Use of HEP software for Medical Applications

E. Lopez Torres
CEADEN, Habana, Cuba
F. Fauci, R. Magro
Università di Palermo and INFN, Catania, Italy
L. Ramello
Università del Piemonte Orientale, Alessandria, and INFN, Torino, Italy
M. E. Fantacci
Università di Pisa and INFN, Pisa, Italy
U. Bottigli, G. L. Masala, P. Oliva
Università di Sassari and INFN, Cagliari, Italy
S. Bagnasco, P. Cerello
INFN, Torino, Italy

HEP requirements for the next generation of experiments emphasize the importance of a GRID approach to a distributed computing system and the associated data management, where the key subject is a virtual organisation, a group of geographycally distributed users with a common goal and the will to share their resources.
Most of these requirements are satisfied by other kinds of Virtual Organisations, like, for example, groups of Hospitals joining common screening programs for the early diagnosis of breast and lung cancer.
HEP techniques already come into play both in writing the application code, which makes use of neural networks for the image analysis and proved to be very effective in the diagnosis. In addition, even though the raw data model is different in this case (several distributed data sources rather than a single data acquisition system), the GRID approach will be extremely useful: it would allow remote image analysis and interactive online diagnosis, with a relevant reduction of the delays in the diagnosis presently associated to screening programs because of the lack of radiologists.
The approach adopted by the INFN/GPCALMA project, in collaboration with EU-Mammogrid, based on ROOT for the application code, on PROOF for the parallel remote data analysis and on *AliEn* for the distributed data management, will be discussed. Thanks to the PROOF functionality, this approach will avoid data replication for all the images with a negative detection (about 95% of the sample), while it will allow a real time detection for the 5% of images with high cancer probability.

1. Introduction

The GPCALMA (*Grid Platform for Computer Assisted Library in MAmmography*) Collaboration involves several italian departments of physics and hospitals, coordinated and supported by the I.N.F.N. (*National Institute of Nuclear Physics*). The aim of this collaboration is the development of tools that would help in the early detection of breast cancer: the availability of suitable tools for computer assisted detection would significantly improve the prospects for mammographic screening, by quickly providing reliable information to the radiologists. During its research program, started in 1998, the Collaboration set up a large distributed database of digitised mammographic images (about 5500 images corresponding to 1650 patients) and developed a CAD (Computer Aided Detection) software which is now integrated in a station that can also be used for the acquisition of new images, as archive and to perform statistical analysis.

The images ($18x24\ cm^2$, digitised by a CCD linear scanner with a 85 μm pitch and 4096 gray levels, corresponding to 12 bits) are completely described: pathological ones have a consistent characterization with radiologist's diagnosis and histological data, non pathological ones correspond to patients with a follow up at least three years.

The GPCALMA tools perform several analysis. A texture analysis, i.e. an automated classification on adipose, dense or glandular texture, is provided by the system. Also, the GPCALMA software allows the classification of pathological features, in particular massive lesions (both opacities and spiculated lesions) and microcalcification clusters.

The detection of pathological features is made using neural network based algorithms that provide a selection of areas with a given "suspicion level" (i.e., probability over threshold) of lesion occurrence.

The performance of the GPCALMA system will be presented in terms of the ROC (Receiver Operating Characteristic) curves: microcalcifications and massive structures can be identified with a sensitivity (i.e., efficiency) of about 92% and 94%, respectively, while keeping the fraction of false positives to a few percent.

While the GPCALMA CAD tools were being developed, in view of the huge distributed computing effort required by the CERN/LHC collaborations, several GRID projects were started, with a relevant INFN involvement. It was soon understood that the application of GRID-like technologies to mammographic screening would represent a significant step towards

