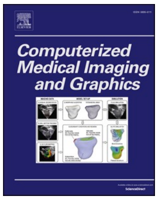
Computerized Medical Imaging and Graphics 110 (2023) 102310

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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



On-cloud decision-support system for non-small cell lung cancer histology characterization from thorax computed tomography scans

Selene Tomassini [a], Nicola Falcionelli [a], Giulia Bruschi [a], Agnese Sbrollini [a], Niccolò Marini [b], Paolo Sernani [c], Micaela Morettini [a], Henning Müller [b], Aldo Franco Dragoni [a], Laura Burattini [a,*]

[a] Department of Information Engineering, Università Politecnica delle Marche (UNIVPM), Ancona, Italy
[b] Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[c] Department of Law, University of Macerata (UNIMC), Macerata, Italy

ARTICLE INFO

Keywords:
Clinical decision-support system
Cloud computing
Convolutional long short-term memory
Non-small cell lung cancer histology characterization
Supervised deep learning
Thorax computed tomography

ABSTRACT

Non-Small Cell Lung Cancer (NSCLC) accounts for about 85% of all lung cancers. Developing non-invasive techniques for NSCLC histology characterization may not only help clinicians to make targeted therapeutic treatments but also prevent subjects from undergoing lung biopsy, which is challenging and could lead to clinical implications. The motivation behind the study presented here is to develop an advanced on-cloud decision-support system, named LUCY, for non-small cell LUng Cancer histologY characterization directly from thorax Computed Tomography (CT) scans. This aim was pursued by selecting thorax CT scans of 182 LUng ADenocarcinoma (LUAD) and 186 LUng Squamous Cell carcinoma (LUSC) subjects from four openly accessible data collections (NSCLC-Radiomics, NSCLC-Radiogenomics, NSCLC-Radiomics-Genomics and TCGA-LUAD), in addition to the implementation and comparison of two end-to-end neural networks (the core layer of whom is a convolutional long short-term memory layer), the performance evaluation on test dataset (NSCLC-Radiomics-Genomics) from a subject-level perspective in relation to NSCLC histological subtype location and grade, and the dynamic visual interpretation of the achieved results by producing and analyzing one heatmap video for each scan. LUCY reached test Area Under the receiver operating characteristic Curve (AUC) values above 77% in all NSCLC histological subtype location and grade groups, and a best AUC value of 97% on the entire dataset reserved for testing, proving high generalizability to heterogeneous data and robustness. Thus, LUCY is a clinically-useful decision-support system able to timely, non-invasively and reliably provide visually-understandable predictions on LUAD and LUSC subjects in relation to clinically-relevant information.

1. Introduction

Lung cancer is one of the cancers with the highest incidence and mortality rate (Tomassini et al., 2022a), accounting over 1.8 million new cases in the world and 1.4 million deaths every year (Prabhu et al., 2022). According to the differentiation based on lung cancer cell size from an histological point of view, about 85% of all lung cancers is Non-Small Cell Lung Cancer (NSCLC) (Naik and Edla, 2021; Marentakis et al., 2021). NSCLC is a malignant lung mass generally located in the mediastinum (Rivera et al., 2013). It can be categorized into three histological subtypes: LUng ADenocarcinoma (LUAD), LUng Squamous Cell carcinoma (LUSC) and LUng Large Cell carcinoma (LULC) (Naik and Edla, 2021; Guo et al., 2020; Suster and Mino-Kenudson, 2020). LULC is diagnosed by exclusion. It represents a rare form (less than 10% of NSCLC histological subtypes) and its diagnosis is restricted to surgically-resected lung cancers with dubious immunohistochemical or morphological differentiation (Travis et al., 2015). Hence, LUAD and LUSC account for about 90% of NSCLC histological subtypes (Kriegsmann et al., 2020; Han et al., 2021; Liu et al., 2021). LUAD originates in the submucosal glands and is generally located along the outer edges of the lungs, not rarely presenting a star-like contour (Tomassini et al., 2022a; Marentakis et al., 2021). In the majority of cases, it appears like a malignant lung mass smaller than 3 cm (Panunzio and Sartori, 2020). LUSC originates in the squamous cells and is generally located in the middle of the lungs, often presenting a central necrosis and a wall thickness larger than 1.5 cm (Tomassini et al., 2022a; Marentakis et al., 2021). In the majority of cases, it appears like a malignant lung mass bigger than 4 cm (Panunzio and Sartori, 2020).

* Corresponding author.
E-mail address: l.burattini@univpm.it (L. Burattini).

https://doi.org/10.1016/j.compmedimag.2023.102310
Received 18 July 2023; Received in revised form 25 October 2023; Accepted 3 November 2023
Available online 10 November 2023
0895-6111/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Lung cancer diagnosis at an early stage is extremely important to increase the subjects' survival rate by improving therapeutic treatments (Monkam et al., 2019; Halder et al., 2020). As diagnostic imaging techniques, clinicians recommend exams, such as Computed Tomography (CT), Positron Emission Tomography (PET) and Magnetic Resonance (MR) (Zhang et al., 2018). PET and MR show limitations in detecting lung masses. MR, in particular, is highly susceptible to miss smaller ones (Naik and Edla, 2021; Thakur et al., 2020). CT, especially low-dose CT, is the most sensitive to small, calcified lung masses (Tomassini et al., 2022a; Cao et al., 2020; Adiraju and Elias, 2021). CT has high spatial resolution, low noise and low distortion, and it is also rapid, non-invasive, widely available and quite affordable. CT allows to obtain a 3D visualization of the thorax, as each lung mass is detected and additional information about other lung structures can be assessed (Zhang et al., 2018; Pereira et al., 2021). Nevertheless, lung cancer is among the most frequently misdiagnosed diseases by CT scan visualization only (Halder et al., 2022). Radiologists make substantial efforts to examine CT scans, search for lung masses and determine whether they are malignant based on their size, shape and texture, as interpretation is subjective (i.e., it depends on professional experience) (Shen et al., 2019; Rubin, 2015). Under ideal circumstances, radiologists spend about 5 min per lung mass (Tomassini et al., 2022a; Rubin, 2015). With the presence of confounders, such as distraction, fatigue, inter-observer variability and intra-observer variability, radiologists could take wrong or even missing decisions, as they may overlook potentially-malignant lung masses (Winkels and Cohen, 2019; Zhao et al., 2013; Pinsky et al., 2013).

Lung cancer histology characterization is fundamental (Naik and Edla, 2021; Marentakis et al., 2021), as the effectiveness of therapeutic treatments as well as the risk of complications are different for each lung cancer histological subtype (Bębas et al., 2021; Cong et al., 2020). NSCLC histological subtypes are characterized by peculiar alterations that allow to quite easily differentiate them at the molecular level (Fig. 1) (Suster and Mino-Kenudson, 2020). LUAD presents acinar, lepidic, cribriform, papillary, micropapillary and solid histopathological growth patterns, whereas LUSC presents abundant eosinophilic cytoplasm, keratin pearl formation and clusters of polyhedral cells (Prabhu et al., 2022). Conversely, the unambiguous differentiation of NSCLC histological subtypes on the basis of morphological features only is still a challenge (Fig. 1) (Han et al., 2021; Liu et al., 2021). As a result, in subjects suspected of having lung cancer based on CT findings, it is recommended to undergo lung biopsy for accurately characterizing the histological subtype (Bębas et al., 2021; Planchard et al., 2018). By inspecting CT scans, potentially-malignant lung masses are detected. Subsequently, lung biopsy is performed and the microscopic structure of the excised tissue sample is analyzed (Han et al., 2021). Lung biopsy, mainly in the form of transthoracic fine needle biopsy, is the first choice for lung cancers located in peripheral lung structures (Planchard et al., 2018). For lung cancers located in proximity to airways or blood vessels as well as for those located in deepest lung structures, performing lung biopsy is very challenging and may also lead to clinical implications (Han et al., 2021; Zhang et al., 2019). Thus, lung biopsy is strongly discouraged in subjects with complex clinical conditions (Guo et al., 2020; Han et al., 2021). Moreover, excising a small tissue sample may not exactly characterize the potentially-malignant lung mass entirely because of its heterogeneous nature (Tomassini et al., 2022a; Marentakis et al., 2021). As a consequence, oncologists could fail in characterizing lung cancer histological subtypes (Moitra and Mandal, 2020). Therefore, developing non-invasive techniques for lung cancer histology characterization may not only help clinicians to make targeted therapeutic treatments in time but also prevent subjects from undergoing lung biopsy (Guo et al., 2020; Han et al., 2021).

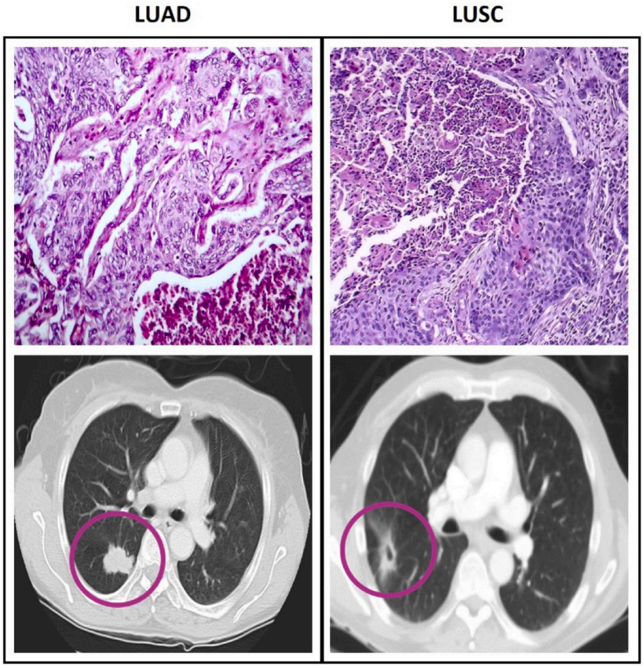


Fig. 1. LUAD (left) and LUSC (right) from a microscopic (above) and CT (below) image.

To the authors' best knowledge, Cloud-YLung, the framework published by Tomassini et al (Tomassini et al., 2022b), is the first leveraging on a Convolutional Long Short-Term Memory (ConvLSTM)-based neural network for NSCLC histology characterization prior to lung biopsy. However, in Cloud-YLung as well as in all the other state-of-the-art frameworks (Section 2), a challenging evaluation protocol to prove the generalization capability to heterogeneous data coming from different sites was not used, neither clinically-relevant information was included in the analysis nor visually-understandable outcomes were dynamically generated. Thus, the motivation behind the study presented here is to propose a fully automatic procedure dedicated to efficient learning of lung mass-related features while overcoming the main limitations raised in (Tomassini et al., 2022b) by developing an advanced on-cloud decision-support system, named LUCY, for non-small cell LUng Cancer histologY characterization directly from thorax CT scans. The main scope is to make available a clinically-useful decision-support system able to non-invasively and reliably provide visually-understandable predictions on LUAD and LUSC subjects in relation to clinically-relevant information (i.e., NSCLC histological subtype location and grade), and with the potential to be not only extended to other pulmonary pathologies by keeping the anatomy of the lungs unaltered but also integrated in any other system for real diagnostic purposes thanks to its machine-independent nature and visually-understandable outcomes. Furthermore, the main source code of LUCY is available, under copyright, on a GitHub repository[1] to promote transparency and reproducibility in the scientific community.

