Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Future Generation Computer Systems 23 (2007) 475–484

FGCS
FUTURE GENERATION COMPUTER SYSTEMS

www.elsevier.com/locate/fgcs

Distributed medical images analysis on a Grid infrastructure

R. Bellotti[a,p], P. Cerello[b], S. Tangaro[a,*], V. Bevilacqua[n], M. Castellano[n,**], G. Mastronardi[n], F. De Carlo[a], S. Bagnasco[b], U. Bottigli[d], R. Cataldo[i,m], E. Catanzariti[g], S.C. Cheran[c], P. Delogu[e], I. De Mitri[h,m], G. De Nunzio[i,m], M.E. Fantacci[e], F. Fauci[f], G. Gargano[a], B. Golosio[d], P.L. Indovina[g], A. Lauria[g], E. Lopez Torres[j], R. Magro[f], G.L. Masala[d], R. Massafra[a], P. Oliva[d], A. Preite Martinez[e], M. Quarta[k,m], G. Raso[f], A. Retico[e], M. Sitta[l,b], S. Stumbo[d], A. Tata[e], S. Squarcia[o], A. Schenone[o], E. Molinari[o], B. Canesi[o]

[a] *Dipartimento di Fisica, Università di Bari, and Sez. INFN di Bari, Italy*
[b] *Sez. INFN di Torino, Italy*
[c] *Dipartimento di Informatica, Università di Torino and Associazione per lo Sviluppo del Piemonte, Italy*
[d] *Struttura Dipartimentale di Matematica e Fisica, Università di Sassari, and Sez. INFN di Cagliari, Italy*
[e] *Dipartimento di Fisica, Università di Pisa, and Sez. INFN di Pisa, Italy*
[f] *Dipartimento di Fisica e Tecnologie Relative, Università di Palermo and Sez. INFN di Catania, Italy*
[g] *Dipartimento di Scienze Fisiche, Università di Napoli and Sez. INFN di Napoli, Italy*
[h] *Dipartimento di Fisica, Università di Lecce, Italy*
[i] *Dipartimento di Scienza dei Materiali, Università di Lecce, Italy*
[j] *CEADEN, Havana, Cuba*
[k] *Dipartimento di Matematica, Università di Lecce, Italy*
[l] *Dipartimento di Scienze e Tecnologie Avanzate, Università del Piemonte Orientale, Italy*
[m] *Sez. INFN di Lecce, Italy*
[n] *Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, Italy*
[o] *University of Genova, Department of Communication, Computer and System Science, Genova, Italy*
[p] *Center of Innovative Tecnologies for Signal Detection and Processing (TIRES), Bari, Italy*

Received 1 February 2006; received in revised form 20 April 2006; accepted 2 July 2006
Available online 1 September 2006

Abstract

In this paper medical applications on a Grid infrastructure, the MAGIC-5 Project, are presented and discussed. MAGIC-5 aims at developing Computer Aided Detection (CADe) software for the analysis of medical images on distributed databases by means of **G**RID Services. The use of automated systems for analyzing medical images improves radiologists' performance; in addition, it could be of paramount importance in screening programs, due to the huge amount of data to check and the cost of related manpower. The need for acquiring and analyzing data stored in different locations requires the use of Grid Services for the management of distributed computing resources and data. Grid technologies allow remote image analysis and interactive online diagnosis, with a relevant reduction of the delays presently associated with the diagnosis in the screening programs. The MAGIC-5 project develops algorithms for the analysis of mammographies for breast cancer detection, Computed-Tomography (CT) images for lung cancer detection and Positron Emission Tomography (PET) images for the early diagnosis of Alzheimer Disease (AD). A Virtual Organization (VO) has been deployed, so that authorized users can share data and resources and implement the following use cases: screening, tele-training and tele-diagnosis for mammograms and lung CT scans, statistical diagnosis by comparison of candidates to a distributed data-set of negative PET scans for the diagnosis of the AD. A small-scale prototype of the required Grid functionality was already implemented for the analysis of digitized mammograms.

Keywords: GRID; Virtual organization; CAD; Mammography; Medical applications

* Corresponding address: Dipartimento di Fisica dell'Università di Bari, and Sez. INFN di Bari, via Amendola 173, 70126 Bari, Italy. Tel.: +39 80 5443244.

** Corresponding address: Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, Via Orabona, 4, 70126 Bari, Italy. Tel.: +39 80 5963709; fax: +39 80 5963410.

E-mail addresses: sonia.tangaro@ba.infn.it (S. Tangaro), castellano@poliba.it (M. Castellano).

0167-739X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2006.07.006

1. Introduction

Screening programs are of paramount importance for early cancer diagnosis in asymptomatic subjects, which is the first factor for reduction of the mortality rate. The development of CADe systems [1–5] would significantly improve the effectiveness of screenings programs, by working as second reader to support the physician's diagnosis. However, the image collection in a screening program intrinsically creates a distributed database, as it involves many hospitals and/or screening centers in different locations. In addition, the amount of data generated by such periodical examinations would be so large that it would not be efficient to concentrate them in a single computing center. As an example, let us consider a mammographic screening program to be carried out in Italy: it should check a target sample of about 6.8 millions women in the 49–69 age range at least once every two years, thus implying a data flux of 3.4 millions mammographic exams/year. For an average data size of 50 MB/exam (4 images), the amount of raw data would be in the order of 160 TB/year. Moreover, this quantity linearly increases with time and a full transfer over the network from the collection centers to a central site would be large enough to saturate the available connections. On the other hand, making the whole database available to authorized users, regardless of the data distribution, would provide several advantages. For example, a CADe system could be trained on a much larger data sample, with an improvement of its performance in terms of both sensitivity and specificity. The CADe algorithms could be used as real time selectors of images with high cancer probability, with a remarkable reduction of the delay between acquisition and diagnosis; moreover, data associated to the images, also known as *metadata*, would be available to select the proper input for epidemiology studies or for the training of young radiologists.

The full-scale implementation of such a framework will certainly take advantage of the use of Grid technologies to manage distributed databases and to allow real time remote diagnosis [6–9].

The INFN-funded MAGIC-5 Project aims at the development of Grid-compliant medical applications along three main lines:

- distributed analysis of mammograms, continuing along the scientific program of GPCALMA (Grid Platforms for Computer Aided Library for MAmmography) [10] which is the MAGIC-5 parent project;
- implementation and testing of new algorithms for the analysis of lung CT scans;
- implementation of a Grid-based framework for the use of SPM [11] in the early diagnosis of the AD.

