Radiol med (2008) 113:477–485
DOI 10.1007/s11547-008-0282-5

BREAST RADIOLOGY
SENOLOGIA

MAGIC-5: an Italian mammographic database of digitised images for research

MAGIC-5: un database mammografico italiano di immagini digitalizzate per scopi di ricerca

S. Tangaro[1] • R. Bellotti[1,2,3] • F. De Carlo[1,2] • G. Gargano[1,2] • E. Lattanzio[4] • P. Monno[4] R. Massafra[5] • P. Delogu[6] • M.E. Fantacci[6] • A. Retico[6] • M. Bazzocchi[7] • S. Bagnasco[8] P. Cerello[8] • S.C. Cheran[8,9] • E. Lopez Torres[10] • E. Zanon[11] • A. Lauria[12] • A. Sodano[13] D. Cascio[14] • F. Fauci[14] • R. Magro[14] • G. Raso[14] • R. Ienzi[15] • U. Bottigli[16] • G.L. Masala[17] P. Oliva[17] • G. Meloni[18] • A.P. Caricato[19] • R. Cataldo[20]

[1]Istituto Nazionale di Fisica Nucleare, Via Amendola 173, 70126 Bari, Italy
[2]Università di Bari, Dipartimento di Fisica, Italy
[3]TIRES, Center of Innovative Technologies for Signal Detection and Processing, Bari, Italy
[4]SARIS, Bari, Italy
[5]ASL, Bari, Italy
[6]Dipartimento di Fisica dell'Università di Pisa e INFN Sezione di Pisa, Italy
[7]ASL Pisa, Italy
[8]INFN, Sezione di Torino, Torino, Italy
[9]INFN, Sezione di Genova, Genova, Italy
[10]CEADEN, Habana, Cuba
[11]Ospedale Valdese di Torino and INFN, Sezione di Torino, Torino, Italy
[12]Università di Napoli "Federico II", Dipartimento di Scienze Fisiche, e Istituto Nazionale di Fisica Nucleare, Sezione di Napoli, Napoli, Italy
[13]ASL Napoli
[14]Università di Palermo, Palermo, and INFN, Sezione di Catania, Italy
[15]ASL Palermo, Italy
[16]Università di Siena, Siena, e Istituto Nazionale di Fisica Nucleare, Sezione di Cagliari, Italy
[17]Università di Sassari, e Istituto Nazionale di Fisica Nucleare, Sezione di Cagliari, Italy
[18]ASL Sassari, Italy
[19]Dipartimento di Fisica Università di Lecce and INFN, Lecce, Italy
[20]Dipartimento di Scienza dei Materiali, Università di Lecce e INFN, Lecce, Italy
Correspondence to: S. Tangaro, Tel.: +39-080-5443244, Fax: +39-080-5442470, e-mail: sonia.tangaro@ba.infn.it

Received: 1 August 2007 / Accepted: 21 September 2007 / Published online: 10 June 2008

Abstract
The implementation of a database of digitised mammograms is discussed. The digitised images were collected beginning in 1999 by a community of physicists in collaboration with radiologists in several Italian hospitals as a first step in developing and implementing a computer-aided detection (CAD) system. All 3,369 mammograms were collected from 967 patients and classified according to lesion type and morphology, breast tissue and pathology type. A dedicated graphical user interface was developed to visualise and process mammograms to support the medical diagnosis directly on a high-resolution screen. The database has been the starting

Riassunto
In questo lavoro viene discussa l'implementazione di un database immagini mammografiche digitalizzate. Le immagini sono state raccolte dal 1999 da un gruppo di fisici in collaborazione con radiology di alcuni ospedali italiani, come primo passo dello sviluppo e implementazione di un sistema di Computer Aided Detection (CAD). I 3369 mammogrammi appartengono a 967 pazienti e sono classificati secondo I tipi e la morfologia delle lesioni, il tessuto mammario e i tipi di patologie. Una interfaccia grafica opportunamente progettata è stata sviluppata per la visualizzazione e l'elaborazione delle mammografie digitalizzate al fine di

point for developing other medical imaging applications, such as a breast CAD, currently being upgraded and optimised for use in a distributed environment with grid services, in the framework of the Instituto Nazionale di Fisicia Nucleare (INFN)-funded Medical Applications on a Grid Infrastructure Connection (MAGIC)-5 project.

Keywords Database · Mammography · Medical image processing · Grid

poter supportare direttamente una diagnosi medica su monitor ad alta risoluzione. Il database ha rappresentato il punto di partenza per lo sviluppo di altre applicazioni di imaging medicale come il CAD mammografico costantemente ottimizzato e aggiornato con l'uso di un ambiente distribuito che dispone di servizi GRID, nel framework del progetto MAGIC-5, finanziato dell'INFN.

Parole chiave *Database · Mammografia · Elaborazione di immagini mediche · Grid*

Introduction

A medical image data set is the starting point for important epidemiological and statistical studies. Usually, it is used to develop and test algorithms for computer-aided detection (CAD) systems but it is used also for teaching and training medical students or as an archive of rare cases. In 1995, Osuch et al. [1] proposed a mammography database for a national mammography audit and to monitor patients through a centralised system. More recently, the Radiological Society of North America Medical Imaging Resource Center (RSNA MIRC) project [2] proposed a highly generic system to store and publish medical images, primarily for research and teaching support, that can federate a large number of remote databases and make them accessible as a single one. Technological improvements in digitising scanners now make it possible to digitise radiographic films, with no significant loss of information. Presently, many large data sets of digitised mammograms are available on the Web [3, 4]. Other grid-based databases are described in the literature [5, 6]. The development of a CAD system is strictly tied to collection of a large data set of selected images.

