Lung Cancer Detection and Characterisation through Genomic and Radiomic Biomarkers

Luca Brunese*, Francesco Mercaldo† §, Alfonso Reginelli‡, Antonella Santone§
*Department of Medicine and Health Sciences "Vincenzo Tiberio" , University of Molise, Campobasso, Italy
luca.brunese@unimol.it
†Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy
francesco.mercaldo@iit.cnr.it
‡Department of Precision Medicine, University of Campania "Luigi Vanvitelli", Napoli, Italy
alfonso.reginelli@unicampania.it
§Department of Biosciences and Territory, University of Molise, Pesche (IS), Italy
{francesco.mercaldo, antonella.santone}@unimol.it

***Abstract*—Medical image bio-markers of cancer are expected to improve patient care through advances in precision medicine. Compared to genomic bio-markers, bio-markers obtained directly from medical images provide the advantages of being a non-invasive procedure, and characterizing a heterogeneous tumor in its entirety, as opposed to limited tissue available for biopsy. In this paper, with the aim to demonstrate that non-invasive features can obtain better performances if compared to invasive ones in lung cancer detection and characterisation, we propose a method to discriminate between different lung cancers (i.e., *Adenocarcinoma* and *Squamous Cell Carcinoma*) by adopting both invasive (genomic) and non-invasive (radiomic) bio-markers, by building supervised machine learning models exploiting both invasive and non-invasive features. Experiments on a data-set of 130 patients show that radiomic bio-markers obtain better performances (with an f-measure equal to 0.993) if compared to the ones obtained by considering genomic ones (reaching an f-measure equal to 0.929) in lung cancer detection and characterisation.**

***Index Terms*—radiomics, genomics, lung cancer, MRI, machine learning, neural network, supervised learning, classification**

I. Introduction and Related Work

Lung cancer represents the deadliest and the most-costly cancer in the world [1]. Its mortality rate is three times higher than deaths of prostate cancer and nearly twice higher than deaths of breast cancer in women. Lung cancer currently accounts for 32% of cancer deaths in men and 20% of cancer deaths in women[1].

According to the American Cancer Society expert in the United States every three and a half minutes someone will die from lung cancer, accounting for about one in four cancer deaths: in 2019, more than 228,000 people will be diagnosed with lung cancer only in United States[2].

About 80% to 85% of lung cancers are Non-small cell lung cancer (NSCLC). The main subtypes of NSCLC are *Adenocarcinoma* and *Squamous Cell carcinoma*[3]. These cancers, which start from different types of lung cells are grouped together as NSCLC because their treatment and prognoses are often similar. In detail Adenocarcinomas start in the cells that would normally secrete substances such as mucus. This lung cancer occurs mainly in current or former smokers, but it is also the most common type of lung cancer seen in non-smokers. It is more common in women than in men, and it is more likely to occur in younger people than other types of lung cancer. Adenocarcinoma is usually found in the outer parts of the lung and is more likely to be found before it has spread. The Squamous cell carcinomas start in squamous cells, which are flat cells that line the inside of the airways in the lungs. They are often linked to a history of smoking and tend to be found in the central part of the lungs, near a main airway (bronchus).

After the patient is diagnosed with NSCLC, doctors will try to figure out if it has spread, and if so, how far. This process is called staging. The stage of a cancer describes how much cancer is in the body. It helps to determine how serious the cancer is and the best way to treat it. Doctors also use a cancer's stage when talking about survival statistics. In particular an important indicator is represented by the spread to nearby lymph nodes (i.e., the *N* parameter), indicating whether the cancer spread to nearby lymph nodes. If the cancer is not thought to have spread to nearby lymph nodes the lung cancer is marked with the *N0* label, if the lymph nodes are on the same side as the cancer the lung cancer is marked with the *N1* label, if the lymph nodes are on the same side as the main lung tumor the lung cancer is marked with the *N2* label. In the worst case the lung cancer is marked with the *N3* label, symptomatic that the cancer has spread to lymph nodes near the collarbone on either side of the body, and/or has spread to hilar or mediastinal lymph nodes on the other side of the body from the main tumor.

The current way to diagnose a lung cancer is represented by the biopsy. A biopsy is a procedure performed to remove tissue or cells from the body for examination under a microscope [2]. A lung biopsy is a procedure in which samples of lung tissue are removed (with a special biopsy needle or during surgery) to determine if lung disease or cancer is present.

Like all medical procedures, a lung biopsy does carry a

[1] https://tinyurl.com/yh63d8ka

[2] https://www.lung.org/assets/documents/research/ALA-SOLC-2019-Key-Findings.pdf

[3] https://www.cancer.org/cancer/lung-cancer/about/what-is.html

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

small risk of complications, such as a pneumothorax. This is when air leaks out of the lung and into the space between your lungs and the chest wall [3]. This can put pressure on the lung, causing it to collapse.

Clearly, the clinician doing the biopsy will be aware of the potential risks involved. In fact, during the procedure, they will monitor the patient to check for symptoms of a pneumothorax, such as sudden shortness of breath [4]. If a pneumothorax does happen, it can be treated using a needle or tube to remove the excess air, allowing the lung to expand normally again [5].

