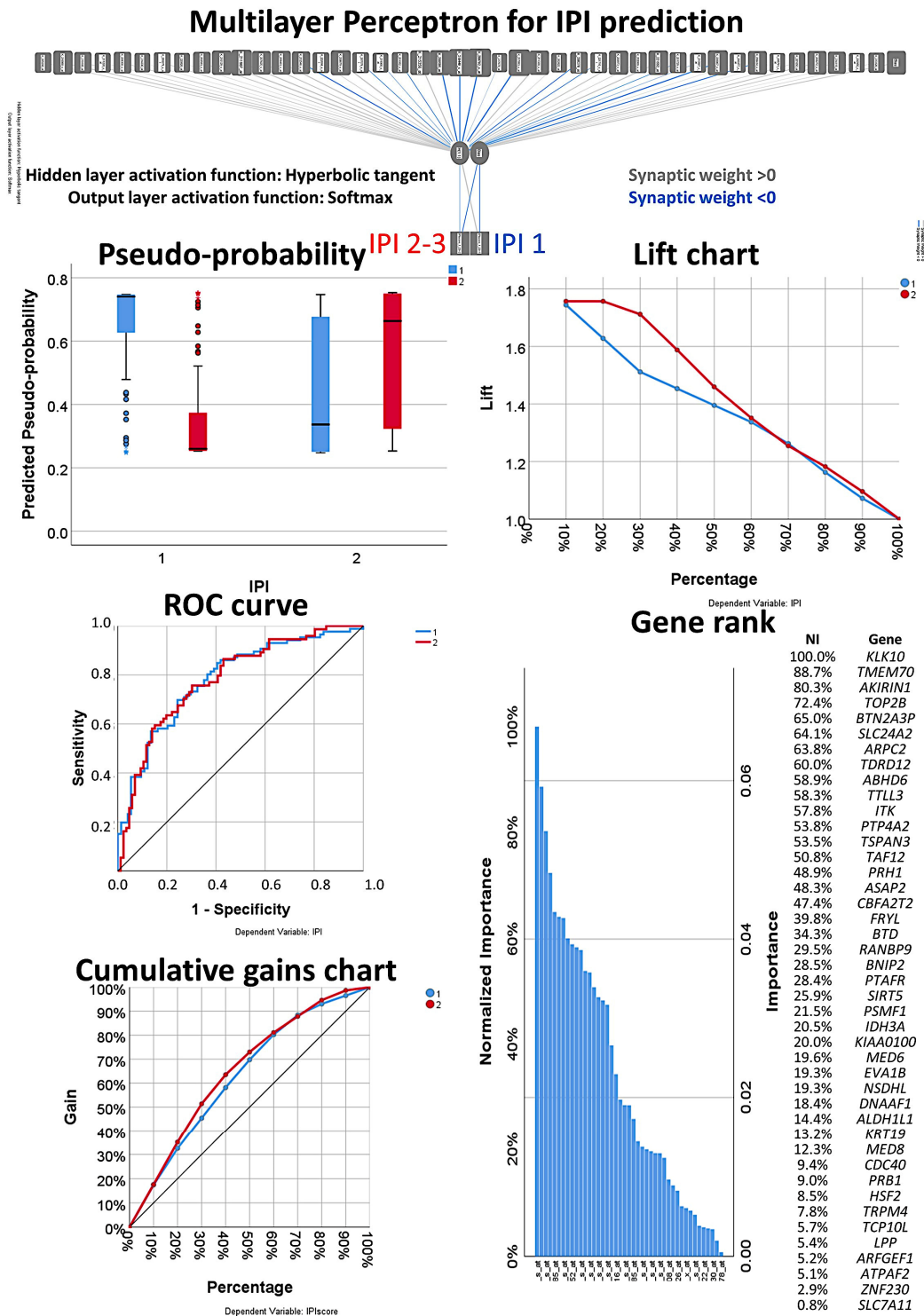
**Figure 11.** Multilayer perceptron analysis based on the set of 43 genes. This MLP analysis includes only the final set of 43 genes as input variables and predicts the international prognostic index (IPI) of the patients. The quality of the neural network analysis can be confirmed with the ROC curve, and the rank of the genes is shown according to their normalized importance for prediction.