Concerning the Grid aspect, it is based on a model where the input data are not moved and the analysis is performed interactively on different nodes. Other projects [12] are following similar approaches, with different data models and algorithms.

From this point of view, the collaboration can be seen as a VO, with common services (Data and Metadata Catalogue, Job Scheduler, Information System) running on a central server and a number of distributed nodes (Clients) providing computing and storage resources.

Medical application entails the following constraints:

1. some of the cases require interactivity;
2. network conditions do not allow transfer of the full data sample;
3. due to privacy reasons and data ownership, local nodes (hospitals) rarely agree on the raw data transfer to other nodes.

Given these constraints, the MAGIC-5 approach to the implementation of a prototype relies on two basic tools: *AliEn* $v2$–8 version [13] for the management of the common services, and *ROOT/PROOF* [14] for the interactive analysis of remote data.

A dedicated *AliEn* Server for the MAGIC-5 Project was configured, with a central Server running common services and several (about 10) Clients connected to it. Fig. 1 shows a screenshot from the *AliEn* WEB Portal. General information about the status of the VO Common Services and connected Clients can be accessed.

Images can be acquired in any site available to the Project: data are stored on local resources and recorded on a common service, known as *Data Catalogue*, together with the related information (*metadata*) required to select and access them in the future. The result of a query can be used as input for the CADe analysis algorithms, which are executed on nodes that are always local to the image and usually remote to the user, thanks to the *ROOT/PROOF* facility. A selection of the cancer candidates can be quickly performed and only images with high cancer probability would be transferred to the diagnostic sites and interactively analyzed by the radiologists. This approach avoids data transfers for images with a negative CADe response and allows an almost real time diagnosis for the images with high cancer probability.

The medical applications developed until now are reviewed in Section 2, together with the topics related to new-born activities. Section 3 is focused on the implementation of the Grid prototype including a brief description of CADe station functionality and the Graphic User Interface (GUI) developed to drive access to the image analysis. Present status and future plans of the Project will be described in Section 4.

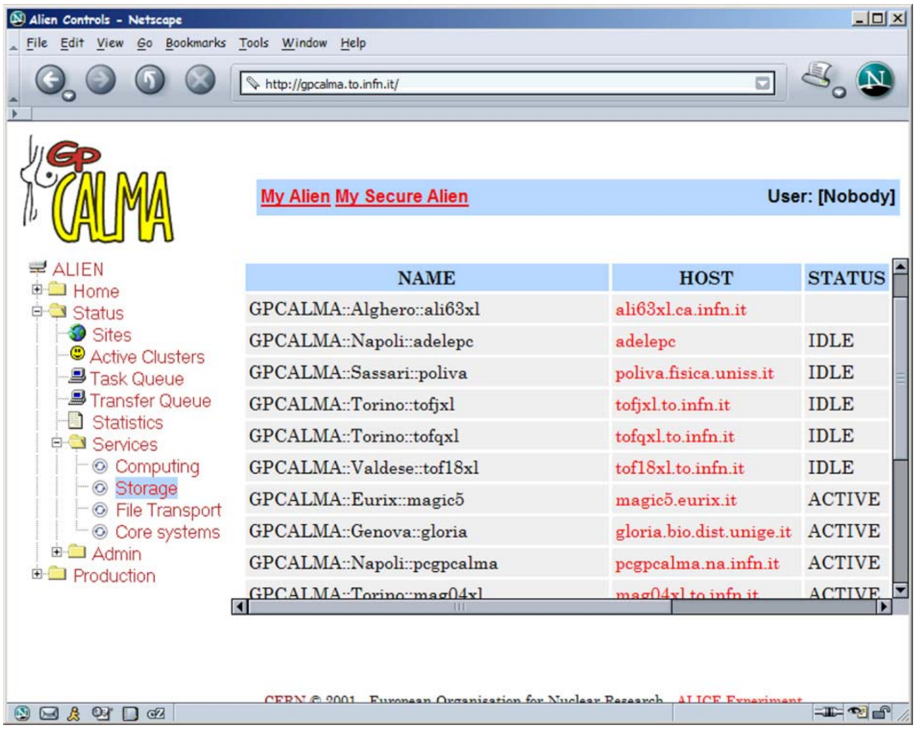
2. The medical applications

As already mentioned, the medical applications of the MAGIC-5 Project cover three main fields:

1. breast cancer detection in mammographic images;
2. nodule detection in lung CT images;
3. early diagnosis of the AD on a Grid.

2.1. Mammographic CADe systems

A database of about 3500 mammographic images, acquired in the hospitals participating in the project, is available. Pathological images have been diagnosed by experienced



NAME	HOST	STATUS
GPCALMA::Alghero::ali63xl	ali63xl.ca.infn.it	
GPCALMA::Napoli::adelepc	adelepc	IDLE
GPCALMA::Sassari::poliva	poliva.fisica.uniss.it	IDLE
GPCALMA::Torino::tofjxl	tofjxl.to.infn.it	IDLE
GPCALMA::Torino::tofqxl	tofqxl.to.infn.it	IDLE
GPCALMA::Valdese::tof18xl	tof18xl.to.infn.it	IDLE
GPCALMA::Eurix::magic5	magic5.eurix.it	ACTIVE
GPCALMA::Genova::gloria	gloria.bio.dist.unige.it	ACTIVE
GPCALMA::Napoli::pcgpcalma	pcgpcalma.na.infn.it	ACTIVE
GPCALMA::Torino::mag04xl	mag04xl.to.infn.it	ACTIVE

Fig. 1. A Screenshot from the *AliEn* Portal. The site can be navigated through the left side frame. General information about the *AliEn* project, the installation and configuration guides, the status of the VO Services can be accessed. The list of the core services is shown on the main frame.

radiologists and confirmed by histological exam. A full description of the pathology including radiological diagnosis, histological data, type and location is available and provides the reference the CADe results will be compared with. Images with no sign of pathology were considered as healthy and included in the database after a follow up of three years.

The images were digitized by means of a Linomed CCD scanner with 85 μm pitch and 12 bits per pixel. Each image is thus described by 2657 × 2067 pixels with $G = 2^{12} = 4096$ grey level tones.

Two different kinds of structures could mark the presence of a breast neoplasia: massive lesions (ML), and microcalcification clusters (MC). ML are rather large (diameter of the order of cm) objects with very different shapes, showing up with faint contrast (see Fig. 2(a)). MC consist of groups of rather small (approximately from 0.1 to 1.0 mm in diameter) but very brilliant objects (see Fig. 2(c)). The database composition is reported in Table 1.