In this paper, a full description of the Medical Application on a Grid Infrastructure Connection (MAGIC)-5 database, whose collection started in a previous project, Grid Platform for Computer Aided Library in Mammography (GPCALMA), is given.

Materials and methods

Images were acquired in various mammographic centres using different mammographic screen/film systems and settings (all with molybdenum anode) in the framework of different applications, including both clinical routine carried out on symptomatic women and screening programmes addressing asymptomatic women. Moreover, many images come from an archive of particularly meaningful clinical cases collected in the previous years at the Bari hospital. A workstation, composed of a personal computer (PC) run-

Introduzione

Un dataset di immagini cliniche è la base per importanti studi epidemiologici e statistici: di norma, esso è usato per sviluppare e testare algoritmi per sistemi CAD, ma anche per l'insegnamento e l'addestramento di studenti di medicina o come archivio di casi rari. Nel 1995 Osuch et al. [1] proposero un database mammografico per una indagine mammografica a livello nazionale e per monitorare i pazienti attraverso un sistema centralizzato. Più recentemente il progetto Medical Imaging Resource Center (MIRC) dell'RSNA [2] ha proposto un sistema molto generale per immagazzinare e pubblicare immagini mediche, innanzitutto per supporto di ricerca e di studio, che potesse mettere insieme un gran numero di database remoti e renderli accessibili come un tutt'uno. I miglioramenti tecnologici negli scanner per digitalizzazione rendono ora possibile digitalizzare film radiografici con nessuna perdita significativa di informazione. Attualmente molti grandi dataset di mammografie digitali sono disponibili sul web [3, 4]. Altri database basati su Grid sono descritti in letteratura [5, 6]. Lo sviluppo di un sistema CAD è strettamente legato alla raccolta di un ampio dataset di immagini selezionate.

In questo lavoro viene data una completa descrizione del database di MAGIC-5 (Medical Application on a Grid Infrastructure Connection), la cui raccolta è cominciata in un progetto precedente (GPCALMA, Grid Platform for Computer Aided Library in Mammography).

Materiali e metodi

Le immagini sono state acquisite in vari centri di mammografia usando differenti sistemi e settaggi mammografici di tipo screen/film (tutti con anodo di molibdeno), nell'ambito di varie applicazioni, incluse sia procedure cliniche su donne sintomatiche, sia programmi di screening rivolti a donne asintomatiche. Inoltre, molte immagini provengono da un archivio di casi clinici particolarmente significativi, raccolti negli anni passati all'Ospedale di Bari. Una stazione di

ning the Linux operating system, a film scanner and a dedicated disk, was installed at each site involved in the programme. The parameters of the charge-coupled device (CCD) scanners used were [7] a pixel size of 85 μm and a 12-bit depth (4,096 grey levels). The typical scan time is 20 s. Each image was initially saved in a file with a special format (called Calma) consisting of a header including the information on row and column number and a sequence of bytes with pixel intensity. These numbers are used to transform the vector in a matrix: each pixel of the image can be represented by a triplet (R, C, I) – where R is the row number, C is the column number and I is the intensity of the pixel – ranging from 0 (black) to 4,095 (white). Such workstations have been continuously operative in various collaboration sites for several years without problems. In sites where clinical studies were performed, the PC was connected to a high-resolution and high-luminosity black-and-white liquid crystal display (LCD) monitor. All mammograms in the Calma database that were digitised with a scanner are available in Digital Imaging and Communications in Medicine (DICOM) file format [8]. Data are not stored as DICOMDIR, as we want to preserve the granularity at the level of the single image to be able to process each image individually and to run parallel analysis on distributed data. Each exam is stored in a directory, which contains one DICOM file per image, including a collection of tags, values, types, length and value fields describing patient information, imaging procedure information and other image related information. The diagnosis is provided according to classification criteria proposed by the American College of Radiology [9] and, if required, is available as a text file.

The patient related information includes an identification string – UID – gender and birth date. The name is cleared on purpose to conform to privacy requests. Correlation between the exam UID and patient identity is kept in a private local database on each site.

The image-related information includes image type (scanned/digital), study date, study identification (ID), series description, image laterality, view position, series number, pixel pitch and name and address of the institution where taken. When available, information about the device manufacturer, acquisition and calibration parameters (model name, tube voltage and current, anode features) is also provided.

The database is still growing, mostly including new cases that are now mainly from digital mammograms: in that case, the information tags of the DICOM file are all those provided by the acquisition device.

