Supplementary Material A: Description of 10-fold CV for the ANN model

Figure S1 shows a detailed representation of the 10-fold CV pipeline for the ANN model. A point by point description is as follows:

1. The algorithm started with the preparation of the dataset for 10-folds CV (grey box, Figure S1). The features-matrix, with dimensions Ns x Nft (Ns = number of samples = 113, Nft = number of features = 129), and the labels vector was partitioned into 10 folds (seven folds of 11 samples and 3 folds of 12 samples).
2. For each (i) of the 10 iterations applied during 10-fold CV, the first step of the i-th CV-loop consisted of *feature standardization* (green box, Figure S1), where mean and standard deviation were calculated from the training set. Standardized training set features were then given as input to the hierarchical clustering for the first step of feature selection.
3. The *feature selection* process (orange box, Figure S1) started with hierarchical clustering with correlation among features as the metric (threshold = 0.85). ReliefF algorithm was then applied to each cluster (Nc = number of clusters) as output from hierarchical clustering, and only the feature with the highest score was considered for the subsequent steps. Thus, a preselection of Nc was obtained. As ReliefF is a supervised feature selection method, target labels were also provided. ReliefF algorithm was used a second time as the last step of the feature selection process, in which from the preselected features, only three features with the highest score were selected for the following hyperparameters search and model training.
4. *Hyperparameters search* (yellow box, Figure S1) followed the feature selection process to identify the best regularization lambda value. In the first phase, 10 lambda values in a logarithmic range (*j* from 0.001 to 0.1) were evaluated, applying a 5-fold CV for each of the lambda parameters; the training set belonged to the *i-th* 10-fold CV iteration was thus partitioned in 5 folds of 20 or 21 samples. A temporary best lambda was then identified as that with the best mean AUC from each 5-fold CV (note that a mean AUC was obtained for each of the 10 possible lambda values). For the second phase, a linear range of lambda values near the temporary lambda (*k* from lambda *j-1* to lambda *j+1*) was established and, for each point, a mean AUC was derived, again through a 5-fold CV (a mean AUC was obtained for each of the 10 possible lambda values in the linear range). The best mean-AUC of this second phase, determined the choice of lambda value for the current (*i-th*) 10-fold CV iteration.
5. Once the three best features and the best hyperparameters had been determined, the *i-th* iteration of the 10-fold CV proceeded to ANN training on the *i-th* training set. The iteration ended with testing the model on the current validation set.

