ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Physics (29^{th}cycle)

Development and application in clinical practice of Computer-aided Diagnosis systems for the early detection of lung cancer

By

Alberto Traverso

******

Supervisor(s):

Prof. Michelangelo Agnello

Dott. Piergiorgio Cerello

Doctoral Examination Committee:

Politecnico di Torino

2017

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Alberto Traverso
2017

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Abstract

Lung cancer is the main cause of cancer-related deaths both in Europe and United States, because often it is diagnosed at late stages of the disease, when the survival rate is very low if compared to first asymptomatic stage. Lung cancer screening using annual low-dose Computed Tomography (CT) reduces lung cancer 5-year mortality by about 20% in comparison to annual screening with chest radiography. However, the detection of pulmonary nodules in low-dose chest CT scans is a very difficult task for radiologists, because of the large number (300/500) of slices to be analyzed. In order to support radiologists, researchers have developed Computer-aided Detection (CAD) algorithms for the automated detection of pulmonary nodules in chest CT scans. Despite proved benefits of those systems on the radiologists detection sensitivity, the usage of CADs in clinical practice has not spread yet. The main objective of this thesis is to investigate and tackle the issues underlying this inconsistency. In particular, in Chapter 2 we introduce M5L, a fully automated Web- and Cloud-based CAD for the automated detection of pulmonary nodules in chest CT scans. This system introduces a new paradigm in clinical practice, by making available CAD systems without requiring to radiologists any additional software and hardware installation. The proposed solution provides an innovative cost-effective approach for clinical structures. In Chapter 3 we present our international challenge aiming at a large-scale validation of state-of-the-art CAD systems. We also investigate and prove how the combination of different CAD systems reaches performances much higher than any best stand-alone system developed so far. Our results open the possibility to introduce in clinical practice very high-performing CAD systems which miss a tiny fraction of clinically relevant nodules. Finally, we tested the performance of M5L on clinical data-sets. In chapter 4 we present the results of its clinical validation, which prove the positive impact of CAD as second reader in the diagnosis of pulmonary metastases on oncological patients with extra-thoracic cancers. The proposed approaches have the potential to exploit at best the features of

different algorithms, developed independently, for any possible clinical application, setting a collaborative environment for algorithm comparison, combination, clinical validation and, if all of the above were successful, clinical practice.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.0.1 Motivation

Lung cancer, also known as bronchogenic carcinoma, is one of the main public health issues in developed countries and still represents the main cause of cancer-related deaths both in Europe and United States, accounting for about 20% and 27% deaths in Europe [1] and United States respectively [2]. Also, the average 5-year survival rate is very low, around 10-17% [2]. Most of lung cancers are diagnosed in the late-stage of the disease (usually referred as Stage IV) when the survival rate is very low if compared to the diagnosis in the preliminary stage (referred as Stage I) where the 5-year survival rate is about between 45-50%. In late stages, the progression of the disease can spread to other parts of the body generating the so-called metastases. Stage I corresponds to the asymptomatic phase of the disease, making the early diagnosis a real clinical challenge. Conversely, in late stages the most common symptoms are chronic cough, blood in the saliva, spread pains in the chest, usually shortness of breath and recurring bronchitis [3]. When lung cancer is detected in the early stage, treatment is very successful leading in most of the cases to the complete healing. The research aiming at optimizing the early diagnosis of lung cancer is a very challenging field which brings together very vast disciplines: clinicians and technicians cooperates in order to improve patient care. In fact, computer scientists, medical physicists and radiologists started working both on hardware and software developments in order to support clinicians by providing them with advanced automated tools that could improve the detection process.

1.0.2 Medical Imaging of the lungs

Relevant technological and clinical developments have been achieved in the last decades in order to improve the early diagnosis: among them, the improvement of scanners for the acquisition of images of the chest. Computed Tomography (CT) for the imaging of the body has been available since 1975 [4]. The physical principle underlying this imaging technique is using X-rays to generate density maps (slices) of the anatomy. A CT scan combines a series of X-ray images taken from

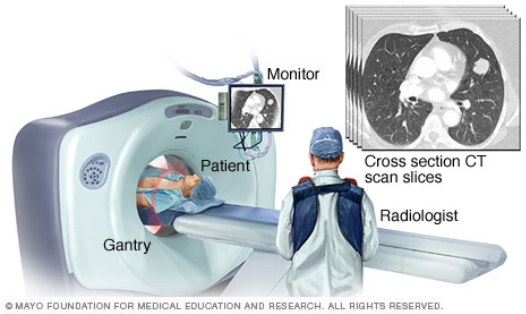


Fig. 1.1 Pictorial view of a CT scanner. The patient is introduced into the CT scanner, while images are acquired and reconstructed directly on the computer in the control room.

different angles and uses computer processing to create cross-sectional images, or slices, of the bones, blood vessels and soft tissues inside the body, as pictorially described in Figure 1. 3D images are formed by combining different information from a big series of 2D X-ray scans from different angles. A CT scan can then be considered as composed by a 3D matrix formed by cuboids called *voxels*. Each *voxel* expresses a particular gray value which represents the density (absorption coefficient) of the corresponding tissue. The standard measurement unit for voxel values is the Hounsfield Unit (HU). It is defined as:

$$HU = 1000 \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}}$$

where μ_{water} and μ_{air} are the linear absorption coefficients of water and air respectively. Sometimes CT scans are performed using a contrast agent injected just before the examination in the blood of the patient. In this case the examination is usually referred as Contrast-Enhanced Computed Tomography (CECT). The role of the agent is usually to increase the density and thus also the HU of the blood. Figure 2 shows a typical example of a CT scan of the lungs from 3 different views: axial, coronal and sagittal. After the acquisition of all the projections, the images

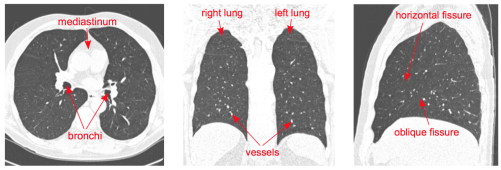


Fig. 1.2 An example of a CT scan of the lungs with some anatomical structures underlined. Three views are usually shown; a) Axial view b) Sagittal view and c) Coronal view of the right lung. The setting of the HU correspond to the typical lung window: white is set to values 50 HU or higher and black is -1450 HU or lower.

are reconstructed using dedicated algorithms, which can be divided into two big families: direct and iterative algorithms. The most famous and still most spread algorithm of the first family is Filtered Back Projection (FBP) [5]. The second family is further divided into algebraic and iterative [6]. This last category includes the most recent algorithms, even though they are still not spread in clinical practice [7]. Improvements on the axial resolution gave to possibility to use in clinical routine High Resolution Computed Tomography (HRCT) scans. Performing scans with axial reconstructed slices of 1 mm thickness allows to investigate anatomical details of the lungs on the same scale of the ones available from pathological specimens [8]. The big revolution in CT imaging was the introduction of multi-detector-row scanners able to acquire up to 64 1-mm slices per rotation at the same time. Furthermore, each rotation can be performed in less than a second. These achievements reduced the acquisition time and consequently motion and breathing artifacts. In parallel to technological developments, software improvements on reconstruction algorithms made the quality of acquired images much better, sensitively decreasing the dose to the patient. On the other side, all previous remarkable improvements brought into clinical practice a major challenge: dealing with an enormous increase in the number and size of digital images. This high-demanding challenge has been called *data explosion* by Rubin [9]. One of the clinical interested challenges was screening potential lung cancer high-risk subjects (such as heavy smokers or former smokers above a certain age) using X-rays or CT scans. The National Lung Screening Trial (NLST), a randomized control trial in the U.S. including more than 50,000 high-risk subjects, showed that lung cancer screening using annual low-dose CT reduces lung cancer mortality by 20% in comparison to annual screening with chest radiography [10]. In 2013, the U.S. Preventive Services Task Force (USPSTF) has given low-dose CT screening a grade B recommendation for high-risk individuals [11] and in early

2015, the U.S. Centers for Medicare and Medicaid Services (CMS) has approved CT lung cancer screening for Medicare recipients. As a result of these developments, lung cancer screening programs using low-dose CT are being implemented in the United States and other countries are expected to follow soon. During a screening campaign, millions of slices need to be analyzed as fast as possible. This situation represents a real burden for radiologists. In fact, the identification of pathological Regions Of Interest (ROIs) in low-dose high resolution CT scans is a difficult and time consuming task for radiologists, mainly due to the high number of slices (on average 200/300 per patient) to be read. Furthermore, it has been proved [12] that radiologist's concentration decreases during the day with a substantial impact on the overall detection sensitivity. Pulmonary nodules are the early manifestation of lung cancer: they show up on CT as a small round or oval-shaped growth in the lung. They may also be called *spots on the lung* or a *coin lesion*. Pulmonary nodules are smaller than three centimeters in diameter. If the growth is larger than that previous dimension, they are called pulmonary mass and they are more likely to represent a cancer. Pulmonary nodules can be distinguished according to the tissue forming them as follows:

- solid pulmonary nodules: a solid nodule completely obscures the surrounding parenchyma; it has homogeneous soft-tissue attenuation and well-defined margins with normal parenchyma. An example of solid pulmonary nodule is shown in Figure 1.3;
- ground glass opacity (GGO or non-solid nodule): it manifests as an area of hazy increased attenuation that does not obliterate the bronchial and vascular margins. They can be due to inflammatory changes, benign lesions, or carcinoma (often bronchioloalveolar). Due to their structure, they are more difficult to be detected by the radiologist, but they usually are more likely to be found malignant [13]. An example of GGO is shown in Figure 1.5;
- semi-solid nodule: it consists of both ground-glass and solid soft-tissue attenuation components. It partially obscures the surrounding parenchyma. An example of semi-solid nodule is shown in Figure 1.4.

Furthermore, pulmonary nodules can be distinguished between malignant and benign. Using only the visual assessment of the nodule there is no certainty about the malignancy of a nodule. Additional investigations such as Positron Emission

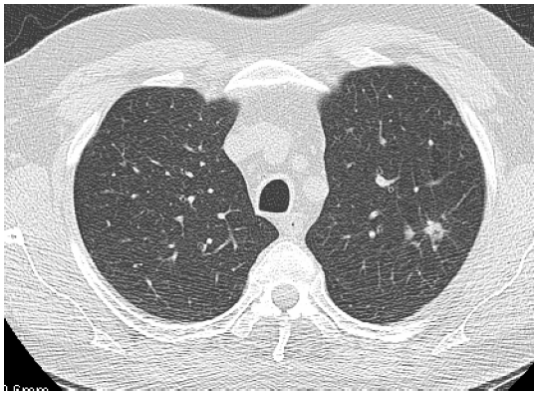


Fig. 1.3 A solid nodule in the upper left lobe (right side) with irregular margins. This is a stage I adenocarcinoma of the lung.

Fig. 1.4 A part solid nodule in the lower right lobe. This is a non-small-cell lung cancer.

Tomography (PET) [14] or biopsy (pathological examination of part of the cells constituting the nodule at the microscope) is required in order to determine the origin of the nodule. In case of malignant nodule, it can be associated to two types of lung cancer: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The classification is made according to the appearance of the tumor cells at the microscope. SCLC is the most aggressive type of cancer, usually related to cigarette smoking and comprises about 10-15% of lung cancers [15]. Conversely, NSCLC is the most common lung cancer, accounting for about 85% of them. [15]. According to the type of cells found in the tumor it is further divided into: adenocarcinomas, squamous cell carcinomas and large cell carcinomas. As mentioned before, the

Fig. 1.5 A ground-glass opacity nodule in the lower right lobe. This is an adenocarcinoma with a bronchioloalveolar component.

visual assessment of a pulmonary nodule is not enough to discriminate between malignant and benign lesions. Conversely, the visual assessment of the properties of a pulmonary nodule is fundamental in order to determine the correct monitoring action for the detected lesion. In fact, the main objective of lung screening is not just to detect nodules, but to determine whether or not immediate follow-up action or monitoring is required. Different strategies and guidelines for the management of pulmonary nodules have been proposed in literature [16] [17]. Most of them are based on some properties of the nodule, such as type of nodules or size. Depending on these features and their combination with other factors such as smoking history, the therapy is determined. It can include both additional follow-up CT surveillance up to 1 year if the nodule is not likely to be malignant or immediate therapy (surgery, chemotherapy, radiotherapy). From the point of view of the radiologist, it is important to detect all the nodules, including small nodules, in order to analyze and monitor their growing rate, which it is a good indicator of malignancy [18]. As already stressed, the detection of pulmonary nodules in chest CT scans is far from being a trivial task for radiologists. In addition, most measurements (such as the quantification of the volume of a nodule) are performed manually by the radiologist, by delimiting the contour of the nodule slice by slice. This represents a very time-consuming task. It is becoming clear that the possibility to provide clinicians with tools for the automated detection and quantification of properties of pulmonary nodules not only can improve the early diagnosis of lung cancer, but also would optimize the cost-effectiveness of screening campaigns. In addition, quantitative

measurements of the properties of nodules are more reliable and can be used as risk predictor factors or discriminant features between malignant or benign lesions. The approach of introducing quantitative automated measurements can be referred to as the field of *quantitative imaging analysis*.

1.0.3 Computer-Aided Detection systems

Computer vision and medical imaging techniques can play a fundamental role in this optic and are fundamental in order to facilitate CT interpretation. In fact, the rapid developments in the imaging acquisition techniques have been accompanied by research on imaging analysis techniques. The algorithms for the automated detection of pathological ROIs are called Computer-Aided Detection or diagnosis (CAD) systems. The usual structure of a CAD system is: 1) pre-processing, 2) candidate detection, and 3) false positive reduction. A pictorial view of the typical setup of a CAD system is shown in Figure 6. The pre-processing stage is used to restrict the

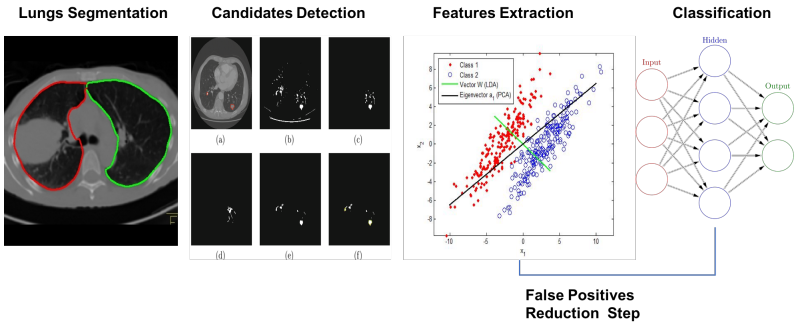


Fig. 1.6 Typical setup of a classical CAD system. The input is represented by a CT scan which pre-elaborated during the pre-processing stage. The input image is segmented in order to extract the binary mask of the lungs. Following stages are: candidate detection and false positive reduction. This last step can be considered as formed by two sub sequential steps: feature extraction and classification. The final output of a CAD system is a list of CAD marks which is sent to the radiologist for review.

search space to the lungs and to limit noise and image artifacts. One of the most important parts of the pre-processing step is the so-called segmentation. In particular, great importance is dedicated to the segmentation of the lungs. Since the manual delineation of the contour of the lungs slice by slice is a very time consuming task, the aim of the segmentation is to perform this task automatically. It has been proved [19] that an accurate segmentation is a key point in the pre-processing stage: as study

[19] shows that a fraction between 5%-17% of lung nodules were missed because of a wrong pre-processing lung segmentation. Most of the segmentation algorithms rely on the fact that the lungs can be considered as a bag full of air, so that they appear as a uniform dark region in the CT scan. Most of the algorithms performing the segmentation task are based on gray-level thresholding and connected component analysis [20] [21]. These methods still suffer from one issue: they are optimized to work on lungs without any pathological or anatomical deformity, while they can fail or produce wrong results when applied on lungs presenting anatomical deformities. The obtained final segmentation of the lungs is usually saved as a binary mask and represents the input for the following step of a CAD system: the candidate detection. The aim of this step is to detect nodule candidates, inside the lung mask. The nodule detection algorithms must provide a very high detection sensitivity, but a very low specificity, which comes with a high number of false positives. Candidate detection algorithms can be based on the widest range of classical medical imaging analysis techniques. One of the most spread and common is the usage of multiple gray-level thresholding techniques [22] [23] [24]: trying to find connected components which present comparable gray level intensities and remove attached anatomical structures such as vessels. Other techniques [25] [26] make use of mathematical morphological operations and additional dedicated filters. The list of candidates, which as mentioned usually includes a very large number of false positives, represents the input for the next step: false positive reduction. The false positive reduction procedure can be seen as a classical binary class machine learning problem: separating non-nodules (negative class) from true nodules (positive class). The classical approaches to solve this problem define a list of features associated to each candidate. The most commonly used features are: gray-level, shape, spatial features and morphological or texture properties. The list of features represents then the input for the classifier. In order to reduce the computational time of the classification stage, some false positives are already discarded using some filtering based on spatial properties. The available CAD algorithms explore a vast range of classifiers: neural networks [27] [28][29], linear classifiers [30] [31], Markov random fields [32] and Bayesian classifiers [33]. The classification step generates the final list of CAD marks which is then given to the radiologist for review. Some classification algorithms also perform a characterization of pulmonary nodules, classifying them according to their type or providing a malignancy score.

1.0.4 Introduction of CAD in clinical practice

CAD systems can be introduced into clinical practice in three different ways: as first reader, as concurrent reader and as second reader. In the first reading mode, all the scans are submitted to the analysis of the CAD before being read by the radiologist. Only slices with CAD marks are then presented to the radiologist. This method can have the appeal to sensibly reduce the reading time, but it requires a system with both high sensitivity and specificity in order not to miss any clinically relevant nodule. In the third reading mode, CAD marks are available to the radiologist only after he/she has finished the first unassisted reading. She / He can then access CAD ROIs and mark them as irrelevant findings (e.g. other pulmonary lesions or not nodules), false positives or true positives originally missed during the un-assisted reading. This solution can have the potential issue to increase the average reading time of a CT scan. The second reading mode is a compromise solution in which CAD marks are shown simultaneously to the radiologists while he/she is inspecting the slices. It is difficult so far to assess which solution should be adopted in clinical practice due to the scarce availability in literature of studies comparing these modalities.

1.0.5 Contribution of this work

The main contribution of this work is aimed at the development and application in clinical practice of CAD systems for the early detection of lung cancer. The first part of the work has been focused on the investigation of available state-of-art CAD systems and their application in clinical practice. This analysis has raised some issues which have been tackled as developments during the research project. In particular, main attention has been focused on:

- developing a complete CAD system which can be accessed and used without requiring any Software or Hardware installation to the clinical facility;
- investigating the impact of combining different CAD systems on the overall detection sensitivity and largely validating new up-to-date candidate detection and false positive algorithms;
- clinically validating the developed system.