2. Literature review

In the literature, the approaches developed for characterizing NSCLC histological subtypes are mostly focused on the processing of microscopic images (Li et al., 2021). Nevertheless, NSCLC histological subtype characterization directly from radiological data may have significant implications for diagnostic decisions and therapeutic treatments (Chaunzwa et al., 2021). For instance, Shen et al (Shen et al., 2017), Zhu et al (Zhu et al., 2018), Liu et al (Liu et al., 2019) and Yang et al (Yang et al., 2020) explored the potential of radiomics-based Machine Learning (ML) algorithms in NSCLC histological subtype characterization directly from CT data. However, radiomics-based ML algorithms

[1] https://github.com/S3l11/LUCY

rely upon predefined radiomics features. Radiomics feature analysis and comparison is strongly influenced by variability. By adopting different radiomics features, clinicians may find difficult to select the appropriate feature set. Thus, the application of such algorithms may be not straightforward in the real clinical practice (Liu et al., 2021). The major drawback of radiomics, indeed, is the lack of standardization and reproducibility in the feature extraction process (Thawani et al., 2018). In particular, CT-derived radiomics features and their applicability heavily depend on the gray level number and voxel size (Reiazi et al., 2021). Hence, a more automatic feature extraction procedure may positively impact the learning process.

Among the plethora of Deep Learning (DL) algorithms, the exploitation of Convolutional Neural Networks (CNNs) in the characterization of NSCLC histological subtypes directly from CT data is increasing (Tomassini et al., 2022a). Generally, CNNs take single slices in input, as done by Chaunzwa et al (Chaunzwa et al., 2021). One of the main limitations of these approaches (i.e., slice-based approaches) is that NSCLC histological subtypes are analyzed by looking at the bidimensional space only. However, slices of the same CT scan form a spatial sequence and each NSCLC histological subtype has a certain depth that cannot be captured by analyzing each slice independently (Tomassini et al., 2022b). Therefore, it is necessary to integrate the volumetric information by designing and developing a framework that includes a neural network able to digest a sequence of slices as a whole (Tomassini et al., 2022a, 2022b). These latter approaches (i.e., scan-based approaches) have the potential to analyze the volumetric information in more detail (Tomassini et al., 2022a). Scan-based approaches comprehend 3D CNNs but also less computationally-expensive algorithms, such as special recurrent neural networks (like ConvLSTMs) used independently and time-distributed 2D CNNs coupled with recurrent layers. Recurrent modules, in fact, are able to exploit their internal state to process sequences of slices and connect the previous information to the present one (Marentakis et al., 2021).

Until now, still few studies aimed to non-invasively characterize NSCLC histological subtypes. The majority of them targeted private data collections, as the one of Guo et al (Guo et al., 2020) where a 3D CNN was developed and trained from scratch, and unprocessed thorax CT scans (554 LUAD and 175 LUSC) were analyzed. Only three studies targeted openly-accessible data collections. Marentakis et al (Marentakis et al., 2021) investigated the potential of four automatic procedures to characterize NSCLC histological subtypes directly from thorax CT scans. They processed 48 LUAD and 54 LUSC scans belonging to NSCLC-Radiomics[2], and developed (1) two radiomics-based ML algorithms, (2) four pretrained 2D CNNs with fine tuning, (3) one pretrained time-distributed 2D CNN combined with a Long Short-Term Memory (LSTM)-based neural network and (4) two joint models. Algorithms (1) and (2) ignored the volumetric information, whereas algorithms (3) and (4) considered it. The one that reached the highest performance was algorithm (3), with a test ACCuracy (ACC) of 74% and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) of 78%. A notable finding of their study was that adding radiomics to the best-performing algorithm did not show any further performance improvement. Tomassini et al (Tomassini et al., 2022b) considered the same openly-accessible data collection to fulfill the same task. Specifically, they processed 50 LUAD and 50 LUSC scans from NSCLC-Radiomics[2], and accomplished both automatic feature extraction and classification by means of a ConvLSTM-based neural network trained from scratch on a scalable GPU cloud service. Their framework, Cloud-YLung precisely, achieved a test ACC of 75% and AUC of 84%, outperforming the best-performing algorithm proposed by Marentakis et al (Marentakis et al., 2021). A different openly-accessible data collection was taken into account by Moitra et al (Moitra and Mandal, 2020), which built a framework that combined a pretrained time-distributed 2D CNN and a bidirectional LSTM (biLSTM)-based neural network to non-invasively characterize NSCLC histological subtypes as LUAD, LUSC or not otherwise specified. To accomplish this task, they processed PET/CT scans of 211 subjects from NSCLC-Radiogenomics[3]. In its best guise, their framework gained a test ACC of 96% and AUC of 99%.

3. LUCY

With the objective to develop an advanced on-cloud decision-support system for NSCLC histology characterization directly from thorax CT scans, LUCY was implemented. Fig. 2 depicts the workflow of LUCY, whereas data and methodological details are reported in the following Subsections.

3.1. Data

A total of 368 unprocessed thorax CT scans (182 LUAD and 186 LUSC subjects) belonging to different openly-accessible data collections of The Cancer Imaging Archive (TCIA)[4] were taken into account, along with biopsy-confirmed labels (Table 1) (Clark et al., 2013). Specifically, thorax CT scans were selected from four openly-accessible data collections: NSCLC-Radiomics[2], NSCLC-Radiogenomics[3], NSCLC-Radiomics-Genomics[5] and TCGA-LUAD[6]. The data selection procedure was conducted under the following criteria: inclusion of CT as imaging modality, thorax as anatomical site, a number of slices per scan higher than 40 and lower than 360, and 512 pixels × 512 pixels as original slice resolution; exclusion of phantoms and 3rd-party results. In case of multiple scans per subject, only the first meeting the afore-mentioned criteria was selected to avoid an intra-subject bias.

3.1.1. Data from NSCLC-Radiomics

A total of 200 unprocessed thorax CT scans (50 LUAD and 150 LUSC subjects) were selected from NSCLC- Radiomics[2]. Each scan was made up of a variable number of slices ranging from 87 to 297 with a resolution of 512 pixels × 512 pixels. All selected scans belonged to anonymized subjects ranging from 45 to 88 years in age at histological diagnosis. Among them, 18 LUAD and 40 LUSC scans belonged to women, whereas 32 LUAD and 110 LUSC scans belonged to men. All thorax CT scans were acquired with Siemens and CMS scanners. Once selected, they were stored as Digital Imaging and Communications in Medicine (DICOM) files.

3.1.2. Data from NSCLC-Radiogenomics

A total of 22 unprocessed thorax CT scans (all LUAD subjects) were selected from NSCLC-Radiogenomics[3]. Each scan was made up of a variable number of slices ranging from 56 to 348 with a resolution of 512 pixels × 512 pixels. All selected scans belonged to anonymized subjects ranging from 24 to 80 years in age at histological diagnosis. Among them, 15 LUAD scans belonged to women, whereas 7 LUAD scans belonged to men. All thorax CT scans were acquired with Siemens scanners. Once selected, they were stored as DICOM files.

3.1.3. Data from NSCLC-Radiomics-Genomics

A total of 78 unprocessed thorax CT scans (42 LUAD and 36 LUSC subjects) were selected from NSCLC-Radiomics-Genomics[5]. Each scan was made up of a variable number of slices ranging from 50 to 356 with a resolution of 512 pixels × 512 pixels. All selected scans belonged to anonymized subjects with no information about age at histological

[2] https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics

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

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

[5] https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Genomics

[6] https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD

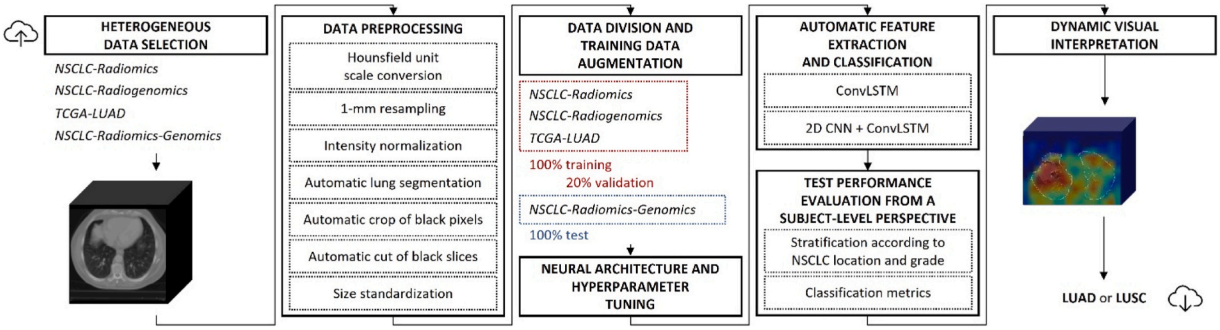

HETEROGENEOUS DATA SELECTION
NSCLC-Radiomics
NSCLC-Radiogenomics
TCGA-LUAD
NSCLC-Radiomics-Genomics
DATA PREPROCESSING
Hounsfield unit scale conversion
1-mm resampling
Intensity normalization
Automatic lung segmentation
Automatic crop of black pixels
Automatic cut of black slices
Size standardization
DATA DIVISION AND TRAINING DATA AUGMENTATION
NSCLC-Radiomics
NSCLC-Radiogenomics
TCGA-LUAD
100% training
20% validation
NSCLC-Radiomics-Genomics
100% test
NEURAL ARCHITECTURE AND HYPERPARAMETER TUNING
AUTOMATIC FEATURE EXTRACTION AND CLASSIFICATION
ConvLSTM
2D CNN + ConvLSTM
TEST PERFORMANCE EVALUATION FROM A SUBJECT-LEVEL PERSPECTIVE
Stratification according to NSCLC location and grade
Classification metrics
DYNAMIC VISUAL INTERPRETATION
LUAD or LUSC

Fig. 2. The workflow of LUCY.

Table 1
Summary of unprocessed data details.