Database description

The database is composed of 3,369 mammographic images, each including data and clinical information. Images were

lavoro composta da un PC con sistema operativo Linux, uno scanner per film e un disco dedicato, è stato installato in ciascun sito coinvolto nel programma. I parametri degli scanner CCD usati sono stati [7]: la misura del pixel di 85 Ìm e una profondità di 12 bit (4096 livelli di grigio). Il tipico tempo di scansione è di 20 s. Ciascuna immagine è stata inizialmente salvata in uno speciale formato (chiamato formato CALMA), che consiste in un header che include le informazioni su numero di riga e colonna e una sequenza di bytes con l'intensità del pixel. Questi numeri sono usati per trasformare il vettore in una matrice: ciascun pixel dell'immagine può essere rappresentato da una tripletta (R, C, I), dove R è il numero di riga, C il numero di colonna e I l'intensità del pixel, che varia tra 0 (nero) a 4095 (bianco). Tali workstations sono state continuamente operative in vari siti siti della collaborazione per alcuni anni senza problemi. Nei siti ove sono stati condotti i studi clinici, il PC è stato collegato ad un monitor B/W LCD ad alta risoluzione e alta luminosità. Tutte le mammografie nel database CALMA che sono state digitalizzate con uno scanner sono disponibili nel formato DICOM [8]. I dati non sono stati conservati come DICOMDIR poiché volevamo preservare la granularità a livello di singola immagine, in modo tale da poter processare ciascuna immagine individualmente e gestire analisi parallele su dati distribuiti. Ciascun esame è archiviato in una directory, che contiene un file di tipo DICOM per immagine, con un insieme di etichettte, valori, tipi, lunghezze e valori di campi che descrivono informazioni del paziente, informazioni sulle procedure di imaging e altre informazioni sull'immagine. La diagnosi è fornita secondo criteri di classificazione suggeriti dall'American College of Radiology [9], e, se richiesto, è disponibile un file di testo.

L'informazione relativa al paziente include una stringa di identificazione (UID), il sesso e la data di nascita. Il nome è in chiaro solo se lo si richiede allo scopo di uniformarsi alle richieste di privacy: la correlazione tra l'esame UID e l'identità del paziente è conservata in un database locale privato su ciascun sito.

L'informazione relativa all'immagine include il tipo di immagine (scansionata/digitale), la data dell'esame, l'ID dell'esame, la descrizione della serie, il lato dell'organo indagato, la vista, il numero di serie, il passo del pixel, il nome e l'indirizzo dell'istituto dove è stato effettuato l'esame. Quando disponibile, vengono forniti anche l'informazione sulla casa costruttrice dello strumento, i parametri di acquisizione e calibrazione (nome del modello, alimentazione del tubo e corrente, caratteristiche dell'anodo).

Il database è tutt'ora in crescita, vengono inclusi nuovi casi che sono adesso per la maggiore parte provenienti da mammografi digitali: in questo caso le informazioni del file DICOM sono quelle provenienti dallo strumento di acquisizione.

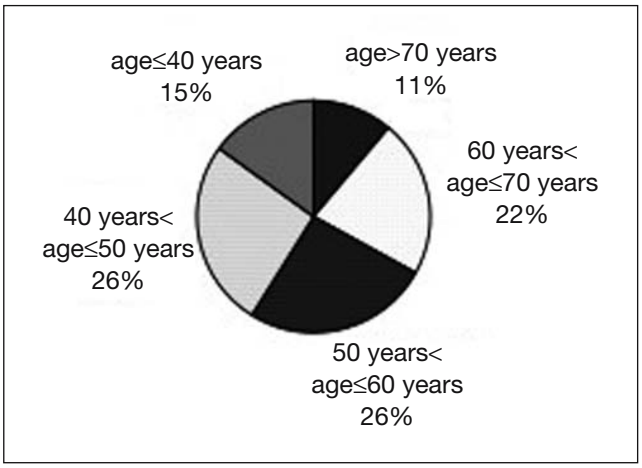

Fig. 1 Age groups of the analysed patients.

Fig. 1 *Gruppi di età dei pazienti analizzati.*

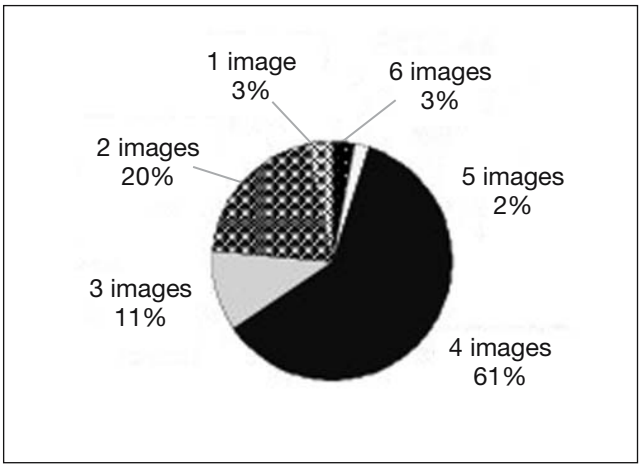

Fig. 2 Number of cases with one to six views.

Fig. 2 *Numero di casi con 1–6 immagini.*

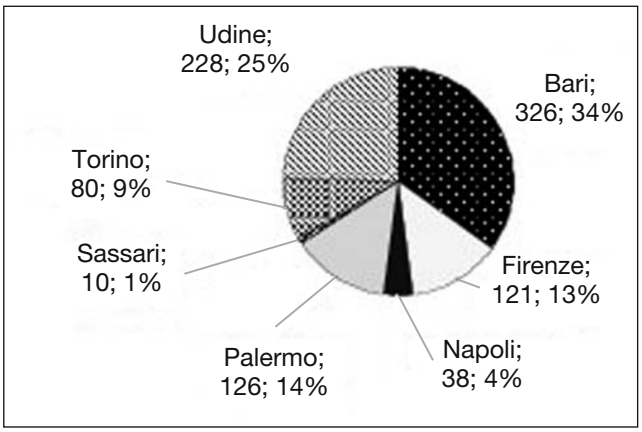

Fig. 3 Geographical provenance of the images within the Medical Applications on a Grid Infrastructure Connection (MAGIC)-5 project.