Starting from these considerations, in this paper we propose a method aimed to distinguish between the *Adenocarcinoma* and *Squamous Cell carcinoma* lung cancers and to characterise it by automatically assigning the *N* spread to nearby lymph nodes. Both of these information are currently obtained by analysing the lung tissue with the biopsy. For this reason, we consider different models to detect the lung cancer type and the detail about the spread to nearby lymph nodes by exploiting a set of invasive (i.e., genomic) and non invasive (i.e., radiomic) bio-markers.

In last years researchers designed methods for the lung cancer detection considering machine learning techniques. For instance, Golan et al. [6] design a framework that train the weights of the CNN by a back propagation aimed to detect lung nodules in the CT image sub-volumes. This system achieved sensitivity of 78.9%.

Sun et al. [7] adopt convolutional neural networks, stat denoising autoencoder and deep belief networks to detect lung cancer exploiting 35 texture and morphological features. The performances obtained from the proposed classifiers are respectively the following: 79%, 81%, and 79%.

Researchers in [8] consider CT image features as 3D volume, tracheal distance, and distance to outer body to determine if an invasive biopsy or a surgical biopsy procedure should considered to diagnose lung cancer in a patient presenting with lung nodules. These features were used to train both logistic regression and random forest models. They best accuracy obtained is 84% with random forests algorithm.

Authors in [9] explore the usage of 2D convolutional neural network, integrating a deconvolution layer aimed to enlarge the feature map and two region proposal networks to concatenate the useful information from the lower layer. The obtained accuracy is equal to 86.42%.

Differently from the cited works, in this paper the main aim is to show the effectiveness of radiomic features, as non invasive bio-markers for lung cancer detection and characterisation (currently inferred by tissue biopsy). To this aim, we experiment several machine learning algorithms on a set of genomic and radiomic bio-markers, showing that the best accuracy results are obtained by exploiting the non invasive bio-markers.

The paper proceeds as follows: Section II describes the proposed method, experimental analysis to demonstrate the effectiveness of the proposed method is discussed in Section III and, finally, in the last section section conclusion and future research directions are drawn.

II. Method

As stated into the introduction, the aim of the following paper is to understand whether it is possible by exploiting a set of non-invasive (radiomic) features to obtain the same information inferred with invasive (genomic) bio-markers, by avoiding to patients an invasive procedure.

Figure 1 shows the proposed methodology.

Two different data repositories are considered: in the first one, the *Genomic Biomarkers* in Figure 1, the data are obtained from genes, a sequence of nucleotides in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein. (i.e., by analysing the lung tissue obtained through a biopsy). The second data, the *Radiomic Biomarkers* in Figure 1, are obtained from magnetic resonance images (i.e., through a non invasive procedure). In detail we consider 14 *Shape* radiomic features:

- *Elongation*: relationship between two largest principal components;
- *Flatness*: relationship between largest and smallest principal components;
- *LeastAxisLength*: yield smallest axis length of the ROI-enclosing ellipsoid;
- *MajorAxisLength*: yield largest axis length of ROI-enclosing ellipsoid;
- *Maximum2DDiameterColumn*: mesh vertices in row-slice plane;
- *Maximum2DDiameterRow*: mesh vertices in the column-slice plane;
- *Maximum2DDiameterSlice*: mesh vertices in row-column plane;
- *Maximum3DDiameter*: mesh vertices;
- *MeshVolume*: volume is obtained using the surface mesh;
- *MinorAxisLength*: second-largest axis length of the ROI-enclosing ellipsoid;
- *Sphericity*: roundness of shape of the tumor region relative to a sphere;
- *SurfaceArea*: the sum of all sub-areas;
- *SurfaceVolumeRatio*: Surface Area to Volume ratio;
- *VoxelVolume*: approximate volume.

In the next, (*Descriptive Statistics* step in Figure 1) we consider descriptive statistics to understand whether the considered radiomic and genomic biomarker populations are able to discriminate between the two considered classes (i.e., between the *Adenocarcinoma* and *Squamous Cell Carcinoma* lung cancers). We depict box-plots, usually considered for graphically depicting groups of numerical data through their quartile. In box-plots outliers may be plotted as individual points. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The more distributions are distinguishable from one another (and therefore not overlapping), the more it will be possible for a classifier to be able to correctly discern between unknown *Adenocarcinoma* and *Squamous Cell Carcinoma* lung cancer genomic and radiomic instances. The rationale behind this analysis is to understand