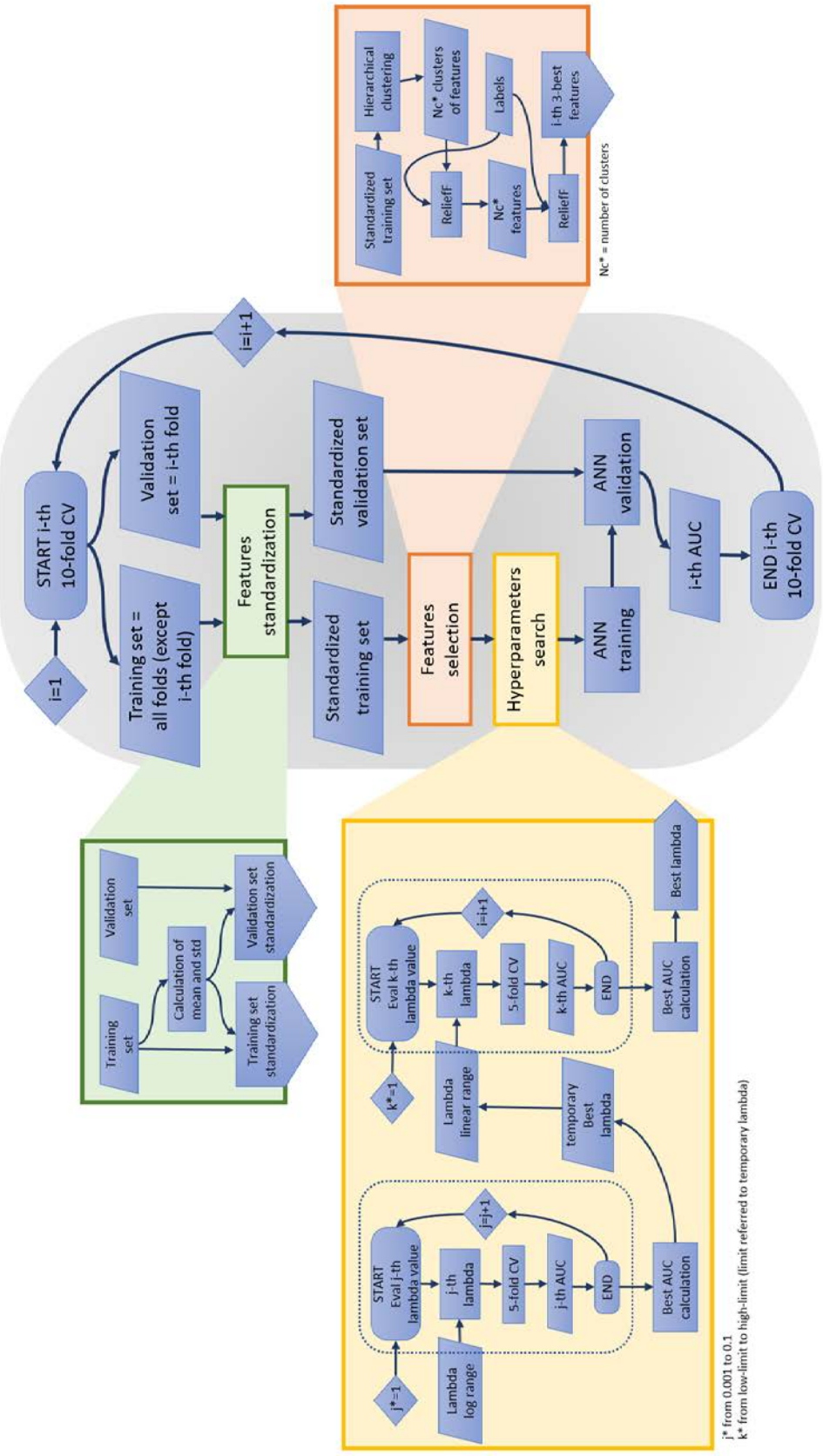

Figure S1. Schematic representation of the 10-fold CV pipeline used for ANN model. Inside the grey box, the 10-fold CV main steps can be found, while on the green, orange and yellow boxes, a more detailed description is given for feature standardization, feature selection and hyperparameters search, respectively.

Supplementary Material B: Justification of network architecture choice

The artificial neural network was defined with a single hidden layer and with the ReLU as activation function. The number of neurons in the hidden layer, the number of hidden layers and the number of input features were experimentally established.

Specifically, starting from a single hidden layer, we trained the network on the training set for each combination of a set of neurons and a set of features, where the number of elements in each set independently ranged from 1 to 20 elements. The heatmap reported in Figure S2 shows performance for each combination in terms of AUC.