Chapter 2

M5L: a Web- and Cloud-based on-demand CAD

2.0.1 Motivation

With the starting of lung cancer CT screening programs, the development of CAD systems has growing rapidly. Considering, for example, the years between 1998 and 2004, the academic production of articles presenting automated nodule detection algorithms has almost doubled each year [34]. In these works, big attention has been devoted to propose new techniques for the detection of pathological ROIs. Other works had the aim to improve exiting CAD systems adding additional improvements or features to existing algorithms. However, scarce effort has been put to analyze the requirements which need to be fulfilled by a system in order to be used daily in clinical practice. This evidence led to a discrepancy between the level of developments reached in the academic research and the level of dissemination of CADs in clinical facilities. It is rare to find hospitals which use those systems as decision support in the diagnosis, but is more rare to find CADs used as *independent reader*. A detailed analysis of issues related to state-of-art CAD systems together with an analysis of clinical requirements for the introduction of a CAD in hospital facilities represents the first step to develop CADs potentially applicable in clinical routine. In addition, the emerging of new technologies (e.g distribute computing) could offer possible solutions to introduce a new paradigm of CAD systems for daily usage in the diagnostic process. A recent study [35] investigates the new informatics tools and developments which can help radiologists to improve the diagnostic process. In

this work a chapter is totally dedicated to the usage of new technologies such as the WEB or Cloud Computing as instruments for daily clinical practice.

2.0.2 CAD systems: state-of-art and challenges

There are several reasons for the inconsistency mentioned in Section 2.0.1: a CAD system needs to be fast and easy to use since radiologists are not used to operate with this tool and it needs to be reliable. It is possible to highlight at least three different kinds of issues which are currently limiting the spreading of CAD systems into clinical practice: software-related, hardware-related and protocol-related.

Software-related issues

The standard approach to make CAD algorithms available in clinical routine of health facilities is the deployment of stand-alone workstations. These workstations are usually equipped with a vendor-dependent Graphic User Interface (GUI) and usually offered by the manufacturer of the scanner with a closed source operating system. Since the algorithms are protected by copyrights, it is not possible to access their internal properties or features. Furthermore, usually there are no scientific publications validating their real performances (including, for example, the validation of the algorithms on external data-sets using an objective and common evaluation metric). In addition, the software usually has very high license costs. Due to its standalone operating systems, additional costs can be charged for the integration of the workstation with software already present in the hospital facility, such as for example the PACS (Picture Archiving Communication System) which manages the storage and transfer of medical scans after their acquisition. Additional impacts on costs are due to the upgrade of the software because of its rapid obsolescence. In fact, considering the rapid developments in the field, more performing CAD systems could be available on the market. Having an up-to-date CAD system is fundamental to benefit from its additional features which could improve its overall detection sensitivity. If these costs can be partially absorbed or accepted by big clinical centers, they can represent a real problem for small clinical facilities.

Hardware-related issues

The computational power required by an algorithm usually increases with the algorithm complexity. This means that high performing algorithms require dedicated powerful hardware, usually sold by the manufacturer. Sometimes running several CAD algorithms in parallel and combining their results can require computational power which is not precisely predictable. In addition, if the CAD tool needs to be used during a massive screening campaign, where hundreds of scans have to be analyzed as fast as possible, a flexibility in the allocation of computing resources becomes fundamental. Considering that screening campaigns can be concentrated in a given period and that the number of CTs which need to be analyzed is much larger than the average number of CTs analyzed during clinical routine, buying a new more powerful hardware only for screening would not be a cost-effective choice. The high computational power will be then only used for screening campaigns, but only a small part of it will be needed for clinical practice. This reason has motivated researchers to optimize the solution for the computation of CAD algorithms. In the past, prototypes for handling the analysis of medical images using a distributed environment were proposed. These solutions were based on the usage of GRID infrastructure, which has been deeply used for high-energy physics projects [36] [37]. Unfortunately, the GRID-approach is not suitable for the majority of Medical Physics applications due to the rigidity and complexity of the infrastructure. In addition, the management of a GRID infrastructure requires dedicated man-power to monitor and maintain the service, which is usually not available (and not cost-effective) in a clinical structure.

Protocol-related issues

Guaranteeing the possibility to share CAD results and medical annotations between radiologists can improve the effectiveness of the diagnosis. The definition of reference standard, and more generally of *true nodule* can vary a lot between radiologists of difference experience. A study [38] analyzed the variability of the *truth* defined by different combinations of experienced radiologists. In addition, the performances of the radiologists were analyzed with respect to different definition of reference standards. The study highlights the importance of forming a panel of different reviewers when annotating a study, to investigate different approaches used by radiologists

in the diagnostic process. A more detailed study [39] showed how review panels composed of 8 or 10 experts achieved a sensitivity greater than that of the most experienced radiologist in the panel. These studies stressed the importance of sharing and combining annotations by different radiologists of different level of expertise to increase the diagnostic sensitivity. Unfortunately, the choice of using stand-alone workstations makes almost impossible to share medical annotations and CAD results between clinicians from different institutions. Developing and providing a CAD system which can be accessed by radiologists all over the world seems to be a mandatory challenge to be faced to improve the diagnosis and detection of pathological ROIs between radiologists. In addition, small clinical facilities sometimes do not have a daily pool of resident experienced radiologists. Giving to these facilities the possibility to ask immediately for a remotely consultation by an experienced radiologist from other structures could be fundamental to improve the diagnosis and reduce costs.

2.0.3 The algorithms

M5L is the combination of two independent CAD systems running in parallel: the Channeler Ant Model (lungCAM) and the Voxel-Based Neural Approach (VBNA). The common step for all the algorithms is the segmentation of the lungs, which is performed using a 3D region growing [40] which includes also the elimination of the trachea. The lung mask is then used as input for the next steps of the algorithms.

lungCAM

This algorithm is based on the simulation of the life-cycle of colonies of virtual ants [41]. They can be released within a seed point inside the lungs and they move inside the volume of the lungs. During their motion they release pheromone according to some predefined rules. Voxels visited by ants are removed and the new ant colony is deposed in the unvisited voxels. The key point of this algorithm is to interpret the CT voxel intensity as the amount of food available, which is progressively reduced during the ant feeding. With this approach, brighter structures, more likely to be nodules, are more likely to be visited by the ants. The probability P_{ij} that a destination is

visited is defined as:

$$P_{ij}(v_i \rightarrow v_j) = \frac{W(\sigma_j)}{\sum_{n=1,26} W(\sigma_n)}$$

where $W(\sigma_j)$ depends on the amount of pheromone in the voxel v_j. The candidate detection algorithm ends when all the ants in colonies die. The output is a list of segmented 3D objects. False positive reduction is performed using a classifier based on 13 features including spatial, intensity and morphological features. The relative small number of features wants to keep the classifier more general as possible avoiding the possibility of over-training. The classifier is a feed-forward artificial neural network (FFNN) with the following architecture: 13 input neurons, 1 hidden layer with 25 neurons and 1 neuron as output layer (giving the probability of the finding to be a real nodule).

VBNA

This algorithm makes use of two different procedures to detect nodules inside the lung parenchyma [42] [43] and attached to the pleura [44]. Nodules candidates inside the parenchyma are found using a dot-enhancement filter plus a multi-scale approach [42] in order to keep into consideration the variability of nodule sizes. The candidates are extracted as the local maxima of the image after the application of the filter. Nodule candidates attached to the pleura are found building normals to the lung wall. The number of normals passing through the voxels are then summed to define a score and the candidates are extracted as the local maxima of the scores. False positive reduction is performed using a Support Vector Machine (SVM) using raw voxels as feature vectors. [43] [44]. The output of the classifier is a list of findings with an associated probability to represent a real nodule.

Combination and Validation

The outputs from the two previous algorithms are then combined. First, a spatial matching between the findings is performed: if the Euclidean distance between two findings is less than 1.5 times their mean diameters, they are merged and considered as a single finding. Since the two classifiers give different probabilities, the average probability is associated to the weighted for different performances of the two algorithms [45]. The M5L CAD has been largely validated in terms of FROC (Free

Response Receiver Operating) curves as described in [46]. The overall sensitivity reaches the value of 80% at 4 False Positive findings per scan. This is a remarkable result considering the size (more than 1000 scans) and the heterogeneity (including both screening and clinical scans with different kind of resolutions) of the validation data-set.

2.0.4 The WEB front-end

The first issue we decided to tackle was to make the M5L results available to radiologists without requiring any software installation in their clinical facilities. One of the desired requirements was to permit to the radiologists not only to access and visualize CAD results, but also to insert their medical annotations of the submitted cases. In addition, the system had to allow the possibility to operate with the CAD in the reading approaches presented in the previous chapter. We followed these requirements when designing and implementing the WEB-based interface, which is accessible at (http://m5l.to.infn.it). The M5L service is available as WEB tool and can then be accessed with any browser using desktop, tablets and mobile devices. The core of the interface infrastructure has been built using the open source tool DRUPAL [47]. DRUPAL has the advantage to present a modular core with a system of hooks and callbacks that can be accessed using Application Programming Interface (API) [48]. This solution allows to use DRUPAL's standard features (e.g users and contents management) *out-of-the-box*. Conversely, it also provides a customizable platform with the possibility to build third-party contributed modules without changing its core code. The development of modules can be easily performed using PHP. The main advantages of DRUPAL can be summarized as follows:

- highly scalable: it can manage high-traffic conditions;
- mobile scalable: all the contents of DRUPAL can be visualized also on mobile devices. The optimization of the design and the visualization is automatically performed according to the device accessing the site;
- security: it has a very detailed system of managing the security of the contents. The site admin can define different sets of users with different kinds of grants and then control the access to protected contents;

- content as a service: it has structured data models allowing to display content with multiple views according to the requests of the users.

For our project, four dedicated custom modules have been developed:

- DICOM information retrieval;
- submission;
- annotation;
- review.

In order to operate with the features of previous modules, three different user profiles were created:

- submitter: associated to a technician who uploads the study case and selects the radiologist reviewing the study;
- reviewer: associated to a radiologist who can review the study cases and access CAD results;
- administrator: the manager of the WEB site with full grants.

DICOM Information Retrieval Module

This module is a dedicated PHP class which allows to retrieve information about the submitted case based on the DICOM properties of the image itself. DICOM (Digital Imaging and Communication in Medicine) is the international standard for medical images and related information (ISO 12052) [49]. It is used to define the formats of medical images which can be exchanged and the quality of clinical data. DICOM is implemented in almost every radiology, cardiology imaging, and radiotherapy device. When a medical image is stored as a data, not only it contains the image data, but also the meta data belonging to the patient (e.g. age, sex, name) and to the examination (e.g. date of acquisition, modality). The developed class reads some of the most important DICOM tags (e.g. properties of the image) of the submitted study and stores them. This class also has a dedicated method which can check the conformity of a DICOM study, such as for example the quality of the images (e.g.

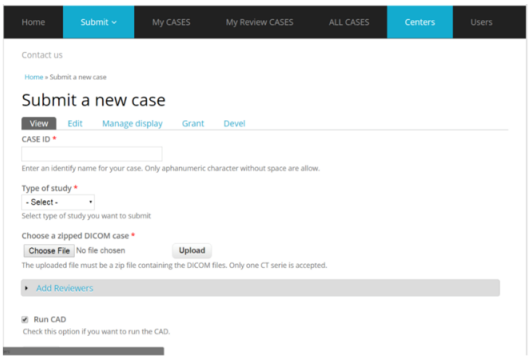


Fig. 2.1 Screenshot of the M5L submission form as seen by a submitter logging in with proper credentials.

minimum number of slices). In addition, the tool checks if the study is anonymized and if not, provides a full anonimization of the case, removing any data which can be associated to sensitive information of the patients.

Submission Module

This module has been conceived to be used by a technician. Its main goal is to submit study cases for CAD computations and select one or more radiologists for the review of the submitted case. A screenshot of the module is shown in Figure 2.1. The user, who has the grants to act as a submitter, connects with proper credentials to the WEB service. The available fields of the submission form are here explained:

- CASE ID: this is a unique identifier (alphanumeric characters without space) which is associated to the case. If the chosen ID is already present in the server, an error message is displayed and a new ID is asked, in order to avoid data corruption and duplication;
- Type of study: the user can choose between Screening study or Clinical study. It can be expanded with additional developments including the possibility to relate the submitted study to previous exams of the same patient;
- Upload browser: the uploaded file must be a compressed zip file containing only one valid DICOM study. In case of wrong format upload, a warning is returned to the user. In case of uploaded file with more than one DICOM study, only the first DICOM study will be processed, while the others will be skipped;

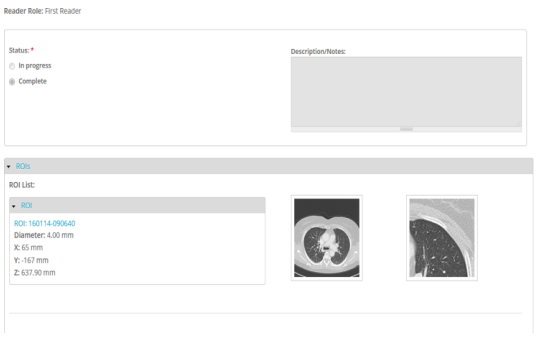


Fig. 2.2 Overview of the annotation module with an example of selected ROI. Screenshot on the left is the global axial view of the slice corresponding to the finding, while the screenshot on the right is the cropped zoomed view of the finding. Additional information about properties of the ROIs can be displayed clicking on the ROI name.

- Reviewers panel: the reviewer can choose, from the list of registered users, one or more radiologists who will review the case.

Reviewers can be chosen not necessarily belonging to the same institution, allowing the possibility to have multicentre annotations. In addition, the submitter has to chose the reading role of the reviewer: first reader or second reader (as explained in previous chapter). After the case have been submitted and the checks on anonymization and quality of the images are fulfilled, the study case will be analyzed by the CAD in a dedicated back-end which will be presented in Section 2.0.5. An e-mail will alert the submitter when the cases will have been processed. Consequentially all the selected reviewers receive an e-mail with the link to insert the medical annotations of the associated cases. With this asynchronous procedure the submitter can upload bunches of cases without waiting for CAD processing completion, while reviewers are finishing their medical reports.

Annotation Module

This module has been conceived to be used by a radiologist. Its main goal is to allow medical annotations of submitted cases by the radiologists through a dedicated WEB form. The radiologist, connecting with proper credentials, can access the list of associated cases and insert his / her annotation. A screen shot of the module is shown in Figure 2.2. The WEB form was built in order to give the possibility to include a detailed annotation, with details about morphological properties of the pathological ROIs. This level of details is a bit simpler than the usual protocol of

annotations followed in clinical practice with the annotation provided as a free-text description. In our approach we followed the guidelines listed by the Lung Image Database Consortium (LIDC) [50], a big initiative aiming at collecting a large-scale of annotated clinical data. The main reasons supporting our choice were:

- definition of a common protocol for the annotations: since there are no strict guidelines available for medical annotations of CT scans, we decided to propose to radiologists a common structured quantitative panel for annotations with the aim of homogenizing the medical report;
- facilitating data retrieval and queries: retrieving data from a free text box can really represents a difficult and time consuming task. The usage of quantitative scores and numbers for the properties of the findings facilitates the possibility to perform queries and statistics. In addition, it makes clinicians and researchers able to use these raw data for additional analyses (including for example the feeding of risk predictor algorithms);
- collecting a data-set of annotated clinical data: precious information that allows to train and validate the classifiers. The scarce availability of annotated anonymized clinical data still represents an issue. Collecting and sharing those data can improve the developments in the field.

A description of the fields available for annotation is then provided:

- Diameter: this is the nodule diameter (in mm) measured by the radiologist. If the finding does not have a perfect spherical structure (e.g. it is an ellipsoid), the diameter is taken as the length of the major axis;
- Position: 3D spatial positions (x,y,z) of the finding in mm;
- Subtlety: it is a measure of how difficult the nodule was to be detected according to the experience of the radiologist. Its score goes from 1 (Extremely Subtle) to 5 (Obvious);
- Internal Structure: it is used to classify the internal properties of the nodule: it goes from 1 (Soft tissue) to 4 (Air);
- Sphericity: it is used to classify if the shape of the nodules deviates from a symmetric spherical shape: it goes from 1 (Linear) to 5 (Round);

- Lobulation: it is used to rank the degree of lobulation, defined as an appearance resembling lobules: it goes from 1 (None) to 5 (Marked);
- Texture: it is used to classify the type of nodule: it goes from 1 (Non solid / GGO) to 5 (Solid);
- Calcification: it is used to highlight if the nodule presents a calcification pattern: it goes from 1 (Pop-corn) to 6 (Absent). Calcified pulmonary nodules are less likely to be malignant;
- Margin: it is a description of how well the margin of the nodule is defined: it goes from 1 (Poorly) to 5 (Sharp);
- Spiculation: it is used to quantify the edges of the nodule: it goes from 1 (None) to 5 (Marked). Typically, benign nodules have well-defined borders while malignant nodules are irregular or elongated;
- Malignancy: it is defined as the subjective assessment of the likelihood of malignancy, assuming the scan originated from a 60-year-old male smoker: it goes from 0 (Highly Unlikely to be cancer) to 4 (Higly Suspicious to be cancer). It is worth to notice that this is a qualitative assessment of the malignancy of the finding based on its visual properties. Only additional investigations can provide a proof of malignancy. Conversely, it is interesting to see how visual properties of the nodule influence the radiologist in the perception of malignancy.

Review Module

This module has been conceived to be used by a radiologist. Its main goal is to allow the review of CAD marks by the radiologist. The radiologist, connecting with proper credentials, can access the list of CAD marks of associated cases in two ways. If he / she was selected as first-reader, CAD marks will be automatically made available only when the first unassisted reading will be declared complete. If he / she was selected as second-reader, CAD marks are directly available during the annotation. A preliminary check is performed in order to compare CAD marks to annotated finding. If a CAD mark is spatially matched with any finding in the list, it is automatically marked as True Positive and it is assigned the same malignancy of the corresponding annotated finding. The remaining findings need to be marked by the radiologist

Fig. 2.3 Screenshot of the review module. Connecting with proper credentials, the radiologist has marked CAD findings as True Positive, False Positives, or Irrelevant.

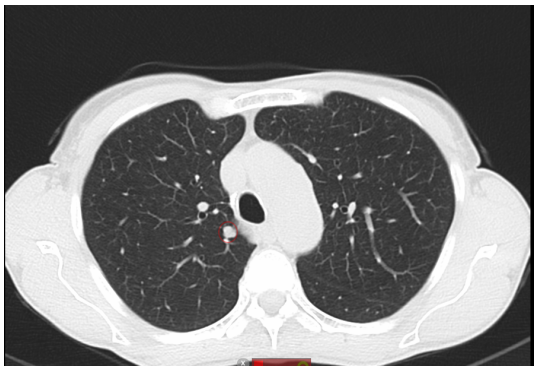


Fig. 2.4 Screenshot of the Osirix plugin with underlined a CAD mark in the circle. The radiologist can scroll through the slices and reject or accept the CAD mark. In this last case the mark is added on radiologist's annotation.

as: False Positive, Irrelevant (according to definition provided in [45]) and True Positive. In this last case, the radiologist is asked to specify the malignancy score. True Positives represent the nodules originally missed by the radiologist in the first unassisted reading and when the reviewing process is completed, they are added to the final annotation.

A screen shot of the the reviewing module with some examples is shown in Figure 2.3. CAD results can be downloaded in several formats (PDF report, XML and DICOM Structured Report) or visualized with a dedicated plugin for Osirix viewer [51]. Osirix is a DICOM viewer for iOS systems which is quite spread in clinical facilities. The plugin allows, after having downloaded the CAD results in XML format, to visualize and scroll CAD marks directly on the images. It allows the possibility to pre-load the list of findings from the radiologist and the list of CAD marks. A screenshot of the plugin is shown in Figure 2.4.