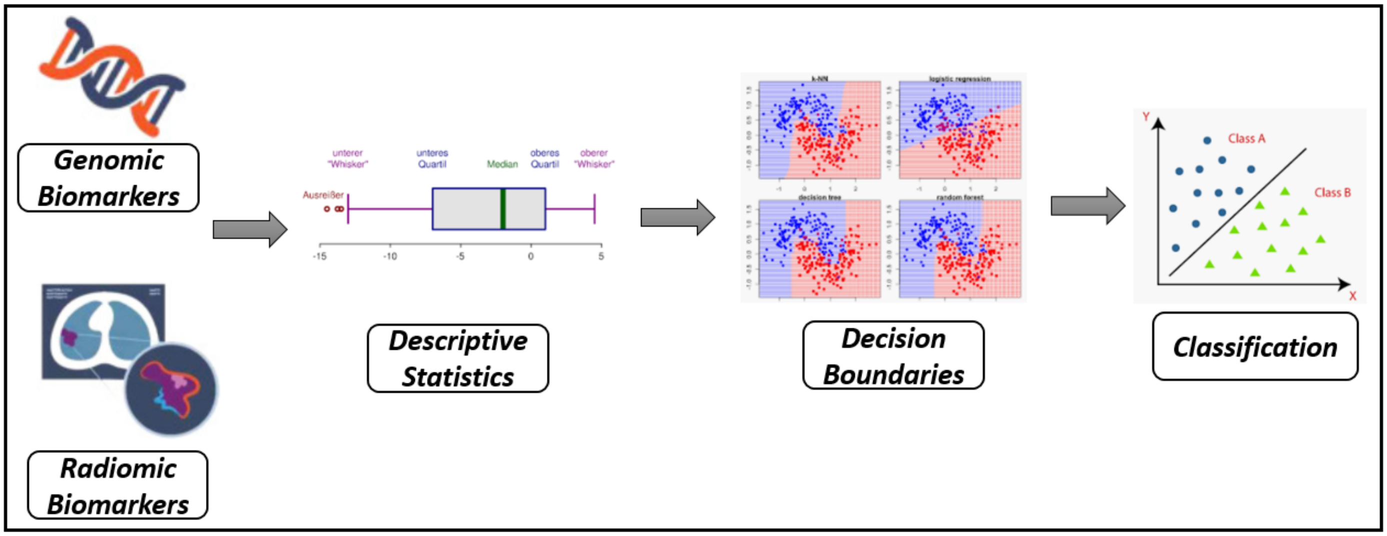

Fig. 1. The proposed method.

whether the genomic and the radiomic features can be helpful to discriminate between *Adenocarcinoma* and *Squamous Cell Carcinoma* lung cancer.

The next step is the *Decision Boundaries*: in a statistical-classification problem with two classes, they represent a hypersurface that partitions the underlying vector space into two sets, one for each class. The classifier will classify all the points on one side of the decision boundary as belonging to one class and all those on the other side as belonging to the other class. A decision boundary is the region of a problem space in which the output label of a classifier is ambiguous. If the decision surface is a hyperplane, then the classification problem is linear, and the classes are linearly separable. Decision boundaries are not always clear cut. That is, the transition from one class in the feature space to another is not discontinuous, but gradual. This effect is common in fuzzy logic based classification algorithms, where membership in one class or another is ambiguous. The rationale behind this analysis is to understand the best classification algorithm able to build efficient model to distinguish between *Adenocarcinoma* and *Squamous Cell Carcinoma* lung cancer. To this aim different classification algorithms are considered, chosen from the most widespread machine learning algorithms usually considered for classification task [10], [11].

In detail we consider the following algorithms:

- *Linear SVM* [12]: the Support Vector Machine linear classifier predicts, for each given input, which of the trained possible classes the input is a member of. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on;
- *RBF SVM* [13]: this algorithm represents the non-linear SVM classifier with Radial Basis Function (RBF) kernel. The kernel is basically a measure of similarity that in this context is reflecting into the similarity between instances under analysis. RBFs are means to approximate multivariable functions by linear combinations of terms based on a single univariate function (i.e., the RBF);
- *Gaussian Process* [14]: this classification algorithm is based on the Laplace approximation: the Laplace approximation is a way of approximating Bayesian parameter estimation and Bayesian model comparison. The data points have associated latent variables which are drawn from a Gaussian Process prior, and the labels are modeled as stochastic functions of the latent variables;
- *RF* [15]: the Random Forest algorithm considers several decision trees for the training phase. The built "forest" is an ensemble of Decision Trees, this is the reason why it builds multiple decision trees and merges them together to get a more accurate and stable prediction;
- *Neural Network* [16]: it consists of a number of nodes in the input layer (equal to the number of features in the input data-set). Each input node is multiplied with a weight (typically initialized with some random value) and the results are added together. The sum is then passed through an activation function. In particular we consider the MultiLayer Perceptron, consisting in three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function;
- *QDA* [17]: the Quadratic classifier separates instance in classes of objects through a quadratic surface. In the Quadratic Discriminant Analysis (QDA) it is assumed that the measurements from each class are normally distributed;
- *Logistic* [18]: this classification algorithm considers the Logistic function, also called *sigmoid function*: a curve

that can take any real-valued number and map it into a value between 0 and 1.

The last step of the proposed method is the classification one (*Classification* in Figure 1). A supervised approach is considered: in the *training* step the model is built, while in the *testing* step the effectiveness of the model is evaluated.