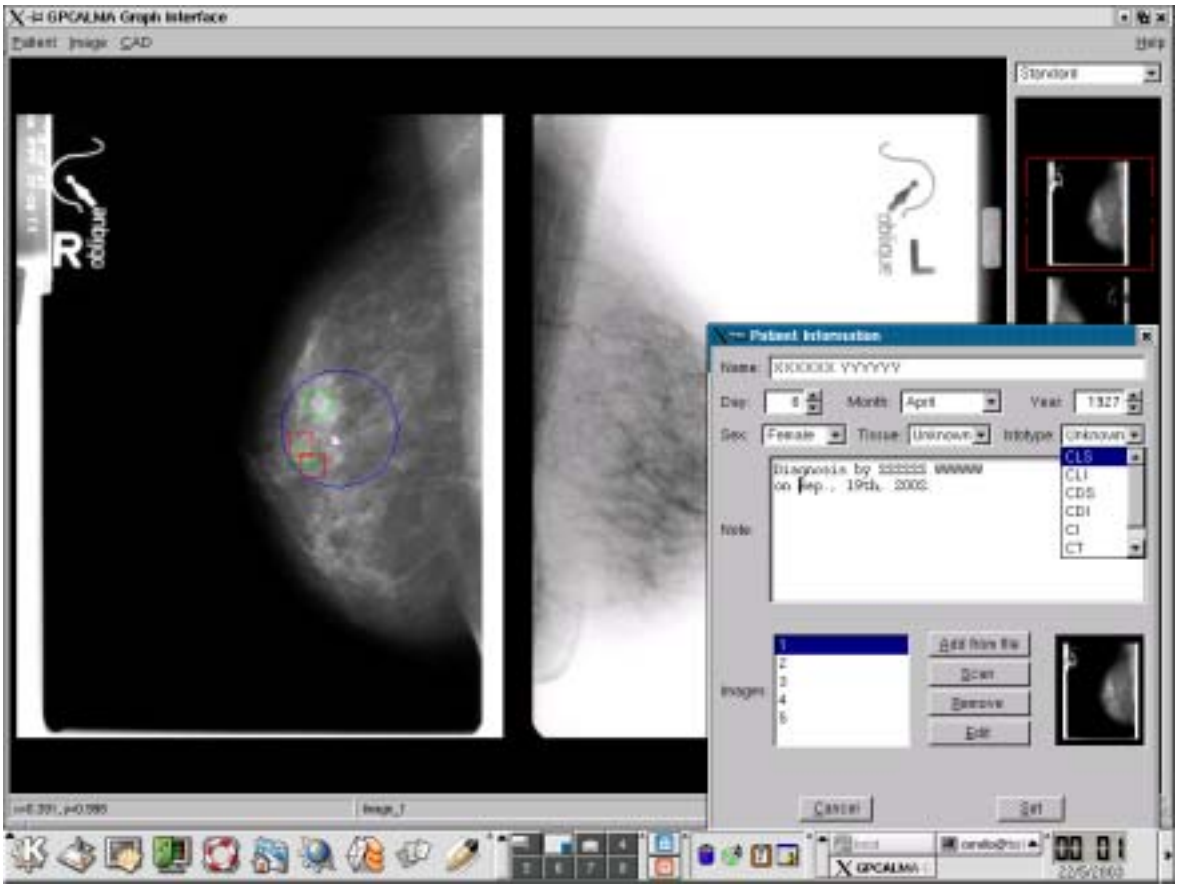

Figure 1: The GPCALMA Graphic User Interface. Part of the functionality is shown. Three menus are available, corresponding to the Patient, the Images, and the CAD diagnosis levels. On the left, the CAD results for microcalcifications and masses are shown in red squares and green circles, together with the radiologist's diagnosis (blue circle). On the right, the image colours are inverted. The widget allows to update the patient and image related metadata.

a large-scale screening program, as well as a medium term application of GRID technologies (and therefore a good feedback for the longer-term and larger scale LHC-GRID).

In fact, the data collection in a mammographic screening program will intrinsically create a distributed database, involving several sites with different functionality. One can imagine to distinguish between data collection sites and diagnostic sites, i.e. access points from where radiologists would be able to query/analyze the whole (or part of the) distributed database.

The problem scale, from the GRID point of view, is pretty similar to that of LHC projects. Considering Italy as an example, a full mammographic screening program would act on a target sample of about 6.8 million women, thus generating 3.4 millions mammographic exams/year. With an average size of 60 MB/exam, the amount of raw data would be in the order of 200 TB/year.

Therefore, a screening program on the European scale would be a data source comparable to one of the LHC experiments (1 $PB/year$).

GPCALMA was proposed in July 2001, with the purpose of developing a GRID application, based on technologies similar to those adopted by the CERN/ALICE Collaboration. These technologies, thanks to the use of the ROOT and PROOF tools [1], would allow the remote analysis of images: in other words, the algorithm for the image analysis would be shipped to the remote site, rather than moving the images to the radiologist's sites. Therefore, a preliminary selection of cancer candidates could be quickly performed and only mammograms with cancer probabilities higher than a given (dynamically fixed) threshold would be transferred to any of the diagnostic sites and interactively analysed by one or more radiologists.

The distributed database is being implemented through the connection of all the hospitals and research centers in GRID technology. In each hospital, local patients digital images are stored in the local database. The GPCALMA Virtual Organisation will allow each node to work transparently on all the distributed database data as well as local database data.