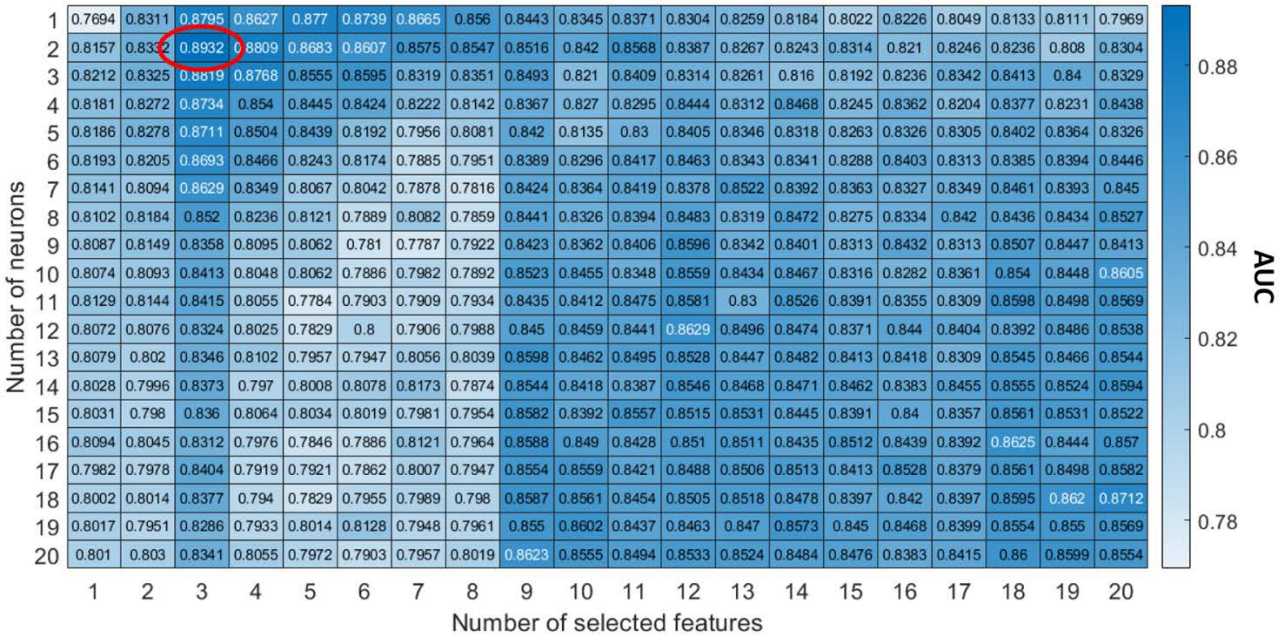

Figure S2. In the construction of the single-layer ANN, the number (1 - 20, vertical axis) of hidden neurons and number (1 - 20, horizontal axis) of selected features was decided on the basis of providing the best AUC (highlighted in red) in the training set.

The same evaluation was performed considering an ANN with two identical hidden layers. Different models were trained varying the number of neurons in the range 2-20 and the number of input features in the range 1-10. In Figure S3, the resulting AUCs show the model performance in the considered analysis.