The model is built starting from a labelled data-set i.e., *genomic* and *radiomic* bio-markers. Relating to the *radiomic* bio-markers a set of medical images are considered, and from these images we extract the set of 14 radiomic features. Subsequently, the radiomic features with the corresponding label obtained from the medical report are parsed in a feature vector. Each feature vector is representing an instance. Once all the instances are obtained, it is possible to input the classification algorithm that will output the *Model*.

Once trained the models, we have to verify whether the built models are able to discriminate between *Adenocarcinoma* and *Squamous Cell Carcinoma* lung cancer and lung cancers with different lymph node spread.

For this reason we consider a set of *genomic* and *radiomic* bio-markers and after gathered the radiomic features and generated the feature vector, we input the *Model* that will output the label.

We adopt the cross validation approach: the full data-set is splitted in two parts: the first one, the training data-set is considered to generate the model, while the second one (i.e., the testing data-set) is considered to evaluate the model effectiveness. We adopt a fair partitioning between training and testing data-set (i.e., 50% training and 50% testing). With the aim to consider all the instances belonging to the genomic and radiomic data-set in both the training and testing phase, we consider two different classifications: in the first classification the first 50% of the (genomic and radiomic) instances are considered to generate the model, while the second 50% is used to evaluate the model; while in the second classification the second 50% is used to train the model, while the first 50% this time is considered to evaluated the model (2-fold cross validation). The final performance values are averaged between the performances obtained in these two classifications.

We recall that we generate two different models with each involved classification algorithm and with each bio-marker category (i.e., genomic and radiomic): the first one aimed to discriminate between *Adenocarcinoma* and *Squamous Cell Carcinoma* lung cancer, while the second one to predict whether the lung cancer does not exhibit lymph nodes diffusion (i.e., *NO*) or exhibit lymph nodes diffusion (i.e., *N1*, *N2* or *N3*): we recall that both these analysis are currently performed by doctors with an invasive tissue biopsy.

III. The Experimental Analysis

In this section we discuss the results of the experimental analysis of the genomic and radiomic bio-markers in the detection and characterisation of lung cancer. We first describe the data-set involved in the study. In the next, reflecting the study design presented in previous section, we discuss the descriptive statistics, the decision boundary analysis and the classification results.

A. Data-set

A real-world data-set was obtained from the Cancer Imaging Archive[4]. In detail, the *NSCLC Radiogenomics* data-set[5] contains NSCLC cohort related to 130 subjects, 98 afflicted by *Adenocarcinoma* and 32 by *Squamous Cell Carcinoma*. With regard to lymph nodes diffusion, 104 patients do not exhibit lymph nodes diffusion (i.e. *N0* stage), while the remaining 26 exhibit lymph nodes diffusion (i.e. *N1*, *N2* or *N3* stage). The data-set comprises Computed Tomography (CT) with correspondent slice segmentation. Moreover imaging data are also paired with gene mutation, RNA sequencing data (i.e., 22127 genes considered as genomic bio-markers) from samples of surgically excised tumor tissue, and clinical data.

B. Descriptive Statistics

Below we show several box-plots we obtained from the *Adenocarcinoma* and *Squamous Cell Carcinoma* genomic and radiomic populations. For space reasons we represent box-plots related to two genomic and two radiomic features, but similar considerations can be made also for the remaining (genomic and radiomic) bio-markers.

Figure 2 shows the box-plots for the *A1BG* gene[6].

As emerges from the box-plots in Figure 2 there is a substantial overlapping between the *Adenocarcinoma* and *Squamous Cell Carcinoma* populations. In fact, the median for the *Adenocarcinoma* population is equal to 2.94144 while the median for *Squamous Cell Carcinoma* is equal to 2.34653. This can be reflected in a not a good discriminant bio-marker between the two populations.

Figure 3 shows the box-plots for the *A4GALT* gene[7].

If compared with the previous box-plots (related to the *A1BG* gene), the *A4GALT* gene exhibits a more discriminate ability: in fact the median of the *Adenocarcinoma* population is in this case 2.43833 while the *Squamous Cell Carcinoma* population exhibits a value equal to 11.9247.

With regard to the radiomic bio-markers, Figure 4 shows the box-plots related to the *MajorAxisLength* radiomic feature.

The median for the *Adenocarcinoma* population is equal to 465.095 while the *Squamous Cell Carcinoma* instances exhibit a median equal to 401.554. From the box-plots is emerges a tiny overlap between the two distributions.

The second radiomic bio-marker box-plots we show is represented in Figure 5.