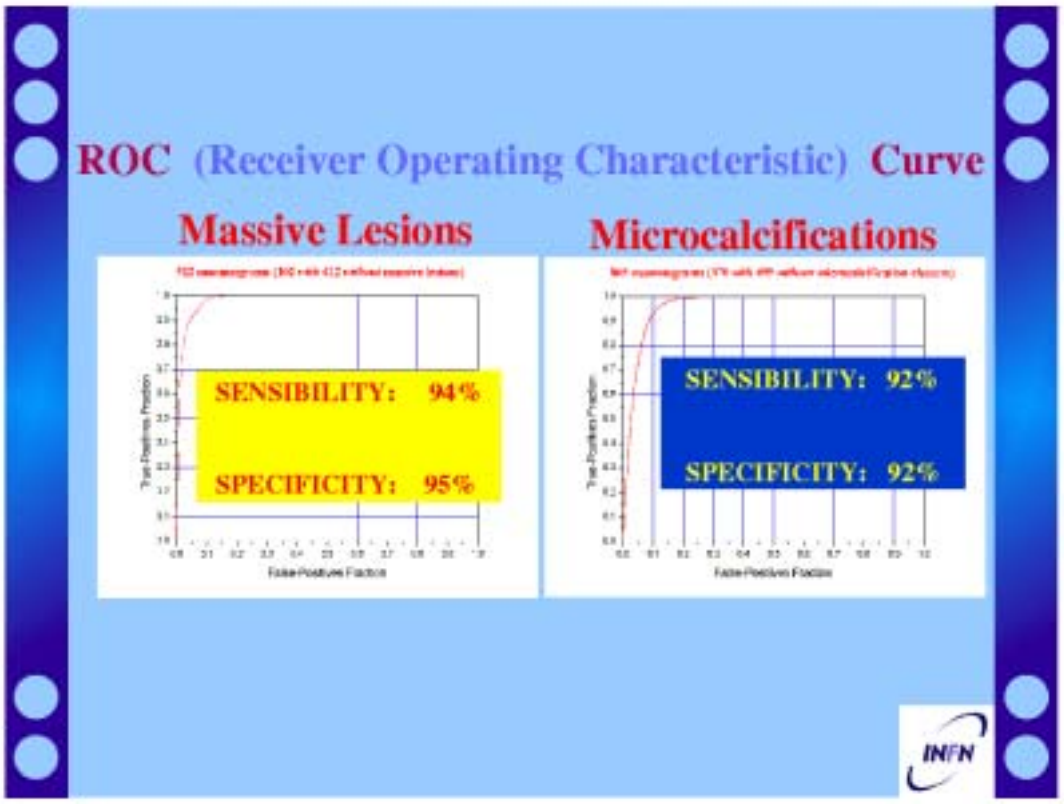

Figure 2: ROC curves for opacities and spikes (left, tuned on 515 images, 102 of them containing opacities) and microcalcifications (right, out of 865 images, 370 of them containing microcalcification clusters).

The database will be managed by the *AliEn* tool [2], which is being developed within CERN/ALICE and will also be used by the EU-MammoGrid project. The data associated to mammograms and stored in the database (also known as metadata) could also be used to define an input sample for any kind of epidemiology study.

Presently, a working version of the GPCALMA application is already available for local analysis. In parallel, a PROOF cluster of several PCs was already configured and successfully tested for the remote analysis of a given set of mammograms. As soon as an *AliEn* managed database will be available (a dedicated *AliEn* Server has already been configured [3]), the GPCALMA GRID-application will be able to dynamically select the input, making use of the *AliEn*-ROOT application program interface, whose first prototype was recently made available.

2. Mammographic Screening Programs

The early detection of breast cancer in asymptomatic women makes possible a reduction of breast cancer mortality [4]. At this moment an early diagnosis can be obtained thanks to screening programs, which consist in a mammographic examination performed for 49-69 years old women. It has been estimated that radiologists fail to detect up to approximately 25% of breast cancers visible on retrospective reviews and that this percentage increases if minimal signs are considered [5, 6]. Other studies show that the sensitivity (percentage of pathologic images correctly classified) and the specificity (percentage of non pathologic images correctly classified) increase if the images are independently analysed by two radiologists [7]. Therefore, independent double reading is now strongly recommended as it allows the reduction of the rate of false negative examinations by 5-15% [8, 9]. Recently, a number of Computer Aided Detection (CADe) systems [10] were developed: among them, the GPCALMA CADe. Its performance in the automated search of pathological masses and microcalcification clusters, as well as its comparison to a commercial CAD, will be discussed.

3. Computer Assisted Detection Tools: CALMA

The hardware requirements for the GPCALMA CAD Station are very simple: a PC with SCSI bus