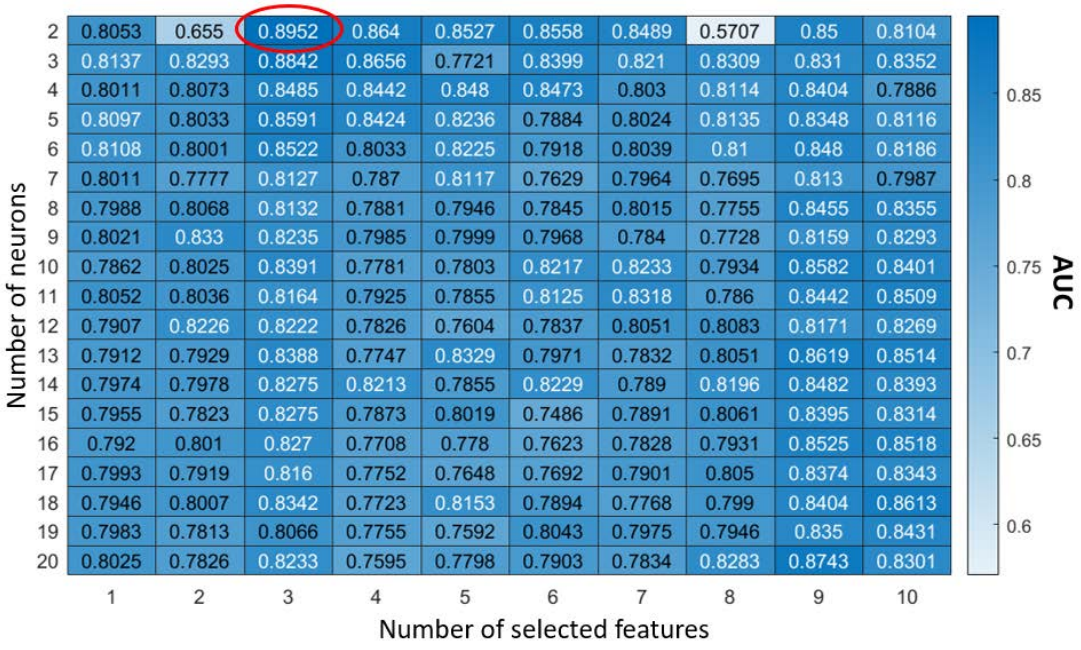

Figure S3. Different number of neurons and input features were tested also for an ANN with two hidden layers. Highlighted with a red circle, the best AUC which is comparable with the single-hidden layer ANN also in terms of number of neurons and number of input features.

The two hidden layer and single hidden layer ANNs produced nearly identical best AUCs (0.8952 vs 0.8932 respectively) with respectively 1 and 2 neurons, and both with three features (Figure S2). The final three candidate features were found the same for both one-hidden layer ANN and two-hidden layer ANN. We chose the single hidden layer ANN on the basis that increasing the complexity of the model adding a hidden layer did not produce substantial improvement.

Supplementary C: list of significant features

For both datasets, Cohort-1 and Cohort-2, Wilcoxon test (alpha=5%) was applied to each of the 129 features to evaluate their statistical power in distinguishing benign from malignant nodules. The test was applied to non-harmonized features as well as to the harmonized feature sets. In the following table (Table S1), the list of features which resulted significant is reported with the correspondent p-values. Features that were significant for only one of the two datasets are marked (*). The features selected in ANN model are in italics and in bold, underlined and in bold features represent candidate features of SVM-LASSO model, whereas features only in bold are those selected on both ANN and SVM-LASSO models.

Table S1. List of features which resulted significant with the correspondent p-values.

	Feature type	**p-value [cross-validation, no-harmonized features]**	**p-value [external validation, non-harmonized features]**	**p-value [cross-validation, harmonized features]**	**p-value [external validation, harmonized features]**
PhysicalSize	Shape	0,0002	1,5 e-05	0,0002	1,5 e-05
BoundingBoxVolume	Shape	1,6 e-05	2,2 e-06	1,6 e-05	2,2 e-06
BoundingBoxSize1	Shape	0,04	9,8 e-06	0,04	9,8 e-06
BoundingBoxSize2	Shape	0,03	1,2 e-06	0,03	1,2 e-06
BoundingBoxSize3	**Shape**	**5,4 e-10**	**3,4 e-06**	**5,4 e-10**	**3,4 e-06**
OrientedBoundingBoxVolume	Shape	7,8 e-05	6,3 e-06	7,8 e-05	6,3 e-06
OrientedBoundingBoxSize1*	Shape	0,11	1,07 e-05	0,11	1,07 e-05
OrientedBoundingBoxSize2*	Shape	0,06	5,6 e-06	0,06	5,6 e-06
OrientedBoundingBoxSize3	Shape	3,2 e-09	8,2 e-06	3,2 e-09	8,2 e-06
EquivalentEllipsoidDiameter1	Shape	0,006	3,4 e-05	0,006	3,4 e-05
EquivalentEllipsoidDiameter2	Shape	0,005	1,3 e-05	0,005	1,3 e-05
EquivalentEllipsoidDiameter3	Shape	1,9 e-05	2,1 e-05	1,9 e-05	2,1 e-05
EquivalentSphericalPerimeter	Shape	0,0002	1,5 e-05	0,0002	1,5 e-05
EquivalentSphericalRadius	Shape	0,0002	1,5 e-05	0,0002	1,5 e-05
FeretDiameter	Shape	0,0001	7,6 e-06	0,0001	7,6 e-06
NumberOfLines	Shape	2,1 e-06	7,02 e-06	2,1 e-06	7,02 e-06
NumberOfPixels	Shape	0,0005	1,5 e-05	0,0005	1,5 e-05
Perimeter	Shape	0,001	7,05 e-06	0,001	7,05 e-06
PrincipalAxes9*	Shape	0,6	0,03	0,6	0,03
PrincipalMoments1	Shape	0,003	1,6 e-05	0,003	1,6 e-05
PrincipalMoments2	Shape	0,003	1,02 e-05	0,003	1,02 e-05
PrincipalMoments3	Shape	1,6 e-05	2,1 e-05	1,6 e-05	2,1 e-05
Eccentricity*	Shape	6,07 e-06	0,11	6,07 e-06	0,11
Elongation*	Shape	0,003	0,14	0,003	0,1
Roundness*	Shape	0,9	2,7 e-07	0,9	2,7 e-07
WeightedPrincipalAxes4*	**Shape+ Intensity**	**0,03**	**0,2**	**0,03**	**0,2**
WeightedPrincipalAxes9*	Shape+ Intensity	0,49	0,008	0,4	0,008
WeightedPrincipalMoments1	Shape+ Intensity	0,02	0,01	0,02	0,01