Table 1
Composition of the MAGIC-5 mammographic image database

Images with ML	Images with MC	Healthy images
1153	287	2322

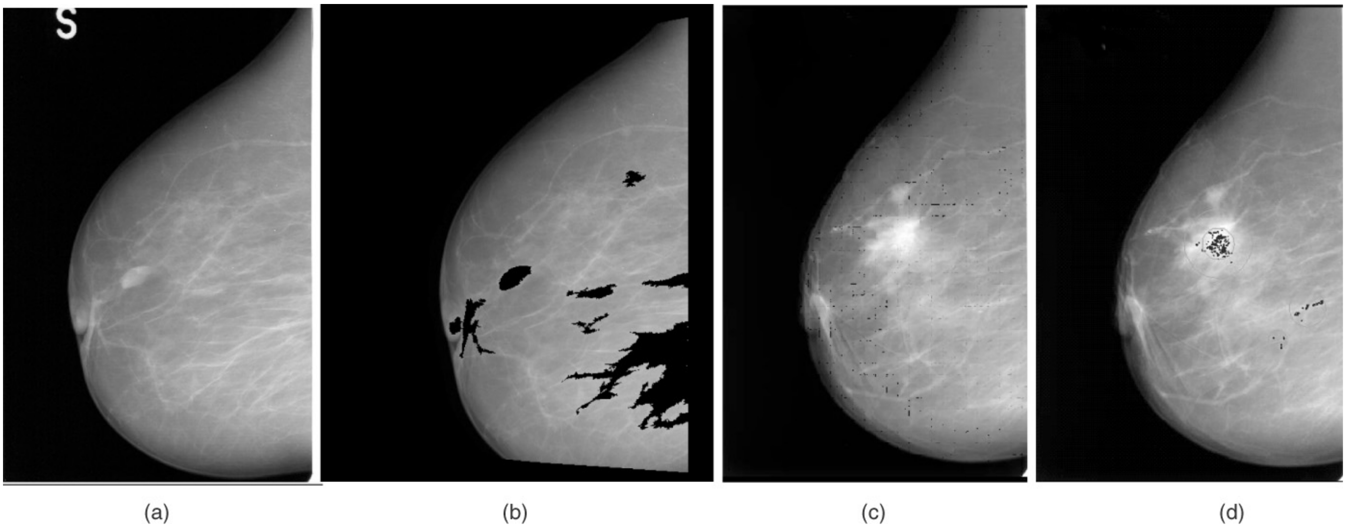


Fig. 2. ML CADe segmentation: original image (a) and result (b); MC CADe segmentation: original image (c) and result (d).

Table 2
Results of the mammographic CADe systems

	Sensitivity	FP/image	Analyzed images
ML CADe	80%	3	3475
MC CADe	96%	0.3	278

Different CADe systems have been developed for ML and MC detection. In both cases, the algorithms consist of a sequence of three main steps:

1. *segmentation*: to perform an efficient detection in a reasonable amount of time, a reduction of the image size is required, without missing any pathology; to this purpose, some regions of interest (ROIs), most likely to contain the pathology, are selected with an efficiency which should be as close as possible to 100%. An example of the segmentation processing is shown in Fig. 2 for both a ML (b) and a MC (d) detection. The original images are reported in (a) and (c), respectively;
2. *feature extraction*: each ROI extracted by the segmentation step is characterized by a proper set of features;
3. *classification*: the feature vectors associated to the ROIs are used as input to a supervised two-layered feed-forward neural network whose output provides a degree of suspicion about the corresponding region.

A detailed description of the CADe algorithms for ML and MC detection is given in [16–18]. In Table 2 the results are shown in terms of sensitivity (fraction of correctly detected pathologies with respect to the total number of images diagnosed by the radiologist) and false positive regions per image (FP/image, number of misclassified healthy ROIs per image), together with the total number of analyzed images for both cases.

Other approaches could be exploited taking into account the biomedical indicators obtained through biopsy [19–21].

2.2. Lung nodule detection in CT scans

The automated identification of small nodules in Computed Tomography (CT) lung scans represents a recently started activity of the MAGIC-5 project. CT has been shown to be the best imaging modality for the detection of small pulmonary nodules [22], particularly after the introduction of the helical technology. The first Italian Randomized Controlled Trial has recently started (*Italung-CT*) [23] to study the potential impact of a screening-based low-dose helical CT on the high-risk population. In this framework the MAGIC-5 Collaboration is developing some automated approaches for the identification of small pulmonary nodules. The database currently available consists of about 100 low-dose (screening setting: 120 kV, 20 mA) CT scans (LDCT), but the data acquisition is still in progress. The average number of slices per scan is about 300, each slice being 512 × 512 voxels with a voxel size in the range [0.5 ÷ 0.7] mm about on the axial plane and a 1 mm reconstructed slice thickness, and 4096 grey level intensity values. The CT scans are stored in the DICOM (Digital Imaging and Communications in Medicine) format. Pathological images have been diagnosed by experienced radiologists participating in the screening program. A full description of the pathology including radiological diagnosis, type, and location is available and provides the reference automated analysis results should be compared with. Images with no sign of pathology were considered as healthy and included in the database. Small pulmonary nodules are quite spherical objects characterized by very low CT values and/or low contrast. Difficulties in detecting such structures arise because they may have CT values in the same range as those of blood vessels and airway walls or may be strongly connected to them. Various approaches are currently being developed and tested:

1. detection of nodule candidates by means of a dot-enhancement filter [25,26];
2. detection of nodule candidates by mean of region growing;
3. detection of non-pathological structure by means of artificial life model [27,28];

A pre-segmentation module, useful for all the approaches, has been developed to identify and extract the pixels belonging to the lung parenchyma. The module consists of a combination of bi-dimensional and a three-dimensional algorithms: the former is based on a mixture of image processing techniques (threshold-based segmentation, morphological operators, border detection) and different geometrical rules (typical organ positions and sizes) to distinguish the pulmonary parenchyma from other low-density chest/abdomen tissues and organs (trachea, stomach, intestine). The 3-D algorithm takes into consideration the position of the various organs in each CT scan slice, to get hints of their position in what follows.

This algorithm depends on few parameters which have been set by an optimization procedure carried out on a number of CT scans. The pre-segmentation procedure is thus totally non-interactive and provides correct results on 80–85% of thirty CT scans with both low-and high-resolution. For other cases, the results should be refined.