Fig. 3 *Provenienza geografica delle immagini del progetto MAGIC-5.*

collected from 967 patients. The age groups are reported in Fig. 1. About 60% of patients in this database is more 50 years old. Each exam contains one to six views according to the distribution shown in Fig. 2. The repartition of the database in left/right breast images is 1,835 (51%) and 1,734 (49%), respectively, whereas for the craniocaudal/oblique/lateral views, it is 1,601 (48%), 1,456 (43%) and 312 (9%), respectively. Image size is 2,067×2,657 pixels, 85 μm of pitch (300 dots/in.); each image file size, in Calma format, is about 11 Mbytes. All mammographic images with extra information related to the patient (follow-up, age, and interesting cases) were collected in the Italian hospitals involved in the collaboration from 1997 to 2002. The geographical provenance of the images is shown in Fig. 3.

Prior to being processed, all images were anonymised. All the images containing one (or more) lesions were classi-

Descrizione del database

Il database è composto da 3369 immagini mammografiche, ciascuna delle quali include dati e informazioni cliniche. Le immagini provengono da 967 pazienti. I gruppi di età sono riportate in Fig. 1. Circa il 60% dei pazienti del database ha più di 50 anni. Ciascun esame contiene da una a sei immagini la cui distribuzione è riportata in Fig. 2. La ripartizione delle immagini in destre/sinistre è 1835 (51%) and 1734 (49%) rispettivamente, mentre la ripartizione cranio-caudale/obliqua/laterale è 1601 (48%), 1456 (43%) e 312 (9%) rispettivamente. La dimensione delle immagini è 2067×2657 pixels, 85 μm di passo (300 punti/pollice); la dimensione di ogni file, nel formato CALMA, è circa 11 MBytes. Tutte le immagini mammografiche, con le relative informazioni extra (follow-up, età del paziente e casi di interesse), sono state raccolte dal 1997 al 2002 negli ospedali italiani coinvolti nella collaborazione. La dislocazione geografica della provenienza delle immagini è riportata in Fig. 3.

Prima di essere analizzate tutte le immagini sono state rese anonime. Tutte le immagini del database, contenenti una o più lesioni sono state classificate secondo il tipo di lesione (massa o microcalcificazione), il suo grado di malignità, il tipo di tessuto della mammella, ecc.

Nel database ci sono 306 (32%) immagini provenienti da soggetti definiti normali (BI-RADS R1) ovvero senza alcuna evidenza di lesione (confermato da tre anni di follow-up radiologico). Le rimanenti immagini provengono da 661 (68%) soggetti patologici: quando una lesione viene trovata dal radiologo, questa è classificata come sospetta, benigna o maligna. Per tutte le lesioni maligne sono disponibili anche i risultati degli esami citologici e istologici. In questo caso dettagliate annotazioni radiologiche delle anormalità sono riportate nel database come note. La distribuzione re-

Table 1 Histotypes as classified in the database

Histotype	Number
Invasive lobular carcinoma	17
Lobular carcinoma in situ	5
Ductal invasive carcinoma	127
Ductal in situ carcinoma	27
Papilloma intraductal	1
Dysplasia	2
Fibrosis	6
Fibroadenoma	7
Fibrocystic disease	2
Sclerosing adenosis	2
Epithelial hyperplasia	3
Adenosis	5
Tubular carcinoma	10
Mucinous carcinoma	5
Total	219

Tabella 1 *Istotipi presenti nel database*

Istotipo	*Numero*
Carcinoma lobulare invasivo	*17*
Carcinoma lobulare in situ	*5*
Carcinoma duttale invasivo	*127*
Carcinoma duttale in sito	*27*
Papilloma intraduttale	*1*
Displasia	*2*
Fibrosi	*6*
Fibroadenoma	*7*
Malattia fibrocistica	*2*
Adenosi sclerosante	*2*
Iperplasia epiteliale	*3*
Adenosi	*5*
Carcinoma tubulare	*10*
Carcinoma mucinoso	*5*
Totale	*219*

fied according to the kind of lesion (mass or microcalcification), malignancy grade, breast texture type, etc.

In this database, there are the images from 306 (32%) patients who were defined as normal [Breast Imaging Reporting and Data System (BI-RADS) R1] when there was no evidence of any lesion (confirmed by 3 years of radiological follow-up). The remaining images are from 661 (68%) "abnormal" patients: when a suspicious lesion was found by the radiologist in these images, it was classified as suspicious, benign or malignant. For all malignant lesions, cytological or histological results are also available. Detailed radiological annotations of abnormalities are included in the database as notes. The relative distribution of malignancy grade is 560 (35%) suspicious lesions, 468 (29%) benign lesions and 592 (37%) malignant lesions. Table 1 shows the histotypes related only to some patients. Images that contain at least one mass or a cluster of microcalcifications, as diagnosed by an expert radiologist, are considered abnormal.

There are 1,062 images containing at least one region of interest (ROI) with a massive lesion and 304 images containing at least one ROI with microcalcifications. In total, there are 1,296 (38%) abnormal images containing at least one lesion (massive or microcalcification or both) and 2,073 (62%) normal images with no lesions. Each image can also contain more than one lesion, so the total number of ROIs is 1,620 (1,236 massive and 384 microcalcification).