Data collection	LUAD + LUSC (#)	Age at diagnosis (avg y.o.)	Women + men (#)	Slices per scan (avg #)	Slice resolution (x × y)
NSCLC-Radiomics	50 + 150	67	58 + 142	192	512 × 512
NSCLC-Radiogenomics	22 + 0	52	15 + 7	202	512 × 512
NSCLC-Radiomics-Genomics	42 + 36	n. s.	26 + 52	203	512 × 512
TCGA-LUAD	68 + 0	n. s.	n. s.	60	512 × 512
Total	182 + 186	n. a.	n. a.	164	512 × 512

n. s.: not specified, if data details are not specified by data providers
n. a.: not applicable, if it cannot be computed due to missing fields

diagnosis. Among them, 19 LUAD and 7 LUSC scans belonged to women, whereas 23 LUAD and 29 LUSC scans belonged to men. All thorax CT scans were acquired with Siemens, GE Medical Systems and Philips scanners. Once selected, they were stored as DICOM files.

3.1.4. Data from TCGA-LUAD

A total of 68 unprocessed thorax CT scans (all LUAD subjects) were selected from TCGA-LUAD[6]. Each scan was made up of a variable number of slices ranging from 48 to 71 with a resolution of 512 pixels × 512 pixels. All selected scans belonged to anonymized subjects with no information about either age at histological diagnosis or sex. All thorax CT scans were acquired with Siemens scanners. Once selected, they were stored as DICOM files.

3.2. Data preprocessing

Since the lungs fill a small fraction of a thorax CT scan, all selected scans were preprocessed with the purpose of removing all disrupting information. First, Hounsfield Unit (HU) scale conversion was accomplished to describe voxel values, considering the intensity window ranging from − 1024 HU to 400 HU. Next, 1-mm resampling was performed to make slices within scans spatially homogeneous. Such spatial homogeneity was described by the x, y and z parameters, and produced a remarkable effect not within each slice but between slices. The x and y parameters, which refers to intra-slice physical arrangement, had a smaller variance than the z parameter, which refers to inter-slice distance. As a result, when the z parameter was bigger than 1, some artificial slices were generated by interpolation; when the z parameter was smaller than 1, some slices were discarded. After 1-mm resampling, intensity normalization was done to smooth out the intensity variation caused by the use of different scanners or scanning parameters during the acquisition, such that the pixel values ranged from 0 to 1. At that point, automatic lung parenchyma segmentation was fulfilled for removing all non-lung tissues by exploiting UNet(R231), a pretrained UNet-like neural network developed and publicly released by Hofmanninger et al (Hofmanninger et al., 2020). Then, each scan was automatically cropped by eliminating all non-informative voxels. Likewise, non-informative slices were automatically cut from each scan. After automatically cropping non-informative voxels and cutting non-informative slices, scan shapes resulted to be all different. Hence, all scans were resized to 250 pixels × 190 pixels × 270 pixels, which was the weighted average shape of the cropped-and-cut scans. For what concerns the z axis, when the number of slices per scan was bigger than 250, the lung volume was centered by discarding initial and final slices in equal number; when the number of slices per scan was smaller than 250, zero padding was performed. Eventually, eighteen preprocessed LUAD and LUSC scans were manually discarded, as most of their slices were dark and, therefore, not informative. Thus, a total of 350 preprocessed thorax CT scans (172 LUAD and 178 LUSC subjects), whose details are summarized in Table 2, were kept for subsequent automatic feature extraction and classification. Fig. 3 shows how a thorax CT scan appears after the afore-mentioned preprocessing steps.

3.3. Data division and augmentation

The choice on how to divide data for training, validation and testing was made according to both cardinality of the samples for each class and presence of clinically-relevant information (i.e., NSCLC histological subtype location and grade) in the corresponding Comma-Separated Values (CSV) file. Accordingly, 100% of preprocessed thorax CT scans belonging to NSCLC-Radiomics[2], NSCLC-Radiogenomics[3] and TCGA-LUAD[6] was used as training dataset, 20% of which served as validation dataset. As test dataset, 100% of preprocessed thorax CT scans belonging to NSCLC-Radiomics-Genomics[5] was used, in order to make the evaluation protocol challenging by computing the performance of LUCY on scans belonging to a different data collection from the ones used in both training and validation phases.

To face the scarcity of training data and also mitigate the overfitting effect (Shorten and Khoshgoftaar, 2019), the training dataset, already balanced in terms of class prevalence (i.e., both classes shared exactly the same number of samples), was augmented through 15° left/right rotation and random in/out zoom ranging from 0.8 to 1.2. Rotation and zooming were chosen as augmentation techniques because they do not alter the real appearance of the lungs, which for instance horizontal flipping and vertical flipping do.

Table 2
Summary of preprocessed data details.

Data collection	LUAD + LUSC (#)	Age at diagnosis (avg y.o.)	Women + men (#)	Slices per scan (avg #)	Slice resolution (x × y)
NSCLC-Radiomics	50 + 150	67	58 + 142	251	274 × 199
NSCLC-Radiogenomics	22 + 0	52	15 + 7	268	272 × 188
NSCLC-Radiomics-Genomics	32 + 28	n. s.	20 + 40	195	265 × 186
TCGA-LUAD	68 + 0	n. s.	n. s.	274	278 × 191
Total	172 + 178	n. a.	n. a.	250	270 × 190

n. s.: not specified, if data details are not specified by data providers
n. a.: not applicable, if it cannot be computed due to missing fields

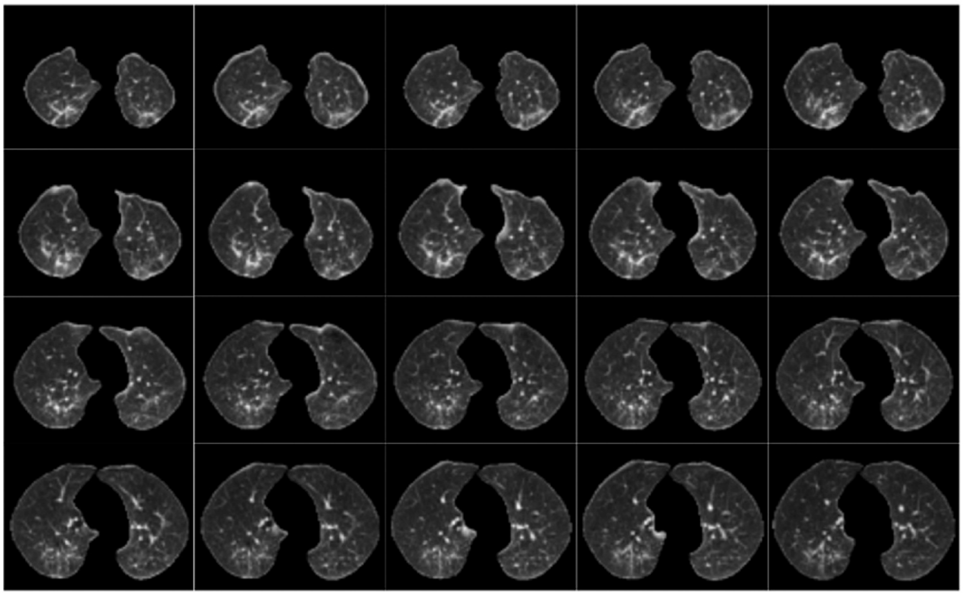


Fig. 3. Appearance of a bunch of slices (20 out of 250 total slices) of a preprocessed thorax CT scan. Slice progression goes from left to right.

3.4. Neural architecture and hyperparameter tuning

Two end-to-end neural networks, a ConvLSTM-based Neural Network (CLSTM-NN) and a Time-Distributed 2D CNN combined with a ConvLSTM-based Neural Network (TDCNN-CLSTM-NN), were designed and compared. CLSTM-NN was implemented because it demonstrated to outperform other scan-based approaches in NSCLC histology characterization (Tomassini et al., 2022b). TDCNN-CLSTM-NN was chosen as term of comparison because time-distributed 2D CNNs combined with recurrent layers demonstrated to be effective in non-invasively characterizing NSCLC histological subtypes, as addressed in Section 2, but no one has ever used a ConvLSTM layer as recurrent layer before.

The final neural architecture of both end-to-end neural networks was customized after a set of preliminary experiments, where the computational cost was kept low and the choice of the optimal hyperparameter configuration was driven by the Bayesian optimization algorithm, as recognized useful to maximize the neural network performance (Snoek et al., 2012; Wu et al., 2019). The Bayesian optimization algorithm was used for both training hyperparameters (i.e., number of epochs, learning rate and batch size) and neural network hyperparameters (i.e., number of filters and dropout rate). Specifically, all the possible combinations between the following values, chosen as they proved to ensure both computational lightness and good speed of training in a preliminary experimental evaluation, were investigated:

- Training hyperparameters:

1. Number of epochs: [40, 50, 60];
2. Learning rate: [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05];
3. Batch size: [1].

- Neural network hyperparameters:

1. Number of filters: [8, 16, 32, 64, 128, 256, 512];
2. Dropout rate: [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8].

The hyperparameter combination that led to the highest validation ACC was selected.

ConvLSTM was chosen as core layer of both end-to-neural networks for taking into account both spatial temporal NSCLC features. ConvLSTM is a special kind of recurrent neural network that exploits convolution filters both input-to-state and state-to-state transitions (Fig. 4). Doing so, it is able to model the long-term interactions while exploring the volumetric information (Tomassini et al., 2022b). As for the "Conv" part, characterizing features are automatically extracted through convolution mechanisms different from the ones exploited by standard CNNs. During each convolution operation, the neural network learns which filters have to be activated when seeing a particular feature at a specific spatial position in the input (Tomassini et al., 2022b). As for the "LSTM" part, the main function of LSTM is to ensure the preservation of the back-propagation error. In a LSTM hidden unit, each sequence is analyzed in its entirety and the acquired information is stored in a gated memory cell. The gated memory cell decides about what to store and when to allow the reading and updating of the information through its input, forget and output gates (Tomassini et al., 2022b). As pointed out by Shi et al (Shi et al., 2015) and Tomassini et al (Tomassini et al., 2022c), the output h_t at time point t is regulated by (1), where * and $\odot$ denote the convolution operator and the Hadamard product, i_t, f_t, o_t and