The first approach for automated detection of pulmonary nodules is based on a dot-enhancement filter for the selection of nodule candidates and a neural-based module for the reduction of the amount of false-positive (FP) findings per scan. The selection of nodule candidates is provided by filter enhancing spherical-shaped objects. To this end we followed the approach proposed by Li et al. [24], where nodules are modelled as fuzzy dots in a 3D space and a dot-enhancement filter gives the maximum response in correspondence of nodule-like objects, while suppressing elongated and planar-shaped objects. This filter attempts to determine the local geometrical characteristics of each voxel, by computing the eigenvalues of the Hessian matrix and evaluating a likelihood function that was purposely built to discriminate among local morphologies of linear, planar, and spherical objects. A simple peak-detection algorithm (i.e. a local maximum detector) is then applied to the filter output to detect the filtered-signal peaks. Since most FP findings are provided at the cross of blood vessels, we attempted to reduce the amount of FP/scan by applying a voxel-based approach (VBA). According to this method, each voxel of a ROI is characterized by the gray level intensity values of its

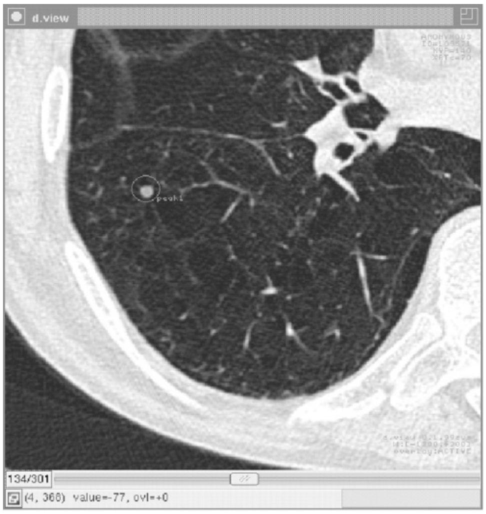


Fig. 3. An example of a correct nodule detection through the application of the dot-enhancement filter and a threshold-based algorithm for peak-detection to the 3D matrix of voxel data.

neighborhood. The CT values of a 3D neighborhood of each selected voxel are rolled down into vectors of features to be analyzed by a neural classification system. Each voxel of the 3D array is flagged by the neural classifier with the appropriate class membership, e.g. voxels belonging to a nodule or to the normal lung parenchyma. The dot-enhancement filter provides 100% sensitivity with 67 FP/scans on a dataset of 20 scans, 8 containing 12 internal nodules. When the VBA is applied the number of FP/scans decreases to 14 (see, for example, the Fig. 3) [25,26].

The second nodule detection method is based on a region-growing algorithm. The algorithm builds nodule candidates starting from seeds in the 3D matrix of voxel data: a seed voxel is found in the CT, according to the region-growing inclusion rule, then starting from this seed a nodule candidate is grown and removed from the CT; this procedure is repeated until no new seed voxel is found. At the present the algorithm has been completely developed and performances tested including with rules on about twenty CT scans. The efficiency in detecting nodule and micronodule is 85% ± 10% with 6 ± 2 FP per CT scan.

The third approach is based on Sworm Intelligence. CT undergoes the above described parenchyma pre-segmentation phase to build a confined area where ants are deployed. The ant-based approach called Ant Colony Reconstruction, will introduce two different models where ants move in all the directions, either starting from random positions (the wander ant model) or from a well defined anthill (the channel ant model) under the influence of a pheromone trail and a directional bias probability. This approach can be used to remove from the lung CTs non pathological structures as the bronchial and the vascular tree. This application is not yet tested but its efficiency is shown in Fig. 4.

2.3. Diagnosis of Alzheimer disease

AD is the leading cause of dementia in elderly people. Clinically, AD is characterized by a progressive loss of cognitive abilities and memory. One of the most widely used tools for the analysis of medical imaging volumes in neurological applications is the SPM (Statistical Parametric Mapping) software [11] which has been developed by the Institute of Neurology at University College in London. SPM provides a number of functionalities related to image processing and statistical analysis, such as segmentation, co-registration, normalization, parameter estimation and statistical mapping. The quantitative comparison of PET images from suspected AD patients with the ones included in a database of normal cases is the base of statistical analysis by means of SPM. For this reason, the access to a large number of normal cases is of paramount importance to obtain good results in diagnosis. Nevertheless, hospitals without PET facilities could be interested in working with SPM in order to produce reliable diagnosis of AD. Our work concerned the development and the deployment of a Grid environment able to remotely access images acquired at large PET facilities on normal subjects. Due to security and privacy issues in medical institutions, this is a crucial step in making SPM analysis available for an increasing number of neurologists. Actually, within our Grid environment,

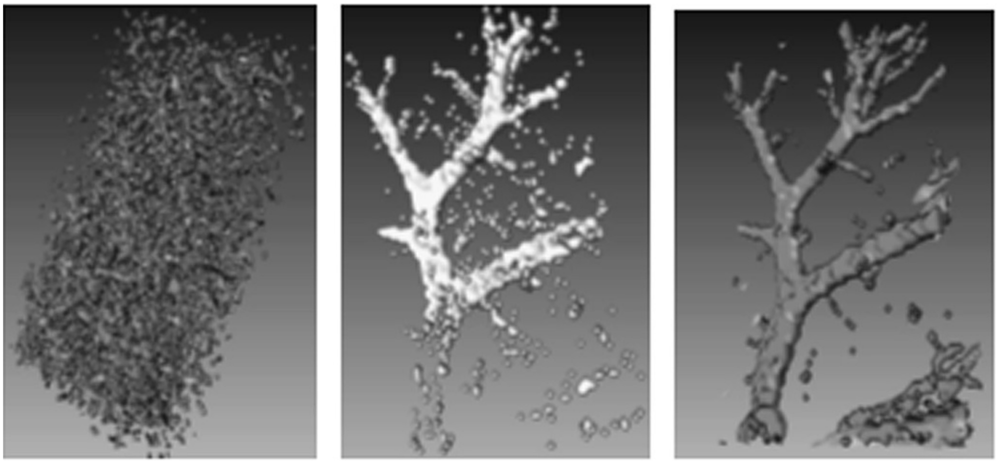
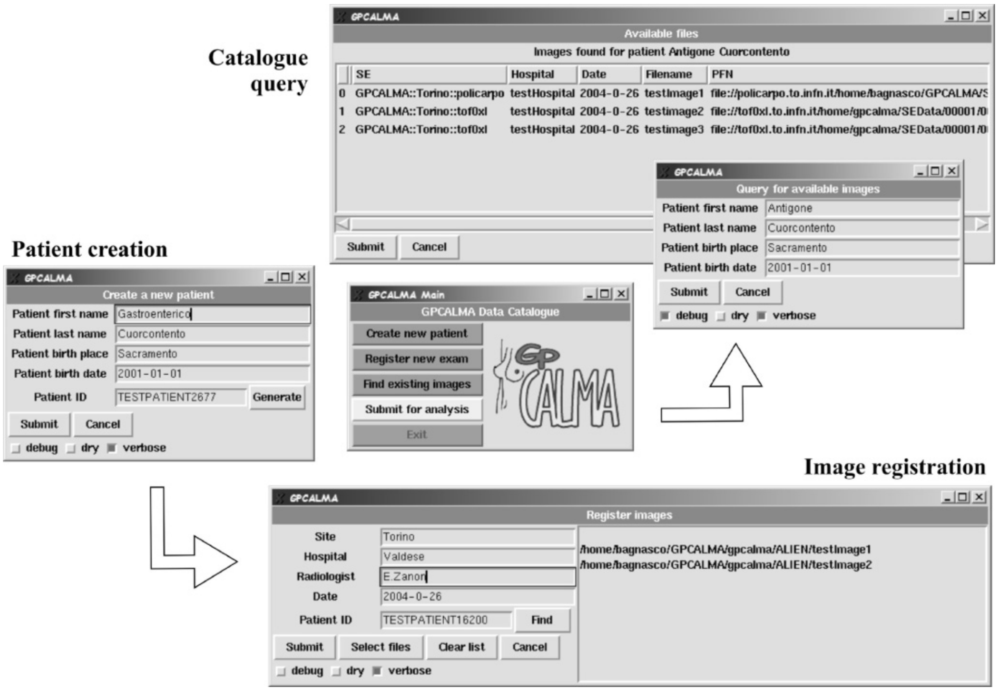