2.0.5 The Cloud back-end

As mentioned at the beginning of this Section, previous attempts to introduce parallel computing solutions to process CAD algorithms mainly failed due to the high rigidity and complexity of the infrastructure. The advent of Cloud Computing seems to solve previous issues. Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [52]. The main characteristics of Cloud Computing are:

- on-demand capabilities: the customer has access to the service maintained by the provider without having to take care of the infrastructure. In addition, the user can access the services directly on-line with an internet connection and can have the capability to change directly on-line the level of the service, or ask the provider to take care of the changes. This is typically referred as *pay-for-what-you-use* scenario, in which there can be a monthly subscription to have full access to the services or an on-demand option only when services are needed;
- broad network access: the services can be accessed everywhere since there is full compatibility with smartphones, laptops and tablets. This features become fundamental to allow users to work remotely on their projects from every location;
- resource pooling: it is possible to access (and monitor) computing resources from everywhere, having at disposal a flexible computational power adapting automatically to the required workload;
- rapid elasticity: it is possible not only to have available elastic computing resources, but also a real-time management of software features, users according to real-time business needs;
- storage and backup: it is possible to have a dedicated multi-site protected storage in which there is no risk of loosing data.

Cloud Computing solutions can be differentiated in three different kind of services:

- Infrastructure as a Service (IaaS): the provider offers to the consumer processing, storage and computational resources. The user deploys and runs the software making use of these resources. The customer is totally *blind* about the underlying Cloud infrastructure, whose control is managed by the provider. Conversely, the user has the control of the application running within this infrastructure;
- Software as a Service (SaaS): the provider gives the user its applications running on the Cloud infrastructure. The applications can be accessed by the user, usually through a WEB browser. The user does not have control on both applications and Cloud infrastructure, excluding some customization on some user-specific configuration settings;
- Platform as a Service (PaaS): the provider has the infrastructure, but it acquires applications developed by the customer. The customer does not have control on the infrastructure, but has total control on the applications. This solution is mainly used by developers.

Cloud Computing has entered with great force our daily life in recent years. An example of IasS is represented for example by Microsoft Azure [53]. Computational resources are provided to the users with different kinds of services, including for example storage. An example of SaaS is represented by Amazon, which offers to the users e-marketing services running on their private Cloud. An example of PaaS is represented by App Engine [54] launched by Google with the intent to offer a platform for app developers to run, test and share their code on a dedicated Cloud infrastructure. When looking at medical physics projects, Cloud Computing seems to be the best solution to tackle and solve mentioned issues with parallel computing. In particular, one of the key points of Cloud Computing is the elasticity of computational resources when allocated. The resources can automatically scale up or down in case of an increase or a decrease of the workload. This feature fits perfectly in the situation of a hospital facility, where resources can be adapted to the number of the submitted cases for CAD computations. Considering for example a scenario during a screening campaign: due to the high number of acquired CT scans we will face a peak in the demand of the computational resources. Conversely, there will be a decreasing during clinical routine, where the number of submitted CTs is much lower than during screening campaigns. Adopting a IaaS solution allows to free the users from buying static hardware to run CAD algorithms. Conversely,

Fig. 2.5 Pictorial view of the M5L on-demand system. CT are submitted directly as soon as acquired from the WEB front-end. While the user is submitting other CTs or inserting medical annotations, previous scans are copied to the Cloud back-end and analyzed by the CAD. When the computations are completed, CAD results are copied back in the front-end and users are alerted with an e-mail containing the links to retrieve CAD results.

adopting a SaaS solution allows to free the users from buying additional software to access CAD results. Our proposed solution was called *CAD on-demand*, since it combines both the IaaS and the SaaS approaches. The SaaS approach is represented by the WEB front-end and the M5L CAD, while the IaaS approach is represented by the dedicated Cloud back-end where the CAD algorithms run. A pictorial view of the M5L on-demand system is shown in Figure 2.5. The user submits study cases and selects reviewers through the WEB front-end. As soon as submitted, cases are processed by our CAD in the dedicated Cloud back-end. While the cases are being processed, the user can submit other cases or the reviewers can insert their medical annotations. A mail-based system will notify the users at the end of the processing. When the cases have been processed the results are copied back from the Cloud to the WEB server and the computational resources in the Cloud are free again to accept additional cases. The M5L CAD on-demand is hosted by the INFN Torino Computing Center. This center has a long story of computing applications, since it is a Tier-2 of the Large Hadron Collider (LHC) Computing Grid [55]. The main motivations for setting up a Cloud facility were to support a large number of applications (not necessarily of particle and nuclear physics) and to reduce the man-power required to manage a Tier-2 infrastructure. To reach this goal a Cloud infrastructure was deployed [56]. We decided to use all open sources tool to manage and control the facility. The main tool is Open Nebula [57], which is a free and open-source Cloud Management Platform. Its basic features are the possibility to manage hardware and infrastructure, but also the capability to manage virtual machine life-cycle. The basic idea of Open Nebula is to orchestrate storage, virtualization, monitoring and security in order to deploy computing facilities as

virtual machines on distributed infrastructures. The M5L version can be considered as formed by:

- one physical host used as WEB server;
- different virtual machines (VMs) representing the computational power.

We deploy the VMs using a specific sandbox (as prescribed by the IaaS approach) within a private network with a router with a public address which allows external access and connection with the WEB server. Presently, we can deploy up to 18 VMs and a total number 48 of cores (processing units). The storage is a 100 GB persistent ISCSI (Internet Small Computer System Interface) which is exported within the different VMs using NFS (Network File System). The disk is used for temporary storage of CT scans and CAD results, which are then copied back to the WEB server. In order to have the possibility to scale resources, an elastic cluster was created based on CERNVM Online system [58]. This system gives the possibility to create clusters formed by a head node (where the monitoring tools are installed) and many workers based on micro-CERNVM OS [59]. Two additional tools are used: HTCondor and Elastiq. HTCondor [60] can be considered as a batch system for managing compute-intensive jobs: taking as input a serie of parallel jobs, it is responsible for the correct management of job queuing, progress monitoring and notifying the user after the completion. Elastiq is a dedicated daemon written in Python which makes possible for resources in a cluster to scale up and down automatically. The basic idea underlying all our infrastructure is to limit the waste of computing resources. If some submitted CTs are waiting in the queue of HTCondor more than a predefined time (which can be tuned by the user), Elastiq deploys new workers so that they can be analyzed. Conversely, if some workers remain idle for a certain time, they are released and made available to other users of the Cloud. This is in line with the model of the Cloud called *multi-tenant model*, where resources are shared between different projects. There is no risk that users from different projects can access or create damages outside their projects, since an automatic systems of grants and securities manages the privileges of the users. Considering that M5L is the combination of two independent sub-systems, we set up two different jobs in HTCondor, so that the algorithms can run in parallel. HTCondor was configured to run a maximum of two jobs per worker at the same time. This configuration is needed in order to keep some cores free allowing multi-threading (i.e. splitting an algorithm into several processes on different processing units) and speed up the analysis.

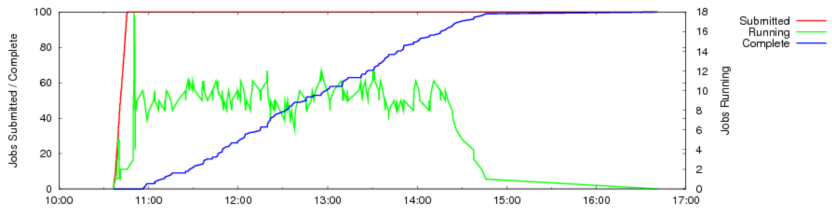


Fig. 2.6 Distribution of jobs and their completion time during the first phase of stress test: 100 exams submitted in one bunch.

Stress Test Configuration and Results

We decided to test the performances and the robustness of the system doing a stress-test. This stress-test was also motivated by the aim of simulating workloads comparable to daily clinical practice. We prepared a dedicated script written in Python which can submit cases to M5L in two phases: a peak submission and then progressively submission from different centers. The first scenario was used to simulate the case where a center sends all the studies in one bunch (it can be thought as for example due to a night batch routine process). 100 exams have been submitted in one bunch to M5L in less than 10 minutes, filling all the available slots of virtual machines. In the second scenario, we simulated three different medical centers with different parameters:

- large center: submission of an exam every 10 minutes. Total number of exams submitted: 100;
- medium center: submission of an exam every 20 minutes. Total number of exams submitted: 30;
- small center: submission of an exam every 30 minutes. Total number of exams submitted: 10.

We defined the size and the submission rates after investigating the activities of radiology departments in Italian hospitals. A total of 240 exams were successfully analyzed by M5L. The average number of slices per exam was 280. In average, each exam was processed in 19 minutes. This corresponds to the computational time of the slower algorithm (lungCAM) plus the time for the combination of results which can be considered as negligible (less than 5 seconds). Figure 2.6 and Figure 2.7 show the number of submitted, running and completed jobs during the first scenario

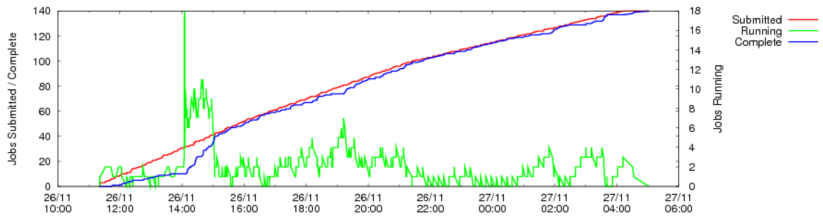


Fig. 2.7 Distribution of jobs and their completion time during the second phase of stress test: 140 exams with a slow and steady submission rate.

and the second scenario respectively. As we can notice from the first Figure, all the 100 cases were processed in approximately 4 hours. Conversely, in the second Figure the completion rate almost matches the submission rate. In both the Figures it is possible to notice a gap between the start of the submission and the peak in the processed jobs. We can explain this referring as the waiting time due to the re-allocation of computational resources which were currently being used by other applications. Since, as mentioned, the infrastructure is shared with several other projects, it is possible that there is a certain waiting time till the resources are released. The results are satisfactory, the system reacted as expected to the workloads and there were no cases which failed. Furthermore, the computational time is acceptable for an application in clinical practice, although it could be optimized

2.0.6 Conclusion and Discussion

After having developed and validated a complete CAD system for the automatic detection of pulmonary nodules, we have faced the problem of wide-spreading the usage of CAD among the biggest community of clinicians and of simplifying the access to the M5L CAD (or any other CAD system). We have developed a dedicated infrastructure based on a WEB front-end and a Cloud back-end. This infrastructure introduces a new paradigm of CAD in clinical practice: the CAD on-demand. The solution frees the users from buying additional software or hardware to access CAD results. This can have potential impact on reducing costs in hospital facilities, while introducing the benefit of supporting the radiologists with CAD systems in the detection process. The stress tests showed how this research prototype could be ready to expand its network of users potentially for a large scale service. In addition, the system allows multi-centre annotations, making possible also to small clinical facilities (which sometimes do not have a resident experienced pool

of radiologists) to benefit from the expertise of other structures. The Saas approach allows to combine several CADs, which have been demonstrated to have benefits on the overall performance [45] (the topic of combining CAD systems will be treated in detail in Chapter 3) with limited effort. The work remains open for additional developments both on the front-end and the back-end side. The front-end WEB could include a detailed viewer Java based (so that accessible from laptops and mobiles) providing a detailed and robust on-line diagnosis workstation to be applied for example in remote health projects. The parameters on the Cloud have not been tuned, but they can include for example the possibility to assign a priority code to the submitted cases, so that urgent and difficult cases can be processed as soon as possible. Only a daily usage of the system in clinical facilities will provide clinical requirements for additional developments. In addition, further features of the M5L CAD could be improved. One of the most desired features would be the possibility to compare different exams of the same patients taken sequentially along time (i.e, longitudinally). The goal of this development is to provide to radiologists the possibility to automatic follow and compare the evolution with time of the detected pathological ROIs (usually referred as *Longitudinal Analysis*). These developments on the algorithms should then be followed by parallel developments on the front-end infrastructure in order to provide to clinicians a real on-demand decision support platform for the diagnosis.

Chapter 3

Combining and Comparing CAD systems on large-scale datasets: the LUNA16 challenge

3.0.1 Motivation

A large number of CAD algorithms for the automated detection of pulmonary nodules have been discussed in the literature. In addition, the number of available commercial CAD systems has rapidly increased. However, comparing these systems in terms of performances is far from being a trivial task. In fact, it is not possible to compare directly the sensitivities of the systems, since they can vary tremendously. The main reason underlying this inconsistency, is that the evaluation can strongly depend on the different datasets used for training and evaluation. It would not be fair and objective to compare CAD systems evaluated on a different datasets, since this operation cannot exclude possible biases. A detailed large-scale evaluation and comparison of state-of-art developed CAD systems has not been published in literature yet. A step toward an objective evaluation of CAD systems was the ANODE09 study [45]. 7 different CAD systems were evaluated on a common data-set of 50 scans using an objective metric. A really important improvement of this work is to significantly increase the evaluation dataset up to 1 order of magnitude together with including recent state-of-the-art algorithms. This last point becomes even more important when considering recent techniques, for example, for classification and false positive reduction such as Convolutional Neural Networks (CNNs) [61]. These techniques

have been deeply used in the field of imaging recognition, but only recently have been applied by machine learning researchers to the field of medical imaging. Another interesting study introduced in [45] investigated how combining different CAD systems influences the overall performance. Results showed how the combination of different CAD systems performs better than single standalone CAD systems. It becomes now interesting to see the results of the combination of classical candidate detectors with state-of-the-art false positive reduction systems (e.g CNNs). All these reasons motivated the organization of an international medical challenge aiming at validating, comparing and combining CAD algorithms on a large heterogeneous dataset. In addition, it is interesting to see how the combination of mentioned systems can introduce clinical benefits. A recent article [62] states that a review of 4 CAD systems underlined that about 20% of cancers originally annotated by the radiologists, were missed by the automated systems. It is worth to repeat a similar experiment using the combination of different CAD systems. The work of this Chapter has been carried on within the Radiology Department of the Radboud University Medical Center in Nijmegen, the Netherlands, during a 1-year period spent by the author as visiting researcher.

3.0.2 Data

The dataset used for this challenge is taken from the biggest publicly available dataset of CT scans: the LIDC-IDRI database [50], which originally contains a total of 1018 CT scans. All the scans have one Extensible Markup Language (XML) file which contains annotations of pulmonary nodules. The annotations represent the final output of a two-phase annotation procedure where the scans were reviewed by four experienced radiologists. The initial phase was called *blinded phase*. In this phase each radiologist annotated independently all the scans. The reading protocol required to annotate: nodules larger than 3 mm, nodules smaller than 3 mm and any other kinds of abnormality in the lung, but not interpreted as possible pulmonary nodule. In addition to 3D spatial coordinates, further information about morphological features of the findings were provided. In the second reading phase (*unblinded reading phase*) the anonymized previous results of all the radiologists were shown to each of the radiologists. Then, each radiologist reviewed the marks of the colleagues with no consensus forced. The final result is a list of marks accepted by at least 1 to 4 radiologists. The dataset is the most suitable to be used for evaluating CAD

Manufacturer	Model name	Number
GE MEDICAL SYSTEMS	LightSpeed16	197
GE MEDICAL SYSTEMS	LightSpeed Ultra	162
GE MEDICAL SYSTEMS	LightSpeed QX/i	97
GE MEDICAL SYSTEMS	LightSpeed Pro 16	79
GE MEDICAL SYSTEMS	LightSpeed VCT	61
GE MEDICAL SYSTEMS	LightSpeed Plus	56
GE MEDICAL SYSTEMS	LightSpeed Power	10
Philips	Brilliance 16P	54
Philips	Brilliance 64	49
Philips	Brilliance 40	9
Philips	Brilliance16	5
SIEMENS	Sensation 16	95
SIEMENS	Sensation 64	5
SIEMENS	Definition	3
SIEMENS	Emotion 6	1
TOSHIBA	Aquilion	5
Total		888

Fig. 3.1 Distribution of manufacturers and scanner models of the scans used in our study.

algorithms, because it is very heterogeneous. In fact, it includes both scans from clinical and screening acquisitions. In addition, a wide range of different acquisition parameters is represented: different scanner models, reconstruction kernels and slice thickness. Figures from 3.1 to 3.3 summarize some characteristics of the scans used in our study. For our study we decided to remove from the data-set all the scans with a slice-thickness larger than 3 mm, as suggested by [63]. An additional check in order to exclude scans with missing slices or inconsistent slice spacing was performed. The final list is formed by 888 scans. We provided participants with the data-set in terms of couple MetaImage (.mhd) / Raw (.raw) images. Participants were able to download the full data-set directly from the web site of our challenge (http://luna16.grand-challenge.org/). Together with the images, the participants were able to download the list of annotated nodules. In our study we considered as reference standard all the nodules larger than 3 mm annotated by at least 3 out of 4 radiologists. The reference standard then consisted of 1186 nodules. The considered dataset is also very heterogeneous in terms of size and type of the nodules. Figure 3.4 shows some screenshots of the nodules in the reference standard. A wide range of type of nodules is represented: from big solid nodules (third, fourth and fifth column) to part-solid (first column) and GGOs (second-column). Figure 3.5 shows the distribution of the frequency of the nodules divided according to their size (defined as the average size of the measurements of the radiologists). Also in this case, the size of the nodules in the reference standard covers a wide range. Other

Section thickness	Number
0.6	7
0.75	30
0.9	2
1	58
1.25	343
1.5	5
2	123
2.5	320
Total	888

Fig. 3.2 Distribution of section thicknesses of the scans used in our study.

findings (nodules annotated by less than 3 out of 4 radiologists, nodules smaller than 3 mm and not nodules) were considered as irrelevant findings and not considered in the performance evaluation.

3.0.3 LUNA16 challenge framework

The challenge has been called Lung Nodule Analysis 2016 (LUNA16) and was proposed and accepted in the challenge section of the International Symposium On Biomedical Imaging 2016 in Prague. The aim of the challenge was to ask participants to develop CAD systems for the automated identification of pulmonary nodules in chest CT scans. LUNA16 was conceived as a completely open challenge. In fact, there is no distinction between training and testing dataset: the evaluation of the algorithms is performed on the same data-set. In order to prevent possible biases due to training and testing on the same dataset, the participants were asked to perform a cross-validation. A dedicated web framework was created, so that participants were able to upload their results. The submissions were automatically evaluated as soon as received and presented on the website in a leader-board. The evaluation code runs in a dedicated Amazon AWS [64] virtual machine specifically set up for this challenge. More details about the challenge are now provided.