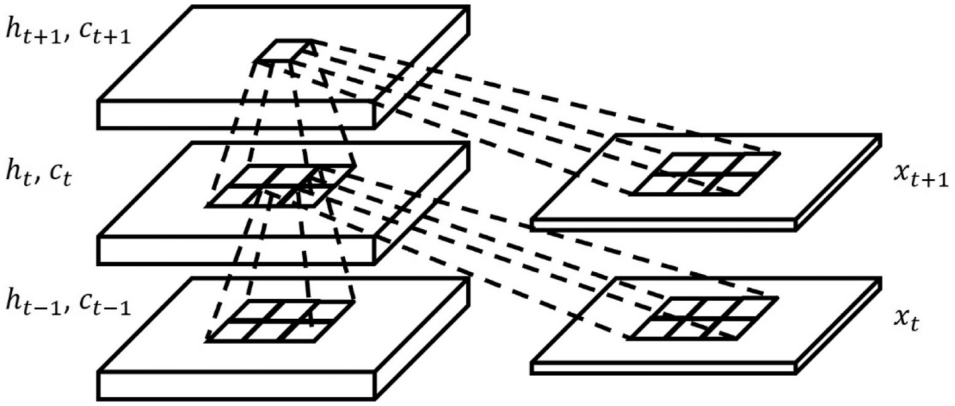


Fig. 4. The inner structure of ConvLSTM.

c_t are the activation vectors of the three gates and of the gated memory cell at time point *t*, x_t is the current input, σ is the sigmoid activation function, *tanh* is the hyperbolic tangent activation function, *W* are the weight matrices, *b* is the bias of the gated memory cell and of its gates.

$$\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \odot c_{t-1} + b_i),\\
f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \odot c_{t-1} + b_f),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c),\\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \odot c_t + b_o),\\
h_t &= o_t \odot \tanh(c_t).
\end{aligned} \tag{1}$$

3.4.1. *ConvLSTM-based neural network*

CLSTM-NN includes a sequential six-layered ConvLSTM-based neural network, whose neural architecture is reported in Table 3. The first layer is a ConvLSTM layer with 8 convolution filters and 3 × 3 kernels to take into account both spatial and temporal NSCLC features. The second layer is a Dropout layer (dropout rate of 70%) to mitigate the overfitting effect. The third layer is a Flatten layer to flatten all the automatically-extracted features into a 1D tensor. The fourth layer is a Dense layer with 128 neurons and Rectified Linear Unit (ReLU) activation function to help the neural network take into account non-linear interactions. The fifth layer is a Dropout layer (dropout rate of 30%). The last layer is a Dense layer with 2 neurons and SoftMax activation function to assign probabilities to each class by producing real values between 0 and 1, with sum equal to 1.

3.4.2. *Time-distributed 2D CNN combined with ConvLSTM-based neural network*

TDCNN-CLSTM-NN includes a sequential twelve-layered 2D CNN followed by a sequential six-layered ConvLSTM-based neural network, whose neural architecture is reported in Table 4. The first layer is a Conv layer with 32 convolution filters, 3 × 3 kernels and ReLU activation function to take into account spatial NSCLC features only. The second layer is a Max Pooling layer (2 × 2 pool and 2 × 2 strides) to downsample the input along its spatial dimensions. The third layer is a Conv layer with 64 convolution filters, 3 × 3 kernels and ReLU activation function. The fourth layer is another Max Pooling layer (2 × 2 pool and 2 × 2 strides). The fifth layer is a Conv layer with 128 convolution filters, 3 × 3 kernels and ReLU activation function. The sixth layer is another Max Pooling layer (2 × 2 pool and 2 × 2 strides). The seventh, eighth and ninth layers are Conv layers with 256 convolution filters, 3 × 3 kernels and ReLU activation functions. The tenth layer is a Batch Normalization layer (momentum of 0.99 and epsilon of 0.001) to maintain the mean output close to 0 and its standard deviation close to 1. The eleventh layer is an Activation layer with ReLU activation function. The twelfth layer is another Max Pooling layer (2 × 2 pool and 2 × 2 strides). The time-distributed layer allows to apply the described 2D CNN to every temporal slice of each thorax CT scan. The ConvLSTM-based neural network is made up by a ConvLSTM layer with 64 convolution filters and 3 × 3 kernels, a Dropout layer (dropout rate of 60%), a Flatten layer, a Dense layer with 128 neurons and ReLU activation function, a Dropout layer (dropout rate of 50%) and a last Dense layer with 2 neurons and SoftMax activation function.

Table 3
CLSTM-NN neural architecture.

Layer	Specifications	Output shape	Parameters
ConvLSTM	8 filters, 3 × 3	(None, 188, 268, 8)	2624
Dropout	0.7	(None, 188, 268, 8)	0
Flatten	-	(None, 403072)	0
Dense	128 neurons, ReLU	(None, 128)	51593344
Dropout	0.3	(None, 128)	0
Dense	2 neurons, SoftMax	(None, 2)	258

Table 4
TDCNN-CLSTM-NN neural architecture.

Layer	Specifications	Output shape	Parameters
Conv	32 filters, 3 × 3, ReLU	(None, 190, 270, 32)	320
Max Pooling	2 × 2	(None, 95, 135, 32)	0
Conv	64 filters, 3 × 3, ReLU	(None, 95, 135, 64)	18496
Max Pooling	2 × 2	(None, 47, 67, 64)	0
Conv	128 filters, 3 × 3, ReLU	(None, 47, 67, 128)	73856
Max Pooling	2 × 2	(None, 23, 33, 128)	0
Conv	256 filters, 3 × 3, ReLU	(None, 23, 33, 256)	295168
Conv	256 filters, 3 × 3, ReLU	(None, 23, 33, 256)	590080
Conv	256 filters, 3 × 3, ReLU	(None, 23, 33, 256)	590080
Batch Normalization	0.99, 0.001	(None, 23, 33, 256)	1024
Activation	ReLU	(None, 23, 33, 256)	0
Max Pooling	2 × 2	(None, 11, 16, 256)	0
Time Distributed	-	(None, 250, 11, 16, 256)	1569024
ConvLSTM	64 filters, 3 × 3	(None, 9, 14, 64)	737536
Dropout	0.6	(None, 9, 14, 64)	0
Flatten	-	(None, 8064)	0
Dense	128 neurons, ReLU	(None, 128)	1032320
Dropout	0.5	(None, 128)	0
Dense	2 neurons, SoftMax	(None, 2)	258

3.5. *Environmental setup and training strategy*

The Pro version of Google Colab cloud service was used as environmental setup, selecting high system RAM (34 GB) and GPU hardware acceleration (NVIDIA Tesla P100 with 16 GB of video RAM) settings.

The Keras library built on a TensorFlow backend (version 2.6.0) was also used.

LUCY was trained from scratch up to a maximum of 50 epochs, using the early stopping callback with a patience of 5 epochs to stop the training at the point where the validation loss reached the minimum. The learning rate was fixed to 0.001 and the batch size to 1. Since, during training, the optimization based on a stochastic gradient is crucial to minimize the loss function while assuring higher efficiency (Tomassini et al., 2022b), the stochastic gradient descent was chosen as optimizer. The binary cross entropy was selected as loss function. During training, the weights that led to the lowest validation loss were saved, as the validation loss captures exactly the divergence between the predicted output and the desired one. Eventually, they were used to evaluate the performance of LUCY on test dataset.

3.6. Performance evaluation

To investigate both generalization capability and robustness of LUCY, its performance was evaluated from a subject-level perspective on 100% of NSCLC-Radiomics-Genomics[5], taken as test dataset, in relation to NSCLC histological subtype location and grade. According to the clinically-relevant information provided in NSCLC-Radiomics-Genomics[5], subjects' predictions were stratified into:

- Five NSCLC histological subtype location groups, namely Left Lower Lobe (LLL), Left Upper Lobe (LUL), Right Lower Lobe (RLL), Right Middle Lobe (RML) and Right Upper Lobe (RUL);
- Four NSCLC histological subtype grade groups, namely Grade Not Available (GNA), Grade 1 (G1), Grade 2 (G2) and Grade 3 (G3).

As classification metrics, the ACC (2), PREcision (PRE) (3), SENsitivity (SEN) (4), F1-score (5), ROC and respective AUC were taken into account, setting the discrimination threshold to 0.5 for computing the ACC, PRE, SEN and F1-score. LUAD was chosen as positive (1;0) class and LUSC as negative (0;1) class because of the higher incidence of the first NSCLC histological subtype (Tomassini et al., 2022a).

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{2}$$

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{3}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{4}$$

$$\mathrm{F1}-\mathrm{score} = \frac{\mathrm{TP}}{\mathrm{TP}+\frac{1}{2}(\mathrm{FP}+\mathrm{FN})}. \tag{5}$$

In the equations:

- *TP* stands for True Positive and it is the case where LUAD subjects are correctly classified as subjects affected by LUAD;
- *TN* stands for True Negative and it is the case where LUSC subjects are correctly classified as subjects affected by LUSC;
- *FP* stands for False Positive and it is the case where LUSC subjects are wrongly classified as subjects affected by LUAD;
- *FN* stands for False Negative and it is the case where LUAD subjects are wrongly classified as subjects affected by LUSC.

Among the afore-mentioned classification metrics, higher attention was paid to the AUC because it highlights the ability to discriminate between LUAD and LUSC classes without depending on any discrimination threshold. For each NSCLC histological subtype location and grade group as well as for the entire test dataset, the AUC of CLSTM-NN was compared with the AUC of TDCNN-CLSTM-NN by means of the DeLong's test (DeLong et al., 1988). Statistical level of significance (P) was set to 0.05.

3.7. Dynamic visual interpretation

To highlight the lung voxels that influenced the automatic decision-making process the most, Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) was included in the learning process of LUCY (TDCNN-CLSTM-NN only). Grad-CAM was chosen as visualization module because it exploits the gradient information flowing into the last convolutional layer to understand the importance of each neuron (Selvaraju et al., 2017). In this study, Grad-CAM was adapted to work with sequences of slices in place of independent images. In order to generate heatmaps, a sub-model was created to map the input image to the last convolutional layer activations. Next, another sub-model was created to map the last convolutional layer activations to the prediction layer (i.e., the last Dense layer). At that point, the gradient of the top predicted class for the input data with respect to the last convolutional layer activations was computed. Since each thorax CT scan was constituted by 250 stacked slices, one heatmap frame was generated for each slice, normalized between 0 and 1, and colorized by using the jet color map. As a result, one heatmap video constituted by 250 stacked slices running at 30 fps was obtained for each thorax CT scans and superimposed to it. This way, it was dynamically displayed where LUCY focused its attention most strongly in the non-invasive NSCLC histology characterization, as the voxel intensity, from blue (low) to red (high), corresponds to the measure of how much it was responsible for a certain prediction.

