Grid Computing for Breast Cancer CAD. A Pilot Experience in a Medical Environment

Raul Ramos, José Miguel Franco, Jorge Sevilla, Naimy González, Noel Pérez, Mario Vaz, Joana Loureiro, Isabel Ramos, Miguel Ángel Guevara

inegi Instituto de Engenharia Mecânica e Gestão Industrial. Universidade do Porto. Portugal

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

U.PORTO
FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS
CETA Ciemat

FEDER
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa

Motivations

- Distributed computing (Grid) environments present several facilities that are worthy for **digital repositories**, such as: strong security contexts, the data federation, the information sharing and the availability of large computing and storage capacity.
- Based on WHO **Breast cancer is a major concern** and the second-most common and leading cause of cancer deaths among women.
- In Portugal, each year, are estimated (diagnosed) 4500 new cases of breast cancer and 1600 women deaths from this disease.

Objectives

1. To develop a Grid platform to **store, retrieval and manage medical digital image repository.**

2. To create of a suitable framework for **evaluating massively medical image analysis algorithm and methods**, based on:
   - **Digital image processing**
   - **Patter recognition techniques**
   - **Artificial intelligence techniques**

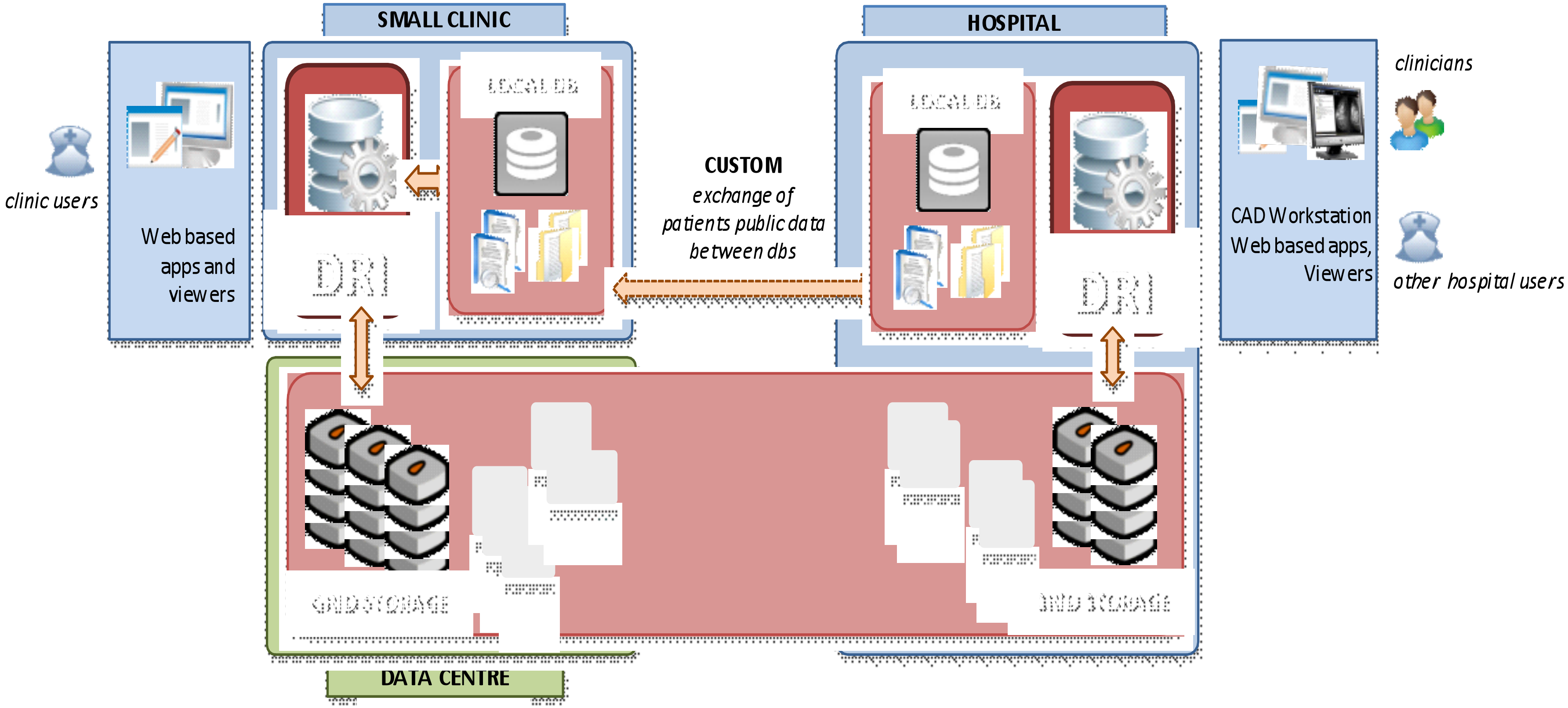
3. To build robust tools for manipulating **medical data to create machine learning classifiers**