Manufacturer and reconstruction kernel	Type	Number
GE MEDICAL SYSTEMS - BONE	Enhancing	220
GE MEDICAL SYSTEMS - LUNG	Overenhancing	70
GE MEDICAL SYSTEMS - STANDARD	Standard	372
Philips - B	Standard	21
Philips - C	Enhancing	7
Philips - D	Overenhancing	45
SIEMENS - B20s	Soft	1
SIEMENS - B30f	Standard	102
SIEMENS - B31f	Standard	1
SIEMENS - B45f	Enhancing	30
SIEMENS - B50f	Enhancing	2
SIEMENS - B70f	Overenhancing	12
TOSHIBA - FC03	Standard	2
TOSHIBA - FC10	Soft	3
Total		888

Fig. 3.3 Distribution of the reconstruction kernels of the scans used in our study.

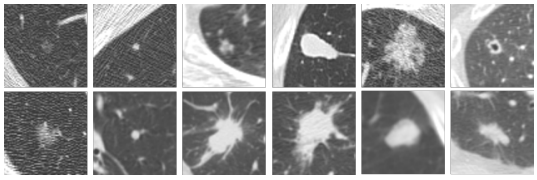


Fig. 3.4 Screenshots of some example of nodules available in the reference standard. A very wide range of type of nodules is intentionally included.

Challenge Tracks

The challenge has been divided into two separate tracks: 1) complete nodule detection (NDET); 2) false positive reduction (FPRED). In the first track participants were required to develop a complete CAD system. The input for the participants to this track are only the CT images. In the second track, participants were required to propose solutions for the false positive reduction stage. Participants to this track are given a list of candidates (composed using existing nodule candidates detection algorithms). The second track can be imagined as a typical two-class machine learning problem: giving a list of false positive and true positives, provide the classification. We decided to include this second track in our challenge for one main reason: involve and attract machine learning groups which have not been working directly on medical imaging projects, but with the expertise on the most recent and newest classification techniques.

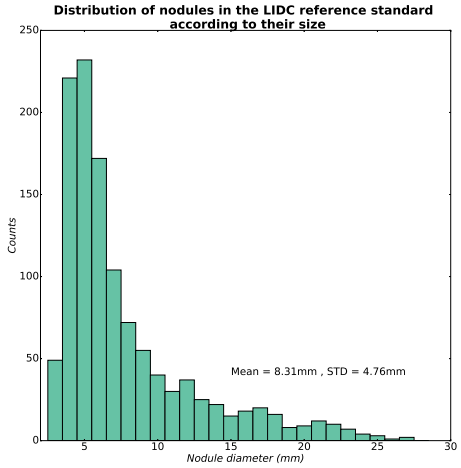


Fig. 3.5 Distribution of the frequency of the nodules divided according to their size (defined as the average size of the measurements of the radiologists). Minimum size: 3 mm, maximum size 33 mm.

Cross-validation

As mentioned, in order to avoid possible biases within an open-challenge, participants need to be instructed on how to train and test their algorithms. It was mandatory for participants to perform a 10-fold cross validation. For this purpose, we randomly split the data-set into ten subsets. The rules to perform 10-fold cross validation are the following (for fold N):

- splitting the data-set in training and testing data-set. So that if subset N is using as testing data-set, the remaining sub-sets are used as training;
- for the FPRED, extracting the training and testing candidates on the corresponding testing and training data-set;
- training the algorithm using the training data-set;
- testing the algorithm on the testing data-set and producing the result file;
- merging the result files after iterating the process for all the folds.

Evaluation

We asked participants to submit their results in terms of a Comma Separated Value (csv) file. Accessing with proper credentials, the result files can be uploaded directly on the web site of the challenge. The mentioned csv files contain the list of marks produced by the CAD systems. For each mark, the position (x,y and z coordinate), the reference to the corresponding image and the corresponding nodule probability are provided. Each CAD mark in the list is considered as a True Positive when the euclidean distance d between the mark and each nodule in the reference standard is smaller than half the diameter of the reference nodule. In the case in which a nodule in the reference standard corresponds to different CAD marks, the mark with the highest probability is selected. If a CAD mark is matched with any nodule in the list of the irrelevant findings, it is discarded from the evaluation and it is not considered as False Positive (FP). All the other CAD marks not falling into two previous categories are considered as FPs. The results are evaluated through the FROC analysis [65]. In this curve, the sensitivity is plotted as a function of the average number of FPs per scan. The sensitivity is defined as the fraction between detected TPs and the number of nodules forming the reference standard. Points on the FROC curve are obtained looking at the probabilities for each CAD mark in the result file. An iterative process considers only CAD marks whose probability p is above a certain threshold t. Then, from this sub-list, TPs and FPs are selected according to the criterion presented above. All the points in the FROC curves are determined by the corresponding values of p. In addition, we also computed the 95% confidence interval of the FROC curve using a bootstrapping technique with 1000 bootstraps as explained in [66]. In order to extract a single score for each system from the FROC curve, 7 values of sensitivity at seven predefined false positive rates per scan are extracted: $\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8$. The overall score, called Competition Performance Metric (CPM), is defined as the average of these seven scores. The worst possible system will have a score of 0, while the best possible system will have a score of 1. It is worth noticing that most of the available CAD systems are tuned to operate in the range between 1-4 FP/scan. We decided to make the evaluation more challenging including in the evaluation very low FP rates (and the corresponding sensitivities). The main motivation underlying this decision was to evaluate the capability of CAD systems to detect clinically relevant nodules already at a very low FPs rate, in order to investigate the possibility to use those systems as *autonomous*

readers. The evaluation script is free for download, so that participants can test their algorithms before the submission of the results.

Methods: candidate detector algorithms

We provide now a description of the best algorithms taking part in the challenge. We first describe the candidate detector algorithms which were used to generate the list of candidates for the false positive reduction track:

- ISICAD: it is based on the methods described in [67]. Reshape of the image is performed as first step in order to obtain an isotropic resolution. The candidate detection algorithm is mainly based on the computation of two quantities: the Shape Index (SI) and the Curvedness (CV). These two quantities are computed for each voxel composing the volume of the lungs, using the principal curvatures k_1 and k_2. In particular, $SI = \frac{2}{\pi}\arctan(\frac{k_1+k_2}{k_1-k_2})$ and $CV = \sqrt[2]{k_1^2+k_2^2}$. Seed points for the candidates are then obtained applying a threshold on the SI and CV values. Seed points are then expanded to form voxel clusters. The center of the mass of the cluster is considered to be the candidate position;

- SUBSOLIDCAD: this algorithm has been built with the dedicated purpose to detect sub-solid nodules [68]. It is based on double thresholding on the mask of the lungs. Sub-solid nodules are usually found to be between the range of 750 and -300 HU. To remove other anatomical structures which can possibly remain attached to candidates, morphological operations and connected component analyses are applied to obtain the final list of candidates;

- LARGECAD: this algorithm has been built with the dedicated purpose to detect nodules with a diameter larger than 10 mm. Due to their big extension and shape properties, solid and sub-solid nodules are not tuned to detect these big structures. The algorithm starts applying a threshold of -300 HU (usual value associated to solid nodules), combined with morphological operations. Clusters are determined using connected component analysis, discarding the clusters with a diameter smaller than 8 mm and larger than 40mm;

- ETROCAD: this algorithm is described in [69]. It is based on nodule and vessel enhancement filters. Three different kind of filters [42] are applied to detect: isolated, juxtavascular and juxtapleural nodules. The maxima of the

divergence of the normalized gradient (DNG) of the image $k = \div(\rightarrow w)$ is used to the reduce the FP rate and to better estimate the center of the nodules. Where $\rightarrow w = \frac{\Delta L}{||\Delta L||}$ and L is the intensity of the image. To include the variability in the size of the nodules, DNG and filters are computed at different scales. Cluster merging is performed to ensure that a nodule is represented by a single mark;

- M5LCAD: the features of this algorithms have been presented in Chapter 2.

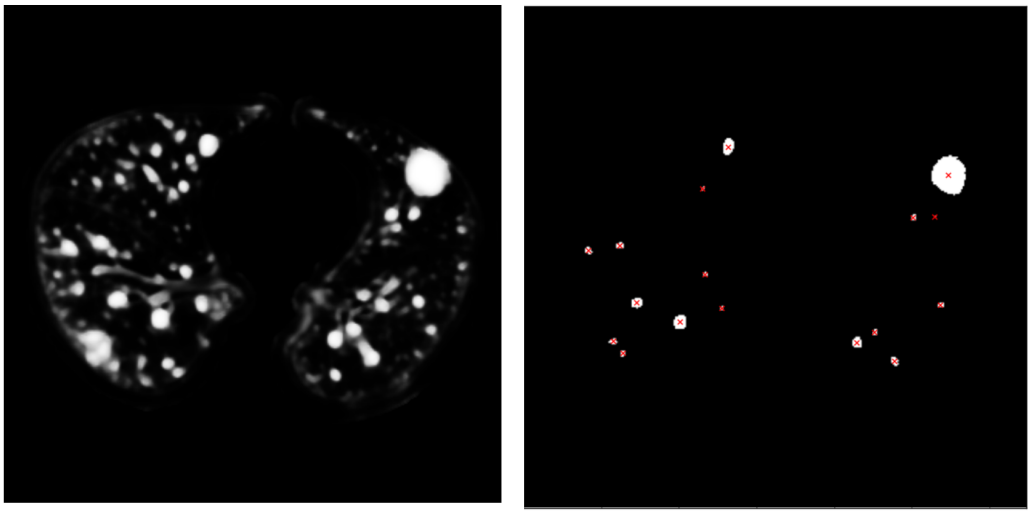
The top 5 systems submitted in the NDET track are then explained:

- ZNET: this system makes use of CNNs for both candidate detection and false positive reduction (a brief overview of main topics of CNNs is available in Appendix **??**). The candidate detection part uses the U-Net [70]. The input image is changed to 512x512 to obtain isotropic resolution. In order to avoid a possible over-training, dropout has been added between the convolutional layers of the network. The candidates are extracted looking at the probability map, which is the output of the net. The slice corresponding to the candidate is then thresholded and eroded. Applying connected component analysis allows to group the candidates. The coordinates of the candidates are taken as the center of mass of the components. A screenshot of the output of the network and the candidates is available in Figure 3.6. False positive reduction stage will be explained later in a dedicated section;

- LUNAAIDENCE: this is a commercial system developed by the AIDENCE (http://aidence.com) company. Since the system is protected by copyright we did not receive any additional information about the algorithms. The owners stated that the system makes use of end-to-end ConvNets;

- VISIACTLUNG: this a commercial system developed by MeVis Medical Solutions AG, Bremen in Germany. The system has also obtained an FDA approval, to be used as a support for radiologists to detect ROIs initially overlooked. Since the system is protected by copyright we did not receive any additional description by the owners;

- ETROCAD: this system is described in [69]. We already described its candidate detector algorithms previously. False positive reduction is performed



(a) (b)

Fig. 3.6 View of the candidate detection procedure of the ZNET system: (a) raw output from the UNET CNN, b) image after thresholding and erosion. The red circle indicates the coordinates of each candidate, corresponding to the center of mass of each connected component.

extracting some dedicated features from the candidates and applying a classification step. Some of those features, chosen to be invariant on a 3D gauge coordinates system are: shape features, regional features. The classification is then performed using the feature vector as input for an artificial neural network (ANN) based on genetic algorithms, called FD-NEAT. A genetic algorithm (GA) is a method for solving optimization problems that is based on natural selection, the process that drives biological evolution.

- M5LCAD: the system has already been described in Chapter 2.

We now describe the the best four methods submitted for the false positive reduction track:

- CUMEDVIS: this false positive reduction algorithm is based on the work presented in [71]. A summarized view of the different architectures is shown in Figure 3.7. It makes use of multi-level contextual 3D CNNs. The improvements of this algorithm with respect to the previous work is to tackle the problem to deal with different size and properties of pulmonary nodules. The adopted solution uses three different CNN architectures (Archi I, Archi

II, Archi III). Each architecture presents different receptive fields applied to the image. This procedure allows to capture different levels of contextual information. Archi I is formed by: a receptive field of 20x20x6, three convolutional layers with 64 kernels of 5x5x3,5x5x3,5x5x1 respectively. The final part of the architecture is composed by a fully connected layer with 150 hidden neurons and then the usual soft-max layer. Archi II is formed by: a receptive field of 30x30x10, a first convolutional layer with 64 kernels 5x5x3, a max-pooling layer of kernel 2x2x1. There are also two additional convolutional layers with 64 kernels of 5x5x3 each. Then there is the usual fully-connected layer of 250 hidden neurons and the softmax layer. Archi III is characterized by a very big receptive field of 40x40x26. It is then followed by a first convolutional layer with 64 kernels of 5x5x3 and a max-pooling layer of 2x2x2 kernels. Two additional layers of 64 5x5x3 kernels are added. Finally, there is the fully-connected layer with 250 hidden neurons and the usual softmax layer. Since each softmax layer of the different architectures returns a separate value of the probability, the final prediction probability is obtained fusing all the previous one. The fusion is performed weighting a linear combination of previous probabilities. A pre-processing step is applied to the input images: clipping of the voxel intensities in the interval from -1000 HU to 400 HU, normalization of clipped intensities between 0 and 1. Since the number of the false positives in the candidate list was much greater than the number of true positives (*class imbalance*) data augmentation to available images has been applied. Data augmentation is a technique which applies geometric transformations to the original images, so that additional images are generated, increasing the data-set. It is worth using this technique when the number of samples in the positive class is much smaller than the number of samples in the negative class. This data augmentation included: translation (one voxel along each axis) and rotation (90,180,270 degrees) in the transverse plane. Weights of the network were initialized using a Gaussian distribution as common practice in CNNs. The optimization of the weights was performed using the standard technique of back-propagation with momentum [72], while to avoid over-training Dropout [73] was applied. All the system has been implemented using the Python library Theano [74]. The training has been performed using a CPU equipped with a Graphic Processing Unit (GPU) NVIDIA TITAN Z for acceleration.

- DIAGCONVNET: this method has been described in [75]. It is based on multi-view CNNs. The first part of the algorithm is the extraction of patches of 65x65 from the input images. Patches are extracted using 9 different views, which correspond to the nine planes of symmetry of a cube. The architecture of the CNN is formed by: 3 consecutive convolutional layers and a max-pooling layer. First convolutional layer is formed by 24 5x5 kernels, second convolutional layer by 32 3x3 kernels, third convolutional layer by 48 3x3 kernels. Max pooling layer is used to reduce the size of the patches by a factor 2. The last layer is represented by a fully-connected layer with 16 neurons using Rectified Linear Units (ReLU) as activation function of the neurons. The different CNNs are fused using the method presented in [76] and [77]. To deal with the problem of data imbalance, data augmentation to the data-set was applied. Applied data augmentation: random zooming [0.9.1.1] and random rotation between -20 and 20 degrees. The system has been implemented using the Python library Theano. The training has been performed using a CPU equipped with a GPU NVIDIA TITAN X for acceleration. A schematic overview of the architecture is shown in Figure 3.8;

- ZNET: this algorithm makes use of the most-recent state-of-the-art CNNs: wide residual networks [78]. The first step is to extract 64x64 patches of sagittal and coronal views of the input images. Each slice is processed separately by a residual network. The final prediction is obtained averaging the output values of the different slices. The architecture is formed by 4 consecutive sets of convolutional layers. First set: one convolutional layer of 16 3x3 kernels. Second set: 10 convolutional layer of 96 3x3 kernels. Third set: 10 convolutional layers of 192 3x3 kernels. Fourth set: 10 convolutional layers of 384 3x3 kernels. The second to forth convolutional layers are then followed by 2x2 max pooling layers. Last layer is connected to a max-pooling layer of 1 8x8 kernel. Weights were initialized using the Xavier Method [79]. Optimization was performed using the ADAM technique [80]. To deal with the problem of data imbalance, data augmentation was applied. It included: flipping, rotation, zooming and translation. Data augmentation was applied not only during the training (as common practice), but also on the test data-set, in order to improve the test set scores. The system has been implemented using the Python libraries Theano and Lasagne [81]. The training has been performed using a CPUs

on different clusters, equipped with different GPUs: Tesla K40M, NVIDIA TITAN X, GTX 980, GTX 970, GTX 760 and the GTX 950M;

- CADIMI: this algorithm makes use of a custom 2D CNNs. Patches from three different views are extracted from input images: axial, sagittal and coronal. The patches are extracted at three different locations: the plane in the exact location of the candidate, 2 mm on both directions on the remaining free axis. Patches are then concatenated together and stored as three dimensional arrays, in such a way that the final patch is centered around the candidate location with a dimension of 52x52x3 mm. The architecture is formed by three convolutional layers and one max pooling layer. First layer: 24 5x5 convolutional kernels. Second layer: 32 3x3 convolutional kernels. Third layer: 48 3x3 convolutional kernels. The max pooling layer is connected to a fully connected layer with 512 neurons with ReLU as activation function. As usual, at the end of the architecture, there is a soft-max layer with 2 output units. In order to reduce over-fitting, batch normalization has been applied. Weights were initialized using HE uniform initialization [82]. The training was performed using 80 epochs with a random sample of 20.000 negative samples and using a mini batch size of 128. Due to data imbalance, data augmentation was performed with vertical / horizontal flipping and random cropping. The system has been implemented using the Python libraries Theano and Lasagne [81].

Archi-a			Archi-b			Archi-c		
layer	kernel	channel	layer	kernel	channel	layer	kernel	channel
C1	5×5×3	64	C1	5×5×3	64	C1	5×5×3	64
M1	1×1×1	64	M1	2×2×1	64	M1	2×2×2	64
C2	5×5×3	64	C2	5×5×3	64	C2	5×5×3	64
C3	5×5×1	64	C3	5×5×3	64	C3	5×5×3	64
FC1	-	150	FC1	-	250	FC1	-	250
FC2	-	2	FC2	-	2	FC2	-	2

C: convolution, M: max-pooling, FC: fully-connected

Fig. 3.7 Summarized views of the architectures of the CUMEDVIS CNN.

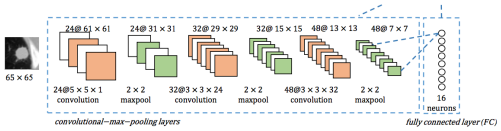


Fig. 3.8 Summarized view of the architecture of the DIAGCONVNET system.