The box-plots in Figure 5 are related to the *Adenocarcinoma* and *Squamous Cell Carcinoma* populations for the *Maximum2DDiameterSlice* bio-markers. In this case the medians are respectively equal to 478.531 for the *Adenocarcinoma* and 388.471 for the *Squamous Cell Carcinoma* population. As

[4] https://www.cancerimagingarchive.net/

[5] https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics

[6] https://www.genecards.org/cgi-bin/carddisp.pl?gene=A1BG

[7] https://www.genecards.org/cgi-bin/carddisp.pl?gene=A4GALT

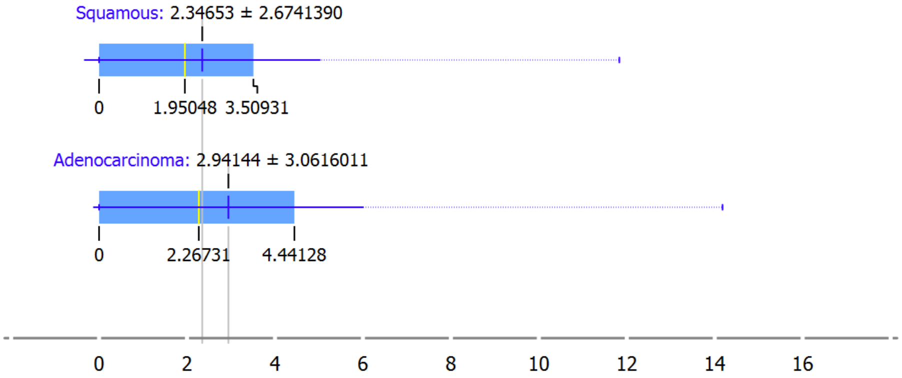

Fig. 2. Box-plots for *A1BG* gene.

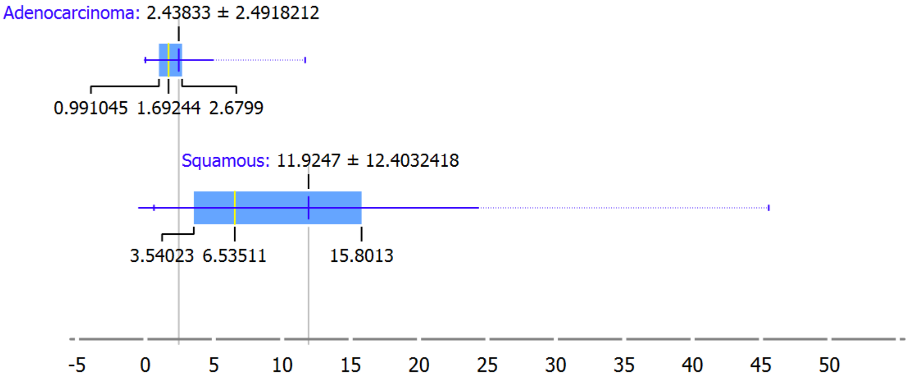

Fig. 3. Box-plots for *A4GALT* gene.

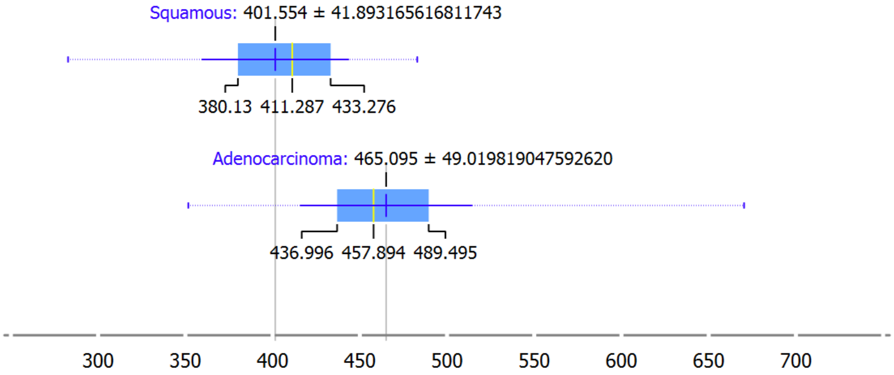

Fig. 4. Box-plots for the *MajorAxisLength* radiomic feature.

shows from the box-plots, there is no overlapping between two distribution. Moreover, there is a significant numeric distance between the third quartile of the *Squamous Cell Carcinoma* distribution (i.e., 420.54) and the first quartile of the *Adenocarcinoma* one (i.e., 477.581), symptomatic that the value of this bio-marker can be really effective in the lung cancer discrimination.

C. Decision Boundaries

Figure 6 shows the decision boundaries for the seven classification algorithms involved in the study, where the top boxes of Figure 6 is related to decision boundaries for the genomic bio-markers, while the down ones for the radiomic ones.

In decision boundaries, the areas where the classifier is able to predict a certain class are identified with different colors (in particular the blue area is related to *Squamous Cell Carcinoma* while the red one are related to *Adenocarcinoma* class). Also the instances are colored according to their class. In particular, the most colored areas correspond to areas in the plane where the prediction is carried out with high accuracy, while the

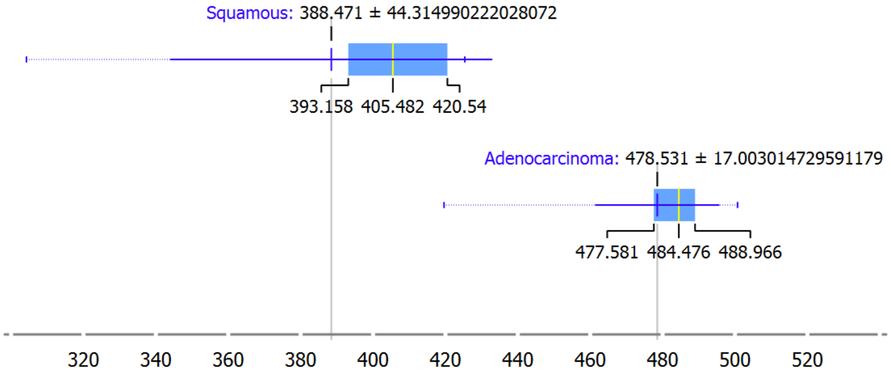

Fig. 5. Box-plots for the *Maximum2DDiameterSlice* radiomic features.