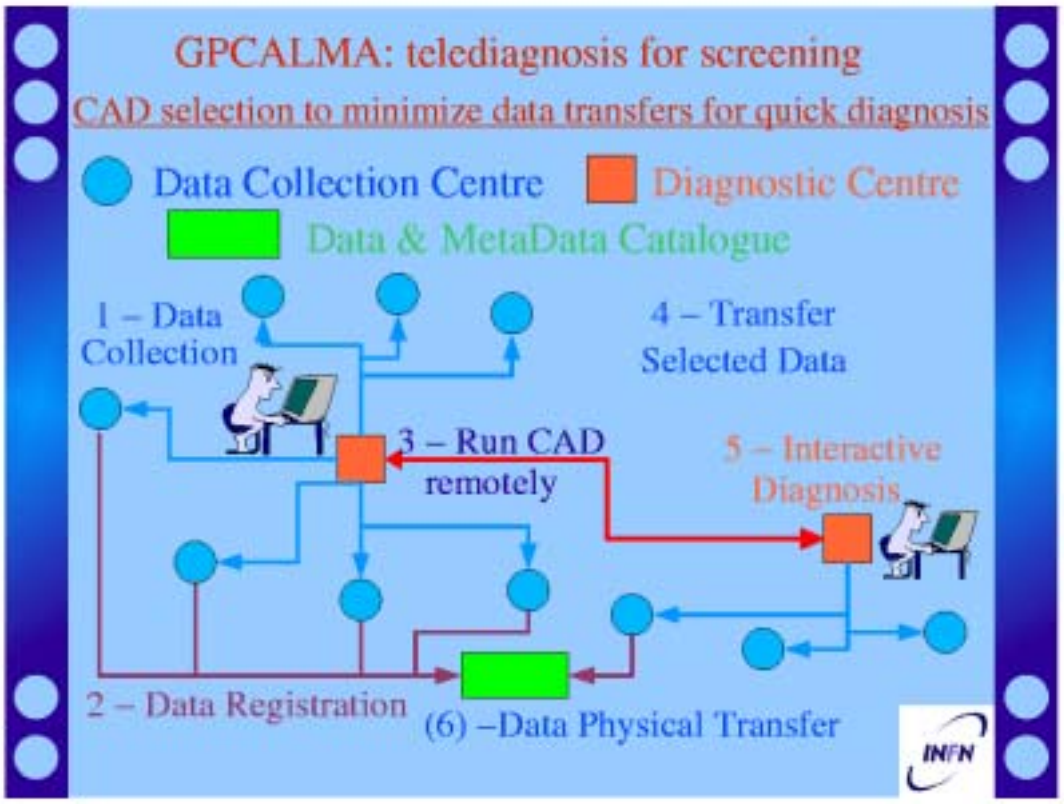

Figure 3: The screening use case: Data Collection Centres collect, store and register the images and the associated in the *AliEn* Data and MetaData Catalogue. Radiologists, from Diagnosis Centres, start the CAD remotely, without raw data transfer, making use of *PROOF*. Only the images corresponding to cancer probability larger than the selected threshold are moved to the Diagnosis Centre for the visual inspection. Eventually, the small fraction of undefined cases can be sent to other radiologists. The full data set must be stored and, later on, asynchronously analysed by radiologists locally or sent to a Diagnostic Centre on a DVD support.

connected to a planar scanner. Given the size of information to be stored (about 60 MB/mammographic exam), big disks and a CD/DVD-ROM recorder are recommended. A high resolution monitor helps the human radiological diagnosis. The station can process mammograms directly acquired by the scanner and/or images from file and allows human and/or automatic analysis of the digital mammogram. The images ($18x24$ cm^2), digitised by a CCD linear scanner with a 85 μm pitch and 4096 gray levels, are stored in 10.5 MB data files.

The software configuration for the use in local mode requires the installation of ROOT and GPCALMA, which can be downloaded either in the form of source code from the respective CVS Servers or in the form of a zipped tar file for Linux from the WEB sites.

The installation and configuration is fairly easy, as well as the update, whenever required.

The functionality is usually accessed through a Graphic User Interface, or, for developers, through the ROOT interactive shell.

3.1. Graphic User Interface

The Graphic User Interface (fig. 1) allows the acquisition of new data, as well as the analysis of existing ones. There are three main menus: two of them drive the creation of (access to) datasets at the patient and the image level; the third one is used to start the CADe algorithms and show their results on the image or to record a new diagnosis by a radiologist.

The images are displayed according to the standard format required by radiologists: for each of them, it is possible to insert or modify diagnosis and annotations, manually select the ROIs corresponding to the radiologists geometrical indication. An interactive procedure allows zooming, either continously or on a selected region, windowing, gray levels and contrast selection, image inversion, luminosity tuning.

The human analysis produces a diagnosis of the breast lesions in terms of kind, localization on the image, average dimensions and, if present, histological type.

The automatic procedure finds the regions of interest (ROIs) on the image which have a probability