Index

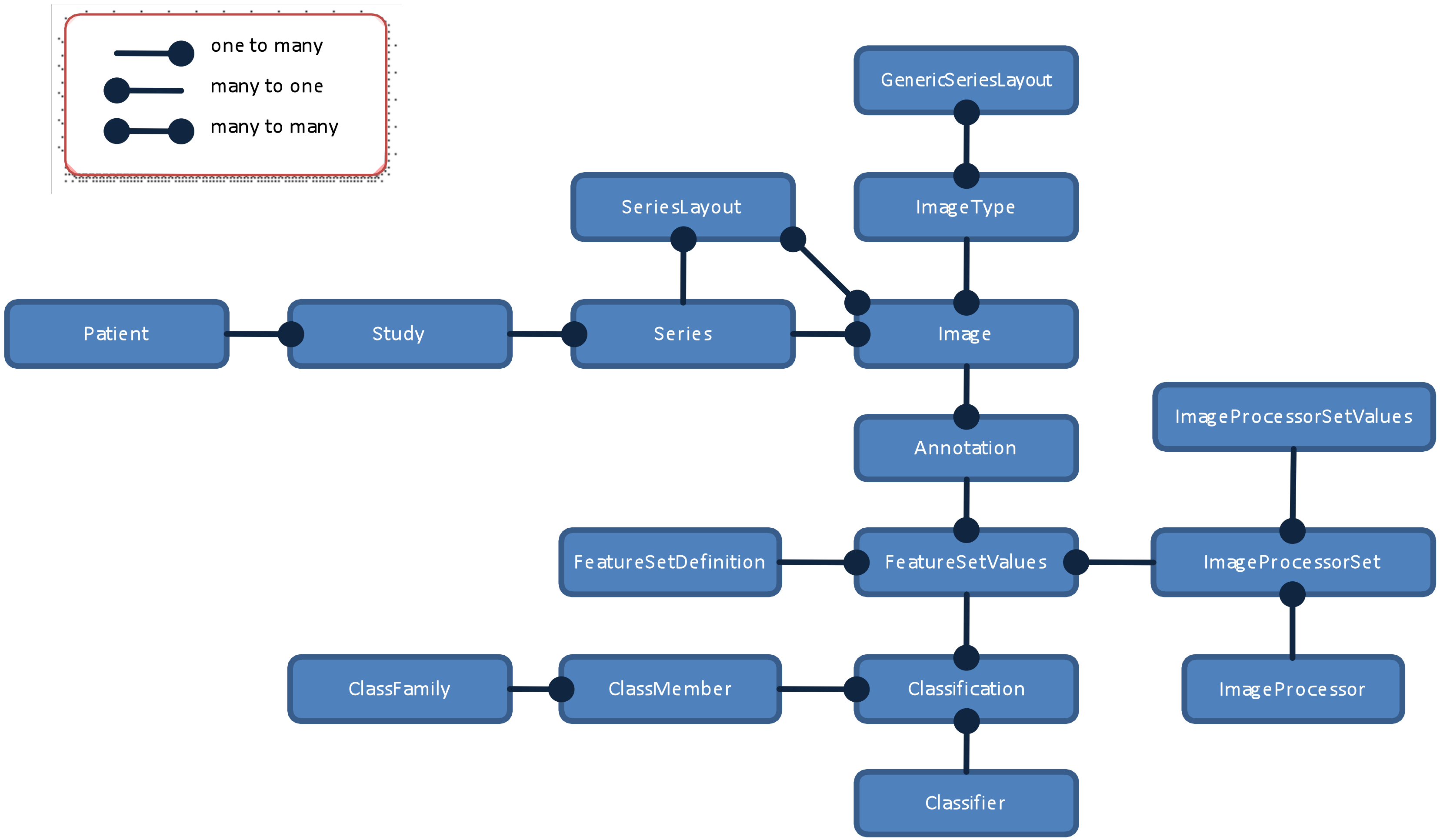
1. Mammography Digital Image Repository
2. Computer-Aided Detection/Diagnosis (CAD)
3. Data Training and Analysis Framework
4. Discussions and results
5. Conclusions