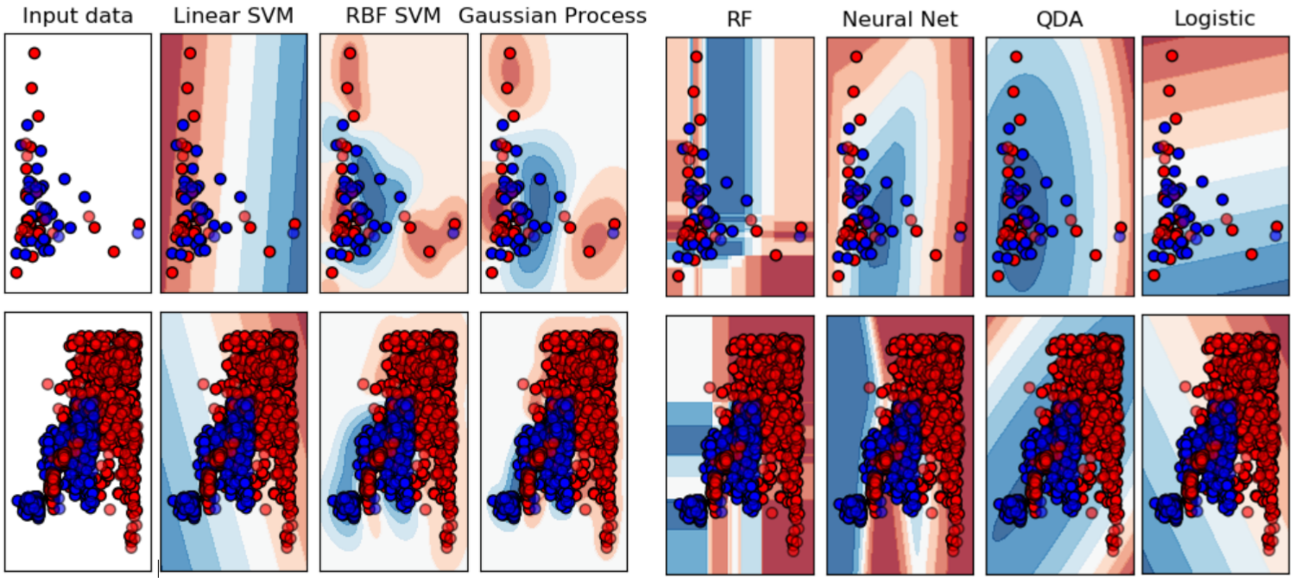

Fig. 6. Decision Boundaries for the genomic and radiomic biomarkers.

less colored areas correspond to spaces in the plane where the classifier is able to make the prediction, but with a less accuracy. The remaining areas i.e., the blank areas, represent the areas in which the classifier is not able to make a prediction for the instances falling in these areas.

From the decision boundaries analysis in Figure 6 emerges that several algorithms exhibits white areas where they are not able to make the prediction for instance, by analysing the genomic bio-markers it emerges that the *Linear SVM*, the *Gaussian Process*, *RF* and *Logistic* algorithms exhibit several white areas. With regard to the radiomic bio-markers decision boundaries in addition to these algorithms also the *RBF SVM* and the *QDA* one exhibits white areas. Only the *Neural Net* algorithm does not exhibit white areas, for this reason the models built whit this algorithms (using the genomic and the radiomic bio-makers) should be able to obtain better performances if compared with the ones obtained from other algorithms.

D. Classification Results

With the aim to understand whether we can confirm the outcomes of descriptive statistics and decision boundaries analysis, below we present the classification results related to the seven algorithms considered in this study built using both the genomic and the radiomic bio-markers.

We consider following metrics in order to evaluate the performance results: False Positive, Precision, Recall, F-Measure and Roc Area.

Table I shows the classification results for the models built by exploiting the genomic and the radiomic bio-markers: with the *H* letter we refer to the lung cancer detection (i.e., *Adenocarcinoma* and *Squamous Cell carcinoma*), while with the *N* to the lung cancer characterisation.

From the results shown in Table I emerges that the models to discriminate between *Squamous Cell Carcinoma* and *Adenocarcinoma* built by exploiting the genomic bio-markers obtain an f-measure ranging from 0.667 obtained with the *Gaussian Process* algorithm to 0.929 reached with the *Neural Net* one.

TABLE I
GENOMIC AND RADIOMIC BIO-MARKERS CLASSIFICATION RESULTS.