To provide a better perspective of the dynamic visual interpretation outcomes, a novel analysis procedure was conducted for each NSCLC histological subtype location and grade group as well as for the overall test dataset. First, a threshold value was established to distinguish the red content from the non-red one in each heatmap video. To do so, a Hue Saturation Value (HSV) colormap was created for quick access to unique colors. By referring to the HSV colormap, the appropriate shade of red was determined and, thus, used to create a mask for the extraction of the red content. At that point, a binarization was performed to make all extracted red content appear as white and the rest as black, and the resultant binarized red mask videos were converted into arrays of voxel values (i.e., 0 for black, 255 for white). Eventually, each binarized red mask array was superimposed to the corresponding lung mask array, itself binary, and the match of white voxels was evaluated to determine the correctness of the dynamic visual interpretation outcomes. Fig. 5 graphically synthesizes the afore-mentioned analysis procedure.

4. Results

Fig. 6 and Fig. 7 display the confusion matrices of LUCY (CLSTM-NN and TDCNN-CLSTM-NN, respectively) in classifying LUAD and LUSC test subjects in relation to NSCLC histological subtype location and grade. Table 5 and Table 6 report the ACC, PRE, SEN and F1-score values of LUCY (CLSTM-NN and TDCNN-CLSTM-NN, respectively) in classifying LUAD and LUSC test subjects in relation to NSCLC histological subtype location and grade. Table 7 reports the AUC values of LUCY (both CLSTM-NN and TDCNN-CLSTM-NN) in classifying LUAD and LUSC test subjects in relation to NSCLC histological subtype location and grade, together with the P value when statistically comparing the performance of the two end-to-end neural networks by means of the DeLong's test. In Table 5, Table 6 and Table 7, classification metrics were not computed for the RML and G1 groups, because statistics cannot be performed in case of too few (less than 3) samples per class.

Table 8 reports the results of the analysis procedure applied on the dynamic visual interpretation outcomes in relation to NSCLC histological subtype location and grade. Fig. 8 displays a bunch of heatmap frames of a correctly-classified test scan of a LUAD subject belonging to the LUL and GNA groups. Fig. 9 displays a bunch of heatmap frames of a

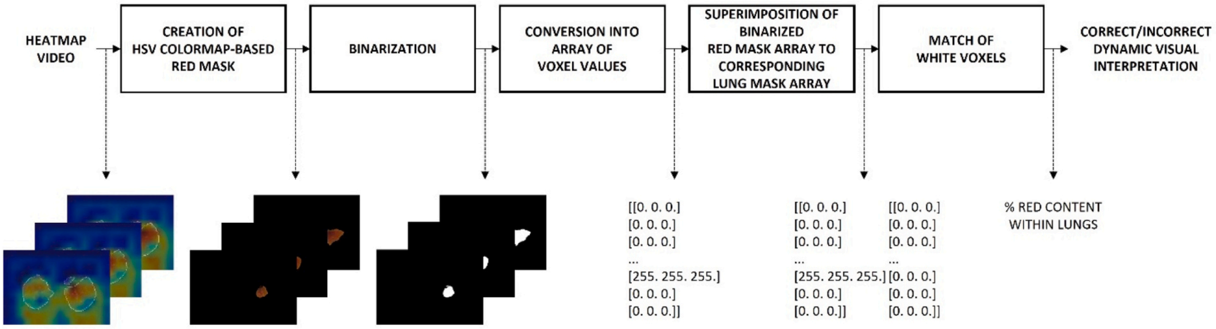


Fig. 5. Analysis procedure applied on the dynamic visual interpretation outcomes.

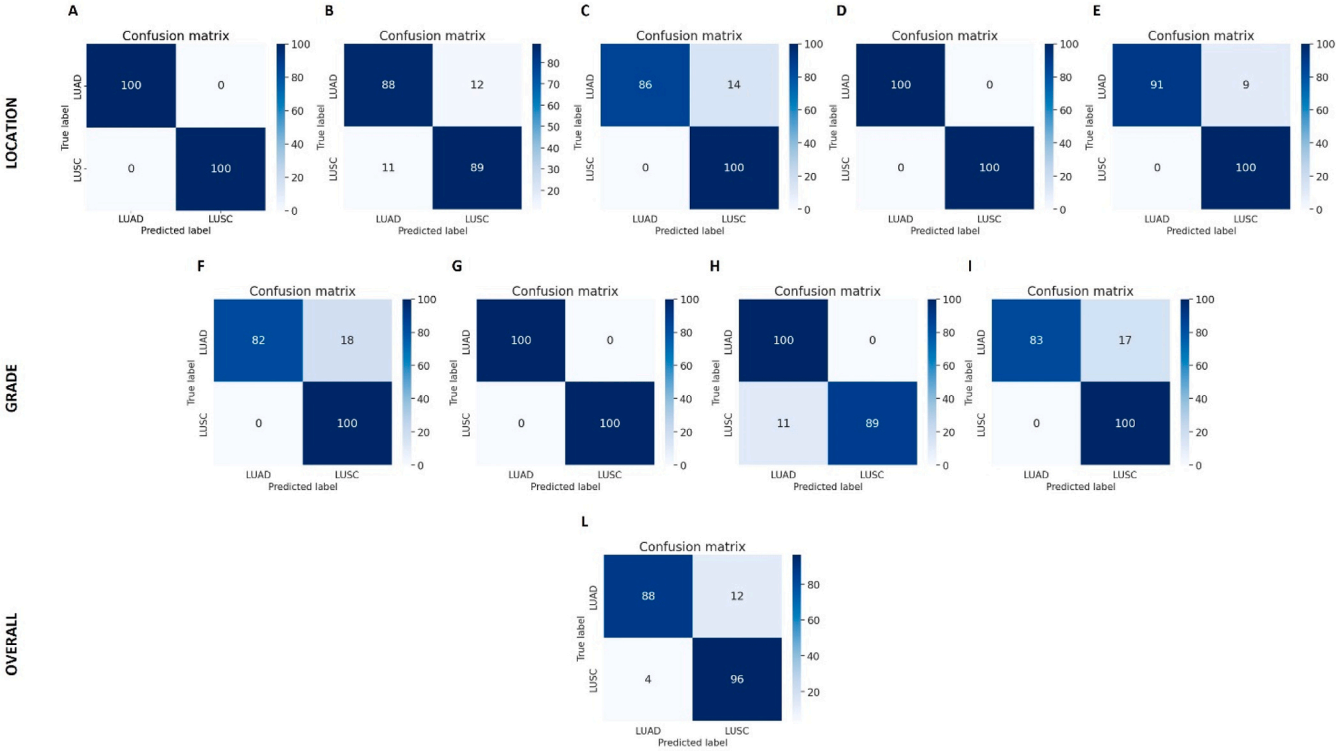


Fig. 6. Confusion matrices of LUCY (CLSTM-NN) for LLL (**A**), LUL (**B**), RLL (**C**), RML (**D**), RUL (**E**), GNA (**F**), G1 (**G**), G2 (**H**) and G3 (**I**) groups, and for overall test dataset (**L**), respectively.

misclassified test scan. In both Fig. 8 and Fig. 9, red spots correspond to the lung voxels where LUCY focused its attention the most in non-invasive NSCLC histology characterization.

Above-reported findings are discussed in Section 5.

5. Discussion

This study aimed to develop LUCY in the cloud for NSCLC histology characterization directly from thorax CT scans. This aim was pursued by:

- Processing heterogeneous thorax CT scans, selected from four openly-accessible data collections;
- Using a challenging evaluation protocol to prove the generalization capability to heterogeneous data coming from different sites by reserving three data collections (NSCLC-Radiomics, NSCLC-Radiogenomics and TCGA-LUAD) as training and validation datasets, and one entire data collection (NSCLC-Radiomics-Genomics) as test dataset;
- Implementing and comparing two end-to-end neural networks, the core layer of whom is a ConvLSTM layer;
- Computing the performance on the dataset reserved for testing from a subject-level perspective by providing a stratification according to NSCLC histological subtype location and grade;
- Making the achieved outcomes visually interpretable by producing and analyzing one heatmap video for each thorax CT scan.

Although using a challenging evaluation protocol to prove the generalization capability to heterogeneous data coming from different sites in such an highly-demanding task, LUCY demonstrated to be highly generalizable and robust. In fact, it reached test AUC values above 77% in all NSCLC histological subtype location and grade groups, and a best AUC value of 97% on the entire dataset reserved for testing. No statistical significance is observed between AUC values of CLSTM-NN and TDCNN-CLSTM-NN for each NSCLC histological subtype location and grade group as well as for the overall test dataset (Table 7). This means that both end-to-end neural networks, the core layer of whom is a

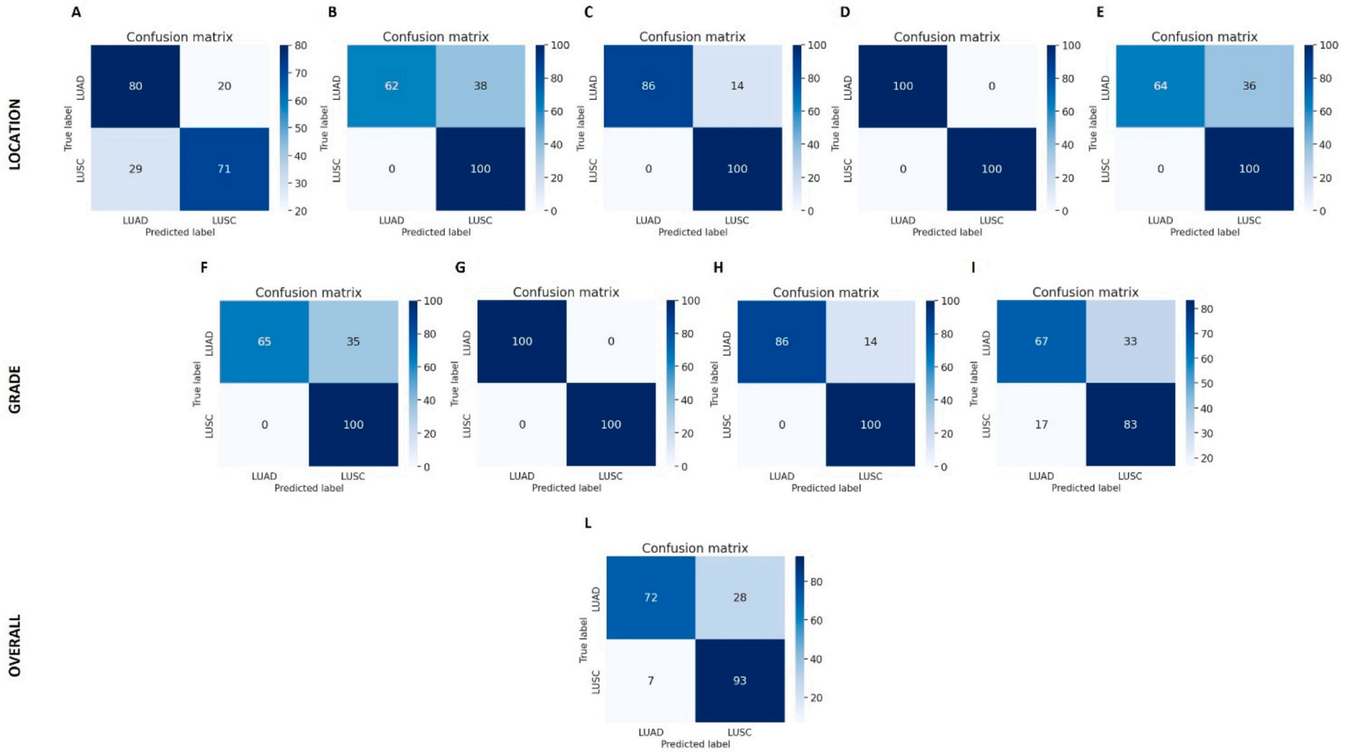


Fig. 7. Confusion matrices of LUCY (TDCNN-CLSTM-NN) for LLL (**A**), LUL (**B**), RLL (**C**), RML (**D**), RUL (**E**), GNA (**F**), G1 (**G**), G2 (**H**) and G3 (**I**) groups, and for overall test dataset (**L**), respectively.