Each of these main classes of lesions (microcalcification clusters and massive lesions) is further classified according to the morphological characteristics of the lesion. We adopted the scheme of Lattanzio and Guerrieri [10], which has been recognised as a satisfactory reference framework by a national panel of radiologists, with more than 20 years of experience in mammography, who identified and localised each lesion according to this classification. Each abnormal image comes with a description of the lesion, as shown in

lativa del grado di malignità delle lesioni è: 560 (35%) sospette, 468 (29%) benigne e 592 (37%) maligne. La Tabella 1 riporta gli istotipi relativi ad alcuni pazienti. Le immagini contenenti almeno una massa o un cluster di microcalcificazioni, secondo la diagnosi del radiologo, sono considerate anormali.

Il database contiene 1062 immagini con almeno una regione di interesse (ROI) con una lesione massiva e 304 images con almeno una ROI contenente un cluster di microcalcificazioni. In totale ci sono 1296 (38%) immagini patologiche contenenti almeno una lesione (massa o microcalcificazione o entrambi) e 2073 (62%) immagini normali, senza lesione alcuna; un'immagine può contenere anche più di una lesione, per cui il numero totale di ROI è 1620 (1236 masse e 384 micro calcificazioni).

Ciascuna di queste principali classi di lesioni (cluster di microcalcificazioni e lessioni massive) viene ulteriormente classificata secondo le caratteristiche morfologiche. Nel nostro database viene adottato lo schema di Lattanzio e Guerrieri [10], che è stato riconosciuto come un valido schema di riferimento da un certo numero di radiologi italiani, con più di 20 anni di esperienza, che identificano e localizzano le lesioni secondo tale classificazione. Ciascuna immagine patologica contiene una descrizione delle lesioni come mostrato nelle Tabelle 2 e 3, in cui è riportata la suddivisone delle ROI per diversi tipi di massa e di microcalcificazione, con il corrispondente numero di immagini che presentano quel dato tipo di lesione.

La posizione e la grandezza della massa sono definite da un cerchio tracciato dal radiologo, con le coordinate del centro {Xrad; Yrad} e il raggio {Rrad}, contenente per intero la massa. I raggi delle masse variano da 3,1 mm a 47,2 mm, con una media di 11,7 mm, mentre il raggio delle microcalcificazioni varia da 1 mm a 72,8 mm con una media di 11,9 mm.

Table 2 Different kinds of masses present in the database; "Others" includes a combination of the types mentioned in the text. The table indicates the number of images containing each kind of lesion

Massive lesions Type	Regions of interest	Images
Irregular roundish opacity	406	369
Spiculated opacity	294	261
Regular roundish opacity	289	210
Parenchymal distortion	111	109
Blurred roundish opacity	47	41
Fibroadenoma	29	29
Others	58	43
Total	1,236	1,062

Tabella 2 *Differenti tipi di masse presenti nel database; "altre" include una combinazione dei tipi citati. È riportato il numero di immagini contenente ciascun tipo di massa*

Masse Tipo	*Regioni di interesse*	*Immagini*
Opacità rotonda irregolare	*406*	*369*
Opacità spiculata	*294*	*261*
Opacità rotonda regolare	*289*	*210*
Distorsione parenchimale	*111*	*109*
Opacità rotonda sfumata	*47*	*41*
Fibroadenoma	*29*	*29*
Altre	*58*	*43*
Totale	*1236*	*1062*

Table 3 Different kinds of microcalcifications present in the database. The table indicates the number of images containing each kind of lesion

Microcalcifications Type	Regions of interest	Images
Glandular	163	124
Mixed	99	73
Lobular	9	8
Scattered	57	45
Ductal	10	10
Teacup	37	37
Eggshell	6	4
Tubular	3	3
Total	384	304

Tabella 3 *Differenti tipi di micro-calcificazioni presenti nel database. È riportato il numero di immagini contenente ciascun tipo di lesione*

Microcalcificazioni Tipo	*Regioni di interesse*	*Immagini*
Ghiandolare	*163*	*124*
MIsto	*99*	*73*
Lobulare	*9*	*8*
Sparso	*57*	*45*
Duttale	*10*	*10*
Teacup	*37*	*37*
Eggshell	*6*	*4*
Tubulare	*3*	*3*
Totale	*384*	*304*

Tables 2 and 3, in which the partition of the ROIs for different kind of massive lesions and microcalcifications, with the corresponding number of images from which each kind of lesions comes, is reported.

Mass location and size is defined by a radiologist-drawn circle, characterised by centre coordinates (Xrad; Yrad) and radius (Rrad), which fully contains the mass. Mass radii range from 3.1 mm to 47.2 mm, with an average value of 11.7 mm, whereas the radius of the microcalcification clusters ranges from 1 mm to 72.8 mm, with an average value of 11.9 mm.