WeightedPrincipalMoments2	Shape+ Intensity	0,05	5,9 e-05	0,05	5,9 e-05
WeightedPrincipalMoments3	Shape+ Intensity	0,0001	1,5 e-05	0,0001	1,5 e-05
Kurtosis	Intensity	0,001	0,007	0,001	0,007
Maximum*	Intensity	0,22	0,02	0,2	0,02
Mean	Intensity	0,0008	0,004	0,0008	0,004
Median	Intensity	0,0006	0,001	0,0006	0,001
Skewness	Intensity	0,002	0,006	0,002	0,006
StandardDeviation*	Intensity	0,001	0,5	0,001	0,56
Sum	Intensity	6,2 e-07	0,002	6,2 e-07	0,002
Variance*	Intensity	0,001	0,5	0,001	0,56
2DPhysicalSize	Shape	0,04	7,8 e-06	0,04	7,8 e-06
2DEquivalentEllipsoidDiameter1*	Shape	0,05	1,2 e-05	0,05	1,2 e-05
2DEquivalentEllipsoidDiameter2*	Shape	0,05	1,3 e-05	0,05	1,3 e-05
2DEquivalentSphericalPerimeter	Shape	0,04	7,8 e-06	0,04	7,8 e-06
2DEquivalentSphericalRadius	Shape	0,04	7,8 e-06	0,04	7,8 e-06
2DFeretDiameter	Shape	0,03	8,4 e-06	0,03	8,4 e-06
2DNumberOfLines	Shape	0,01	4,5 e-06	0,01	4,5 e-06
2DNumberOfPixels*	Shape	0,06	7,8 e-06	0,06	7,8 e-06
2DPerimeter	Shape	0,01	5,2 e-06	0,01	5,2 e-06
2DPrincipalAxes1*	Shape	0,1	0,01	0,1	0,01
2DPrincipalAxes4*	Shape	0,1	0,01	0,1	0,01
2DPrincipalMoments1	Shape	0,05	9,9 e-06	0,05	9,9 e-06
2DPrincipalMoments2	Shape	0,04	1,1 e-05	0,04	1,1 e-05
2DRoundness	Shape	0,01	3,5 e-05	0,01	3,5 e-05
2DWeightedPrincipalMoments2*	Shape+ Intensity	0,2	0,006	0,2	0,006
2DKurtosis	Intensity	0,02	0,004	0,02	0,004
2DMean*	Intensity	0,03	0,13	0,03	0,13
2DMedian	Intensity	0,02	0,08	0,02	0,08
2DSkewness	Intensity	0,04	0,02	0,04	0,02
2DStandardDeviation*	Intensity	0,005	0,2	0,005	0,27
2DSum	Intensity	0,001	0,004	0,001	0,004
2DVariance*	Intensity	0,005	0,27	0,005	0,279
MeanOfEnergy*	Texture	9,2 e-05	0,4	9,2 e-05	0,4
MeanOfEntropy*	Texture	0,14	0,01	0,14	0,01
MeanOfCorrelation	Texture	2,5 e-06	0,0006	2,5 e-06	0,0006
MeanOfInverseDifferenceMoment	Texture	0,001	0,0002	0,001	0,0002
MeanOfInertia	Texture	4,6 e-07	0,0004	4,6 e-07	0,0004
MeanOfClusterShade*	**Texture**	**0,0003**	**0,5**	**0,0003**	**0,53**
StandardDeviationOfEnergy*	**Texture**	**0.0024**	**0.426**	**0.0024**	**0.426**
StandardDeviationOfEntropy	Texture	0,09	0,01	0,09	0,01
StandardDeviationOfInverseDifferenceMoment*	Texture	0,5	1,4 e-06	0,5	1,4 e-06