Table 5
Performance of LUCY (CLSTM-NN) in classifying LUAD and LUSC on test dataset stratified according to NSCLC histological subtype location and grade. Number of LUAD and LUSC subjects, ACC, PRE, SEN and F1-score values are given in percentage (%).

	LUAD (%)	LUSC (%)	ACC (%)	PRE (%) LUAD/ LUSC	SEN (%) LUAD/ LUSC	F1-score (%) LUAD/ LUSC
LLL	8	12	100	100/100	100/100	100/100
LUL	13	15	88	88/89	88/89	88/89
RLL	12	5	90	100/75	86/100	92/86
RML	2	2	n. a.	n. a.	n. a.	n. a.
RUL	18	13	95	100/89	91/100	95/94
GNA	28	7	86	100/57	82/100	90/73
G1	3	5	n. a.	n. a.	n. a.	n. a.
G2	12	15	94	88/100	100/89	93/94
G3	10	20	94	100/92	83/100	91/96
Overall	53	47	92	97/87	88/96	92/92

n. a.: not applicable, if there is not statistics due to too few samples per class

Table 6
Performance of LUCY (TDCNN-CLSTM-NN) in classifying LUAD and LUSC on test dataset stratified according to NSCLC histological subtype location and grade. Number of LUAD and LUSC subjects, ACC, PRE, SEN and F1-score values are given in percentage (%).

	LUAD (%)	LUSC (%)	ACC (%)	PRE (%) LUAD/ LUSC	SEN (%) LUAD/ LUSC	F1-score (%) LUAD/ LUSC
LLL	8	12	75	67/83	80/71	73/77
LUL	13	15	82	100/75	62/100	77/86
RLL	12	5	90	100/75	86/100	92/86
RML	2	2	n. a.	n. a.	n. a.	n. a.
RUL	18	13	79	100/67	64/100	78/80
GNA	28	7	71	100/40	65/100	79/57
G1	3	5	n. a.	n. a.	n. a.	n. a.
G2	12	15	94	100/90	86/100	92/95
G3	10	20	78	67/83	67/83	67/83
Overall	53	47	82	92/74	72/93	81/83

n. a.: not applicable, if there is not statistics due to too few samples per class

ConvLSTM layer, are effective in non-invasively characterize NSCLC histological subtypes. Despite non-statistically different AUC values, the ACC, PRE, SEN and F1-score values show that there was no improvement by coupling the ConvLSTM-based neural network with the time-distributed 2D CNN (Table 5 and Table 6). Indeed, CLSTM-NN gained equal or even higher ACC values than the ones achieved by TDCNN-CLSTM-NN. Moreover, according to the PRE, SEN and F1-score values, TDCNN-CLSTM-NN underestimated one of the two classes in the majority of NSCLC histological subtype location and grade groups, whereas CLSTM-NN did not. Thereby, a ConvLSTM-based neural network used independently is even better suited than a time-distributed 2D CNN in automatically-extracting the most salient features for non-invasively characterizing NSCLC histology, thanks to the convolution mechanisms that are different from the ones of CNNs. Nevertheless, CLSTM-NN did not allow the inclusion of visualization modules (e.g., Grad-CAM) as they work for CNNs only. Conversely, TDCNN-CLSTM-NN allowed it, making possible to visually interpret the cases where LUCY succeeds or not in classifying LUAD and LUSC test subjects. In Fig. 8, the generated heatmaps are well confined (i.e., there is a unique, circumscribed red spot on one lung) in the upper lobe of the left lung. This is the case where LUCY correctly characterize the NSCLC histological subtype, being able to focus the attention most strongly exclusively on the area where the lung mass is actually present, which is exactly the upper lobe of the left lung. In Fig. 9, instead, the generated heatmaps are less confined (i.e., there are multiple red spots on both lungs). This is the case where LUCY failed in non-invasively characterizing the NSCLC histological subtype. The reason of the misclassification is that LUCY was unable to focus the attention exclusively on a single area, the one where the lung mass is

Table 7

Performance of LUCY (both CLSTM-NN and TDCNN-CLSTM-NN) in classifying LUAD and LUSC on test dataset stratified according to NSCLC histological subtype location and grade. AUC values are given in percentage (%). P value is also reported.

	AUC (%)		P
	CLSTM-NN	TDCNN-CLSTM-NN	
LLL	100	77	0.1486
LUL	93	100	0.4093
RLL	91	100	0.3711
RML	n. a.	n. a.	n. a.
RUL	100	100	1
GNA	97	100	0.4093
G1	n. a.	n. a.	n. a.
G2	95	100	0.3657
G3	97	92	0.5071
Overall	97	95	0.7078

n. a.: not applicable, if there is not statistics due to too few samples per class

Table 8

Results of the analysis procedure applied on the dynamic visual interpretation outcomes stratified according to NSCLC histological subtype location and grade. Number of LUAD and LUSC subjects, number of correctly-classified test scans and red content within the lungs are given in percentage (%).

	LUAD (%)	LUSC (%)	Correctly classified (%) LUAD/LUSC	Red content within lungs (%) LUAD/LUSC
LLL	8	12	80/71	88/83
LUL	13	15	63/100	74/97
RLL	12	5	86/100	90/98
RML	2	2	100/100	97/98
RUL	18	13	64/100	75/99
GNA	28	7	65/100	75/96
G1	3	5	100/100	97/97
G2	12	15	86/100	90/100
G3	10	20	67/83	78/88
Overall	53	47	72/93	85/94

present, thus automatically extracting and learning features not strictly related to the specific NSCLC histological subtype. To bolster the credibility of LUCY, dynamic visual interpretation outcomes were also analyzed for each NSCLC histological subtype location and grade group as well as for the entire dataset reserved for testing (Table 8), thus, providing a more robust and comprehensive evaluation. In the heatmap videos of all NSCLC histological subtype location and grade groups, the red content that lies inside the lungs exceeds 74% and 83% for LUAD and LUSC test subjects, respectively, while it reaches 85% and 94% in the entire test dataset. From Table 8 it can be also noticed that, overall, LUCY succeeded in classifying the 72% of LUAD test subjects and the 93% of LUSC test subjects, reaching 100% in case of RML location and G1 grade groups.

In the literature, a common choice is to exploit 3D CNNs to process volumetric data, such as thorax CT scans. Although 3D CNNs have the ability to preserve inter-slice context information, they come with a high computational cost mainly due to the 3D convolution mechanism and the abundant number of parameters. Conversely, by using less computationally-expensive algorithms like the ones reviewed in Section 2, it is possible to simultaneously process multiple slices of the same scan, preserving their spatial correlation in terms of anatomy, while ensuring good performance and improved execution times. In this study, a performance comparative analysis with the state-of-the-art studies was not provided because of the use of different data. By using different data, a comparative analysis restricted to the achieved performance would be non-objective. To establish the novelty and superiority of this study with respect to the state-of-the-art ones, a qualitative methodological comparative analysis was provided, instead. To simultaneously process multiple slices of the same scan, both Moitra et al (Moitra and Mandal, 2020) and Marentakis et al (Marentakis et al., 2021) used transfer learning by exploiting pretrained 2D CNNs coupled with recurrent layers (LSTM and biLSTM, respectively). However, pretraining 2D CNNs on large natural data collections (e.g., ImageNet) may cause the learning process to be invariant to scale variations (Graziani et al., 2021). Such invariance can be detrimental in medical applications because scale carries valuable information. Moreover, the use of adapted neural networks pretrained on natural images may not yield clinically satisfactory outcomes, as medical images are generally more difficult to handle due to unique challenges, such as high inter-class similarity (Chen et al., 2022). Thus, training from scratch is preferable to introduce the desired invariances in the automatically-learned features (Graziani et al., 2021). Accordingly, LUCY was trained from scratch. Since approaches trained from scratch does not require an input tensor of a fixed shape, it was also possible to resize all scans to the ad-hoc shape, obtained by weighting the average shapes of the cropped-and-cut scans (Subsection 3.2). One of the most important caveats of the state-of-the-art frameworks is the single-site origin of CT data that limits the generalizability of findings. In light of this, one of the main novelties introduced in LUCY was to analyze heterogeneous data from different sites, thus acquired with

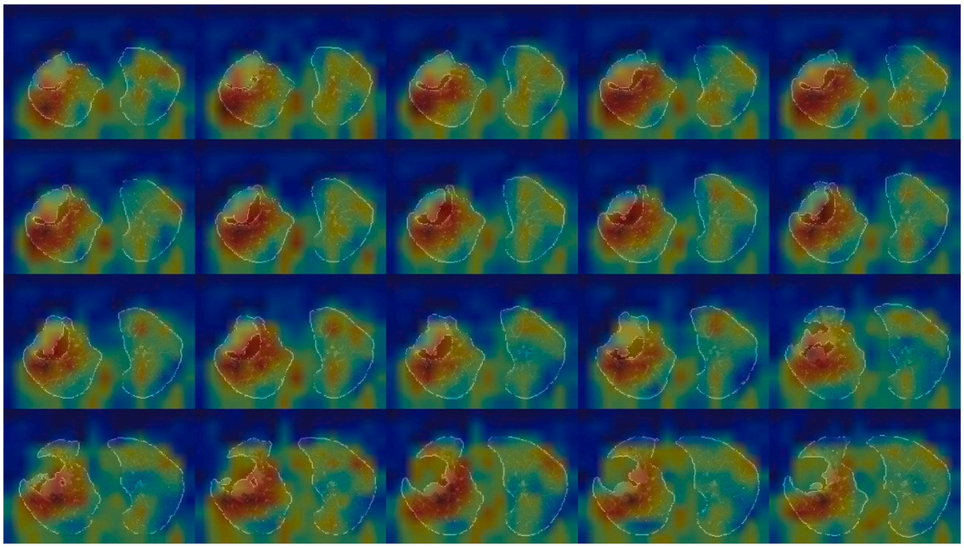


Fig. 8. A bunch of heatmap frames (20 out of 250 total heatmap frames) of a correctly-classified test scan. Heatmap frame progression goes from left to right.