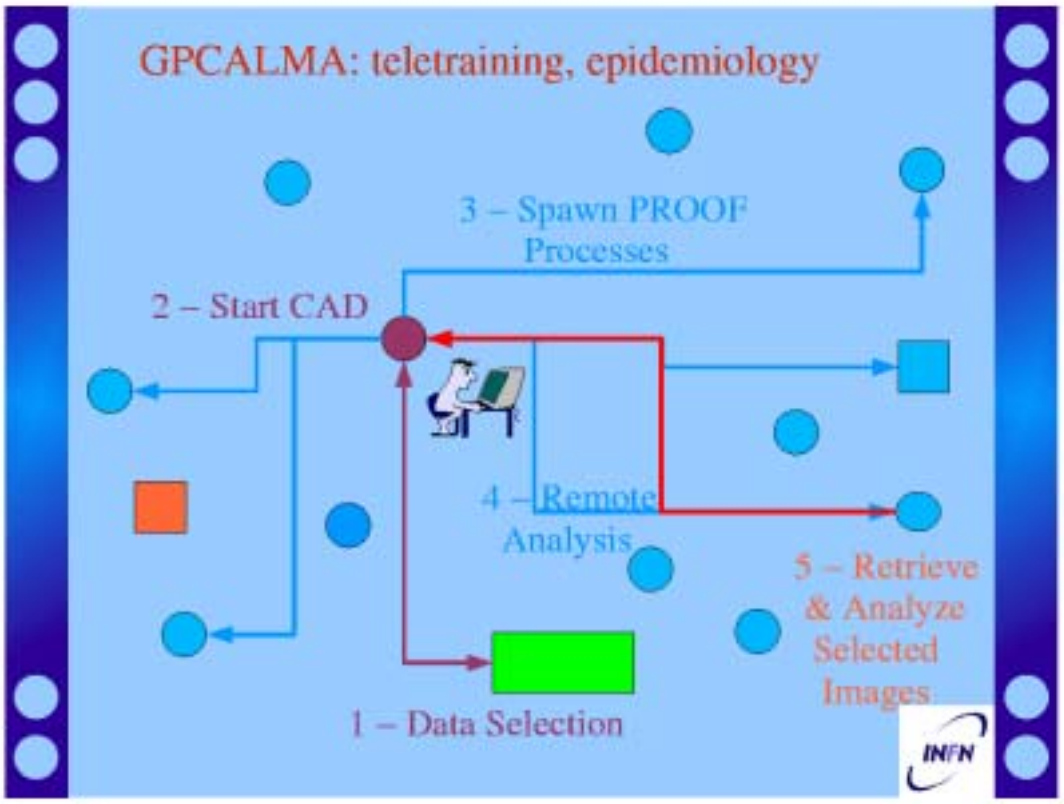

Figure 4: The teletraining and epidemiology use case. The CAD algorithm can be started remotely on a data set selected making use of the associated MetaData. In case the visual analysis is required (teletraining), the selected data are sent to the requesting site; otherwise, the CAD is run and the results are analysed for epidemiology studies.

(greater than a pre-selected threshold value) of containing an interesting area.

The operator can start the CADe analysis by choosing, from another window, the threshold for ROIs selection for microcalcifications or opacities. The station allows also for queries and statistical studies on the local database.

As soon as the distributed configuration will be fully operational, it will be possible to register new data in the Virtual Organisation Data Catalogue and to access the data already registered.

3.2. Opacities and Spiculated Lesions

Masses are rather large objects with very different shapes and show up with a faint contrast, slowly increasing with time. In the GPCALMA database, the average diameter of such lesions, as indicated by our radiologists, is about 2.1 *cm*. The Collaboration developed algorithms for the recognition of opacities in general and specifically for spiculated lesions, which present a particular spike shape.

The interesting areas are selected by the construction of a structure with concentric rings centered on local intensity maxima, until the average pixel value reaches a fixed threshold, thus identifying ROIs consisting of circles of radius R. As a further step, for the search for spiculated lesions, a spiral is unrolled around each maximum. For opacities, features are extracted by calculating the average intensity, variance and skewness (index of asymmetric distribution) of the pixel value distributions in circles of radius 1/3 R, 2/3 R and R, respectively. In the case of spiculated lesions, the number of oscillations per turn is calculated and processed by means of a Fourier Transform. These features are used as input to a feed-forward neural network (FFNN) which performs the final classification: the network has an output neuron whose threshold (i.e. a number between 0 and 1) represents the degree of suspiciousness of the corresponding ROI.

The FFNN previously described was trained with a set of 515 images (102 containing opacities, 413 without) and tested on a different set, also composed of 515 images (again 102 containing opacities and 413 without. The results of such classification are reported in the ROC (Receiver Operating Characteristics) curve (fig. 2), in which the sensitivity (true positives fraction) is shown as a function of 1−specificity (false positives fraction) for different threshold values. The best results correspond to values of 94% for the sensitivity