Combination of the systems

It is interesting to investigate how the combination of different CAD systems influences the overall detection performance. In particular it is worth verifying if systems can be complementary: when multiple systems focus on different techniques to approach and solve a problem, a proper combination of those techniques can produce a performance which might be much better than that of the stand-alone systems. Defining a method for combining different systems without accessing the properties of the algorithms (e.g. the features of the classifier) is far from being a trivial task. Since in our challenge we could only access the list of CAD marks and their probabilities for each system, we defined a method for the combination which makes use only of those information. In addition, to build a correct and fair combination, systems with higher performances should be weighted more than systems with a lower performance. Let's define with $p_i = 1,..,n$ the probability of each CAD mark. For every value of p it is possible to compute the number of TPs when considering all the findings with $p_i \geq p$ as positive. Discarding irrelevant findings, it is also possible to compute the number of FPs at this threshold. We can define the following quantity:

$$f(p) = \frac{TP}{TP + FP + 1}$$

where we needed to add the factor +1 in the denominator to count for the situation in which all the findings are irrelevant (denominator goes to 0). This quantity represents the probability that a finding in the evaluation data-set with a probability p or higher can be a true nodule. The combination is then done by combining each $f(p)$ for every finding and for every system. Since all the findings are sorted in term of descending probability we have:

$$f_i \geq f_j$$

when $i \leq j$. Starting from $i = 1$, f_j is checked for all the findings if corresponds to f_i. In our combination we considered findings closer than 5 mm to each other the same entity, and we merged them. When two findings correspond, the probability of the merged finding is set to be:

$$f_i \rightarrow f_i + f_j$$

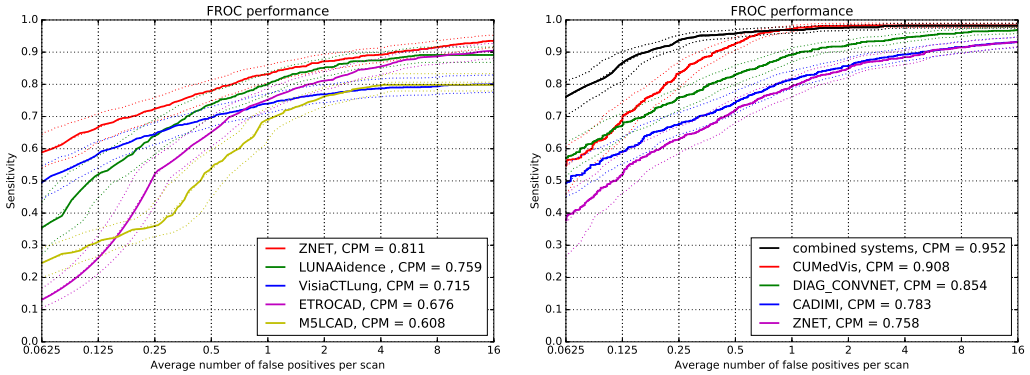
System name	Combination	Total number of candidates	Sensitivity	Best single sensitivity	Difference	Average number of candidates / scan
ISICAD	■□□□□	298 256		0.856		336
SubsolidCAD	□■□□□	258 075		0.361		291
LargeCAD	□□■□□	42 281		0.318		48
M5L	□□□■□	19 687		0.768		22
ETROCAD	□□□□■	295 686		0.929		333
	■■■■■	754 975	**0.983**	0.929	0.054	850
	■■□■■	732 901	0.983	0.929	0.054	825
	■■■□□	750 838	0.980	0.929	0.051	845
	■■□□■	728 162	0.977	0.929	0.048	820
	■□■■■	553 327	0.969	0.929	0.040	623
	□■■■■	551 227	0.969	0.929	0.040	620
	■□□■■	529 404	0.967	0.929	0.038	596
	■■■■□	559 543	0.965	0.856	0.109	630
	□■□■■	524 726	0.965	0.929	0.036	591
	■□■□■	548 523	0.964	0.929	0.035	618
	□■■□■	545 204	0.964	0.929	0.035	614
	■□□□■	524 108	0.959	0.929	0.030	590
	■■□■□	530 942	0.954	0.856	0.098	598
	□■□□■	518 058	0.954	0.929	0.025	583
	□□■■■	326 274	0.954	0.929	0.025	367

Table 3.1 Results of all the candidate detectors standalone and the best 15 combinations from the five systems sorted by the sensitivity. The filled and open squares indicate which systems have and have not been included in the combination. Total number of detected candidates is shown in the third column. The forth column lists the sensitivity, while the fifth column is the best score of any single system included in the combination. The difference between the sensitivity of the combination and the best score of a single system in the combination is given in the sixth column. The seventh columns is the average number of candidates per scan.

Summarizing, systems with lower performances have values of f which are approximately around 0 for almost all the findings. These systems are therefore weighted less in the combination.

Results

We now describe the results of the methods presented in the previous Section. We start from the candidate detector systems. Table 3.1 summarizes the results of all the candidate detector systems and the top performance combinations. It is immediately possible to see how the single sensitivities cover a vast range: from 32% to 93%. Conversely, the combination increases the performance up to 98.3%. We can then compute the FROC curves of the systems for the NDET track, shown in Figure 3.9a. The FROC curves for the systems in the FPRED track are shown in Figure 3.9b. CUMEDVIS is the best system, achieving a CPM score of 0.908. Table 2 presents all the possible combinations of the FPRED systems. The results are quite unexpected: despite all the systems making use of CNNs, the combination is still much better than the single best system.



(a) (b)

Fig. 3.9 FROC curves of the systems in (a) nodule detection track and (b) false positive reduction track. Dashed curves show the 95% confidence interval estimated using bootstrapping

Analysis of false positives

So far we have demonstrated how state-of-the-art CAD systems and false positive reduction systems reach very remarkable performances. In addition, the combination of previous systems significantly improves the overall performances, with results much better than any stand-alone system. However, it is still needed to evaluate how those systems (or the best combination) can influence the detection of clinically relevant nodules. In order to evaluate if there can be some lesions missed during the reading process in the LIDC/IDRI data-set, we performed an observer study. We considered the CAD marks from the best combination of the FPRED systems. We extracted all the marks, labeled as False Positives during the evaluation process, corresponding to the fix operating point of 1 FP/scan. The output of this operation is a list of 888 marks. A preliminary reading was performed by researchers to eliminate from the list marks which were obviously false positives (e.g. vessels). The filtered list was given for reading to four experienced radiologists. Using a dedicate workstation prepared for this study, they had to mark the findings as: false positive, true positive or irrelevant. They were able to visualize the findings in coronal, sagittal and axial views and to use measurement tools in order to determine the size of the nodules. A summary of the observer study is shown in Table 3.3. Among 194 CAD marks, 166, 122, 67, and 30 CAD marks are accepted as nodules $\geq$3 mm by at least 1, 2, 3, or 4 radiologists, respectively; 22 out of 28 remaining CAD marks are

System name	Combination	0.125	0.25	0.5	1	2	4	8	CPM	Best single CPM	Difference
CUMedVis	■□□□	0.677	0.834	0.927	0.972	0.981	0.983	0.983	0.908		
DIAG CONVNET	□■□□	0.669	0.760	0.831	0.892	0.923	0.945	0.960	0.854		
ZNET	□□■□	0.583	0.677	0.743	0.815	0.857	0.893	0.916	0.783		
CADIMI	□□□■	0.511	0.630	0.720	0.793	0.850	0.884	0.915	0.758		
	■■□□	0.831	0.917	0.965	0.979	0.981	0.981	0.981	0.948	0.908	0.040
	■□■□	0.802	0.903	0.948	0.976	0.979	0.979	0.980	0.938	0.908	0.030
	■□□■	0.831	0.927	0.968	0.976	0.979	0.981	0.981	0.949	0.908	0.041
	□■■□	0.550	0.680	0.796	0.869	0.912	0.938	0.959	0.815	0.854	-0.039
	□■□■	0.616	0.737	0.831	0.888	0.931	0.953	0.964	0.845	0.854	-0.009
	■■■□	0.817	0.912	0.954	0.968	0.975	0.979	0.982	0.941	0.908	0.033
	■■□■	0.859	0.937	0.958	0.969	0.976	0.982	0.982	**0.952**	0.908	0.044
	■□■■	0.820	0.907	0.946	0.968	0.976	0.981	0.981	0.940	0.908	0.032
	□■■■	0.635	0.777	0.839	0.888	0.929	0.954	0.965	0.855	0.854	0.001
	■■■■	0.830	0.912	0.947	0.964	0.973	0.979	0.981	0.941	0.908	0.033

Table 3.2 Results of all the false positive reduction systems and all the possible combinations of the four systems. The filled and open squares in the first column indicate which systems have and have not been included in the combination. Columns from 3 to 9 indicate the value of sensitivity for different working points. The average sensitivity (CPM score) is indicated in the tenth row. The eleventh row is the best CPM score of any single system included in the combination. The difference between the CPM score of the combination and the best single CPM score is given in the last column.

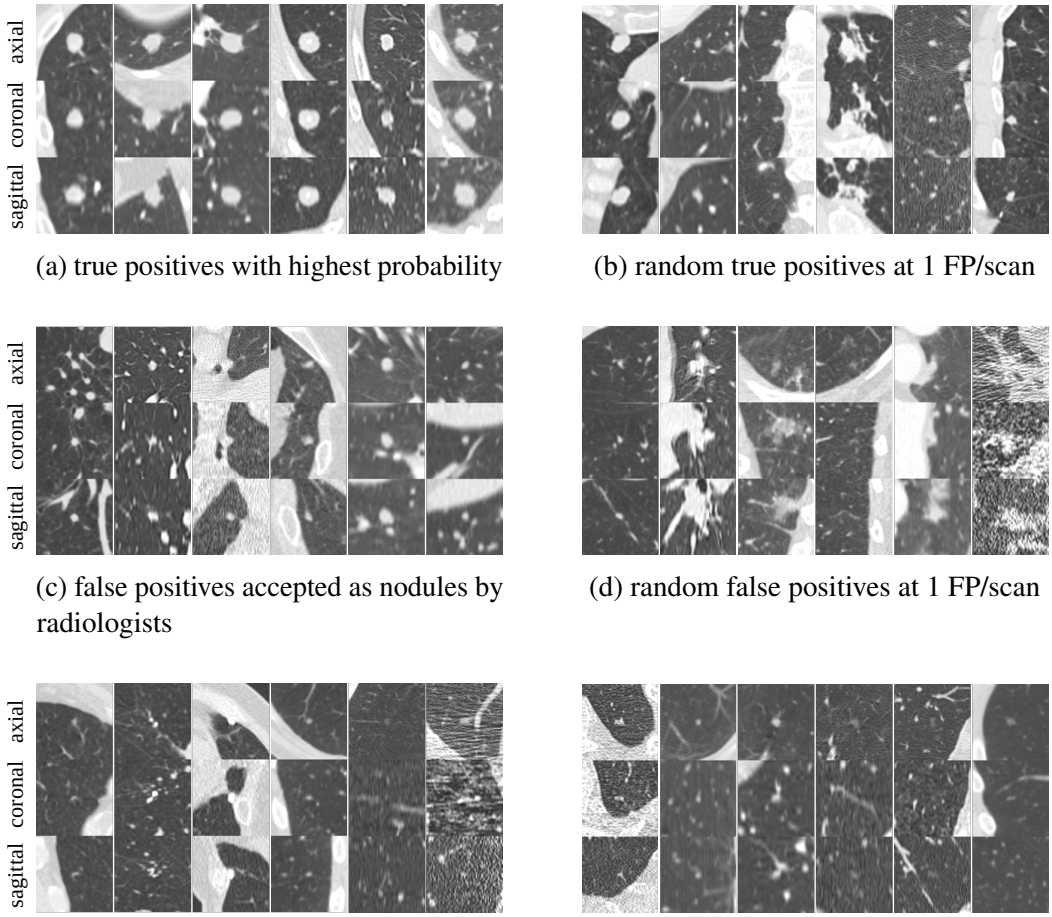
considered as nodule <3 mm. Examples of nodules found in this observer study are shown in Figure 3.10c.

Table 3.3 Overview of the observer study on 888 false positives at 1 FP/scan. The table shows the number of false positives that are accepted by the radiologists as nodules ≥3 mm at different agreement levels. The number of false positives that are not accepted as nodules ≥3 mm, but are accepted as nodules <3 mm, are also included. The number of accepted CAD marks at different range of FPs/scan is shown, where n is the FPs/scan rate.

Category	Number	$n \leq 0.25$	$0.25 < n \leq 0.5$	$0.5 < n \leq 0.75$	$0.75 < n \leq 1.0$
nodule ≥3 mm - at least 1	166	71	41	34	20
nodule ≥3 mm - at least 2	122	60	28	21	13
nodule ≥3 mm - at least 3	67	44	13	8	2
nodule ≥3 mm - at least 4	30	22	4	3	1
not nodule ≥3 mm - nodule <3 mm	22	5	5	8	4

Discussion of the results

We showed how combining different systems leads to a increase in the overall detection sensitivity. In our challenge we highlighted how candidate detectors play an important role. First of all, they were used to give to participants to the FPRED track the list of candidates to perform the classification. The maximum sensitivity which can be achieved by a false positive reduction system is the upperbound performance of the candidate detector. In Table 3.1 we showed that the evaluated candidate detector systems present a wide range of detection sensitivity: from 31.8%



(a) true positives with highest probability

(b) random true positives at 1 FP/scan

(c) false positives accepted as nodules by radiologists

(d) random false positives at 1 FP/scan

(e) false negatives from the candidate detectors

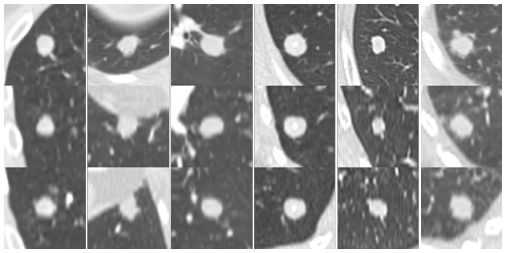
(f) random false negatives at 1 FP/scan

Fig. 3.10 Examples of true positives, false positives, and false negatives from the combined system. Each lesion is located at the center of the 50×50 mm patch in axial, coronal, and sagittal views.

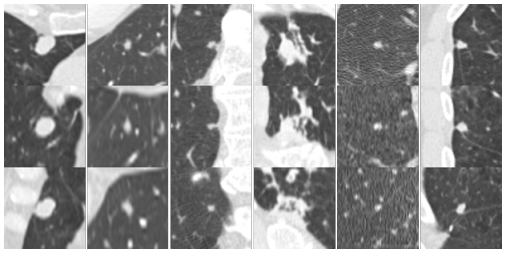
to 92.9%. The low performance of candidate detector systems like SUBSOLIDCAD, LARGECAD with respect to other systems can be explained looking at the nature of these systems. In fact, these systems were built and tuned to detect a particular type of nodules (sub-solid and large solid nodules respectively). Traditional candidate detectors are usually not built to detect particular kinds of nodules, but they use more general techniques. The contribution of mentioned algorithms becomes more clear when looking at the combination of the systems. In fact, adding these particular algorithms to more general ones, increases the global sensitivity. Combining several morphological characteristics from different algorithms allows to capture variations in size and properties of different nodules, with benefits on the detection. The best

combination of the candidate detectors achieves a sensitivity of 98.3 %, around 5% bigger than the sensitivity of the best system in the combination. In the NDET track, five systems were presented. Two of these systems were commercial systems. The sensitivity ranges from 69% to 91% at 1 and 8 false positive per scan. 4 out of 5 systems use traditional imaging techniques. One system introduces the usage of CNNs not only for the false positive reduction stage, but also for the candidate detection part. This system clearly outperforms traditional CAD systems, with a high sensitivity already at a very low false positive rate. These important results demonstrate not only how CNNs are gaining importance as state-of-the-art imaging processing techniques, but opens the possibility to use the system as independent reader in clinical practice. In the FPRED track a total of 4 systems were evaluated. The recent trend of using CNNs as classifiers in the medical imaging field is demonstrated by the fact that all the systems use classifiers based on CNNs. Figure 1b shows a detection sensitivity which ranges from 79.3% to 98.3% at 1 and 8 FP per scans. We could assess that there could be no additional benefits when combining the systems, since all based on CNNs. Results showed in Table 2 totally retracts the previous assessment. Combining multiple CNNs improves the detection performance, as shown by the black curve in Figure 1b. Furthermore, already starting from 2 FP/scan the sensitivity saturates, approaching the maximum achievable sensitivity of 98.3%. A possible explanation to the benefits arising from the combination of conceptually similar systems could be found in the complexity of the systems themselves. Differences in network parameters, such as for example the choice of the architecture or the different methods for the extraction of the patches can produce complementary predictions when combining the systems. To visualize some predictions of the combined system and, moreover, to have an idea of missed nodules and false positives, some screenshots are shown in Figures 3.11-3.13.

When looking at some examples of detected true positives, a wide range of nodules with different morphological characteristics are identified. CNNs are then able to capture and to learn to discriminate different morphological features at a very low FP/scan rate (e.g. 1 FP/scan in the picture). The nodules which are classified as the most suspicious (highest output probability of the classifier) are large solid nodules. Large solid nodules are usually found to be malignant and requires an immediate diagnostic follow-up. For this reason, they do not have to be missed by CAD systems. Conversely, most of false positives detected by CADs are large anatomical structures: large vessels, mediastinal structures, scarring and

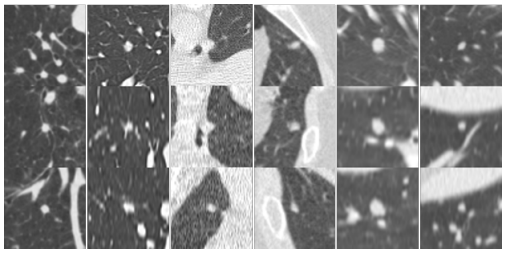


(a) true positives with highest probability

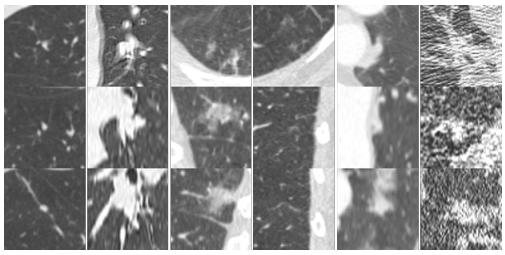


(b) random true positives at 1 FP/scan

Fig. 3.11 Examples of true positives detected by the combined system. Each lesion is located at the center of the 50×50 mm patch in axial view.



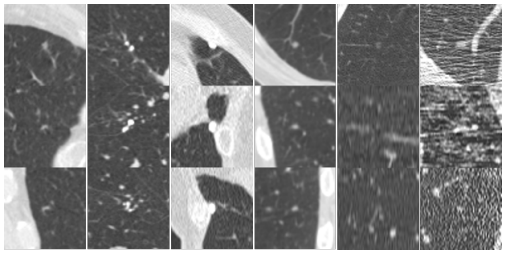
(a) false positives accepted as nodules by radiologists



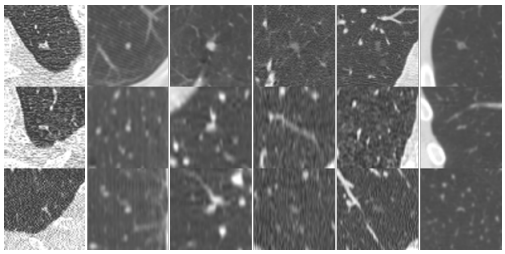
(b) random false positives at 1 FP/scan

Fig. 3.12 Examples of false positives detected by the combined system. Each lesion is located at the center of the 50×50 mm patch in axial view.

spinal abnormalities. It is also very interesting to have a look at missed nodules at 1 FP/scan. They consist of small nodules or nodules presenting an irregular shape. We found out how most of these nodules were missed by candidate detection algorithms.



(a) false negatives missed by the candidate detectors



(b) random false negatives at 1 FP/scan

Fig. 3.13 Examples of nodules missed by the combined system. Each lesion is located at the center of the 50×50 mm patch in axial view.