Bio-markers	**Algorithm**	**Prediction**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**RocArea**
Genomic	LinearSVM	H	0.123	0.908	0.908	0.907	0.892
	LinearSVM	N	0.333	0.857	0.453	0.593	0.560
	RBF SVM	H	0.125	0.903	0.904	0.903	0.888
	RBF SVM	N	0.650	0.711	0.692	0.701	0.521
	Gaussian Process	H	0.414	0.666	0.677	0.667	0.632
	Gaussian Process	N	0.373	0.739	0.492	0.543	0.560
	RF	H	0.126	0.897	0.899	0.898	0.884
	RF	N	0.774	0.737	0.722	0.708	0.701
	Neural Net	H	0.167	0.907	0.951	0.929	0.892
	Neural Net	N	0.750	0.824	0.792	0.808	0.721
	QDA	H	0.542	0.717	0.805	0.759	0.632
	QDA	N	0.598	0.610	0.354	0.412	0.368
	Logistic	H	0.333	0.833	0.976	0.899	0.942
	Logistic	N	1.000	0.815	1.000	0.898	0.500
Radiomic	LinearSVM	H	0.000	1.000	0.887	0.940	0.943
	LinearSVM	N	0.988	0.988	0.988	0.982	0.988
	RBF SVM	H	0.001	0.999	0.888	0.940	0.942
	RBF SVM	N	0.986	0.986	0.986	0.980	0.984
	Gaussian Process	H	0.865	0.719	0.977	0.829	0,571
	Gaussian Process	N	0.717	0.827	1.000	0.905	0.641
	RF	H	0.005	0.991	0.989	0.991	0.997
	RF	N	0.169	0.946	0.945	0.943	0.994
	Neural Net	H	0.003	0.993	0.993	0.993	1.000
	Neural Net	N	0.006	0.985	0.984	0.985	1.000
	QDA	H	0.607	0.722	0.720	0.645	0.658
	QDA	N	0.707	0.665	0.642	0.621	0.601
	Logistic	H	0.005	0.992	0.988	0.990	0.996
	Logistic	N	0.007	0.983	0.981	0.981	0.998

With respect to the spread lymph node detection the f-measure is ranging from 0.412 with the *QDA* algorithm to a value equal to 0.808 obtained with the *Neural Net* model.

With regard to the models built exploiting radiomic biomarkers in the discrimination between *Squamous Cell Carcinoma* and *Adenocarcinoma* the f-measure is ranging between 0.645 with the *Gaussian Process* classifier to 0.993 obtained with the *Neural Net*. For lymph node spread detection, the f-measure is ranging from 0.621 with the *Gaussian Process* classifier to 0.985 reached with the *Neural Net* model.

The classification performances results highlight that the non invasive radiomic bio-markers are more effective that invasive one (i.e., genomic) in lung cancer detection and characterisation.

The best classification algorithm for both the lung cancer detection and characterisation tasks is the *Neural Net* with both the genomic and radiomic bio-markers. This is confirming the outcomes of the descriptive statistics and of the decision boundaries analysis. We highlight the effectiveness of the radiomic bio-markers, as evidenced by the obtained performances: f-measure 0.993 for lung cancer detection (with the genomic features an f-measure equal to 0.929 is obtained). The same consideration can be inferred for the lymph node spread detection, where the *Neural Net* model built with the genomic bio-markers reaches an f-measure of 0.985 while the one built with the radiomic bio-markers obtaines an f-measure equal to 0.993.

Considering that the algorithm obtaining the best performances in terms of f-measure is the neural network, in the follow we evaluate the loss function trends of the model for lung cancer detection. The loss function evaluation is aimed to find the best loss function for the model training. We experiment with seven different loss configurations: three with constant learning rate (i.e., *constant learning rate*, *constant with momentum* and *constant Nesterov's momentum*), three with learning rate not constant (i.e., i.e., *inv-scaling learning-rate*, *inv-scaling with momentum* and *inv-scaling Nesterov's momentum*) and Adam (i.e. the adaptative moment estimation). The best loss function is the one that permits firstly to reach the lowest value (symptomatic that the prediction provided by the neural network is really closed to the real one) and as a second point to reach the lowest value in the shortest time.

As emerging from the plots in Figure 7 for the genomic model *adam* is able to reach the lowest value, while with regard to the radiomic model both the *adam* and the *constant with momentum* are able to reach the lowest value.