## 4. Discussion

Follicular lymphoma (FL) is one of the most frequent subtypes of non-Hodgkin lymphomas, and it is the second most common subtype in Western countries. According to the World Health Organization (WHO) classification of tumors of the hematopoietic and lymphoid tissues, FL is characterized by a proliferation of B lymphocytes of the germinal center of lymphoid follicles, and it almost always shows a histological follicular growth pattern under the microscope [6,7].

The molecular pathogenesis of FL is complex, and it is characterized by a multistep process in which a germinal center B-lymphocyte must acquire several genetic and epigenetic changes that eventually will lead to the malignant transformation. Most of the FL cases are characterized by the t(14;18) that conveys the overexpression of the *BCL2* anti-apoptosis oncogene. Nevertheless, although the overexpression of BCL2 is necessary and seen is most of the cases even in FL "in situ", by itself, it is not capable of malignant transformation. FL is morphologically heterogenous, comprised in morphologic terms as centrocytes (small and large cleaved cells) and centroblasts (small and large non-cleaved cells), and being in different stages of proliferation, maturation, and/or apoptosis. In addition, the tumoral immune microenvironment is rich in T-lymphocytes (including cytotoxic T-cells, follicular T helper cells, macrophages, and dendritic cells) [9,11].

Artificial Intelligence (AI) techniques make use of algorithms that can be defined as a series of instructions that a computer can execute. Many of the AI algorithms can learn from data and enhance themselves. Artificial neural networks are computer systems that try to mimic the biological neural networks. The artificial units or nodes are called neurons. The connections between neurons are called edges. Neurons and edges have a weight that adjusts in time along with the learning process. In this project, we were interested in performing AI analysis of gene expression data of FL because to the best of our knowledge, this had not been attempted before. We used two methods, two types of neural network analyses, the multilayer perceptron (MLP) and radial basis function (RBF), to analyze the gene expression of FL, and we correlated with the prognosis of the patients. In addition, due to the importance of the tumoral immune microenvironment in FL, we also queried the AI system for correlation with the two known immune responses type 1 and 2 that are associated with the prognosis of the patients [12]. Finally, the AI analysis was completed with more conventional statistical analyses including the gene set enrichment analysis (GSEA), univariate and multivariate Cox regression analyses, and the Kaplan–Meier with log rank test for overall survival analysis. Our method was successful, as we managed to simplify from 22,215 gene probes to a final list of 43 genes. Of note, we integrated both the MLP and RBF in the analysis, but we also could compare both types of neural networks, and we found that MLP was more efficient than RBF, as seen by its faster computational time and better areas under the curve of the ROC analysis. Therefore, out data show that MLP fits better the type of data of gene expression in lymphoid neoplasia.

In this project, we have highlighted several genes with prognostic relevance in FL. Some of them were associated to a poor prognosis of the patients, while others were associated to a good prognosis. Among the bad prognosis group, we found that the high gene expression of *EVA1B*, *KRT19*, *BTN2A3P*, *KLK10*, and *TRPM4* was associated to an unfavorable overall survival of the FL patients. EVA1B (Protein eva-1 homolog B) is an integral component of the membrane [14]. Very little information about this marker is found in the literature (including Pubmed), but this gene was found to be highly significantly changed by altered DNA methylation [14]. According to the Human Protein Atlas [15], a high expression of *EVA1B* is associated to a poor prognosis of renal, breast, and liver cancers. KRT19 (Keratin, type I cytoskeletal 19) is involved in the organization of myofibers, as it is a structural constituent of the cytoskeleton. It is also associated to the Notch signaling pathway [16]. In cancer, it has been related to the pathogenesis of breast and colon cancers [15]. According to the Human Protein Atlas [15], high expression is associated to poor prognosis in renal and pancreatic cancer. BTN2A3P (Putative butyrophilin subfamily 2 member A3) regulates the cytokine production; therefore, it regulates immune responses and the T-cell receptor pathway. By genome-wide association and transcriptome studies, this gene was identified as a target gene and risk loci for breast cancer [17]. KLK10 (Kallikrein-10) is a