This outcome motivates the importance of improving candidate detection algorithms. whose sensitivity is an upper bound for a false positive reduction system. We also decided to compare the systems participating in our challenge with the systems which were already developed and partially or totally validated using the LIDC data-set. Table 3.4 summarizes the performances of CAD systems which used the LIDC scans for training or validation. For each CAD system the number of scans in the validation data-set is listed, as well as the nodule inclusion criteria (e.g. if some cuts on the nodule size were applied). It is possible to show, as final results of our challenge, how the combination of classical candidate detectors with state-of-the-art (CNNs) false positive reduction algorithms clearly outperform all the available CAD systems. In addition, these systems also have a positive impact in clinical practice, since they add nodules originally overlooked by radiologists, as shown in the results of our observer study.

3.0.4 Conclusion and Discussion

Several CAD systems have been developed in recent years. Their methods have been vastly published and are available in the literature. Conversely, it is difficult,

Table 3.4 Performance summary of published CAD systems evaluated using LIDC-IDRI data set. Different subsets of scans from LIDC-IDRI data set were used by different research groups over-time. For completeness, number of scans, reference standard criteria, and resulting number of nodules used for evaluation are included in the table. The reported performance at one or two operating points is provided.

CAD systems	Year	# scans	slice thick-ness	nod-ules size (mm)	agree-ment levels	# nod-ules	sensitivity (%) / FPs/scan	
LIDC-IDRI data set								
Combined LUNA16	-	888	≤2.5	≥3	at least 3	1,186	98.2 / 4.0	96.9 / 1.0
[83]	2016	243	-	≥3	at least 1	690	85.9 / 2.5	-
[75]	2016	888	≤2.5	≥3	at least 3	1,186	90.1 / 4.0	85.4 / 1.0
[46]	2015	949	-	≥3	at least 2	1,749	80.0 / 8.0	-
[84]	2015	865	≤2.5	≥3	at least 3	1,147	76.0 / 4.0	73.0 / 1.0
[85]	2014	108	0.5-3	≥4	at least 3	68	75.0 / 2.0	-
[86]	2013	58	0.5-3	3-30	at least 1	151	95.3 / 2.3	-
[87]	2013	360	-	≥3	at least 4	-	83.0 / 4.0	-
[88]	2013	84	0.5-3	5-20	at least 1	103	80.0 / 4.2	-
[89]	2012	84	1.25-3	≥3	at least 1	148	97.0 / 6.1	88.0 / 2.5
[90]	2012	85	1.25-3	≥3	at least 3	111	80.0 / 7.4	75.0 / 2.8
[91]	2011	125	0.75-3	≥3	at least 4	80	87.5 / 4.0	-

based on the available literature, to make a direct and objective comparison of those systems. Since the systems have been trained or validated on different data-sets (e.g. of different sizes) a direct comparison would not be an objective evaluation. Only a large scale validation of CAD systems on a common data-set allows an objective comparison. In addition, new techniques (e.g. CNNs) originally used for machine learning problems outside the field of medical imaging have showing better performances when compared to traditional classifiers (e.g. neural networks). Conversely, these algorithms have only started being applied recently for the automated detection of pulmonary nodules in chest CT scans. As a further point, preliminary results showed potential benefits on the overall detection sensitivity when combining different CAD systems. However no detailed studies investigating the combination of classical candidate detector systems with more recent state-of-the-art false positive reduction systems have been performed. We organized LUNA16, a medical challenge for a large scale evaluation of state-of-the-art nodule detection algorithms using the biggest public available data-set (LIDC/IDRI). In addition, a novel web-based framework was developed to allow participants to the challenge to evaluate their algorithms. We showed how the combination of classical candidate detector systems

with multiple deep learning architecture outperforms exiting and published CAD algorithms. We also provided a new reference standard for the LIDC/IDRI data-set, including nodules found by the CAD and originally overlooked by radiologists. It is worth noticing some limitations and points of improvement of the presented study. First of all the LIDC-IDRI is a public available data-set, including the images and the reference standard. For this reason we had to train participants to perform a cross-validation. This moves a bit apart from the traditional setup of a challenge where an independent test set is provided. References on this test set are usually not available. In this view, we are planning to add additional data-sets for our challenge (e.g. NLST data-set) and split them into training, validation and test set. A next step can be to ask participants to submit their trained algorithms which then will be running on a Cloud platform on test data. This solution is usually referred to as *dock containers*. From the clinical point of view we believe that the next challenge for CAD algorithms will be their capability to discriminate between malignant and benign lesions. Having proved that the combination of several state-of-the-art algorithms allows to achieve excellent performances in the detection task, we could ask participants to tune their algorithms to provide a degree of malignancy associated to the findings. The organization of such a challenge represents a very demanding task. First of all, the reference standard will have to include pathological information about the origin of the nodules. Meanwhile, the LUNA16 challenge will remain open for additional submissions and possible additional data-sets could be included.

Chapter 4

Clinical validation of the M5L on-demand CAD

4.1 Motivation

In Chapter 1 we have described the importance of CAD systems as support for radiologists in the detection process. In Chapter 2 we have presented a new paradigm for introducing those systems in clinical practice: making CAD results available without asking to clinical facilities any additional installation of hardware and software. In addition, we have pointed out how screening high-risk subjects with low-dose CT reduced lung cancer mortality [10]. We also mentioned how several studies proved the positive impact of CAD systems on the radiologists performance. Conversely, we also alerted how CAD systems have not spread and / or not used constantly in clinical practice so far. It is worth inspecting the causes of this discrepancy, with a dedicated analysis of the clinical requirements which a CAD system should fulfill to be used as daily support in clinical routine. First of all, it is fundamental to notice how the majority of the studies have evaluated CADs using datasets coming from screening campaigns. Conversely, lung cancer screening campaigns do not represent the primary source of chest CT scans acquired in a hospital. In fact, clinical studies usually investigate the appearance of a pulmonary nodule as first sign of a metastatic tumor. This diagnostic analysis is usually performed in oncological patients with an extra-thoracic cancer. The early detection of pulmonary nodules (combined with an adequate follow-up) can really improve the survival of the patients. A study [92]

showed how pulmonary metastases could be resected with an improvement up to 30-35% of the five year survival rate. CT seems to be the right diagnostic modality also for the detection of pulmonary metastases. A famous work [93] demonstrated how CT is the optimal or most sensitive screening study in patients with a suspected presence of pulmonary metastases. In particular, CT was able to detect contralateral nodules, or more in general nodules which could be seen with a traditional X-ray tomography only at later time, when they would have already grown and become untreatable. Previous results open the possibility to apply CAD systems to support the detection of pulmonary metastases. In literature, there have been some studies aiming at investigating the best approach to insert CAD in clinical practice. These studies were mainly focused on comparing CAD as concurrent reader or CAD as second reader. A recent work [94] compared the sensitivity of detection of pulmonary nodules and reading time of radiologists when using the CAD as concurrent reader mode or second reader mode. They showed that the concurrent reader approach presents the advantage to significantly reduce the reading time when compared to second reader approach. Conversely, the overall sensitivity was found to be larger in the second-reader mode with respect to the concurrent reader mode. Another study [95] confirmed the results regarding the reading time, but reached different results for the sensitivities. In fact, concurrent reader and second reader mode were found to have the same sensitivities. When looking at these studies it is possible to find some limitations and points of improvement. First of all these studies share a limited data-set. In [94], for example, the observer study was performed using only 50 scans. In addition, the study was limited to solid nodules, excluding the other typologies of nodules. Furthermore, all the studies are retrospective, usually on datasets collected for screening purposes. Due to the nature of being retrospective, the authors can not exclude *memory effects*. In fact, the scans were annotated by the same radiologists twice, with a small interval of time between the two reading phases. All these reasons motivated us to set up an observer study to investigate the impact of CAD in the radiologist performances. We decided to built a prospective observer study on a clinical dataset of oncological patients. The second aim of the study was to clinically validate the M5L on-demand system, since we believe that only a direct usage in clinical practice could give us the possibility to test all the functionalities of the developed system, and build new ones if requested by clinicians. Finally, this observer study has the appeal to collect a dataset of annotated clinical data to be used for further clinical investigations. In particular, a study [96] compared

the appearance of pulmonary metastases as found on CT scans with pathological findings in lung specimens obtained during the autopsy. The research underlines how the nature of a metastatic tumor mainly manifests itself by a growing margin. The properties of the margin seem to have a correlation with pathological characteristics of the tumor. The authors stated that the analysis should be verified with additional datasets and including several characteristics of the nodules. Our observer study opens the possibility to improve studies like the cited one, since for each finding we collect information about different properties of the nodules.

4.2 Material and Methods

The study has been set up in collaboration with the Institute for Cancer Research and Care (IRCCS) in Candiolo, Italy. The Candiolo Cancer Institute is a private non-profit institution founded and supported by the Fondazione Piemontese per la Ricerca sul Cancro-Onlus (FPRC) and operated by the Fondazione del Piemonte per l' Oncologia (FPO: co-founded by FPRC and the Regione Piemonte). The Institute is a recognized IRCCS (Istituto di Ricovero e Cura a Carattere Scientifico) by the Italian Ministry of Health. Its mission is a significant contribution to fight cancer, by understanding the basics, and by providing state-of-the-art diagnostic and therapeutic services. This study was a single-centre cross sectional study. Each participant underwent chest CT clinical examinations. Local institution board approval was obtained to publish the data. This study was a direct collaboration with the Medical Physics Department coordinated by Dr. Michele Stasi and the Radiology Department coordinated by Prof. Daniele Regge.

4.2.1 CT Protocol

Two different kinds of examinations were used for the patients in our study: CT with or without contrast enhancement. Different clinical protocols are prescribed for the preparation of the patient undergoing a scan with contrast enhancement. In particular, in our study, if the patients referred some allergies in the three days before the exam, they were prescribed the following therapy: Omeprazole, Deltacorte, Zirtec. The day of the scan, the patients are delivered the following additional therapy: Trimeton, Flebocortid. The images were acquired only if the patient was

accompanied by a person, fasting and with blood examination for creatinine not older than 6 days. Before underlying the differences in the acquisition parameters, it is worth to describe some of the main parameters defining the properties of the acquisition protocols used in our study:

- Tube potential: the electric potential applied across the X-ray tube in order to accelerate electrons to the target material. It is expressed in unit of kilovolts (kV);
- Tube current-time product: the product of the current of the tube and the exposure time per rotation. It is expressed in units of milliampere x seconds (mAs). In helical scan mode, it is equal to the current of the tube x rotation time;
- pitch: unit-less parameter which is used to describe the table travel during helical CT. It is equal to the table travel (mm) per gantry rotation divided by the total nominal beam width (in mm);
- acquisition time: it corresponds to the time used to acquire a single signal for each rotation. It is measured in seconds;
- configuration of the detectors: on multi-channel CT scanners, there are multiple data channels that route the signals collected on the detector surface to the reconstruction computer. These data channels can be utilized in a flexible manner to produce more than one thin image during one rotation of the gantry, or fewer thicker images during one rotation of the gantry. The combination of the number of active channels and the detector width (in z-direction) assigned to each channel is called the detector configuration and it is measured in mm;
- reconstruction algorithm: represents the algorithm used by the software that, starting from the single raw projections, produces the final image.

Table 4.1 shows the summary of the main characteristics of different protocols used for the acquisition of the scans in our study. The first protocol is called *Basal CT* and it does not include the usage of any contrast enhancement technique. The second protocol is called *Contrast-Enhanced (CE) CT*. Both protocols make use of the same reconstruction algorithm: Filtered-Back Projection (FBP), the most common approach used in clinical practice so far. Before briefly explaining the

CT Protocol	**Thorax-Basal CT**	**CE-CT**
Tube Potential (kV)	120	$140-100$
Tube current-time product (mAs)	130	139
Slice Thickness (mm)	3	3
Reconstruction Algorithm	Filtered Back Projection	Filtered Back Projection
Acquisition Time (s)	0.5	0.5
Pitch	1.2	0.9
Detector Configuration (mm)	32x1	64x0.6

Table 4.1 Summary of the main parameters corresponding to the acquisition protocols used in our study. Thorax-basal uses only one value for the electric potential applied across the x-ray tube. CE-CT is a Dual Energy CT (DECT) making use of two different values of the tube potential in order to acquire a combine different X-ray spectra.

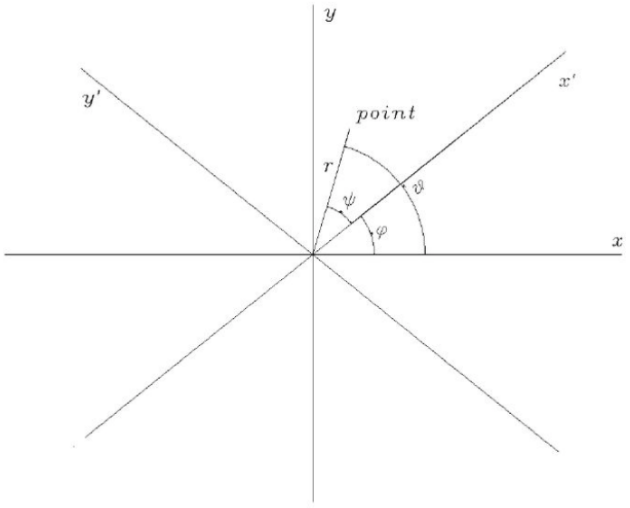


Fig. 4.1 View of the two different coordinate systems: patient coordinates (x,y) and CT coordinates $(x^{'},y^{'})$

theory underlying this algorithm, it is fundamental to define two different systems of coordinates: the patient coordinates system and the CT coordinate system. At the beginning it is assumed that the patient stays still, so it is assigned to the Cartesian coordinate system (x,y). There is also the CT system of coordinates $(x^{'}, y')$ which can rotate around the patient at an angle ϕ. Figure 4.1 gives a schematic view of those systems. After having acquired a sufficient number of projections at different acquisition angles, it is required to fit each projection to its corresponding position in order to obtain a reconstructed image. The algorithm makes use of the so-called Radon transformation. We can define the attenuation of the X-rays at a particular

position $x^{'}$ taken from a projection at angle ϕ as:

$$p(x^{'},\phi) = \log(\frac{I}{I_0}) = -\int \mu(x^{'},y^{'})dy^{'}$$

where μ is the absorption coefficient, I is the intensity of the attenuated ray and I_0 is the initial original intensity. A pictorial view of the Radon transform is shown in Figure 4.2. It it possible to rewrite the previous equation as:

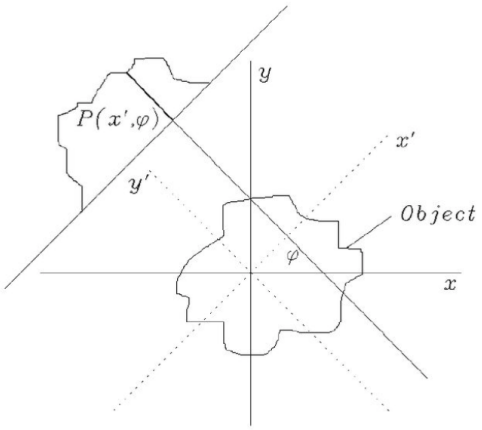


Fig. 4.2 Pictorial view of the Radon transform.

$$p(x^{'},\phi) = \int_{-\inf}^{+\inf}\int_{-\inf}^{+\inf} \mu(x,y)\delta(x\cos(\phi)+y\sin(\phi)-x^{'})dxdy$$

$$= \int_{-\inf}^{+\inf} \mu(x^{'}\cos(\phi)-y^{'}\cos(\phi), x^{'}\sin\phi + y^{'}\cos(\phi))dy'$$

where δ is the point impulse. This last equation is called Radon transform and it represents the integral of each projection all over the angles. Using the Radon transform, the back projection can be defined as:

$$B(x,y) = \int_0^{\pi} p(x^{'},\phi)d\phi$$

In a more computational way the algorithm can be described as shown in Figure 4.3. Coming back to our description, the two protocols share this illustrated reconstruction algorithm, but they adopt two different approaches. In fact, CE CT falls into the category of Dual Energy CT (DECT). In this protocol two CT datasets are acquired

Algorithm 1 Filtered Back-projection algorithm

```
1: for all projection angles do
2:     filter all projection data in the spatial domain
3:     relate CT coordinate system to the patient coordinate system
4:     back-project the corresponding projections
5: end for
```

Fig. 4.3 Schematic view of the steps composing the algorithm of filtered back projection.

using different X-ray spectra. The spectra result different because they are generated using different voltages of the X-ray tube. After the acquisition of the images, all the CT scans are stored in the hospital PACS on an optical support using the DICOM-3 standard. All the scans used in the study are also anonymized before being uploaded for CAD analysis.

4.2.2 Image Interpretation

Two experienced faculty radiologists (range 20-35 years of experience) and one young radiologist (training as resident radiologist, 2 years of experience) took part in our study. All the radiologists work in the Radiology Department of IRCCS Candiolo. First of all, we trained the radiologists on how to use the M5L CAD system before the beginning of the study using 5 test cases for a total of 4 nodules. Each exam is submitted to the M5L front-end as soon as acquired, anonymized and stored in the hospital PACS. To make the procedure faster we allowed the submission of bunches of 10 cases per upload, which were processed using a batch routine every evening. Immediately after the submission, the three radiologists receive an e-mail with the direct link to annotate the case. Since the study was performed with the CAD as second reader, the radiologists first annotate the cases without having access to CAD results. Medical annotations are inserted as usually done in clinical practice: radiologists use the post-processing technique of the Maximum Intensity Projection (MIP). This technique projects all the voxels with the highest attenuation value on every view. Considering each XY coordinate, only the pixel with the highest HU along the z axis is represented. This allows to show in a single bidimensional image all the dense structures in a given volume. This technique is mostly used to detect pulmonary nodules. An example of an axial view with MIP of a chest CT scan in shown in Figure 4.4. For each lesion, the radiologists manually segment the identified lesions and then fill in the on-line report. All the characteristics of the annotation web form were described in Chapter 2. In order to reduce the

Fig. 4.4 Axial view of a chest CT scan after applying MIP. Denser structures appear as static when scrolling through slices.

variability between radiologists in the annotation protocol,during the training session each category present in the report was explained with dedicated examples. Special attention has been dedicated to the malignancy score of a finding. The radiologists were trained to assign the malignancy score only based on visual assessment of the properties of the nodule without considering any additional clinical information on the patient as support for the evaluation. It is worth to remind the definition of the malignancy scores. The degree of malignancy is defined as: subjective assessment of the likelihood of malignancy, assuming that the scan originated from a 60-year-old male smoker. The score goes from 0 (Highly Unlikely to be malignant) to 4 (Highly suspicious to be malignant) as defined in [50]. The dimension of the nodule is reported (in mm) as the measure of the major diameter on the 2D axial images using a window level for the lung parenchyma with depth 1360 and level 530. The main motivation underlying the idea of asking the radiologist to fill such a detailed annotation form was to collect a clinical dataset of annotated scans in order to use it for additional investigations. Finally, in our study, to be consistent with available guidelines in clinical practice [17], we asked the radiologists to annotate only nodules with a diameter equal or larger than 3 mm and discard smaller ones.

After the radiologist has completed the insertion of the ROIs, the annotation is stored and no additional changes can be performed.