Mammography Digital Image Repository

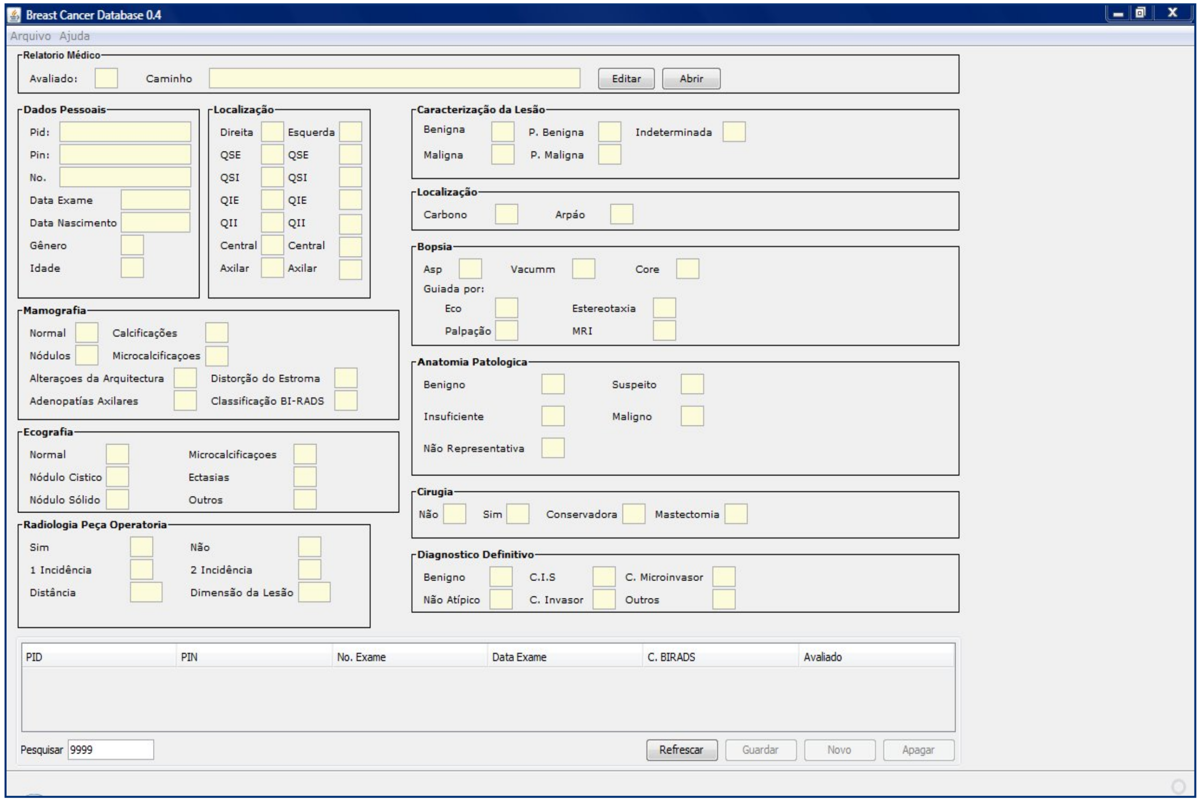
- **Based on DRI Plaftorm**
- **Separates Large Digital Data (to Grid) from Metadata (to DB)**



Reference data model (DICOM based)