serine-type endopeptidase involved in the cell cycle. *KLK10* has a tumor-suppressor role for NES1 in breast and prostate cancer. High expression of this gene is associated to an unfavorable prognosis of the patients in pancreatic and breast cancer [15]. *KLK10* is also involved in other several types of cancer. For example, the *NES1/KLK10* gene represses proliferation, enhances apoptosis, and down-regulates the glucose metabolism of PC3 prostate cancer cells [15]. TRPM4 (Transient receptor potential cation channel subfamily M member 4) functions as a calcium-activated non-selective (CAN) cation channel that mediates membrane depolarization [15]. It has been involved in the adaptive immune response and in many types of cancer. For example, *TRPM4* is highly expressed in human colorectal tumor buds and contributes to the proliferation, cell cycle, and invasion of colorectal cancer cells [18]. Among the good prognosis group, we found that the high gene expression of *TDRD12* and *ZNF230* were associated to a favorable overall survival of the FL patients. TDRD12 (Putative ATP-dependent RNA helicase TDRD12) has RNA helicase activity (nuclei acid binding) and is involved in cell differentiation. Not much information of this gene is found in the literature, but it has been related to salivary gland adenoid cystic carcinoma [19]. ZNF230 (Zinc finger protein 230) may be involved in transcriptional regulation, with DNA and metal ion binding properties. The RNA levels are detected as low levels in many cancers, and the protein level in cancer tissue has not been completed yet [15]. Therefore, using AI, we not only managed to highlight several genes with prognostic relevance in FL, but also, we identified some genes that have a role in other types of cancer. In addition, these genes were not only rated to the prognostic outcome (dead/alive) but also related to the tumoral immune microenvironment, which is a field of importance due to the development of personalized immunotherapies.

The future research directions of this research include the validation of these markers using an independent series of FL from another institution.

## 5. Conclusions

Using Artificial Intelligence, we have analyzed the gene expression of follicular lymphoma and managed to identify a set of 43 genes that correlated with the overall survival of the patients. Therefore, AI including the multilayer perceptron and radial basis function neural networks are useful bioinformatics tools for the identification of biomarkers in follicular lymphoma.

In this research, we used the overall survival endpoint. Nevertheless, overall survival is not the best endpoint for follicular lymphoma analysis due to the protracted and variable nature of the disease; rather, response to therapy, prediction of progression (often to large B-cell lymphoma), and quality of life are additional considerations. This aspect will be assessed in a continuation research.

Among the 43 genes, we highlighted seven genes, five associated to poor and two associated to overall survival. Those five genes also associated to M2-like macrophage markers. There is little scientific evidence about the role of these seven genes in lymphoid neoplasia, but there is some evidence of their association in other types of cancer. Therefore, these genes open a path of potential application in development of clinical trials and targeted therapies.

## References

1. Aunalytics. Artificial Intelligence, Machine Learning, and Deep Learning. Available online: https://www.aunalytics.com/artificial-intelligence-machine-learning-and-deep-learning%E2%81%A0/ (accessed on 8 November 2020).

2. *IBM SPSS Neural Netwoks 25*; IBM Corporation: Armonk, NY, USA, 2018.

3. IBM Corporation. IBM SPSS Neural Networks 20. New Tools for Building Predictive Models. IBM Business Analytics. Available online: https://www.ibm.com/support/pages/node/618179 (accessed on 28 October 2020).

4. Carreras, J.; Hamoudi, R.; Nakamura, N. Artificial Intelligence Analysis of Gene Expression Data Predicted the Prognosis of Patients with Diffuse Large B-Cell Lymphoma. *Tokai J. Exp. Clin. Med.* **2020**, *45*, 37–48. [PubMed]

5. Carreras, J.K.; Kikuti, Y.Y.; Miyaoka, M.; Hiraiwa, S.; Tomita, S.; Ikoma, H.; Kondo, Y.; Ito, A.; Shiraiwa, S.; Hamoudi, R.; et al. 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*, 342–360. [CrossRef]

6. Swerdlow, S.H.; Campo, E.; Pileri, S.A.; Harris, N.L.; Stein, H.; Siebert, R.; Advani, R.; Ghielmini, M.; Salles, G.A.; Zelenetz, A.D.; et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **2016**, *127*, 2375–2390. [CrossRef] [PubMed]

7. Freedman, A.S.; Aster, J.C.; Lister, A.; Connor, R.F. Clinical Manifestations, Pathologic Features, Diagnosis, and Prognosis of Follicular Lymphoma. UptoDate. Available online: https://www.uptodate.com/contents/clinical-manifestations-pathologic-features-diagnosis-and-prognosis-of-follicular-lymphoma?search=lymphomas%20definition&source=search_result&selectedTitle=1~{}150&usage_type=default&display_rank=1 (accessed on 28 October 2020).