Fig. 4. Ants at work: Reconstruction of a part of the bronchial tree after 1 cycle (A), 50 cycles (B), 100 cycles (C).



Images found for patient Antigone Cuorcontento

	SE	Hospital	Date	Filename	PFN
0	GPCALMA::Torino::policarpo	testHospital	2004-0-26	testImage1	file://policarpo.to.infn.it/home/bagnasco/GPCALMA/S
1	GPCALMA::Torino::tof0xl	testHospital	2004-0-26	testimage2	file://tof0xl.to.infn.it/home/gpcalma/SEData/00001/0
2	GPCALMA::Torino::tof0xl	testHospital	2004-0-26	testimage3	file://tof0xl.to.infn.it/home/gpcalma/SEData/00001/0

Patient first name	Antigone
Patient last name	Cuorcontento
Patient birth place	Sacramento
Patient birth date	2001-01-01

☒ debug ☐ dry ☒ verbose

Patient first name	Gastroenterico
Patient last name	Cuorcontento
Patient birth place	Sacramento
Patient birth date	2001-01-01
Patient ID	TESTPATIENT2677

☐ debug ☐ dry ☒ verbose

Site	Torino
Hospital	Valdese
Radiologist	E.Zanon
Date	2004-0-26
Patient ID	TESTPATIENT16200

☐ debug ☐ dry ☒ verbose

/home/bagnasco/GPCALMA/gpcalma/ALIEN/testImage1
/home/bagnasco/GPCALMA/gpcalma/ALIEN/testImage2

Fig. 5. The high level analysis interface: four access points correspond to the data registration functionality (registration of new patient and/or new exams) and to the analysis (Data Catalogue query for input selection and analysis execution with *PROOF*).

remote images are securely accessed by final users through authorization and certification tools intrinsically provided by Grid middleware. Moreover, remote images can be selected through metadata associated with patient data (age, gender) to improve the results of the analysis. As a third step information is remotely extracted from selected images, without moving the images, to produce the data set needed for statistical analysis. Then, remote extracted data are moved to computation nodes where the SPM statistical analysis is performed. This Grid version of the SPM application for the analysis of PET images has been made accessible through the Italian Portal of Neuroscience [29].

3. Grid infrastructure prototype for distributed analysis

As a consequence of the intrinsically distributed nature of medical databases, the non-technical problems related to image replication (e.g. privacy, ownership) and the need to minimize data replication and transfer, our approach is based on the following general assumption: whenever possible data should be stored and analyzed where they are collected. That assumption, which also maximizes parallelization in the analysis of distributed data samples, requires algorithms to be shipped to each site where a fraction of the input images is stored. Such a functionality is available thanks to the *ROOT/PROOF* framework, which provides a C++ interpreter and a system to start remote interactive processes on demand (*PROOF*).

The dependencies on different framework components are minimized by decoupling the data management and data analysis functionalities, implemented by means of *AliEn* and *PROOF*, respectively. The interface to Grid services, driven by the local application GUI, is used in the registration of new data in the catalogue or in the queries to define the input dataset, but not in the actual data analysis. Should the need arise to replace one of the components, such a modular approach would allow a quick replacement without affecting the other part of the system.

The application interacts with the *AliEn* Server via a Perl/TK high level graphic interface, shown in Fig. 5, which accesses the *AliEn* Data Management and Storage Element services. Once the information about the images to be analyzed (site and physical file name on the filesystem) is retrieved from the Data Catalogue, the *PROOF* cluster is dynamically configured and managed via the *ROOT*-based Grid Portal.

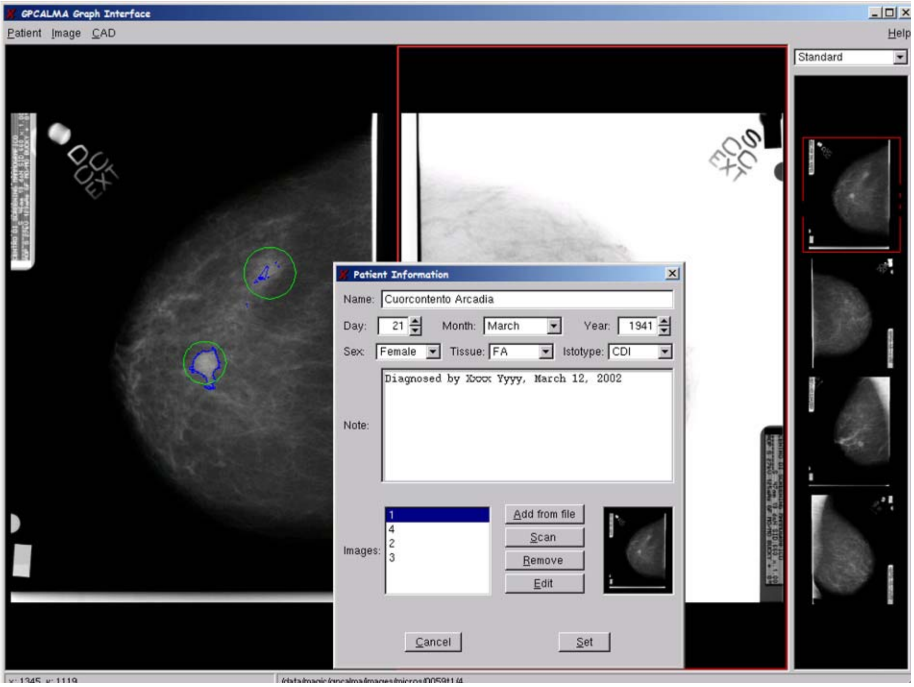
3.1. Interface to Grid services

Interface to Grid Services provides access to the Grid authentication and Data Management Services, while the WEB portal provides a user-friendly view of the present status of the VO resources (i.e. monitoring).