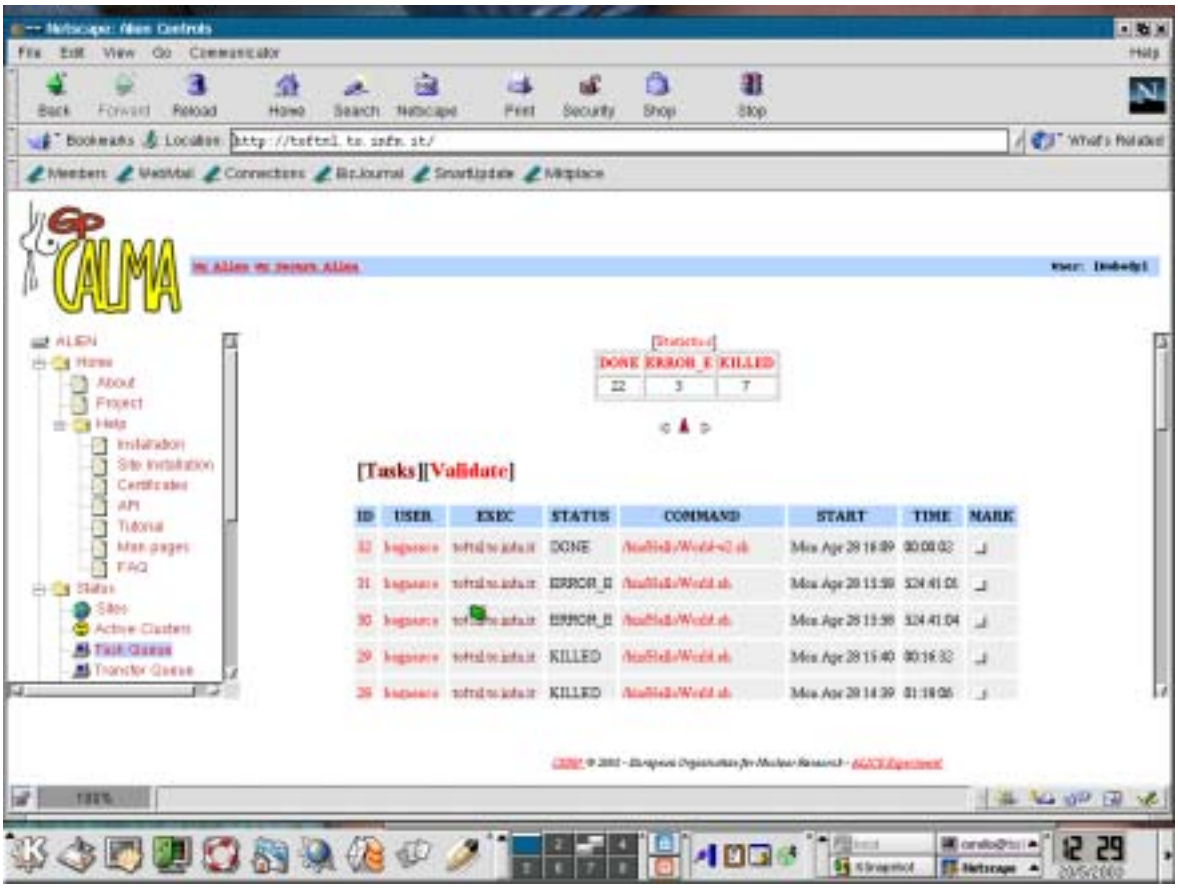

Figure 5: Screenshot from the GPCALMA AliEn WEB Portal. Making use of the left side frame, the site can be navigated. General Information about the AliEn project, the installation and configuration guides, the status of the Virtual Organisation Services can be accessed. The main frame shows, as an example, the status of the AliEn queue, with some sample jobs in different status.

and 95% for the specificity.

3.3. Microcalcifications clusters

A microcalcification is a rather small (0.1-1.0 mm in diameter) but very brilliant object. Some of them, either grouped in cluster or isolated, may indicate the presence of a cancer. In the GPCALMA database, the average diameter of microcalcification clusters, as indicated by our radiologists, is 2.3 cm. The microcalcification cluster analysis is carried on using the following approach:

- the digital mammogram is divided into $60x60$ pixels wide windows;
- the windows outside the breast image are rejected;
- the windows are statistically selected comparing the local and the global maxima;
- the windows are shrunk from $60x60$ to $7x7$ and are classified (with or without microcalcifications clusters) using a FFNN with 49 input, 6 hidden, and 2 output neurons;
- the windows are processed by a convolution filter to reduce the large structures;
- a self-organizing map (a Sanger's neural network) analyses each window and produces 8 principal components;
- the principal components are used as input of a FFNN able to classify the windows matched to a threshold (the response of the output neuron of the neural network);
- the windows are sorted by the threshold; at most three windows are memorized, if their threshold exceeds a given value; the selected windows are zoomed to 180x180 pixels, i.e. 15x15 mm2; the overlapping windows are clusterized.

The previously described procedure was run on a dataset of 676 images containing microcalcifitations and 995 images without microcalcification clusters. In