IV. CONCLUSION AND FUTURE WORK

With the aim to propose a non invasive method for lung cancer detection, in this paper we investigated the effectiveness of supervised machine learning techniques for lung cancer characterisation. In detail we consider a set of invasive genomic bio-markers (extracted by tissue biopsy) and a set of non-invasive radiomic bio-markers and we compare the performances reached by these models in lung cancer detection. The experimental analysis, performed considering genomic and radiomic bio-markers related to 130 patients, demonstrated

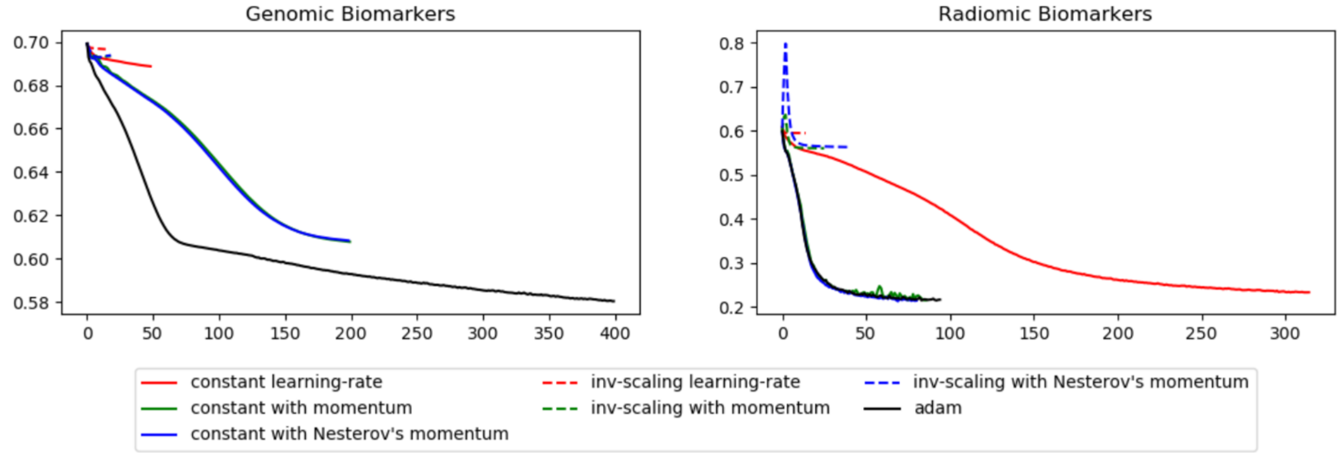

Fig. 7. Loss function trends for the models built with genomic and radiomic bio-markers.

that the models built exploiting radiomic bio-markers are able to reach better performances if compared with the ones trained by using genomic features. As future work we plan to introduce formal verification techniques [19]–[21], [21], [22] with the aim to improve the obtained performances.

References

[1] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Neural networks for lung cancer detection through radiomic features," in *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, 2019, pp. 1–10.

[2] A. Manhire, M. Charig, C. Clelland, F. Gleeson, R. Miller, H. Moss, K. Pointon, C. Richardson, and E. Sawicka, "Guidelines for radiologically guided lung biopsy," *Thorax*, vol. 58, no. 11, pp. 920–936, 2003.

[3] E. A. Kazerooni, F. T. Lim, A. Mikhail, and F. J. Martinez, "Risk of pneumothorax in ct-guided transthoracic needle aspiration biopsy of the lung." *Radiology*, vol. 198, no. 2, pp. 371–375, 1996.

[4] E. Sadot and E. Y. Lee, "Pleural effusion and pneumothorax," in *Imaging in Pediatric Pulmonology*. Springer, 2020, pp. 237–252.

[5] H. Li, K. Shi, Y. Zhao, J. Du, D. Hu, and Z. Liu, "Timp-1 and mmp-9 expressions in copd patients complicated with spontaneous pneumothorax and their correlations with treatment outcomes," *Pakistan Journal of Medical Sciences*, vol. 36, no. 2, 2020.

[6] R. Golan, C. Jacob, and J. Denzinger, "Lung nodule detection in ct images using deep convolutional neural networks," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 243–250.

[7] W. Sun, B. Zheng, and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," in *Medical Imaging: Computer-Aided Diagnosis*, 2016.

[8] Y. Sumathipala, M. Shafiq, E. Bongen, C. Brinton, and D. Paik, "Machine learning to predict lung nodule biopsy method using ct image features: A pilot study," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 1–8, 2019.

[9] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in ct images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, 2019.

[10] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[11] T. M. Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[13] B. Scholkopf, K.-K. Sung, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2758–2765, 1997.

[14] M. Seeger, "Gaussian processes for machine learning," *International journal of neural systems*, vol. 14, no. 02, pp. 69–106, 2004.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] J. M. Zurada, *Introduction to artificial neural systems*. West publishing company St. Paul, 1992, vol. 8.

[17] S. Bose, A. Pal, R. SahaRay, and J. Nayak, "Generalized quadratic discriminant analysis," *Pattern Recognition*, vol. 48, no. 8, pp. 2676–2684, 2015.

[18] F. E. Harrell Jr, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[19] M. G. Cimino, N. De Francesco, F. Mercaldo, A. Santone, and G. Vaglini, "Model checking for malicious family detection and phylogenetic analysis in mobile environment," *Computers & Security*, p. 101691, 2019.

[20] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "A blockchain based proposal for protecting healthcare systems through formal methods," *Procedia Computer Science*, vol. 159, pp. 1787–1794, 2019.

[21] ——, "Prostate gleason score detection and cancer treatment through real-time formal verification," *IEEE Access*, vol. 7, pp. 186 236–186 246, 2019.

[22] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques," *Procedia Computer Science*, vol. 112, no. C, pp. 2519–2528, 2017.