A GUI has been developed (see Fig. 6) to drive the execution of three basic functionalities related to the data management:

1. registration of a new patient, based on the generation of a unique identifier, which could be easily replaced by the *Social Security* identification code;
2. registration of a new exam for an existing patient;
3. analysis of the selected image with algorithms for the search of MC and ML.

Images are displayed according to the standard format required by the radiologists: for each image, it is possible to insert or modify diagnosis and annotations, and to manually select the portion of the mammogram corresponding to the radiologist's



Name: Cuorcontento Arcadia
Day: 21 Month: March Year: 1941
Sex: Female Tissue: FA Istotype: CDI
Note: Diagnosed by Xxxx Yyyy, March 12, 2002
Images: 1 4 2 3

Fig. 6. The graphic user interface. Three menus allow browsing of the Patient, the Image, and the CAD detection levels. On the left and right mammograms, the CADe results for MC and ML (blue polygons) are shown, together with the radiologist's diagnosis (green circles).(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

indication. An interactive procedure allows a number of operations such as zooming, windowing, gray levels and contrast selection, image inversion, luminosity and contrast tuning. The human analysis produces a diagnosis of the breast lesions in terms of kind, localization on the image, average dimensions and, if available, histological type. The CADe procedure finds the ROIs of the image with a probability of containing a pathological area larger than a pre-defined threshold value.

3.2. Security

The access to the VO resources is only granted to users owning a Grid Certificate and registered in the VO Servers as being authorized to access the system. Authentication and authorization services are based on the use of standard X.509 Certificates, thank to the *AliEn* interface to the lower level functionality provided by GLOBUS [30].

Typically, data registered in the Data Catalogue can be classified in two categories: patient-related, including some private information, and exam-related. The issue of privacy is particularly important: in our prototype, we implemented a simple protection based on the association of a random identifier to each registered patient. Private and public data are recorded in different areas of the file system-like structure in such a way that it is possible to identify exams once the patient information is available but not the other way around.

3.3. Data management

Data management issues are important for two basic use cases that must be implemented: registration of a new exam (with related information, including patient data) in the Data Catalogue and retrieval of the information related to one or more exams/patients from the Data Catalogue.

Data (and related metadata) are recorded in several hospitals, stored on their local Storage Elements (SE) and registered with a user-defined logical name (LFN) in the Data Catalogue, which keeps track of the correspondence between the LFN and the system-defined physical name (PFN) on the filesystem. *AliEn* [13] implements these features in its Data Catalogue Central Service, run by the Server: data are registered making use of a hierarchical namespace for their LFNs. In addition, it is possible to attach metadata to each level of the hierarchical namespace. The Data Catalogue can be browsed from the *AliEn* command line as well as from the C++ Application Program Interface (API).

Presently, the Data Catalogue is a Common Service running on the VO server. The new generation of Grid services will probably provide a two-level catalogue, with a central service storing the information on the data location and a distributed service tightly connected to the Storage Element, which will provide the information to physically access the data.

Metadata associated with the images belong to several categories: patient and exam identification data, results of the CADe algorithm analysis, histological or radiologist's diagnosis, and so on. At the exam level, it is possible to define whether image replication is allowed or not. The prototype implements the registration of a few basic metadata, which allow queries for the selection of:

- all the exams associated to a known patient;
- all the images that meet a user-defined set of requirements (e.g. all the images with a positive diagnosis by the radiologist).

Data analysis usually starts with the identification of the subsample of the images recorded in the Data Catalogue that must be analyzed. In order to do that, it is possible to define a query which restricts the answer to the list of images meeting a set of criteria based on the value of one or more metadata. The simplest example is the retrieval of all the images associated with a given patient. In order to simplify the access, a GUI (see Fig. 5) which, according to the metadata values requested by the user, lists the nodes and PFNs of the selected data, was designed. At this point, all the information required to start the remote analysis is available.

3.4. Remote data processing

Both tele-diagnosis and tele-training require interactivity in order to be fully exploited. The *PROOF* (Parallel ROOT Facility) system [14] provides the functionality needed to run interactive parallel processes on a distributed cluster of computers.

According to our Data Model, our preferred use of resources goes through a distributed interactive session managed by a *PROOF* Master node. The assumption that data are analyzed where they are stored requires a local intrinsic balancing between computing and storage resources. With such an approach, resource management is intrinsically implemented: tasks requiring the analysis of input files stored in different sites are split and sent, in parallel, to the corresponding sites by the *PROOF* Master.

A batch mode is also possible, although it is not the main focus of our work. In that case, the above-mentioned assumption is still true. However, job requests in that option are published in the *Server Task Queue* service and pulled by their destination sites whenever free resources are available.

A dedicated cluster of several PCs in five different locations was configured and the remote analysis of digitized mammograms without data transfer was run. As previously discussed, whenever a set of input metadata is selected, the query to the Data Catalogue retrieves the list of physical file names, one per image, consisting of the name of the Storage Element where the image is located, and the physical path on the SE file-system. That information is used to dynamically generate, out of a predefined template, a C++ script that drives the execution of the CADe algorithm and is sent to the remote node. Whenever a user requires the remote execution of an algorithm, a new PROOF session is started and the user node acts as a *PROOF Master*. It forwards the user-defined algorithm to the remote sites where a fraction of the selected input is stored: those sites act as *Slaves* and carry out the task.

Requests are sent via a pre-defined port which must be open on both the *Master* and the *Slave* site. However, generic inbound connectivity is not required on the remote site.

The script output is a list of positions and probabilities corresponding to the image regions identified as pathological by the CADe algorithm (ROIs). Based on that, it is possible to decide whether the image retrieval (if allowed) is required for immediate visualization in the GUI or not.