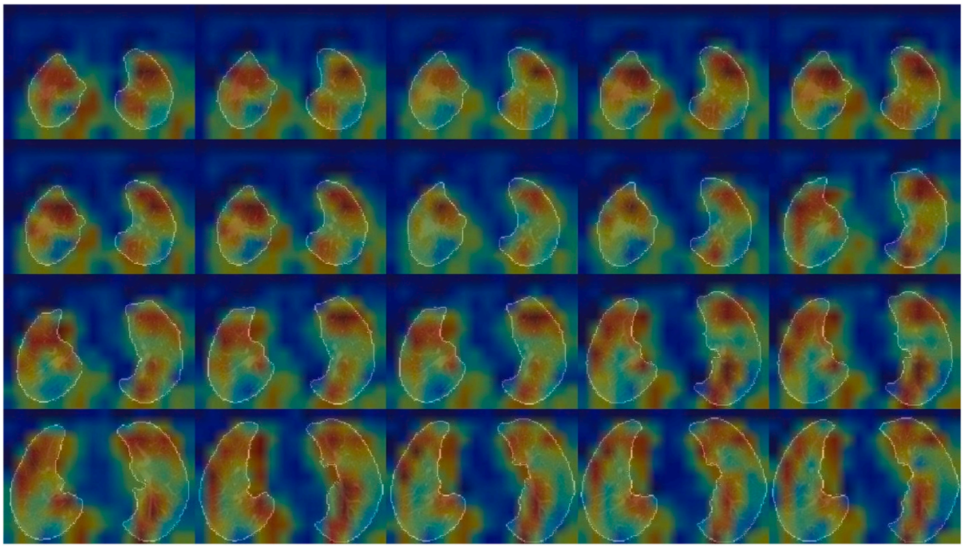


Fig. 9. A bunch of heatmap frames (20 out of 250 total heatmap frames) of a misclassified test scan. Heatmap frame progression goes from left to right.

different acquisition protocols and by using scanners of different manufacturers, in order to reach both reliability and reproducibility of the results. In fact, LUCY is the only that uses three data collections as training and validation datasets, and the 100% of a fourth data collection as test dataset. Doing so was more challenging but also fairer, as it allowed to effectively prove the generalization capability of LUCY. Additionally, entire lung volumes were processed in LUCY to preserve the anatomical integrity of the lungs, whereas lung patches were processed by both Moitra et al (Moitra and Mandal, 2020) and Marentakis et al (Marentakis et al., 2021). By processing patches, the possibility of conducting an analysis on a subject-level perspective is nullified. Conversely, LUCY evaluates the performance on test dataset from a subject-level perspective, consistently with what is done in the real clinical practice. Another important difference from the frameworks developed by Moitra et al (Moitra and Mandal, 2020) and Marentakis et al (Marentakis et al., 2021) is that LUCY was developed entirely in the cloud to ensure a machine-independent reproducibility of the decision-support system, guaranteeing also cost saving and sustainability. To allow so, publicly-available data collections, already compliant with ethical and regulatory issues, were used. The uploading of thorax CT scans in the DICOM format took just few seconds. Although the training of LUCY took approximately 20 h to be completed, only up to 25 s served to generate the predictions on test dataset, which is a perfectly acceptable time in terms of execution efficiency. Moreover, no special network bandwidth requirements were necessary for on-cloud training, validating and testing. In case LUCY would like to be tested on a private data collection, its weights could be exploited to produce predictions in a local environment/private workstation, with no need for anonymization. Computing in the cloud was chosen in this study because one of the main challenges facing medical image analysis is the development of benchmarks that allows algorithms to be compared under common measures and standards (Kagadis et al., 2013). The cloud can contribute to such benchmarks by facilitating their creation and usage. For what specifically concerns radiological data analysis, the use of computer-aided diagnostic systems has been suggested as a way of increasing the detection power. However, these systems have problems, such as high software license costs, rapid obsolescence and the need for powerful hardware, which do not make them cost-effective solutions. The scalable and distributed computational and resource pooling features of cloud computing, instead, have the potential to increase the execution speed while keeping the costs low (Erfannia and Alipour, 2022). Indeed, the pivotal component of the cloud is the analysis platform, which supports a wide spectrum of data queries and cost-effectiveness computational resources, without the surcharge of purchasing and maintaining additional equipment (Kagadis et al., 2013; Erfannia and Alipour, 2022). Eventually, LUCY is the first and only to include clinically-relevant information in the analysis and provide a dynamic visual interpretation of the achieved results, paving the way to understandable scan-based clinical decision-support systems driven by DL.

LUCY achieved encouraging results in non-invasive NSCLC histology characterization, but its performance could be further improved. The main limitation relies on the fact that this study performed binary classification only. NSCLC histological subtypes were classified as LUAD or LUSC, without considering LULC. However, the diagnosis of LULC is restricted to surgically-resected lung cancers with unclear immuno/morphological differentiation (Travis et al., 2015). Moreover, since the considered data collections contain subjects diagnosed with LULC before the clinical implementation of the afore-mentioned guidelines, the annotation regarding LULC could be questionable. For these reasons, LULC was not taken into account in developing LUCY. Another limitation is that the choice of analyzing entire lung volumes made not possible to increase the batch size to a value greater than 1 due to the needed RAM, ending up in disabling the exploitation of the advantages that larger (without exceeding) batch sizes could carry, such as a faster convergence. However, hardware capabilities are now experiencing rapid empowerment also in cloud environments, so it will soon be possible to easily manage this limitation. In addition, it is true that using cloud computing to support clinical decisions in the healthcare sector is a significant opportunity for both researchers and practitioners (Ali et al., 2018), but cloud computing should be rigorously evaluated before its wide adoption. In fact, despite its numerous benefits, a notable issue of cloud computing is related to the challenges associated with the usage of on-cloud systems in the real clinical setting. The main challenge is linked to data security (Ali et al., 2018). There is a long line of research pertaining this challenge, and data encryption is currently the best strategy for protecting data storage and retrieval in the cloud (Mehrtak et al., 2021). However, this is still an area of active research (Ali et al., 2018). The second most important challenge is linked to data integrity, confidentiality and authenticity. Access control and endpoint authentication are valid strategies to handle this challenge (Mehrtak et al., 2021). Another challenge is linked to data anonymity. Subjects' identities must be made anonymous when storing private health data in the cloud. Such anonymization is assessed by removing all elements that could be used to identify the subjects or the subjects' relatives, above all the name, geographical information, phone number and biometrics

(Al-Issa et al., 2019). However, anonymization in healthcare data setting is still a very active area of research. Other challenges are strictly technical, such as the lack of necessary Internet connectivity infrastructure in healthcare institutions to support cloud computing-enabled healthcare projects or the possible interference of such systems with medical equipment (Ali et al., 2018). Identifying these challenges is the first step to tackle them, and future investigations need to provide more feasible solutions to fix such bugs.

As part of future work, LUCY may be refined by coupling thorax CT scans with scans of different imaging modalities (e.g., PET), as done by Moitra et al (Moitra and Mandal, 2020). Although the multi-modality approach is likely to perform better than the single imaging modality approach, using a multi-modality approach is more expensive and time-consuming in the real clinical practice for both subjects and healthcare centers. Despite challenging, a way to extend the analysis to a multi-modality classification will be found. Using parallel neural networks with independent or shared weights for each imaging modality and fusing them at one of the middle layers may be an idea in this regard. Furthermore, it could be interesting to look for novel strategies to make the encouraging results produced by a special recurrent neural network (like ConvLSTM) visually understandable, as done for CNN-like models. LUCY may be refined further by extending its application to the non-invasive classification of additional pulmonary pathologies, as it keeps the anatomy of the lungs unaltered. Experiments may also be carried out by including other meta-learners. Such evaluative studies will develop the usage of prognostic medical image biomarkers with NSCLC histological subtypes.

6. Conclusions

LUCY is an advanced on-cloud decision-support system that effectively characterizes NSCLC histology directly from thorax CT scans. It is the first that makes use of two end-to-end neural networks, the core layer of whom is a ConvLSTM layer, able to non-invasively and reliably provide visually-understandable predictions on LUAD and LUSC subjects in relation to clinically-relevant information. Another important characteristic that has to be emphasized is that LUCY is lung mass segmentation free. Therefore, its performance is not affected by the variability of lung mass margins. Furthermore, it could easily be integrated in any other system for real diagnostic purposes thanks to its machine-independent nature, execution efficiency and visually-understandable outcomes.

CRediT authorship contribution statement

Selene Tomassini: Conceptualization, Investigation, Data curation, Methodology, Software, Visualization, Writing - Original draft preparation. **Nicola Falcionelli:** Methodology, Software, Writing - Review and editing. **Giulia Bruschi:** Software, Writing - Review and editing. **Agnese Sbrollini:** Validation, Writing - Review and editing. **Niccolò Marini:** Validation, Writing - Review and editing. **Paolo Sernani:** Validation, Writing - Review and editing. **Micaela Morettini:** Writing - Review and editing. **Henning Müller:** Writing - Review and editing. **Aldo Franco Dragoni:** Writing - Review and editing. **Laura Burattini:** Writing - Review and editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process

Neither generative AI nor AI-assisted technologies have been used in writing this manuscript.

Declaration of Competing Interest

This manuscript is an honest and transparent account of the research being pursued. No important aspects of the study have been omitted. No relationships with other people or organizations that could inappropriately introduce a bias have been established. Neither this manuscript nor any parts of its content are currently under consideration or published elsewhere in any language.

Data Availability

Data used in this manuscript are openly accessible. Full code is available under request.

References

Adiraju, R.V., Elias, S., 2021. A survey on lung CT datasets and research trends. Res. Biomed. Eng. 1–16.

Ali, O., Shrestha, A., Soar, J., Wamba, S.F., 2018. Cloud computing-enabled healthcare opportunities, issues, and applications: a systematic review. Int. J. Inf. Manag. 43, 146–158.

Al-Issa, Y., Ottom, M.A., Tamrawi, A., 2019. eHealth cloud security challenges: a survey. J. Healthc. Eng. 2019.