8. Byers, R.J.; Sakhinia, E.; Joseph, P.; Glennie, C.; Hoyland, J.A.; Menasce, L.P.; Radford, J.A.; Illidge, T. Clinical quantitation of immune signature in follicular lymphoma by RT-PCR-based gene expression profiling. *Blood* **2008**, *111*, 4764–4770. [CrossRef] [PubMed]

9. Carreras, J.; Lopez-Guillermo, A.; Fox, B.C.; Colomo, L.; Martinez, A.; Roncador, G.; Montserrat, E.; Campo, E.; Banham, A.H. High numbers of tumor-infiltrating FOXP3-positive regulatory T cells are associated with improved overall survival in follicular lymphoma. *Blood* **2006**, *108*, 2957–2964. [CrossRef] [PubMed]

10. Carreras, J.; Lopez-Guillermo, A.; Kikuti, Y.Y.; Itoh, J.; Masashi, M.; Ikoma, H.; Tomita, S.; Hiraiwa, S.; Hamoudi, R.; Rosenwald, A.; et al. High TNFRSF14 and low BTLA are associated with poor prognosis in Follicular Lymphoma and in Diffuse Large B-cell Lymphoma transformation. *J. Clin. Exp. Hematop.* **2019**, *59*, 1–16. [CrossRef] [PubMed]

11. Carreras, J.; Lopez-Guillermo, A.; Roncador, G.; Villamor, N.; Colomo, L.; Martinez, A.; Hamoudi, R.; Howat, W.J.; Montserrat, E.; Campo, E. High numbers of tumor-infiltrating programmed cell death 1-positive regulatory lymphocytes are associated with improved overall survival in follicular lymphoma. *J. Clin. Oncol.* **2009**, *27*, 1470–1476. [CrossRef] [PubMed]

12. Dave, S.S.; Wright, G.; Tan, B.; Rosenwald, A.; Gascoyne, R.D.; Chan, W.C.; Fisher, R.I.; Braziel, R.M.; Rimsza, L.M.; Grogan, T.M.; et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* **2004**, *351*, 2159–2169. [CrossRef] [PubMed]

13. Glas, A.M.; Kersten, M.J.; Delahaye, L.J.; Witteveen, A.T.; Kibbelaar, R.E.; Velds, A.; Wessels, L.F.; Joosten, P.; Kerkhoven, R.M.; Bernards, R.; et al. Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. *Blood* **2005**, *105*, 301–307. [CrossRef] [PubMed]

14. Lim, W.J.; Kim, K.H.; Kim, J.Y.; Jeong, S.; Kim, N. Identification of DNA-Methylated CpG Islands Associated With Gene Silencing in the Adult Body Tissues of the Ogye Chicken Using RNA-Seq and Reduced Representation Bisulfite Sequencing. *Front. Genet.* **2019**, *10*, 346. [CrossRef] [PubMed]

15. Uhlen, M.; Fagerberg, L.; Hallstrom, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347*, 1260419. [CrossRef] [PubMed]

16. Saha, S.K.; Choi, H.Y.; Kim, B.W.; Dayem, A.A.; Yang, G.M.; Kim, K.S.; Yin, Y.F.; Cho, S.G. KRT19 directly interacts with beta-catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties. *Oncogene* **2017**, *36*, 332–349. [CrossRef] [PubMed]

17. Ferreira, M.A.; Gamazon, E.R.; Al-Ejeh, F.; Aittomaki, K.; Andrulis, I.L.; Anton-Culver, H.; Arason, A.; Arndt, V.; Aronson, K.J.; Arun, B.K.; et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat. Commun.* **2019**, *10*, 1741. [CrossRef] [PubMed]

18. Kappel, S.; Stoklosa, P.; Hauert, B.; Ross-Kaschitza, D.; Borgstrom, A.; Baur, R.; Galvan, J.A.; Zlobec, I.; Peinelt, C. TRPM4 is highly expressed in human colorectal tumor buds and contributes to proliferation, cell cycle, and invasion of colorectal cancer cells. *Mol. Oncol.* **2019**, *13*, 2393–2405. [CrossRef] [PubMed]

19. Shao, C.; Sun, W.; Tan, M.; Glazer, C.A.; Bhan, S.; Zhong, X.; Fakhry, C.; Sharma, R.; Westra, W.H.; Hoque, M.O.; et al. Integrated, genome-wide screening for hypomethylated oncogenes in salivary gland adenoid cystic carcinoma. *Clin. Cancer Res.* **2011**, *17*, 4320–4330. [CrossRef] [PubMed]