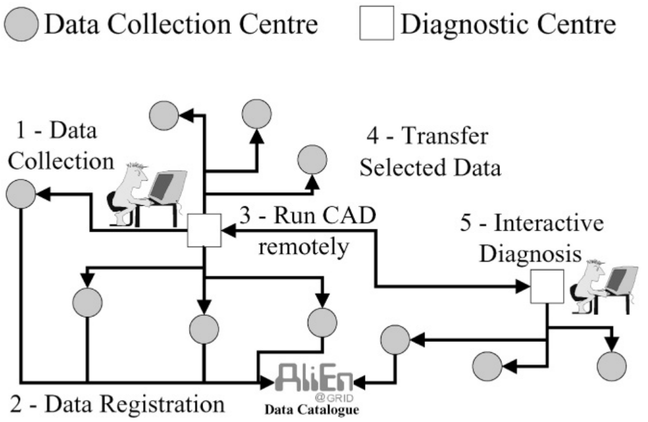


Fig. 7. Prototype deployment. Services running at the central Server include the system configuration, the Data Catalogue, the Users' Database and the WEB interface. Clients run local services, with the Compute and Storage Element, the File Transfer Service, the *ROOT* and *PROOF* deamons and application.

3.5. Prototype deployment and performance

The prototype was deployed with a star-like topology, with an *Alien* Server running common services (magic5.to.infn.it) and several Clients running local services, according to the sketch shown in Fig. 7. The user accesses the required Grid Services only via the Grid Portal, which acts as an interface to the application software and to the distributed system services. As already said, until now the work was mostly focused on the implementation of the required functionality for distributed interactive analysis. The additional time in remote execution mode, in the order of a few seconds per task (independent of the task duration), is certainly acceptable, particularly if compared to the unknown delay associated with a batch submission. A typical analysis job on a single image, running CADe algorithms for the search of MC and ML on a mammogram, would last about 20 s and would analyze 10 MB (the raw data size for a digitized image).

The scalability issue is completely dependent on Grid Services, as the application code does not change when the mode is local or remote, and is mostly related to the Data Catalogue size and the number of concurrent users in the system. Both issues were addressed in the framework of the CERN/ALICE project, in its 2004 Physics Data Challenge [15]. The ALICE *AliEn* Data Catalogue presently contains more than 9 M entries, without any degradation in performance and with an efficiency for registration and query very close to 100%.

The number of concurrent users is a very important parameter for the successful deployment of the system on a large scale. In batch mode, *AliEn* proved to scale very well up to 30 sites and was able to run up to 1000 jobs at the same time [15]. However, interactive access requires the availability of *PROOF* deamons on the selected nodes that accept incoming requests for the algorithm execution. The implementation of that functionality turned into a major restructuring of the *PROOF* architecture, which now maps the two-level architecture of the Compute Element/Worker Node scheme used for the batch mode. The *PROOF* Master now contacts a remote Master running on the site interface machine and the remote Master triggers Slaves on the different CPUs locally connected to it, as recently demonstrated at the SuperComputing 2004 Conference (Pittsburgh, Nov. 2004).

Among the MAGIC-5 goals is the implementation of a prototype service in about 10 hospitals, a scale which does not require the new approach. However, in view of a possible extension to a much larger number of hospitals (i.e. sites), the next generation of our prototype will be implemented with the new *PROOF* functionality. Assuming a large scale deployment, a two-level topology does map well with the expected functionality of sites in a screening program: *Data Collection* centers would host *PROOF* slaves, *Diagnostic Centers* would host sub-masters, while the Grid Provider node would host the central Grid Services as well as the *PROOF* master.

4. Present status and future plans

The Grid approach to the analysis of distributed medical data is very promising. An *AliEn* Server managing the MAGIC-5 VO common services is installed and configured, and about 10 *AliEn* Clients are in use. The remote analysis of mammographic images was successfully implemented, thanks to the *PROOF* facility. Presently, all the functionality required for the implementation of tele-diagnosis and screening use cases is integrated into a prototype system. The Graphic User Interface for the mammogram analysis is now considered satisfactory by the radiologists involved in the project, thanks to the possibility to manipulate the image and set of modify the associated metadata.

The MAGIC-5 Grid strategy for the screening use case relies on the principle that images are collected in the hospitals and analyzed by means of the CADe systems; only the images with a high probability to contain a pathology are moved over the network to the diagnostic centers, where the radiologists can analyze them, almost in real time, by taking advantage of the CADe selection.

Medical applications are continuously under development. Both new algorithms (pulmonary CADe) and improvements of the existing ones (mammographic CADes) are nonstop under study. At the same time, part of the future work will be focused on the collection of a CT image database (at present, a limited number of scans is available) and the implementation of the VO related to the PET image analysis for the early AD diagnosis.

Aknowledgements

We would like to acknowledge Dr L. Battolla, Dr F. Falaschi, and Dr C. Spinelli of the Unità Operativa Radiodiagnostica 2 of the Azienda Universitaria Pisana and Prof. D. Caramella and Dr T. Tarantino of the Diagnostic and Interventional Radiology division of the Dipartimento di Oncologia, Trapianti e Nuove Tecnologie in Medicina of the Pisa University who have made available the full annotated CT lung scans.

References

[1] S. Timp, N. Karssemeijer, A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammograph, Medical Physics 31 (2004) 958–971.

[2] A.H. Baydush, D.M. Catarious Jr., C.K. Abbey, C.E. Floyd, Computer aided detection of masses in mammography using subregion Hotelling observers, Medical Physics 30 (2003) 1781–1787.

[3] G.D. Tourassi, R. Vargas-Voracek, D.M. Catarious Jr., C.E. Floyd Jr., Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information, Medical Physics 30 (8) (2003) 2123–2130.

[4] D.M. Catarious Jr., A.H. Baydush, C.K. Abbey, C.E. Floyd Jr., A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: Preliminary results, in: SPIE Medical Imaging 2003, San Diego, CA San Diego, CA, USA, 2003, p. 1927.

[5] D.M. Catarious Jr., A.H. Baydush, C.E. Floyd Jr., Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system, Medical Physics 31 (6) (2004) 1512–1520.

[6] B. Marovic, Z. Jovanovic, Web-based Grid-enabled interaction with 3D medical data, Future Generation Computer Systems 22 (4) (2006) 385–392.

[7] M. Castellano et al., An Intelligent Grid Resource Selection Mechanism Based on Neural Network, GESTS Intl Trans. Computer Sciente and Engr. 26(1) 181–193.