MIWAD: Mammography Digital Image Repository



Deployment at FMUP

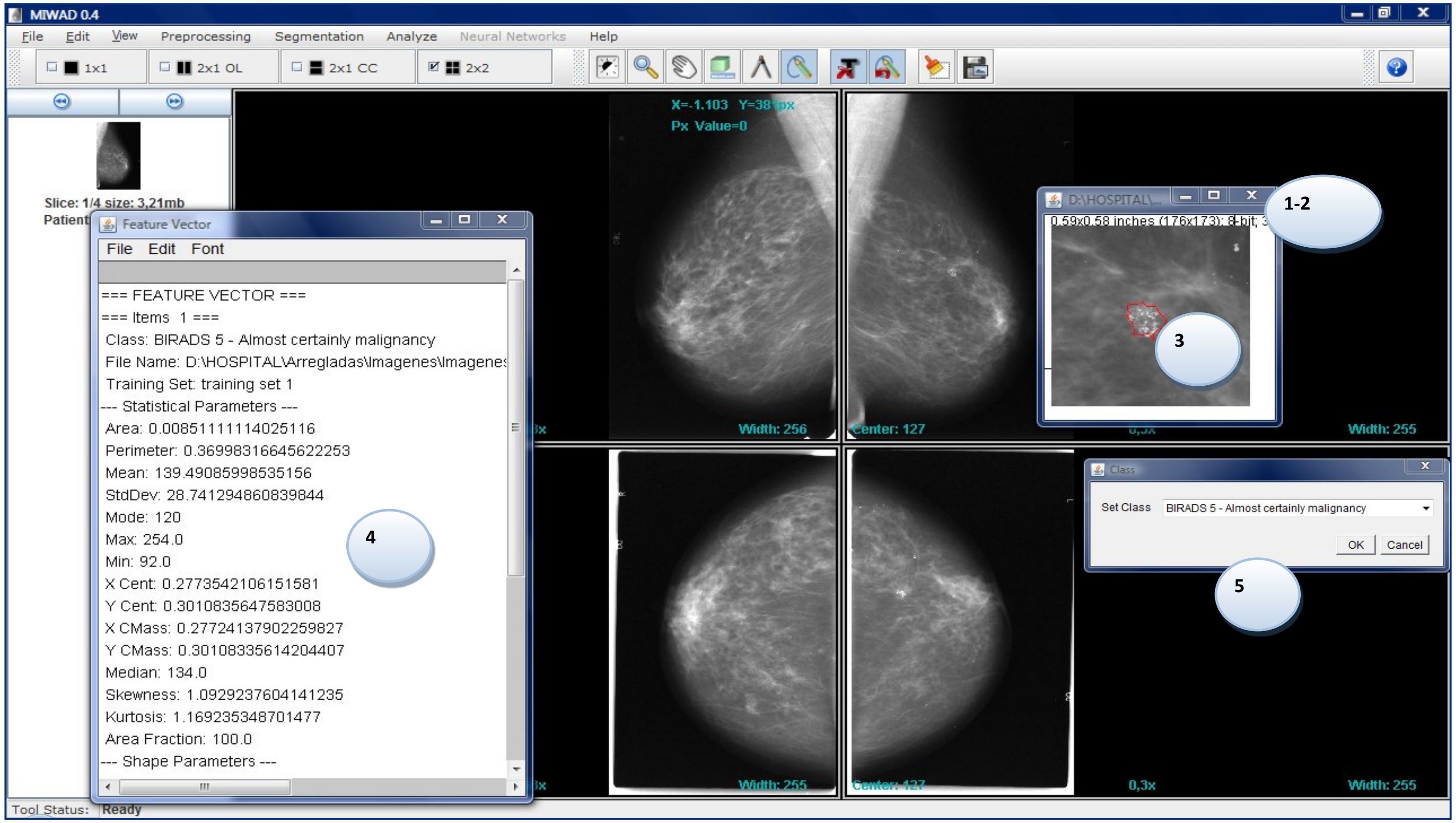
Stand alone pilot

UI validation

Building datasets

Ready to integrate in Grid

MIWAD: Computer Aided Diagnosis



1. ROI SELECTION → **2. IMAGE ENHANCEMENT** → **3. SEGMENTATION** → **4. FEATURE EXTRACTION** → **5. CLASSIFICATION**

1. ROI SELECTION	2. IMAGE ENHANCEMENT	3. SEGMENTATION	4. FEATURE EXTRACTION	5. CLASSIFICATION
GUI based	image processing algorithms (pixel/bit based)	semiautomatic	9 features Initially	ANN Models – based classifiers

Dataset build mode + CADx mode (diagnosis)

Data Training and Analysis Framework

- **Dataset:** vectors of feature sets + class (for supervised training)
- **Binary Dataset/Multiclass Dataset**
- **Classifier:** assigns class to input vector
- **Engine**: A third party classifier (encog ANN, WEKA, etc.)
- **Engine configuration**: Input parameters for an engine (layers, neurons, lrate, pop)
- **Exploration**: Set of engine configurations to train.
- **Jobs**: Explorations are split into jobs to be set to a Grid
- **ROC Az**: Area under the Receiver Operating Characteristic Curve.