StandardDeviationOfInertia	**Texture**	**7,6 e-07**	**0,0001**	**7,6 e-07**	**0,0001**
StandardDeviationOfClusterShade	Texture	0,02	0,004	0,02	0,004
StandardDeviationOfClusterProminence	Texture	0,02	0,0001	0,02	0,0001
StandardDeviationOfHaralickCorrelation	Texture	0,004	0,01	0,004	0,01
MeanOfShortRunEmphasis	Texture	0,08	0,0001	0,08	0,0001
MeanOfGreyLevelNonuniformity	Texture	1,03 e-05	8,2 e-06	1,03 e-05	8,2 e-06
MeanOfRunLengthNonuniformity	Texture	0,0004	7,05 e-06	0,0004	7,05 e-06
MeanOfLowGreyLevelRunEmphasis*	Texture	0,1	0,0001	0,13	0,0001
MeanOfHighGreyLevelRunEmphasis	Texture	0,004	0,02	0,004	0,02
MeanOfShortRunLowGreyLevelEmphasis*	Texture	0,1	0,0001	0,15	0,0001
MeanOfShortRunHighGreyLevelEmphasis*	Texture	0,004	0,4	0,004	0,45
MeanOfLongRunLowGreyLevelEmphasis*	Texture	0,2	0,0006	0,25	0,0006
MeanOfLongRunHighGreyLevelEmphasis*	Texture	0,005	0,6	0,005	0,68
StandardDeviationOfShortRunEmphasis	**Texture**	**0,01**	**5,6 e-06**	**0,012**	**5,6 e-06**
StandardDeviationOfLongRunEmphasis	Texture	0,05	2,7 e-05	0,051	2,71 e-05
StandardDeviationOfGreyLevelNonuniformity	Texture	0,001	4,2 e-05	0,001	4,2 e-05
StandardDeviationOfRunLengthNonuniformity	Texture	0,005	4,03 e-05	0,005	4,03 e-05
StandardDeviationOfLowGreyLevelRunEmphasis*	Texture	0,1	0,0003	0,13	0,0003
StandardDeviationOfHighGreyLevelRunEmphasis*	Texture	0,0001	0,9	0,0001	0,9
StandardDeviationOfShortRunLowGreyLevelEmphasis*	Texture	0,5	0,001	0,53	0,001
StandardDeviationOfShortRunHighGreyLevelEmphasis*	Texture	8,6 e-05	0,4	8,6 e-05	0,4
StandardDeviationOfLongRunLowGreyLevelEmphasis*	Texture	0,2	0,0001	0,22	0,0001
StandardDeviationOfLongRunHighGreyLevelEmphasis	Texture	0,004	0,004	0,004	0,004

* features where significant difference between malignant and benign nodules was found just for cross-validation or external validation set.

Supplementary Material D: Clinical model results and comparison with radiomics-based models

Table S2. *Clinical model prediction results in terms of area under the curve (AUC), accuracy (Acc), false positive rate (FPR) and true positive rate (TPR). Performance on the training set summarizes predictions of the 10x10-fold CV loops. External validation results were instead obtained applying on Cohort-2 the final model trained on Cohort-1.*

	Cross-validation	External validation
AUC (95% CI)	0.76 (0.68-0.85)	0.76 (0.66-0.87)
Acc [%]	71.7	66.7
FPR [%]	27.3	12.9
TPR [%]	70.7	51.2