[8] M. Castellano et al., Evaluation scheduling alghoritms in a data Grid network, in: Proceeding of 10th IEEE International Conference on Software, Telecommunication and Computer Networks, Croazia, October 2002.

[9] V. Bevilacqua, G. Mastronardi, G. Piscopo, SCIA-Evolutionary approach to inverse planning in coplanar radiotherapy, Journal of Image and Vision Computing (in press). Elsevier.

[10] P. Cerello, et al., GPCALMA: A Grid based tool for mammographic screening, Methods of Information in Medicine 44 (2005) 244–248.

[11] R. Frackowiak, et al., Human Brain Function, 2nd ed., Academic Press, 2003.

[12] A. Solomonides, R. McClatchey, M. Odeh, M. Brady, M. Mulet-Parada, D. Schottlander, S.R. Amendolia, MammoGrid and eDiamond: Grids applications in mammogram analysis, in: Proceedings of the IADIS International Conference: e-Society 2003, Lisbon, Portugal, June 2003.

[13] P. Saiz, L. Aphecetche, P. Buncic, R. Piskac, J.E. Revsbech, V. Sego, AliEn-Alice environment on the Grid, Nuclear Instruments and Methods in Physics Research Section A (Accelerators, Spectrometers, Detectors and Associated Equipment) 502 (2–3) (2003) 437–440.

[14] M. Ballintijn et al., The PROOF distributed parallel analysis framework based on ROOT, in: Proceedings of the CHEP2003 Conference, La Jolla, CA, US, 2003.

[15] P. Canal et al., Global distributed parallel analysis using PROOF and AliEn, in: Proceedings of the CHEP2004 Conference, Interlaken, CH, 2004.

[16] F. Fauci, et al., A massive lesion detection algorithm in mammography, Physica Medica XXI (1) (2005) 21–28.

[17] R. Bellotti, et al., A completely automated CAD system for mass detection in a large mammographic database, Medical Physics 33 (8) (2006) 3066–3075.

[18] C.S. Cheran, R. Cataldo, P. Cerello, F. De Carlo, F. Fauci, G. Forni, B. Golosio, A. Lauria, E. Lopez Torres, D. Martello, G. Masala, G. Raso, A. Retico, A. Tata, Detection and classification of microcalcification clusters in digital mammograms, in: Proc. IEEE Medical Imaging Conference, 16–22 October 2004, Rome, Italy.

[19] V. Bevilacqua, G. Mastronardi, F. Menolascina, Hybrid data analysis methods and artificial neural network design in breast cancer diagnosis: IDEST experience, in: Proceeding of IEEE, Int Conf on Computational Intelligence for Modelling, Control and Automation, Vienna, Austria, 2005.

[20] V. Bevilacqua, G. Mastronardi, F. Menolascina, Intelligen information structure investigation in biomedical database: The breast cancer diagnosis problem, in: Proceeding of Int Conf on Intelligent Systems and Control, Cambridge, USA, 2005, pp. 310–314.

[21] V. Bevilacqua, G. Mastronardi, F. Menolascina, P. Pannarale, A. Pedone, A novel multi-objective genetic algorithm approach to artificial neural network topology optimisation the breast cancer classification problem,

in: Proceedings of International Joint Conference on Neural Networks, 16–21 July 2006, Sheraton Vancouver Wall Center, Vancouver, BC, Canada.

[22] S. Diederich, et al., Detection of pulmonary nodules at spiral CT: Comparison of maximum intensity projection sliding slabs and single-image reporting, European Radiology 11 (2001) 1345.

[23] http://www.med.unifi.it.

[24] Q. Li, S. Sone, K. Doi, Selective enhancement filters for nodules, vessels, and airway walls in two and three-dimensional CT scans, Medical Physics 30 (8) (2003) 2040.

[25] S.C. Cheran, P. Delogu, I. De Mitri, G. De Nunzio, M.E. Fantacci, F. Fauci, G. Gargano, E. Lopez Torres, R. Massafra, P. Oliva, A. Preite Martinez, G. Raso, A. Retico, S. Stumbo, A. Tata, Pre-processing methods for nodule detection in lung CT, in: Computer Assisted Radiology and Surgery (Proceedings of the 19th International Congress and Exhibition, Berlin, Germany, 22 – 25 June 2005), in: International Congress Series vol. 1281, pp. 1099–1103.

[26] P. Delogu, M.E. Fantacci, I. Gori, A. Preite Martinez, A. Retico, A. Tata, Lung nodule detection in low-dose and high-resolution CT scans, in: Proceedings of the Frontier Science 2005, 4th International Conference on Frontier Science, Milano, Italy, September 12–17.

[27] M. Castellano, A. Aprile, R. Bellotti, P. Cerello, S.C. Cheran, G. Gargano, E.L. Torres, S. Tangaro, Artificial Life Models in Lung CTs, GESTS Intl Trans. Computer Sciente and Engr. 27(1)159–167.

[28] S.C. Cheran, G. Gargano, Artificial Life Models in Lung CTs, in: Lecture Notes in Computer Science, vol. 3907, 2006, pp. 510–514.

[29] http://www.neuroinf.it.

[30] B. Sotomayor, L. Childers, Globus Toolkit 4, First edition: Programming Java Services, Morgan Kaufmann.

Sabina Tangaro was born in 1972. She received Laurea degree in Physics from University of Pisa (Italy) and Ph.D. in Physics from University of Bari (Italy). Currently she is researcher at National Institute of Nuclear Physics, sez. of Bari. Previously, she has been research fellow at Italian National Council of Researches and Post Doctoral Researcher at University of Bari. Dr Tangaro's research interests include many topics on Image Processing, Computer Vision and Pattern Recognition, Machine Learning, with application in Medicine and on Medical Imaging. In these fields, she authored highquality scientific papers in international journals.

Marcello Castellano was born in 1961. He received "Laurea cum Laude" in Computer Science in 1985 from University of Bari (Italy). Currently he is Assistante Professor at the Department of Electrical and Electronic Engineering of the Polytechnic of Bari, Italy. Previously, he has been staff member researcher at National Institute of Nuclear Physics, and computer specialist at Italian National Council of Researches. He received a scientific associate contract from Center European of Nuclear Researcher and Visiting Researcher at New Mexico State University and Gran Sasso International Laboratory (Italy). He serves as reviewer in several scientific international journals and conferences. Dr Castellano's main research interests are in machine learning, data analysis and mining. In these fields, he authored high-quality scientific papers in international journals.