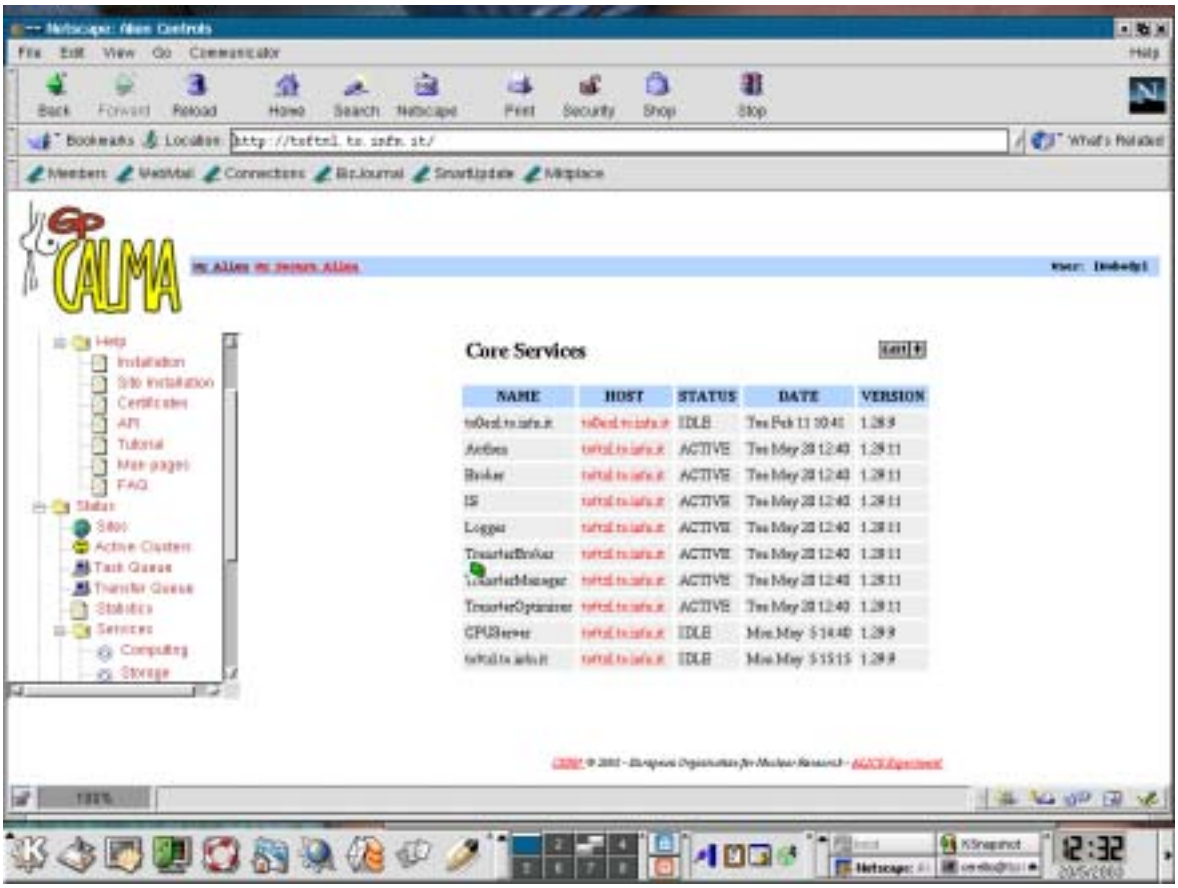

Figure 6: On the main frame, the list of the GPCALMA Virtual Organisation core services is shown, together with their status.

particular, 370 images with microcalcifications clusters and 495 without were used for the training phase and 306 with and 500 without for the test step. The results obtained in this test are summarized in the ROC curve reported in fi. 2, obtained by varying the threshold which drives the sorting of the interesting windows. The best results are 92% for both sensitivity and specificity.

4. Grid Approach

As remarked in the introduction, the amount of data generated by a national (italian) or european screening program is so large that they can't be managed by a single computing centre. In addition, data are generated according to an instrinsically distributed pattern; any hospital participating to the program will collect a small fraction of the data: still, that amount would be large enough to saturate the available network connections.

On the other hand, the availability of the whole database to a radiologist, regardless of the data distribution, would provide several advantages:

- the CAD algorithms could be trained on a much larger data sample, with an improvement on their performance, in terms of both sensitivity and specificity.
- the CAD algorithms could be used as real time selectors of images with high breast cancer probability (see fig. 3). Therefore, radiologists would be able to prioritise their work, with a remarkable reduction of the delay between the data acquisition and the human diagnosis (it could be reduced to a few days).
- data associated to the images (i.e., metadata) and stored on the distributed system would be available to select the proper input for epidemiology studies (fig. 4).
- data associated to the images would allow to select images relevant for the training of young radiologists (fig. 4).

These advantages would be granted by a Grid-like approach: the configuration of a Virtual Organisation, with common services (Data and Metadata Catalogue, Job Scheduler, Information System) and a number of

distributed nodes providing computing and storage resources would allow the implementation of the previously described use cases.

However, with respect to the model applied to High Energy Physics, there are some important differences:

- the network conditions do not allow the transfer of large amounts of data;
- the local nodes (hospitals) do not agree on the raw data transfer to other nodes as a standard procedure;
- some of the use cases require interactivity.

According to these restrictions, the GPCALMA approach to the implementation of a Grid application is based on two different tools: *AliEn* [2] for the management of common services, *PROOF* [1] for the interactive analysis of remote data without data transfer.