Another important parameter to characterise the image is breast tissue type. Collaborating radiologists were asked to identify breast texture for a full-image characterisation. As far as the breast background is concerned, in the MAGIC-5 database, we adopt a tissue classification recognised as a standard by many Italian radiologists [11, 12]:

1. Fibroadipose tissue indicates a fatty breast with little fibrous connective tissue (dense tissue percentage <25%);
2. Glandular tissue indicates the presence of prominent duct patterns (dense tissue percentage <25%–75%)

Un altro importante parametro dell'immagine è il tipo di tessuto della mammella. Ai radiologi della collaborazione è stato chiesto di identificare il tessuto mammario per una completa caratterizzazione dell'immagine mammografica. Per quanto riguarda il tessuto mammario presente nel database MAGIC-5, la classificazione adottata è quella riconosciuta come standard da molti radiologi italiani [11, 12]:

1. *Tessuto fibro-adiposo: indica un seno grasso con poco tessuto connettivo fibroso (percentuale di tessuto denso<25%).*
2. *Tessuto ghiandolare: indica la presenza preminente di ghiandole (percentuale di tessuto denso compreso fra 25% e 75%).*
3. *Tessuto denso: indica un parenchima mammario denso (percentuale di tessuto denso>75%).*

La classificazione del tessuto mammario è basata solo sull'apparenza del parenchima, senza tener conto del tipo di pelle, della vascolarità, della presenza/assenza di masse e/o calcificazioni, dei linfonodi, né della corrispondenza, della storia della malattia, dell'età e della storia familiare.

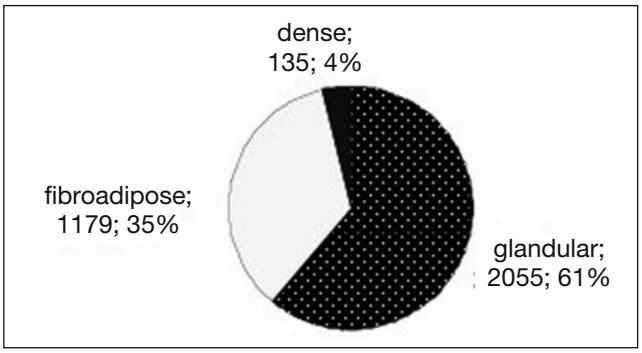

Fig. 4 Breast tissue composition of the database.

Fig. 4 *Composizione del database relativamente al tessuto mammario.*

3. Dense tissue indicates a dense breast parenchyma.(dense tissue percentage >75%)

Breast background classification is based only on the appearance of the parenchyma, without any reference to skin, vascularity, presence/absence of masses, calcifications or lymph nodes or to parity, history of breast disease, age and family history.

Fig. 4 reports the background composition of the database. Most of the images are glandular like: the detection of pathological structures in this kind of image is a relatively hard task, as the target is surrounded by a "noisy" environment. Information on breast background is transferred from each patient to each single view, because the CADs tested on this database analyse single views independently.

The database presents some limitations, especially from an epidemiological point of view. Images were collected in different clinical and screening conditions, so they do not represent a typical distribution of masses and microcalcifications in terms of ratio of benign to malignant cases; moreover, they were collected from different centres and were acquired with different mammography units under different conditions. However, the uniformity of the digitisation makes the database technically suitable for research studies devoted to the application of CAD procedures and also in a epidemiologically heterogeneous scenario.

The database represents the largest Italian sample of digitised mammograms. It has been successfully used for developing and testing the MAGIC-5 CAD system [7, 13, 14], and integrated into a grid environment [15] so that distributed data can be remotely accessed and analysed. It has been the basis of an experimental study about the peculiarities of computer display for reporting [16], and it has been used to test the performance of different CAD systems as second readers [17]. Moreover, it has been successfully used [18] for developing a CAD system for identifying microcalcification clusters in a pan-European-distributed database of mammograms in the framework of the MammoGrid project [19–21], after rescaling the images in terms of pitch and dynamic range, in the training phase of the neural-based classification analysis.

La Fig. 4 riporta la partizione del database sulla base del tessuto mammario. La maggioranza delle immagini sono di tipo ghiandolare, cosa che rende particolarmente difficile la rilevazione delle strutture patologiche dal momento che le stesse sono circondate da un fondo "rumoroso". L'informazione sul tessuto mammario è trasferita da ciascun paziente a ciascuna immagine estratta da quel paziente per l'analisi poiché il CAD testato su questo database analizza ogni singola vista indipendentemente.

Il database presenta alcune limitazioni, specialmente dal punto di vista epidemiologico: le immagini sono state raccolte in condizioni cliniche differenti, per cui non riproducono la tipica distribuzione di masse/microcalcificazioni in termini di rapporto benigne/maligne; inoltre, essendo state raccolte in differenti centri, le immagini sono state acquisite con diversi mammografi e in differenti condizioni. In ogni caso, l'uniformità del processo di digitalizzazione rende il database adatto a studi di ricerca focalizzati all'applicazione di CAD anche in presenza di uno scenario eterogeneo dal punto di vista epidemiologico.

Il database rappresenta il più ampio campione di immagini mammografiche in Italia ed è stato utilizzato con successo per lo sviluppo e il test dei sistemi CAD della collaborazione MAGIC-5 [7, 13, 14], nonché integrato in un ambiente Grid [15], in modo che sia possibile l'accesso e l'analisi in remoto delle immagini e dei dati distribuiti. Il database è stato alla base degli studi sperimentali circa le peculiarità della refertazione a video [16] ed è stato usato per verificare l'efficacia di diversi CAD come secondo lettore [17]. Inoltre è stato utilizzato con successo [18] per lo sviluppo di un sistema CAD per l'identificazione di cluster di microcalcificazioni all'interno di un database distribuito europeo di immagini mammografiche nell'ambito del progetto MammoGrid [19–21], dopo aver riscalato le immagini in termini di passo e range dimanico, nella fase di training di una classificazione neurale.