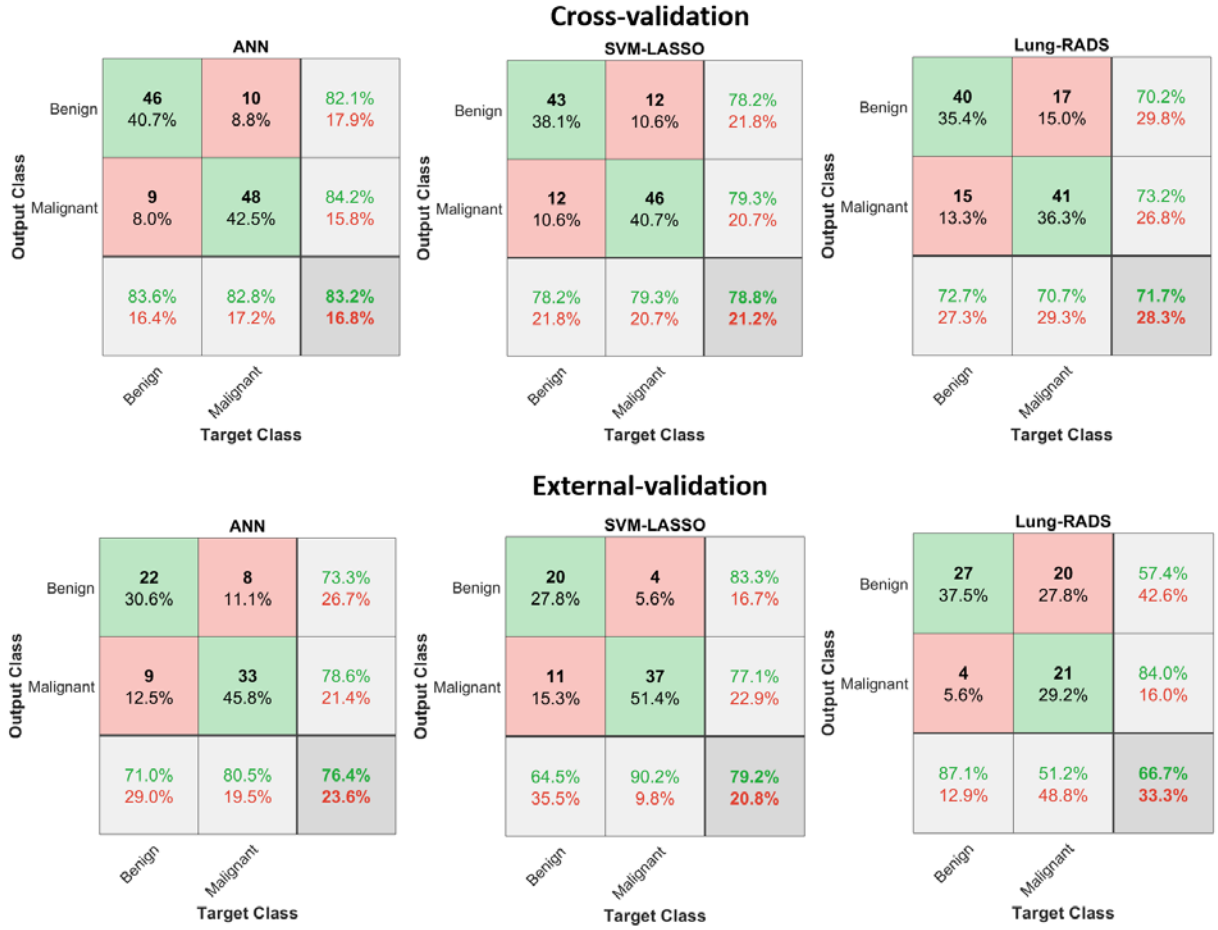

Figure S4. Confusion matrices that summarize the misclassification rate for ANN, SVM-LASSO and clinical model. Matrices on the top are related to the cross-validation, therefore show performance on Cohort-1. On the bottom, matrices are instead related to the external-validation (Cohort-2). In the cross-validation section, matrices of ANN and SVM-LASSO were obtained considering the models when no-harmonization was applied. Accordingly, in the external-validation section, predictions related to Scenario A were considered.