Bębas, E., Borowska, M., Derlatka, M., Oczeretko, E., Hładuński, M., Szumowski, P., Mojsak, M., 2021. Machine-learning-based classification of the histological subtype of non-small-cell lung cancer using MRI texture analysis. Biomed. Signal Process. Control 66, 102446.

Cao, W., Wu, R., Cao, G., He, Z., 2020. A comprehensive review of computer-aided diagnosis of pulmonary nodules based on computed tomography scans. IEEE Access 8, 154007–154023.

Chaunzwa, T.L., Hosny, A., Xu, Y., Shafer, A., Diao, N., Lanuti, M., Christiani, D.C., Mak, R.H., Aerts, H.J., 2021. Deep learning classification of lung cancer histology using CT images. Sci. Rep. 11, 1–12.

Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y., 2022. Recent advances and clinical applications of deep learning in medical image analysis. Med. Image Anal., 102444

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 26, 1045–1057.

Cong, L., Feng, W., Yao, Z., Zhou, X., Xiao, W., 2020. Deep learning model as a new trend in computer-aided diagnosis of tumor pathology for lung cancer. J. Cancer 11, 3615.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 837–845.

Erfannia, L., Alipour, J., 2022. How does cloud computing improve cancer information management? A systematic review. Inform. Med. Unlocked, 101095.

Graziani, M., Lompech, T., Müller, H., Depeursinge, A., Andrearczyk, V., 2021. On the scale invariance in state of the art CNNs trained on ImageNet. Mach. Learn. Knowl. Extr. 3, 374–391.

Guo, Y., Song, Q., Jiang, M., Guo, Y., Xu, P., Zhang, Y., Fu, C.-C., Fang, Q., Zeng, M., Yao, X., 2020. Histological subtypes classification of lung cancers on CT images using 3D deep learning and radiomics. Acad. Radiol.

Halder, A., Dey, D., Sadhu, A.K., 2020. Lung nodule detection from feature engineering to deep learning in thoracic CT images: a comprehensive review. J. Digit. Imaging 33, 655–677.

Halder, A., Chatterjee, S., Dey, D., 2022. Adaptive morphology aided 2-pathway convolutional neural network for lung nodule classification. Biomed. Signal Process. Control 72, 103347.

Han, Y., Ma, Y., Wu, Z., Zhang, F., Zheng, D., Liu, X., Tao, L., Liang, Z., Yang, Z., Li, X., et al., 2021. Histologic subtype classification of non-small cell lung cancer using PET/CT images. Eur. J. Nucl. Med. Mol. Imaging 48, 350–360.

Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G., 2020. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur. Radiol. Exp. 4, 1–13.

Kagadis, G.C., Kloukinas, C., Moore, K., Philbin, J., Papadimitroulas, P., Alexakos, C., Nagy, P.G., Visvikis, D., Hendee, W.R., 2013. Cloud computing in medical imaging. Med. Phys. 40, 070901.

Kriegsmann, M., Haag, C., Weis, C.-A., Steinbuss, G., Warth, A., Zgorzelski, C., Muley, T., Winter, H., Eichhorn, M.E., Eichhorn, F., et al., 2020. Deep learning for the classification of small-cell and non-small-cell lung cancer. Cancers 12, 1604.

Li, Y., Chen, D., Wu, X., Yang, W., Chen, Y., 2021. A narrative review of artificial intelligence-assisted histopathologic diagnosis and decision-making for non-small cell lung cancer: achievements and limitations. J. Thorac. Dis. 13, 7006.

Liu, H., Jing, B., Han, W., Long, Z., Mo, X., Li, H., 2019. A comparative texture analysis based on NECT and CECT images to differentiate lung adenocarcinoma from squamous cell carcinoma. J. Med. Syst. 43, 59.

Liu, H., Jiao, Z., Han, W., Jing, B., 2021. Identifying the histologic subtypes of non-small cell lung cancer with computed tomography imaging: a comparative study of capsule net, convolutional neural network, and radiomics. Quant. Imaging Med. Surg. 11, 2756.

Marentakis, P., Karaiskos, P., Kouloulias, V., Kelekis, N., Argentos, S., Oikonomopoulos, N., Loukas, C., 2021. Lung cancer histology classification from CT images based on radiomics and deep learning models. Med. Biol. Eng. Comput. 59, 215–226.

Mehrtak, M., SeyedAlinaghi, S., MohsseniPour, M., Noori, T., Karimi, A., Shamsabadi, A., Heydari, M., Barzegary, A., Mirzapour, P., Soley-Manzadeh, M., et al., 2021. Security challenges and solutions using healthcare cloud computing. J. Med. Life 14, 448.

Moitra, D., Mandal, R.K., 2020. Prediction of non-small cell lung cancer histology by a deep ensemble of convolutional and bidirectional recurrent neural network. J. Digit. Imaging 33, 895–902.

Monkam, P., Qi, S., Ma, H., Gao, W., Yao, Y., Qian, W., 2019. Detection and classification of pulmonary nodules using convolutional neural networks: a survey. IEEE Access 7, 78075–78091.

Naik, A., Edla, D.R., 2021. Lung nodule classification on computed tomography images using deep learning. Wirel. Pers. Commun. 116, 655–690.

Panunzio, A., Sartori, P., 2020. Lung cancer and radiological imaging. Curr. Radiopharm. 13, 238–242.

Pereira, T., Freitas, C., Costa, J.L., Morgado, J., Silva, F., Negrão, E., de Lima, B.F., da Silva, M.C., Madureira, A.J., Ramos, I., et al., 2021. Comprehensive perspective for lung cancer characterization based on AI solutions using CT images. J. Clin. Med. 10, 118.

Pinsky, P.F., Gierada, D.S., Nath, P.H., Kazerooni, E., Amorosa, J., 2013. National lung screening trial: Variability in nodule detection rates in chest CT studies. Radiology 268, 865–873.

Planchard, D., Popat, S., Kerr, K., Novello, S., Smit, E., Faivre-Finn, C., Mok, T., Reck, M., Van Schil, P., Hellmann, M., et al., 2018. Metastatic non-small cell lung cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann. Oncol. 29, iv192–iv237.

Prabhu, S., Prasad, K., Robels-Kelly, A., Lu, X., 2022. AI-based carcinoma detection and classification using histopathological images: a systematic review. Comput. Biol. Med., 105209

Reiazi, R., Abbas, E., Famiyeh, P., Rezaie, A., Kwan, J.Y., Patel, T., Bratman, S.V., Tadic, T., Liu, F.-F., Haibe-Kains, B., 2021. The impact of the variation of imaging parameters on the robustness of computed tomography radiomic features: a review. Comput. Biol. Med. 133, 104400.

Rivera, M.P., Mehta, A.C., Wahidi, M.M., 2013. Establishing the diagnosis of lung cancer: diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 143, e142S–e165S.

Rubin, G.D., 2015. Lung nodule and cancer detection in CT screening. J. Thorac. Imaging 30, 130.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. IEEE Int. Conf. Comput. Vis. 618–626.

Shen, C., Liu, Z., Guan, M., Song, J., Lian, Y., Wang, S., Tang, Z., Dong, D., Kong, L., Wang, M., et al., 2017. 2D and 3D CT radiomics features prognostic performance comparison in non-small cell lung cancer. Transl. Oncol. 10, 886–894.

Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W., 2019. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst. Appl. 128, 84–95.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Int. Conf. Neural Inf. Process. Syst. 802–810 (p.).

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 1–48.

Snoek, J., Larochelle, H., Adams, R., 2012. Practical Bayesian optimization of machine learning algorithms. Adv. Neural Inf. Process. Syst. 25.

Suster, D.I., Mino-Kenudson, M., 2020. Molecular pathology of primary non-small cell lung cancer. Arch. Med. Res.

Thakur, S.K., Singh, D.P., Choudhary, J., 2020. Lung cancer identification: a review on detection and classification. Cancer Metastasis Rev. 1–10.

Thawani, R., McLane, M., Beig, N., Ghose, S., Prasanna, P., Velcheti, V., Madabhushi, A., 2018. Radiomics and radiogenomics in lung cancer: a review for the clinician. Lung Cancer 115, 34–41.

Tomassini, S., Falcionelli, N., Sernani, P., Burattini, L., Dragoni, A.F., 2022a. Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: a survey. Comput. Biol. Med., 105691

Tomassini, S., Falcionelli, N., Sernani, P., Sbrollini, A., Morettini, M., Burattini, L., Dragoni, A.F., 2022b. Cloud-YLung for non-small cell lung cancer histology classification from 3D computed tomography whole-lung scans. Conf. Proc. IEEE Eng. Med. Biol. Soc. 1556–1560.

Tomassini, S., Sbrollini, A., Covella, G., Sernani, P., Falcionelli, N., Müller, H., Morettini, M., Burattini, L., Dragoni, A.F., 2022c. Brain-on-Cloud for automatic diagnosis of Alzheimer's disease from 3D structural magnetic resonance whole-brain scans. Comput. Methods Prog. Biomed., 107191

Travis, W.D., Brambilla, E., Nicholson, A.G., Yatabe, Y., Austin, J.H., Beasley, M.B., Chirieac, L.R., Dacic, S., Duhig, E., Flieder, D.B., et al., 2015. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. J. Thorac. Oncol. 10, 1243–1260.

Winkels, M., Cohen, T.S., 2019. Pulmonary nodule detection in CT scans with equivariant CNNs. Med. Image Anal. 55, 15–26.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., Deng, S.-H., 2019. Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. 17, 26–40.

Yang, F., Chen, W., Wei, H., Zhang, X., Yuan, S., Qiao, X., Chen, Y.-W., 2020. Machine learning for histologic subtype classification of non-small cell lung cancer: a retrospective multicenter radiomics study. Front. Oncol. 10.

Zhang, G., Jiang, S., Yang, Z., Gong, L., Ma, X., Zhou, Z., Bao, C., Liu, Q., 2018. Automatic nodule detection for lung cancer in CT images: a review. Comput. Biol. Med. 103, 287–300.

Zhang, G., Yang, Z., Gong, L., Jiang, S., Wang, L., Cao, X., Wei, L., Zhang, H., Liu, Z., 2019. An appraisal of nodule diagnosis for lung cancer in CT images. J. Med. Syst. 43, 1–18.

Zhao, B., Tan, Y., Bell, D.J., Marley, S.E., Guo, P., Mann, H., Scott, M.L., Schwartz, L.H., Ghiorghiu, D.C., 2013. Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals. Eur. J. Radiol. 82, 959–968.

Zhu, X., Dong, D., Chen, Z., Fang, M., Zhang, L., Song, J., Yu, D., Zang, Y., Liu, Z., Shi, J., et al., 2018. Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. Eur. Radiol. 28, 2772–2778.