L'approccio Grid

Al crescere delle dimensioni del database la natura distribuita della collaborazione pone un ulteriore problema: analogamente a quanto avviene in un programma di screening, i dati sono raccolti in siti geograficamente separati e generalmente non replicati tra un sito e l'altro. L'approccio scelto per risolvere il problema dell'accesso remoto a dati distribuiti è l'uso di tecniche e strumenti sviluppati per la realizzazione di griglie computazionali [15]. Nel paradigma della Grid ("spostare i programmi invece dei dati") un sito che ospita una frazione del database distribuito deve fornire una certa quantità di potenza di calcolo condivisa oltre alle risorse di storage. Le tecnologie Grid non rendono pos-

The grid approach

The growth of the database and the distributed nature of the collaboration raises a problem. Similarly to what happens in a screening programme, data are collected from several geographically remote sites and are generally not replicated between sites. The approach used to solve the problem of remote access was to use techniques and tools developed for grid computing [15]. In the grid paradigm ("move code rather than data"), a site hosting a fraction of the distributed database should provide not only storage resources but also some shared computational power. Grid technologies thus allow not only interactive online diagnosis, but also image analysis locally with respect to the data, with a relevant reduction of the delays presently associated with the diagnosis in screening programmes. Also, they can, for example, allow researchers with limited access to computational resources or network bandwidth to remotely process a large database, possibly with some degree of parallelism. Privacy policies are easily enforced, as in some use cases, the original images never leave the local network of the site.

This approach sets truly grid-based projects aside from purely database-oriented projects, such as the RSNA MIRC initiative [2], which offer a platform to organise and publish teaching file cases but always require the full download of any file to be examined.

A virtual organisation (VO) has been deployed so that authorised users can share data and resources and implement screening, teletraining and telediagnosis use cases for mammograms. A small-scale prototype of the required grid functionality was already implemented for analysing digitised mammograms [22]. From the grid point of view, it is based on the aforementioned model in which input data are not moved and their analysis is run in parallel on the nodes where they are stored and, if possible, interactively. 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.

Integration of tools for remote disk storage access into the CAD system has successfully been tested: a prototype that makes it possible to share data between the different sites of the research and to run CAD from remote sites has been built [8, 15]. The next step is to transfer the prototype into a clinical environment, involving radiologists collaborating in the project, to implement telediagnosis and telescreening.

Conclusions

The database collected represents a useful archive of digitised mammographic images. It can be a valuable tool to the scientific community for different tasks such as training and

sibile solo la diagnosi interattiva on-line, ma anche l'esecuzione asincrona di analisi di immagini localmente rispetto ai dati, potenzialmente riducendo i tempi di attesa per la diagnosi nei programmi di screening; o possono permettere a ricercatori con limitato accesso a risorse di calcolo o di rete, ad esempio, di processare remotamente un database di grandi dimensioni, eventualmente anche con qualche grado di parallelismo. Inoltre, l'implementazione di criteri di privacy e riservatezza può essere semplificata, dato che in alcuni use case le immagini originali non escono mai dalla rete locale del sito che le ospita.

Questo approccio, che contraddistingue i progetti basati sulle tecnologie Grid, li differenzia dai progetti orientati alla sola raccolta e pubblicazione di database, come l'iniziativa MIRC del RSNA [2], che offre una piattaforma per organizzare e rendere accessibili on-line raccolte di immagini mediche digitali, ma richiede sempre il download completo dei file da esaminare.

Per la realizzazione dell'infrastruttura Grid è stata creata una Virtual Organization (VO), i cui membri possono essere autorizzati a condividere dati e risorse per implementare use case di screening, tele-diagnosi e tele-training in ambito mammografico. Un prototipo su piccola scala, con tutte le funzionalità richieste, era già stato realizzato per l'analisi di mammografie digitalizzate [22]. Il prototipo implementa il menzionato modello in cui a spostarsi da un sito all'altro è il codice eseguibile invece delle immagini, che vengono analizzate in parallelo e, quando possibile, interattivamente, nei diversi siti in cui sono conservate. Da questo punto di vista, la collaborazione può essere vista come una VO con servizi comuni (Data e Metadata Catalogue, Job Scheduler, Information System) ospitati su un server centrale ed un certo numero di nodi distribuiti (Client) che forniscono le risorse di storage e calcolo.

L'integrazione nella stazione CAD delle funzionalità per l'accesso alle risorse Grid è stata realizzata in un prototipo che rende possibile la condivisione di dati tra i siti che compongono l'infrastruttura e l'esecuzione remota da un sito all'altro di algoritmi CAD [8, 15]. Il passo successivo sarà l'installazione del prototipo in ambiente clinico, con il coinvolgimento dei radiologi che partecipano alla collaborazione, per implementare gli use case di tele-diagnosi e telescreening.

Conclusioni

Il database raccolto rappresenta un utile archivio di immagini mammografiche digitalizzate, che può efficacemente essere usato dalla comunità scientifica per svariati scopi come l'addestramento di strumenti di classificazione basati su reti neurali [23–25], per uso didattico e per studi statistici ed epidemiologici.

testing of neural-network-based classification tools [23–25], for retrieval use and for statistics and epidemiology studies.