- **DTAF** allows you to:
  - Create datasets, split into binary datasets, define explorations, use several engines and engine configurations, split exploration into jobs, send them to a grid, plot roc curves, etc.

Data Training and Analysis Framework

- **Currently supported engines**

Table 1. Currently supported ANN engines in DTAF.

Engine name	Description
encog.ffbp	Feedforward with backpropagation training
encog.ffga	Feedforward with genetic algorithms based training
encog.ffsa	Feedforward with simulated annealing based training
encog.ffsaroc	FFSA with WEKA ROC based error evaluation
encog.rb	Radial basis
encog.som	Self-organizing map (unsupervised)

- **Exploration definition**

Table 2. Sample exploration file

```
explore.neurons.input        = 9
explore.neurons.output       = 2
explore.neurons.layer.01     = 18:36
explore.neurons.layer.02     = 9:18
explore.neurons.layer.03     = 5:9

explore.activation.input     = tanh
explore.activation.output    = tanh:sigm
explore.activation.layer.01 = tanh:gaus
explore.activation.layer.02 = tanh
explore.activation.layer.03 = tanh

explore.nblayers.fixed       = yes

explore.trainingsets         = BCW
explore.stop.error           = 0.1
explore.stop.epochs          = 200
explore.trainengines =
                  rb:ffrp:ffga:ffsaroc:ffsa
explore.validation         = asints
```

```
explore.encog.ffga.matepercent  = 0.5
explore.encog.ffga.percentmate  = 0.2
explore.encog.ffga.population   = 100
explore.encog.ffsa.starttemp    = 10:15
explore.encog.ffsa.endtemp      = 2:5
explore.encog.ffsa.cycles       = 100

explore.encog.ffsaroc.starttemp = 10
explore.encog.ffsaroc.endtemp   = 2
explore.encog.ffsaroc.cycles    = 100

explore.encog.rb.gaus.center    = 0.5:0.7
explore.encog.rb.gaus.peak      = 0.5
explore.encog.rb.gaus.width     = 0.5
explore.encog.rb.gaus.spread    = 0.5:1.0

explore.encog.ffbp.learnrate    = 0.2:0.5
explore.encog.ffbp.momentum     = 0.0
explore.numberofjobs            = 50
```

Results

- **MIWAD deployed at FMUP**
  - Mammograms being digitalized
  - Gathering 50 vectors/week aprox.
  - Optimizing interface for HSJ workflow
  - Offline synch with repository hosted at CETA-CIEMAT and INEGI

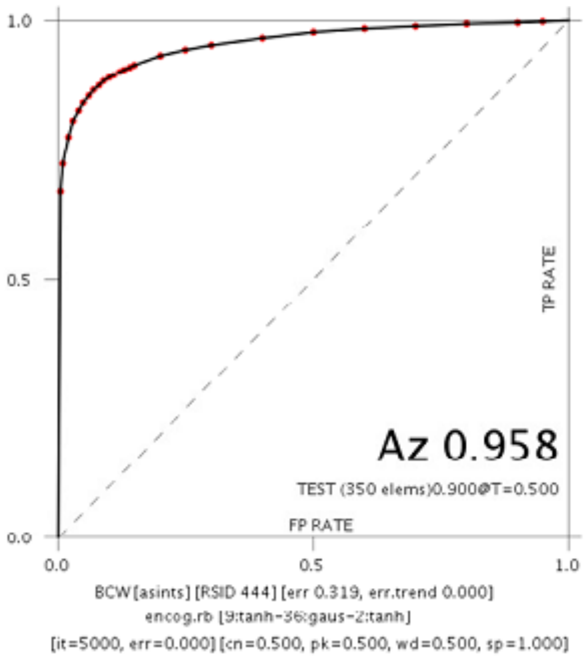
- **DTAF tested with binary datasets**
  - **Breast Cancer Wisconsin UCI dataset (2 classes 699 cases)**
  - **MIAS derived binary datasets (7 classes 113 cases)**