4.2.3 Double reading with CAD

As already mentioned in the beginning, the goal of our study is to investigate the impact of CAD as second reader in the overall sensitivity. For this reason, CAD results are available for review only when the first unassisted reading has been completed and validated. As soon as the radiologists validate the unassissted annotations, an e-mail containing the link to CAD results is automatically sent and the review of CAD results becomes available. In order to reduce the review time, CAD marks are automatically compared by our system to the pathological ROIs in each annotation. A matching algorithm associates two findings when the 3D Euclidean distance between them is smaller than the mean diameter. This procedure allows to directly associate the findings of the CAD which match with the annotated findings. In this case, the CAD finding takes the same properties (e.g, malignancy score) of the corresponding annotated finding. This automated procedure allows radiologists to save time and focus on overlooked findings. The list of unmatched finding must be reviewed by the radiologists. Using again the web-form, the radiologist can mark a finding as: False Positive, Irrelevant or True Positive. In the irrelevant findings are included the definition in [45] and also all the nodules smaller than 3mm or with a malignancy score smaller than 2. For each CAD mark the radiologist has the possibility to display a zoomed view of the finding and the overall axial view of the slice corresponding to the findings. For each finding marked as True Positive, the radiologist is asked to specify its malignancy score. Findings marked as TP represent nodules missed by radiologists in the original annotation. They are automatically added to the original annotation, as soon as the review process has been completed. A comparison between the first unassisted reading and the second assisted reading gives an indication of the CAD contribution of CAD to the overall detection sensitivity.

4.2.4 Reading Time

Reading time for unassisted reading and CAD assisted read (sum of the reading time without CAD and the following time for reviewing the CAD results) were recorded for each case and each reader. Reading time was documented by a person without

professional experience in radiology (the author of this thesis) and not involved in the reading process.

4.2.5 Reference Standard

To build a robust reference standard [38] all the nodules equal or larger than 3 mm annotated by at least one radiologist and all the CAD findings marked as true positives by at least one radiologist have been evaluated again by a pool of two faculty radiologists with more than 10 years of experience using the Osirix plugin presented in Chapter 2. They could visualize the finding (marked with a circle) directly on the CT scans with the possibility to scroll the slices. Additional measurements tool were available to verify and assess the dimension of the finding. The panel members, without knowledge of the source of detection of the nodule candidates, assessed each candidate and arrived at a final decision in consensus as to whether it was a nodule (TP finding) or not (irrelevant finding). The assessment (malignant / benign) of the nodules in the reference standard, monitoring them within at least a 9-month follow-up from the first detection, was investigated by looking at results of available follow-up examinations and / or at additional medical reports (eg. PET-CT results) of the patient. In order to perform a sub-analysis regarding the spatial location of annotated lesions, the findings in the reference standard were classified as: central lesions those lesions next to a bronchus or blood vessel, as peripheral lesions at maximum of 2cm from pleural surface and as intermediate the remaining ones [97].

4.2.6 Statistical Analysis

Per-lesion specificity and sensitivity diagnoses given by each reader, and the average of the three readers, with and without CAD were compared by using the McNemar test [98]. Sensitivity, specificity, and the 95% confidence intervals (CIs) of each were calculated for each reader, for the sum of the three readers and for each reading mode. Reviewing times were compared by using the paired t test. Statistical analyses were performed using software R (The R Foundation for Statistical Computing) and Microsoft Excel 2010. All statistical tests were 2-sided and were considered statistically significant at $P \leq 0.05$. All Confidence Intervals were reported at the 95% level.

4.2.7 Sample Size Estimation

The stand-alone sensitivity of CAD in detecting pulmonary nodules equal or larger than 3 mm in diameter was found to be 80% with a specificity of 4 False Positive findings per scan in a previous validation work [46]. A power calculation (at 5% significance and 80% power) performed on the basis of this estimate suggests that at least 180 patients with a nodule equal or larger than 3 mm in diameter are required to detect a 10% difference in the detection between unassisted and assisted reading. Since, from a preliminary study, the prevalence of target cases was found to be 80% at least 216 patients must be included in the study. The sample size was increased to 237 cases to increase the statistical significance.

4.3 Results

4.3.1 Study Population

Of the 237 patients, 12 were excluded from the analysis for the following reasons: severe pulmonary fibrosis (n =2), diffuse bronchiectasis (n =1), pneumonia (n =6) and massive pleural effusion (n=3). The media age of the population was 60 years (range, 21-90 years) and 108 of 225 (48%) were women. There was no age difference between men and women (P-value < 0.01). The large majority of patients underwent CE-CT (88%). Fifty-eight of the 225 (26%) patients had at least one metastases and the average number of metastases per positive patient was 4 (range 2-6). The most common sites of primaries were the following: colon (54/225, 24%), sarcomas (25/225, 11%), melanomas (22/225, 10%), and breast (21/225, 9%). The percentage of malignant nodules was found to be 30% (64/215). The median diameter of lesions was 8.5 mm (range 4-28 mm). The majority of nodules had a diameter between 3 and 5 mm (77%, 166/215) followed by nodules between 6 and 9 mm (19% 40/215) and by those larger than 9mm (4%, 9/215). Most of the nodules were located at the periphery of the lungs (62%, 133/215), followed by the intermediate area (27% 59/215) and only few nodules (11%, 23/215) were located in the central region. In regard of the tissue characteristics most of the nodules were solid (67%, 143/215), followed by part-solid (15%, 33/215), sub-solid (12%, 25/215) and calcified (6%, 14/215).

	Reader 1		Reader 2		Reader 3		Average of all readers	
	Sensitivity (%)	95% CI	Sensitivity (%)	95% CI	Sensitivity (%)	95% CI	Sensitivity (%)	95% CI
Size								
3 - 5 mm	63% (105/166)	55,71	67% (110/166)	58, 73	54% (89/166)	46,61	61% (304/498)	57,65
6 - 9 mm	70% (28/40)	53,83	83% (31/40)	62,89	83% (33/40)	67,93	77% (92/120)	68,84
>9 mm	93% (8/9)	52,99	100% (9/9)	66,99	93% (8/9)	52,99	93% (25/27)	76,99
Location								
Central	52% (12/23)	31,73	65% (15/23)	43,84	52% (12/23)	31,73	57% (39/69)	44,68
Intermediate	61% (36/59)	47,73	68% (40/59)	54,79	61% (36/59)	47,73	63% (112/177)	56,70
Sub Pleural	70% (93/133)	61,78	71% (95/133)	63,79	62% (82/133)	53, 70	68% (270/399)	63,72
Tissue Type								
Solid	66% (94/143)	57,73	67% (96/143)	59,75	62% (89/143)	54,70	65% (279/429)	60,70
Part Solid	61% (20/33)	42,77	73% (24/33)	54,87	55% (18/33)	36, 72	63% (62/99)	52,72
Sub Solid	60% (15/25)	39,79	72% (18/25)	50,88	52% (13/25)	31,72	61% (46/75)	49,72
Calcified	86% (12/14)	57,98	86% (12/14)	57,98	71% (10/14)	42,92	81% (34/42)	66,91
Total	65% (141/215)	59,72	70% (150/215)	63,76	60% (130/215)	54,67	65% (421/645)	61,69

Table 4.2 Per-lesion sensitivity results for all the readers and the average of all the readers for the unassisted reading.

4.3.2 Stand-Alone CAD performance

Results of per-lesion analysis are shown in Table 4.2 and Table 4.3 for the unassisted and double reading respectively. Mean sensitivity of unassisted reading was 65% (421/645, CI: 61%,69%) for nodules greater than 3mm. Mean sensitivity of CAD assisted reading for the same nodules was 88% (570/645, CI: 86%,91%). There was a statistically significant difference (88% vs 65%, P<0.01) between unassisted and assisted CAD reading. Mean sensitivity for nodules between 3-5mm was 61% (304/498, CI: 57%,65%) and 91% (451/498, CI: 88%,93%) for respectively unassisted and double reading paradigm. There was a statistically significant difference (91% vs 61%, P<0.01) between the two reading modalities. Mean sensitivity for nodules located within the central region was 57% (39/69, CI: 44%,68%) and 89% (61/69, CI: 78%,95%) for unassisted and double reading protocol respectively. There was a statistically significant difference (89% vs 57%, P=0.02) between the two reading modalities. A statistically significance difference (61% vs 91%, P=0.02) was also found between unassisted and double reading modalities for sub-solid nodules. To investigate the relation between the sensitivities and the malignancy score, we have built the ROC curves for all the readers for both unassisted and second reading. On the x axis the values of the malignancy score are reported, while on the y axis there are the sensitivities. Each point on the ROC curve is built considering all the nodules in the reference standard with a malignancy score greater than the corresponding value on the x axis. Results are shown in Figure 4.5. The unassisted sensitivity

	Reader 1		Reader 2		Reader 3		Average of all readers			
	Sensitivity (%)	95% CI	Sensitivity (%)	95% CI	Sensitivity (%)	95% CI	Sensitivity (%)	95% CI	Delta (*)	P-value
Size										
3 - 5 mm	89% (148/166)	83,93	96% (160/166)	92, 99	86% (143/166)	80,91	91% (451/498)	88,93	+30%	<0.01
6 - 9 mm	82% (33/40)	67,93	85% (34/40)	70,94	88% (35/40)	73,96	85% (102/120)	77,91	+8%	0.8
>9 mm	100% (9/9)	66,100	100% (9/9)	66,100	100% (9/9)	66,100	100% (27/27)	87,100	+7%	0.2
Location										
Central	87% (20/23)	66,97	91% (21/23)	72,99	87% (20/23)	66,97	89% (61/69)	78,95	+32%	0.02
Intermediate	89% (53/59)	79,96	92% (54/59)	81,97	86% (51/59)	75,94	89% (158/177)	84,93	+26%	0.04
Sub Pleural	88% (117/133)	81,93	89% (118/133)	82,94	87% (116/133)	80,92	88% (351/399)	84,91	+20%	0.03
Tissue Type										
Solid	89% (126/143)	82,93	89% (127/143)	82,93	87% (125/143)	81,92	88% (378/429)	85,91	+23%	<0.01
Part Solid	87% (29/33)	72,97	91% (30/33)	76,98	82% (27/33)	65,93	87% (86/99)	79,93	+24%	0.03
Sub Solid	88% (22/25)	69,97	92% (23/25)	74,99	92% (23/25)	74,99	91% (68/75)	82,96	+30%	0.02
Calcified	93% (13/14)	66,99	93% (13/14)	66,99	86% (12/14)	57,98	90% (38/42)	66,91	+9%	0.8
Total	89% (190/215)	83,92	90% (193/215)	85,93	87% (187/215)	82,91	88% (570/645)	86,91	+23%	<0.01

Table 4.3 Per-lesion sensitivity results for all the readers and the average of all the readers for the double reading reading. * Delta = (average sensitivities of all readers with Double reading - average sensitivities of all readers unassisted Reading) for each category in the table.

Readers	**Shortest Measured Reading Time (time $\pm$ SD**	**Longest Measured Reading Time (time $\pm$ SD)**
Reader 1 Unasissted	(290 ± 83) s	(305 ± 90) s
Reader 2 Unasissted	(300 ± 90) s	(310 ± 90) s
Reader 3 Unasissted	(290 ± 83) s	(295 ± 70) s
Overall Average Unassisted Reading	(293 ± 85) **s**	(303 ± 85) **s**
Reader 1 + CAD assisted	(319 ± 100) s	(330 ± 95) s
Reader 2 + CAD assisted	(330 ± 70) s	(341 ± 70) s
Reader 3 + CAD assisted	(319 ± 70) s	(330 ± 70) s
Overall Average Double Reading	(323 ± 82) **s**	(334 ± 80) **s**

Table 4.4 Reading time performances per individual readers

grows up choosing a larger malignancy score as cut off point in the ROC curves, Table 4.4 presents the results of individual reading time for all the readers for the two different reading paradigms. Using CAD as second reader, Radiologist 1 experienced a maximum difference in reading time of +40 s and a minimum difference of +14s compared to the unassisted reading time. Radiologist 2 experienced a maximum difference in reading time of +41 s and a minimum difference of +20s compared to the unassisted reading time. Radiologist 3 experienced a maximum difference in reading time of +40 s and a minimum difference of +29s compared to the unassisted reading time. Pooling the individual reading times, the mean reading time of CAD assisted reading and unassisted was found to be 330s (CI: 319,341) and 300s (CI: 290,310), respectively (P value < 0.05).

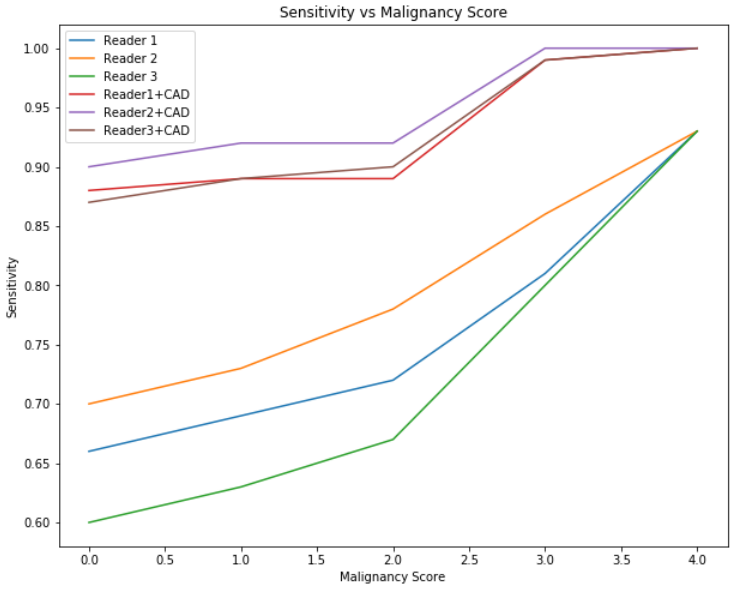


Fig. 4.5 ROC curves divided per malignancy score for each reader for both unassisted and assisted reading modalities

4.3.3 False positive, True positive and False negative analysis

The majority of false positive findings could be attributed to the following categories: rib-vertebral joints and superposition of different vascular structures. The majority of CAD false negatives were Ground Glass Opacities (GGOs) presenting a poorly marked structure with an undefined margin and usually totally embedded into the lung parenchyma or nodules completely attached to a vascular structure. Radiologist Overlooked findings were quite often sub-solid nodules usually smaller than 5 mm in diameters placed next to the pulmonary hilum and over the diaphragm.

4.4 Conclusion and Discussion

According to our knowledge only few CAD systems have been tested specifically in patients with extra-thoracic cancer to detect lung metastases, but in small sample of less than a hundred nodules, since the better part of the works are usually based on screening databases. In our study, per-lesion sensitivity of unassisted interpretation and that of double reading interpretation were founded to be significantly different. In fact, double reading protocol increased radiology stand-alone sensitivity from

65 to 88%, much larger than average radiologists' sensitivity [99]. This finding is consistent with results in previous studies [94] [95], where second-reader CAD was used on a smaller data-sets of only solid nodules. In addition, second reading acts as detection rates equalisers between observers of different level of experience, as also pointed out in [12]. CAD stand-alone sensitivity was in some cases found to be greater than second reading sensitivity, meaning that radiologists sometimes marked as irrelevant of false positive findings correctly detected by the automated system. With our study, we significantly increased the sample (both on size and nodules type) eliminating possible memory effects. Regarding the dimensions, the major contribution of CAD was observed for small nodules (3-5mm), while it was less marked or quite absent for larger ones (>6mm). As expected sub-solid nodules were the more difficult to be detected due to poor density difference with respect to lung parenchyma. In fact, the readers in our study had the lowest detection sensitivity for this type of nodules which remarkably increased after assisted reading (91% vs 61%). When looking at the localization most of the lung nodules in our study population showed a peripheral lung distribution as reported in literature (Seo, et al. 2001) and had smooth margins. The main contribution of the automated system regarded central lesions (89% vs 57% of the unassisted reading) which were more likely to be overlooked by radiologists' because they are attached to vascular or bronchial structures which can totally embed the finding. Despite the malignancy score could be a reliable indicator to discriminate between malignant and benign nodules, since based on a qualitative assessment of the visual properties of the finding, it can be considered as a confidence score. In fact, radiologists' sensitivity grew up when considering a higher malignancy score as cut-off point. The role of the CAD was to increase the number of nodules with lower malignancy scores.

The CAD stand-alone performance was found to be larger than the sensitivity obtained in our previous validation on the LIDC-IDRI dataset (Lopez Torres, et al. 2015). This result proved the capability of the system to achieve good performances also on external datasets, even without an optimization of the algorithms on mentioned databases. Further developments on the algorithms could increase the detection sensitivity and specificity. The sensitivity can be improved by adding to our algorithms dedicated features for the detection of sub-solid nodules, as for example suggested in (Jacobs, et al. 2011). The specificity can be improved by retraining the classifier in order to reduce the most common type of False Positive (rib-vertebral joints and superposition of different vascular structures). As expected,

the second reading with CAD decreased time-efficiency in reviewing time. To tackle this issue, the automated matching system between CAD findings and radiologists findings allowed to reduce the second-reading time, which was found to be only 10% larger than the reading time without CAD. These results are much lower (P-value < 0.05) than the values available in the literature [94] [95]. Finally, the frequency of pulmonary metastases was found to be 25.8% which is compatible to the value (30%) found in the literature. The average number of nodules per metastatic patients was 4. The collected data-sets could allow additional clinical investigations to relate the nodule visual features and the malignancy. It is worth noticing some limitations of our study. First of all, since not all the types of cancer are treated in the structure from which data were acquired, the population does not include all the type of possible cancers. Secondly, the number of readers could be increased (not necessarily belonging to the same institution) in order to build a more robust reference standard. In addition, it is necessary to perform an additional observer study aiming at investigating possible differences (sensitivity / reading time) between concurrent reader and second reader approaches in order to find the most suitable protocol to insert CAD in clinical practice. In conclusion, we proposed and clinically validated our Web- and Cloud-based system for the automated detection of pulmonary metastases. This CAD represents a cost-effective solution for clinical facilities because it allows both inserting medical reports and accessing / reviewing CAD results just using a web browser, while a separated cloud back-end is taking care of CAD computations. Finally, this study showed the positive impact of CAD as second reader opening the possibility to use automated system as clinical decision support.

References

[1] M Malvezzi, P Bertuccio, T Rosso, M Rota, F Levi, C La Vecchia, and E Negri. European cancer mortality predictions for the year 2015: does lung cancer have the highest death rate in eu women. *Ann Oncol*, 26(4):779–786, 2015.

[2] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29, 2015.

[3] Michael A Beckles, Stephen G Spiro, Gene L Colice, and Robin M Rudd. Initial evaluation of the patient with lung cancer: symptoms, signs, laboratory tests, and paraneoplastic syndromes. *Chest Journal*, 123(1_suppl):97S–104S, 2003.

[4] Jiang Hsieh. Computed tomography: principles, design, artifacts, and recent advances. SPIE Bellingham, WA, 2009.

[5] Hui Hu. Multi-slice helical ct: scan and reconstruction. *Medical physics*, 26(1):5–18, 1999.

[6] Alvin C Silva, Holly J Lawder, Amy Hara, Jennifer Kujak, and William Pavlicek. Innovations in ct dose reduction strategy: application of the adaptive statistical iterative reconstruction algorithm. *American Journal of Roentgenology*, 194(1):191–199, 2010.