4.1. Distributed Database Management with AliEn

As previously described, the GPCALMA data model foresees several Data Collection Centres 3 , where mammograms are collected, locally stored and registered in the Virtual Organisation Data Catalogue. In order to make them available to a radiologist connecting from a Diagnostic Centre, it is mandatory to use a mechanism that identifies the Data Set corresponding to the exam in a site-independent way: the data must be selected by means of a set of requirements on the attached metadata and identified through a Logical Dataset Name which must be independent of their physical location. *AliEn* implements these features in its Data Catalogue Services, run by the Server, which allow the description of data in terms of a *Unix-Filesystem*-like hierarchical namespace for Logical Dataset Names and keep track of their association to the actual name of the Physical Dataset. In addition, it is possible to attach metadata to each level of the hierarchical namespace. The Data Catalogue is browsable from the *AliEn* command line as well as from the Web portal; the C++ Application Program Interface (*API*) to ROOT is under development by the *AliEn* and ROOT teams.

The virtual file system will contain the mapping of raw digitised mammograms and their graphical formats to the corresponding Logical Names: the hierarchical structure is being defined, and will take into account the location of data collection, the date, the identification of the person who collected them, etc.. The attached Metadata can be classified in several categories:

- patient identification data;
- exam identification data;
- results of the CAD algorithm analysis;
- radiologist's diagnosis;
- histological diagnosis.

Some of these data will be stored in the catalogue, but some of them may be stored as files and registered in the Data Catalogue: the decision will be made after a discussion with the radiologists.

Meanwhile, a small sample of mammograms will be registered in a prototype of Data Catalogue, in order to allow the testing of queries and data selection.

A dedicated *AliEn* Server for GPCALMA has been configured [3], in collaboration with the *AliEn* development team. Fig. 5 and 6 show screenshots from the WEB Portal, with the side menu allowing the access to the different services, and the main frame showing the configuration of the required system components (task queue, registered nodes, storage and computing elements, transfer queue, etc.).

4.2. Remote Data Processing with PROOF

Tele-diagnosis and tele-training require interactivity in order to be fully exploited, while in the case of screening it would be possible - altough not optimal - to live without. On the other hand, the PROOF *Parallel ROOt Facility* system allows to configure a distributed cluster and run interactive parallel processes on it, thanks to the C++ interpreter distributed with ROOT. A dedicated cluster of several PCs was configured and the remote analysis of a digitised mammogram without data transfer was recently run. As soon as selection through Metadata from the *AliEn* Data Catalogue will be possible, more complex use cases will be deployed. The basic idea is that, whenever a list of input LogicalFileNames will be selected, that will be split into a number of sub-lists containing all the files registered in a given StorageElement and each sub-list will be shipped the the corresponding node, where the process which analyses the mammogram will be started.

5. Present Status and Plans

The project is developing according to the original schedule. The CAD algorithms were rewritten in C++, making use of ROOT, in order to be PROOF-compliant; moreover, the ROOT functionality allowed a significant improvement of the Graphic User Interface, which, thanks to the possibility to manipulate the image and the associated description data, is now considered fully satisfactory by the radiologists involved in the project. The GPCALMA application

code is available via CVS server for download and installation; a script to be used for the node configuration is being developed. The *AliEn* Server, which will describe the Virtual Organisation and manage its services, is installed and configured; some *AliEn* Clients are in use, and they will soon be tested with GPCALMA jobs.

The remote analysis of mammograms was successfully accomplished making use of PROOF.

Presently, all but one the building blocks required to implement the tele-diagnosis and screening use cases were deployed. The only missing part is the implementation of the data selection from the ROOT shell through the *AliEn* C++ API: as soon as that functionality will be available, GPCALMA nodes will be installed in the participating hospitals and connected to the *AliEn* Server, hosted by INFN. Hopefully, that task will be completed by the end of 2004.

Acknowledgments

The authors wish to thank the *AliEn* development team for their support and guidance in the installation and configuration of the GPCALMA server.

References

[1] http://root.cern.ch.
[2] http://alien.cern.ch.
[3] http://gpcalma.to.infn.it.
[4] Lancet 2000, 355, 1822-1823.
[5] N. Karssemejer, "A stochastic method for authomated detection of microcalcifications in digital mammograms" in Information processing in medical imaging, Springer-Verlag New York, 227-238, 1991.
[6] N. Karssmejer, "Reading screening mammograms with the help of neural networks", Nederlands Tijdschriff geneeskd, 143/45, 2232-2236, 1999.
[7] S.A. Feig and M.Yaffe, Radiologic Clinics of North America, Vol.33 n.6, 1205, 1995.
[8] R.E. Bird, "Professional quality assurance for mammographic programs", Radiology 177, 587-592, 1990.
[9] E.L. Thurfjell, K.A. Lernevall, A.A.S. Taube, "Benefit of indipendent double reading in a population based mammography screening program" Radiology 191, 241-244, 1994.
[10] C.J. Viborny "Can computer help radiologists read mammograms?" Radiology 191, 315-317, 1994.