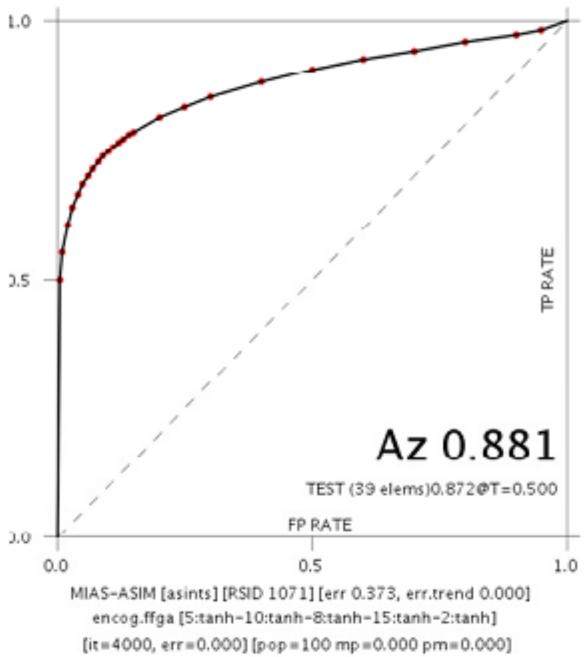
Results

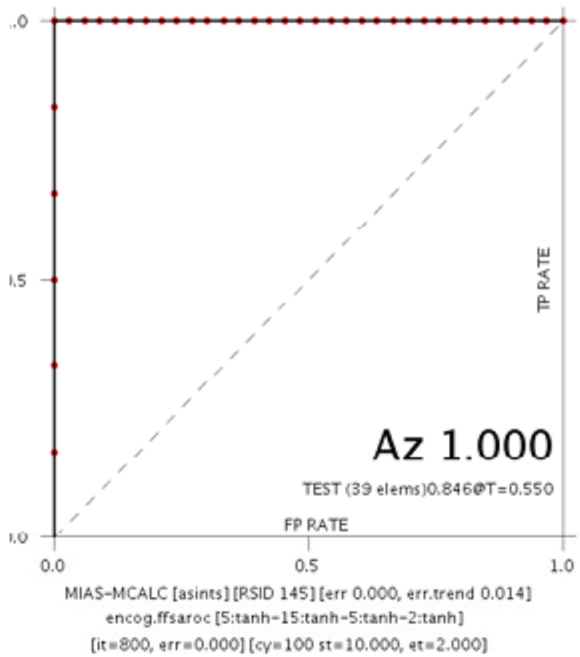
- **DTAF:** 815 ANN configurations explored on gLite, 10 cores, 12 CPU hours

Table 3. Exploration results

Training set	Best test Az		Best test score		Number of configurations	Computing time
BCW	0.906	encof.rb	0.956	encof.ffrb	66	5.16 hrs
MIAS-ASIM	0.871	encof.ffga	0.897	encog.ffsa	101	63.18 min
MIAS-CIRC	0.859	encog.ffsa	0.897	encog.ffsa	104	54.14 min
MIAS-ILLDEF	0.745	encog.ffrp	0.875	encog.ffsa	102	61.09 min
MIAS-MCALC	1.000	encog.ffsaroc	0.923	encog.ffsaroc	104	46.67 min
MIAS-ARCHD	0.824	encog.ffsa	0.875	encog.ffsa	101	63.16 min
MIAS-SPIC	0.843	encof.ffga	0.878	encog.ffsa	132	62.21 min
MIAS-NORM	0.766	encof.ffga	0.825	encof.ffsaroc	105	63.65 min







Conclusions

- **Three developed technologies were developed, which enabling an effective exploitation of Grid resources in the area of medical imaging:**
  - A MDIR hosted on Grid storage based on the CETA-DRI platform.
  - A MIWAD enabling the full CAD lifecycle over content stored in the mammography repository.
  - A DTAF that exploits Grid computer power to explore configurations of machine learning classifiers, which can be used as assistance for lesions diagnos**is.**

Future work

- **Future immediate work will be focused on:**
  - **Tuning the platforms to integrate it completely within medical work flows, enabling systematic construction of federated repositories of mammograms.**
  - **Building training sets and custom made classifiers and integrating them within easy to use workstations.**

This work, present and future, is always done in close collaboration with end-users and professionals from medical environments to ensure its acceptability and validation.

THANKS!!
QUESTIONS?

inegi Instituto de Engenharia Mecânica e Gestão Industrial. Universidade do Porto. Portugal

U.PORTO

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA E INNOVACIÓN

CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS
CETA Ciemat

FEDER
Fondo Europeo de Desarrollo Regional

Una manera de hacer Europa