As in a screening programme, data are collected from geographically remote sites. The growth of the database and the distributed nature of the collaboration raises a problem, however, as images are generally not replicated between remote sites. The approach used to solve the problem of remote access was to use techniques developed for grid computing. Its integration in a grid computing environment makes possible the implementation of several remote-analysis-use cases, in which we demonstrate functionalities that can prove useful in screening programmes and teletraining.

Acknowledgements This work was made in the framework of the MAGIC-5 collaboration. We thank the medical staff involved recently in this study: Prof. A. Carriero, Prof. M. Mazzotta and Prof. M. Torsello with Ospedale "V. Fazzi" (Lecce, Italy).

Analogamente a quanto avviene nei programmi di screening, i dati sono raccolti e conservati in siti geograficamente distinti. Il crescere delle dimensioni del database e la natura distribuita della collaborazione pone un problema poiché le immagini non possono essere replicate tra un sito e l'altro. Il problema dell'accesso remoto ad un database distribuito è stato affrontato facendo uso di tecnologie e strumenti sviluppati nel contesto delle attività di Grid Computing. L'integrazione del database in un ambente Grid ha reso possibile l'implementazione di alcuni use case di analisi remota, con i quali sono dimostrate funzionalità utili in programmi di screening e tele-training.

Ringraziamenti *Questo lavoro è svolto nell'ambito della collaborazione MAGIC-5. Ringraziamo lo staff medico coinvolto recentemente in questi studi: il prof. A. Carriero, il prof. M. Mazzotta e il prof. M. Torsello dell'Ospedale "V. Fazzi" (Lecce, Italia).*

References/*Bibliografia*

1. Osuch JR, Anthony M, Bassett LW et al (1995) A proposal for a national mammography database: content, purpose, and value. AJR Am J Roentgenol 164:1329–1334
2. http://mircwiki.rsna.org/. Accessed 14/04/2008
3. http://marathon.csee.usf.edu/Mammography/Database.html. Accessed 14/04/2008
4. http://peipa.essex.ac.uk. Accessed 14/04/2008
5. Nunes FLS, Schiabel H, Rodrigo H, Benatti RH (2003) A computer system to record and retrieve information from a mammographic images database. Internet World Congress on Medical Physics and Biomedical Engineering, pp 24–29
6. Ertas G, Gulcur HÖ, Aribal E, Semiz A (2001) Development of a secure mammogram database. MEDNET 2001 abstract 54
7. Fantacci ME, Bottigli U, Delogu P et al (2002) Search of microcalcifications Clusters with the CALMA CAD Station. Proc SPIE 4684:1301–1310
8. http://medical.nema.org. Accessed 14/04/2008
9. http://www.acr.org. Accessed 14/04/2008
10. Lattanzio E, Guerrieri A (1998) The mammography report. Radiol Med 96:283–288
11. Wolfe JN (1976) Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer 37:2486–2492
12. Lattanzio V, Simonetti G (2002) Mammography – Guide to interpretation, reporting and auditing mammographic images Re.Co.R.M. Springer-Verlag, Berlin New York
13. Cascio D, Fauci F, Magro R et al (2006) Mammogram segmentation by contour searching and massive lesions classification with Neural Network. IEEE-Transactions on Nuclear Science (TNS) 53:2827–2833
14. Bellotti R, De Carlo F, Gargano G et al (2006) A completely automated CAD system for mass detection in a large mammographic database. Med Phys 33:3066–3075
15. Bagnasco S, Bottigli U, Cerello P et al (2005) GPCALMA: a GRID based tool for mammographic screening. Methods Inf Med 44:244–248
16. Lauria A, Drogo M, Fantacci ME et al (2003) Comparison between different monitors to be used in the reading of digital mammographic images. Proc SPIE 5034:448–452
17. Lauria A, Fantacci ME, Bottigli U et al (2003) Diagnostic performance of radiologists with and without different CAD systems for mammography. Proc SPIE 5034:244–248
18. Delogu P, Fantacci ME, Preite Martinez A et al (2005) A scalable system for microcalcification cluster automated detection in a distributed mammographic database. Nuclear Science Symposium Conference Record, 2005 IEEE 3:1530–1534
19. Solomonides A, McClatchey R, Odeh M et al (2003) MammoGrid and eDiamond: Grids applications in mammogram analysis. Proceedings of the IADIS International Conference: e-Society 2003. Lisbon, Portugal. IADIS Press, Lisbon, pp 1032–1033
20. McClatchey R, Manset D, Hauer T et al (2003) The MammoGrid project. Grids architecture. CHEP 03, San Diego
21. McClatchey R, Estrella F, Rogulin D et al (2004) Resolving clinicians queries across a Grids infrastructure. Proceedings of the 2nd International HealthGRID Conference Clermont-Ferrand, France
22. Bellotti R, Cerello P, Bevilacqua V et al (2006) Distributed medical images analysis on a Grid infrastructure. Special Issue on Life Science Grids for Biomedicine and Bioinformatics 23/3:475–484
23. Masala GL, Tangaro S, Golosio B et al (2007) Comparative study of feature classification methods for mass lesion recognition. Il Nuovo Cimento C 30, pp 305–316
24. Fauci F, Raso G, Magro R et al (2005) A massive lesion detection algorithm in mammography. Physica Medica XXI:21–28
25. Masala G, Tangaro S, Quarta M et al (2006) Classifiers trained on dissimilarity representation of medical pattern: a comparative study. Il Nuovo Cimento C 28:905–912