[7] Xiaochuan Pan, Emil Y Sidky, and Michael Vannier. Why do commercial ct scanners still employ traditional, filtered back-projection for image reconstruction? *Inverse problems*, 25(12):123009, 2009.

[8] Moulay A Meziane, Ralph H Hruban, EA Zerhouni, Paul S Wheeler, Nagi F Khouri, Elliot K Fishman, Grover M Hutchins, and Stanley S Siegelman. High resolution ct of the lung parenchyma with pathologic correlation. *Radiographics*, 8(1):27–54, 1988.

[9] Geoffrey D Rubin. Data explosion: the challenge of multidetector-row ct. *European journal of radiology*, 36(2):74–80, 2000.

[10] National Lung Screening Trial Research Team et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*, 2011(365):395–409, 2011.

[11] Robert A Smith, Deana Manassaram-Baptiste, Durado Brooks, Mary Doroshenk, Stacey Fedewa, Debbie Saslow, Otis W Brawley, and Richard Wender. Cancer screening in the united states, 2015: a review of current american cancer society guidelines and current issues in cancer screening. *CA: a cancer journal for clinicians*, 65(1):30–54, 2015.

[12] Matthew S Brown, Jonathan G Goldin, Sarah Rogers, Hyun J Kim, Robert D Suh, Michael F McNitt-Gray, Sumit K Shah, Dao Truong, Kathleen Brown, James W Sayre, et al. Computer-aided lung nodule detection in ct: Results of large-scale observer test1. *Academic radiology*, 12(6):681–686, 2005.

[13] Claudia I Henschke, David F Yankelevitz, Rosna Mirtcheva, Georgeann McGuinness, Dorothy McCauley, and Olli S Miettinen. Ct screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. *American Journal of Roentgenology*, 178(5):1053–1057, 2002.

[14] Sanjiv Sam Gambhir. Molecular imaging of cancer with positron emission tomography. *Nature Reviews Cancer*, 2(9):683–693, 2002.

[15] William D Travis, Elisabeth Brambilla, Masayuki Noguchi, Andrew G Nicholson, Kim R Geisinger, Yasushi Yatabe, David G Beer, Charles A Powell, Gregory J Riely, Paul E Van Schil, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *Journal of Thoracic Oncology*, 6(2):244–285, 2011.

[16] CT Lung. Screening reporting and data system (lung-rads). *ACR. Available online: http://www. acr. org/Quality-Safety/Resources/LungRADS*, 2014.

[17] Heber MacMahon, John HM Austin, Gordon Gamsu, Christian J Herold, James R Jett, David P Naidich, Edward F Patz Jr, and Stephen J Swensen. Guidelines for management of small pulmonary nodules detected on ct scans: a statement from the fleischner society 1. *Radiology*, 237(2):395–400, 2005.

[18] MH Nathan, VP Collins, and RA Adams. Differentiation of benign and malignant pulmonary nodules by growth rate 1. *Radiology*, 79(2):221–232, 1962.

[19] Samuel G Armato and William F Sensakovic. Automated lung segmentation for thoracic ct: Impact on computer-aided diagnosis1. *Academic Radiology*, 11(9):1011–1021, 2004.

[20] Matthew S Brown, Michael F Mcnitt-Gray, Nicholas J Mankovich, Jonathan G Goldin, John Hiller, Laurence S Wilson, and DR Aberie. Method for segmenting chest ct image data using an anatomical model: preliminary results. *IEEE transactions on medical imaging*, 16(6):828–839, 1997.

[21] Shiying Hu, Eric A Hoffman, and Joseph M Reinhardt. Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images. *IEEE transactions on medical imaging*, 20(6):490–498, 2001.

[22] Samuel G Armato, Maryellen L Giger, Catherine J Moran, James T Blackburn, Kunio Doi, and Heber MacMahon. Computerized detection of pulmonary nodules on ct scans 1. *Radiographics*, 19(5):1303–1311, 1999.

[23] Jane P Ko and Margrit Betke. Chest ct: Automated nodule detection and assessment of change over time—preliminary experience 1. *Radiology*, 218(1):267–273, 2001.

[24] Binsheng Zhao. Automatic detection of small lung nodules on ct utilizing a local density maximum algorithm. *journal of applied clinical medical physics*, 4(3):248–260, 2003.

[25] Toshiharu Ezoe, Hotaka Takizawa, Shinji Yamamoto, Akinobu Shimizu, Tohru Matsumoto, Yukio Tateno, Takeshi Iimura, and Mitsuomi Matsumoto. Automatic detection method of lung cancers including ground-glass opacities from chest x-ray ct images. In *Medical Imaging 2002*, pages 1672–1680. International Society for Optics and Photonics, 2002.

[26] Catalin I Fetita, Francoise Preteux, Catherine Beigelman-Aubry, and Philippe Grenier. 3d automated lung nodule segmentation in hrct. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*, pages 626–634. Springer, 2003.

[27] Xiangwei Zhang, Geoffrey McLennan, Eric A Hoffman, and Milan Sonka. Computerized detection of pulmonary nodules using cellular neural networks in ct images. In *Medical Imaging 2004*, pages 30–41. International Society for Optics and Photonics, 2004.

[28] Manuel G Penedo, Maria J. Carreira, Antonio Mosquera, and Diego Cabello. Computer-aided diagnosis: a neural-network-based approach to lung nodule detection. *IEEE Transactions on Medical Imaging*, 17(6):872–880, 1998.

[29] Kenji Suzuki, Samuel G Armato III, Feng Li, Shusuke Sone, and Kunio Doi. Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. *Medical Physics*, 30(7):1602–1617, 2003.

[30] Yongbum Lee, Takeshi Hara, Hiroshi Fujita, Shigeki Itoh, and Takeo Ishigaki. Automated detection of pulmonary nodules in helical ct images based on an improved template-matching technique. *Medical Imaging, IEEE Transactions on*, 20(7):595–604, 2001.

[31] Aly A Farag, Ayman El-Baz, Georgy G Gimel'farb, Robert Falk, and Stephen G Hushek. Automatic detection and recognition of lung abnormalities in helical ct images using deformable templates. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, pages 856–864. Springer, 2004.

[32] Hotaka Takizawa, Shinji Yamamoto, Tohru Matsumoto, Yukio Tateno, Takeshi Iinuma, and Mitsuomi Matsumoto. Recognition of lung nodules from x-ray ct images using 3d markov random field models. In *Medical Imaging 2002*, pages 716–725. International Society for Optics and Photonics, 2002.

[33] Colin C McCulloch, Robert A Kaucic, Paulo RS Mendonça, Deborah J Walter, and Ricardo S Avila. Model-based detection of lung nodules in computed tomography exams 1: thoracic computer-aided diagnosis. *Academic radiology*, 11(3):258–266, 2004.

[34] Ingrid Sluimer, Arnold Schilham, Mathias Prokop, and Bram Van Ginneken. Computer analysis of computed tomography scans of the lung: a survey. *Medical Imaging, IEEE Transactions on*, 25(4):385–405, 2006.

[35] David S Mendelson and Daniel L Rubin. Imaging informatics: essential tools for the delivery of imaging services. *Academic radiology*, 20(10):1195–1212, 2013.

[36] Roberto Bellotti, Piergiorgio Cerello, Sonia Tangaro, Vitoantonio Bevilacqua, Marcello Castellano, Giuseppe Mastronardi, Francesco De Carlo, Stefano Bagnasco, Ubaldo Bottigli, Rosella Cataldo, et al. Distributed medical images analysis on a grid infrastructure. *Future Generation Computer Systems*, 23(3):475–484, 2007.

[37] Massimo Lamanna. The lhc computing grid project at cern. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 534(1):1–6, 2004.

[38] Samuel G Armato, Rachael Y Roberts, Masha Kocherginsky, Denise R Aberle, Ella A Kazerooni, Heber MacMahon, Edwin JR van Beek, David Yankelevitz, Geoffrey McLennan, Michael F McNitt-Gray, et al. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". *Academic radiology*, 16(1):28–38, 2009.

[39] Nicholas Petrick, Brandon D Gallas, Frank W Samuelson, Robert F Wagner, and Kyle J Myers. Influence of panel size and expert skill on truth panel performance when combining expert ratings. In *Medical Imaging*, pages 49–57. International Society for Optics and Photonics, 2005.

[40] Giorgio De Nunzio, Eleonora Tommasi, Antonella Agrusti, Rosella Cataldo, Ivan De Mitri, Marco Favetta, Silvio Maglio, Andrea Massafra, Maurizio Quarta, Massimo Torsello, et al. Automatic lung segmentation in ct images with accurate handling of the hilar region. *Journal of digital imaging*, 24(1):11–27, 2011.

[41] Piergiorgio Cerello, Sorin Christian Cheran, Francesco Bagagli, Stefano Bagnasco, Roberto Bellotti, Lourdes Bolanos, Ezio Catanzariti, Giorgio De Nunzio, Elisa Fiorina, Gianfranco Gargano, et al. The channeler ant model: object segmentation with virtual ant colonies. In *2008 IEEE Nuclear Science Symposium Conference Record*, pages 3147–3152. IEEE, 2008.

[42] Qiang Li, Hitetaka Arimura, and Kunio Doi. Selective enhancement filters for lung nodules, intracranial aneurysms, and breast microcalcifications. In *International Congress Series*, volume 1268, pages 929–934. Elsevier, 2004.

[43] Alessandra Retico, Pasquale Delogu, Maria Evelina Fantacci, Ilaria Gori, and A Preite Martinez. Lung nodule detection in low-dose and thin-slice computed tomography. *Computers in biology and medicine*, 38(4):525–534, 2008.

[44] Alessandra Retico, Maria Evelina Fantacci, Ilaria Gori, Parnian Kasae, Bruno Golosio, Alessio Piccioli, Piergiorgio Cerello, Giorgio De Nunzio, and Sonia Tangaro. Pleural nodule identification in low-dose and thin-slice lung computed tomography. *Computers in biology and medicine*, 39(12):1137–1144, 2009.

[45] Bram van Ginneken, Samuel G Armato, Bartjan de Hoop, Saskia van Amelsvoort-van de Vorst, Thomas Duindam, Meindert Niemeijer, Keelin Murphy, Arnold Schilham, Alessandra Retico, Maria Evelina Fantacci, et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study. *Medical image analysis*, 14(6):707–722, 2010.

[46] E Lopez Torres, E Fiorina, F Pennazio, C Peroni, M Saletta, N Camarlinghi, ME Fantacci, and P Cerello. Large scale validation of the m5l lung cad on heterogeneous ct datasets. *Medical physics*, 42(4):1477–1489, 2015.

[47] Karen Coombs. Drupal done right. *Library journal*, 134(19):30–32, 2009.

[48] Todd Tomlinson and John VanDyke. *Pro Drupal 7 Development*. Apress, 2010.

[49] Peter Mildenberger, Marco Eichelberg, and Eric Martin. Introduction to the dicom standard. *European radiology*, 12(4):920–927, 2002.

[50] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[51] Antoine Rosset, Luca Spadola, and Osman Ratib. Osirix: an open-source software for navigating in multidimensional dicom images. *Journal of digital imaging*, 17(3):205–216, 2004.

[52] Peter Mell and Tim Grance. The nist definition of cloud computing. 2011.

[53] Marshall Copeland, Julian Soh, Anthony Puca, Mike Manning, and David Gollob. Overview of microsoft azure services. In *Microsoft Azure*, pages 27–69. Springer, 2015.

[54] Jack Schofield. Google angles for business users with'platform as a service'. *The Guardian.*, 2008.

[55] W Pitt Turner IV, JH PE, PE Seader, and KJ Brill. Tier classification define site infrastructure performance. *Uptime Institute*, 17, 2006.

[56] Stefano Bagnasco, Dario Berzano, R Brunetti, S Lusso, and S Vallero. Integrating multiple scientific computing needs via a private cloud infrastructure. In *Journal of Physics: Conference Series*, volume 513, page 032100. IOP Publishing, 2014.

[57] Dejan Milojičić, Ignacio M Llorente, and Ruben S Montero. Opennebula: A cloud management tool. *IEEE Internet Computing*, (2):11–14, 2011.

[58] Predrag Buncic, C Aguado Sanchez, Jakob Blomer, Leandro Franco, Artem Harutyunian, Pere Mato, and Yushu Yao. Cernvm–a virtual software appliance for lhc applications. In *Journal of Physics: Conference Series*, volume 219, page 042003. IOP Publishing, 2010.

[59] Jakob Blomer, Dario Berzano, Predrag Buncic, Ioannis Charalampidis, Gerardo Ganis, Georgios Lestaris, René Meusel, and V Nicolaou. Micro-cernvm: slashing the cost of building and deploying virtual machines. In *Journal of Physics: Conference Series*, volume 513, page 032009. IOP Publishing, 2014.

[60] Condor Team. Htcondor. *HYPERLINK" http://research. cs. wisc. edu/htcondor/htc. html" http://research. cs. wisc. edu/htcondor/htc. html.[Links]*.

[61] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[62] Mingzhu Liang, Wei Tang, Dong Ming Xu, Artit C Jirapatnakul, Anthony P Reeves, Claudia I Henschke, and David Yankelevitz. Low-dose ct screening for lung cancer: Computer-aided detection of missed lung cancers. *Radiology*, page 150063, 2016.

[63] David P Naidich, Alexander A Bankier, Heber MacMahon, Cornelia M Schaefer-Prokop, Massimo Pistolesi, Jin Mo Goo, Paolo Macchiarini, James D Crapo, Christian J Herold, John H Austin, et al. Recommendations for the management of subsolid pulmonary nodules detected at ct: a statement from the fleischner society. *Radiology*, 266(1):304–317, 2013.

[64] AWS Amazon. Cloudhsm, 2015.

[65] HL Kundel, KS Berbaum, DD Dorfman, D Gur, CE Metz, and RG Swensson. Receiver operating characteristic analysis in medical imaging. *ICRU Report*, 79(8):1, 2008.

[66] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[67] Keelin Murphy, Bram van Ginneken, Arnold MR Schilham, BJ De Hoop, HA Gietema, and Mathias Prokop. A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, 13(5):757–770, 2009.

[68] Colin Jacobs, Eva M van Rikxoort, Thorsten Twellmann, Ernst Th Scholten, Pim A de Jong, Jan-Martin Kuhnigk, Matthijs Oudkerk, Harry J de Koning, Mathias Prokop, Cornelia Schaefer-Prokop, et al. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical image analysis*, 18(2):374–384, 2014.

[69] Maxine Tan, Rudi Deklerck, Bart Jansen, Michel Bister, and Jan Cornelis. A novel computer-aided lung nodule detection system for ct images. *Medical physics*, 38(10):5630–5645, 2011.

[70] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[71] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, Defeng Wang, Vincent CT Mok, Lin Shi, and Pheng-Ann Heng. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195, 2016.

[72] Ilya Sutskever, James Martens, George E Dahl, and Geoffrey E Hinton. On the importance of initialization and momentum in deep learning. *ICML (3)*, 28:1139–1147, 2013.

[73] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[74] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.

[75] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169, 2016.

[76] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–253. Springer, 2013.

[77] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[78] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[79] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[80] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.

[81] Sander Dieleman, Jan Schlüter, Colin Raffel, Eben Olson, Søren Kaae Sønderby, Daniel Nouri, Daniel Maturana, Martin Thoma, Eric Battenberg, J Kelly, et al. Lasagne: First release. *Zenodo: Geneva, Switzerland*, 2015.

[82] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[83] Martin Bergtholdt, Rafael Wiemker, and Tobias Klinder. Pulmonary nodule detection using a cascaded SVM classifier. pages 978513–978513. International Society for Optics and Photonics, 2016.

[84] Bram van Ginneken, Arnaud A. A. Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. pages 286–289, 2015.

[85] Matthew S. Brown, Pechin Lo, Jonathan G. Goldin, Eran Barnoy, Grace Hyun J. Kim, Michael F. McNitt-Gray, and Denise R. Aberle. Toward clinically usable CAD for lung cancer screening with computed tomography. 2014.

[86] Wook-Jin Choi and Tae-Sun Choi. Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. *Entropy*, 15:507–523, 2013.

[87] Maxine Tan, Rudi Deklerck, Jan Cornelis, and Bart Jansen. Phased searching with neat in a time-scaled framework: Experiments on a computer-aided detection system for lung nodules. 2013.

[88] Atsushi Teramoto and Hiroshi Fujita. Fast lung nodule detection in chest CT images using cylindrical nodule-enhancement filter. pages 1–13, 2013.

[89] D. Cascio, R. Magro, F. Fauci, M. Iacomi, and G. Raso. Automatic detection of lung nodules in CT datasets based on stable 3D mass-spring models. 2012.

[90] W. Guo and Q. Li. High performance lung nodule detection schemes in CT using local and global information. 39:5157–5168, 2012.

[91] Maxine Tan, Rudi Deklerck, Bart Jansen, Michel Bister, and Jan Cornelis. A novel computer-aided lung nodule detection system for CT images. 38:5630–5645, 2011.

[92] Edward J Beattie, Nael Martini, Gerald Rosen, et al. The management of pulmonary metastases in children with osteogenic sarcoma with surgical resection combined with chemotherapy. *Cancer*, 35(3):618–621, 1975.

[93] Alfred E Chang, Everett G Schaner, David M Conkle, M Wayne Flye, John L Doppman, and Steven A Rosenberg. Evaluation of computed tomography in the detection of pulmonary metastases. a prospective study. *Cancer*, 43(3):913–916, 1979.

[94] F Beyer, L Zierott, EM Fallenberg, KU Juergens, J Stoeckel, W Heindel, and D Wormanns. Comparison of sensitivity and reading time for the use of computer-aided detection (cad) of pulmonary nodules at mdct as concurrent or second reader. *European radiology*, 17(11):2941–2947, 2007.

[95] Sumiaki Matsumoto, Yoshiharu Ohno, Takatoshi Aoki, Hitoshi Yamagata, Munenobu Nogami, Keiko Matsumoto, Yoshiko Yamashita, and Kazuro Sugimura. Computer-aided detection of lung nodules on multidetector ct in concurrent-reader and second-reader modes: a comparative study. *European journal of radiology*, 82(8):1332–1337, 2013.

[96] Keiko Hirakata, H Nakata, and Jyoji Haratake. Appearance of pulmonary metastases on high-resolution ct scans: comparison with histopathologic findings from autopsy specimens. *AJR. American journal of roentgenology*, 161(1):37–43, 1993.

[97] Phillip A Letourneau, Lianchun Xiao, Matthew T Harting, Kevin P Lally, Charles S Cox, Richard J Andrassy, and Andrea A Hayes-Jordan. Location of pulmonary metastasis in pediatric osteosarcoma is predictive of outcome. *Journal of pediatric surgery*, 46(7):1333–1337, 2011.

[98] A Trajman and RR Luiz. Mcnemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations. *Scandinavian journal of clinical and laboratory investigation*, 68(1):77–80, 2008.

[99] Feng Li, Hidetaka Arimura, Kenji Suzuki, Junji Shiraishi, Qiang Li, Hiroyuki Abe, Roger Engelmann, Shusuke Sone, Heber MacMahon, and Kunio Doi. Computer-aided detection of peripheral lung cancers missed at ct: Roc analyses without and with localization 1. *Radiology*, 237(2):684–690, 2005.