Incremental Gaussian Discriminant Analysis based on Graybill and Deal weighted combination of estimators for brain tumour diagnosis

Salvador Tortajada [a,*] Elies Fuster-Garcia [a] Javier Vicente [a]
Pieter Wesseling [b] Franklyn A Howe [c] Margarida Julià-Sapé [d,e,m]
Ana-Paula Candiota [d,e] Daniel Monleón [f]
Àngel Moreno-Torres [d,g] Jesús Pujol [h,d] John R Griffiths [i]
Alan Wright [j] Andrew C Peet [k] M Carmen Martínez-Bisbal [d]
Bernardo Celda [d,ℓ] Carles Arús [d,e,m] Montserrat Robles [a]
Juan Miguel García-Gómez [a]

[a] *IBIME, Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Spain*

[b] *Dept. of Pathology, Radboud University Nijmegen Medical Centre, The Netherlands*

[c] *St George's University of London, UK*

[d] *Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Spain*

[e] *Dept. de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Spain*

[f] *Fundación Investigación Hospital Clínico Valencia (INCLIVA), Spain*

[g] *Research Dept., Centre Diagnòstic Pedralbes, Spain*

[h] *Institut d'Alta Tecnologia-PRBB, CRC Hospital del Mar, Spain*

[i] *Cambridge Research Institute, UK*

[j] *Dept. of Radiology, Radboud University Nijmegen Medical Centre, The Netherlands*

[k] *University of Birmingham and Birmingham Children's Hospital NHS Foundation Trust, UK*

[ℓ] *Dept. Química Física, Universitat de València, Spain*

[m] *Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Spain*

Abstract

In the last decade, machine learning (ML) techniques have been used for developing classifiers for automatic brain tumour diagnosis. However, the development of these ML models rely on a unique training set and learning stops once this set has been processed. Training these classifiers requires a representative amount of data, but the gathering, preprocess, and validation of samples is expensive and time-consuming. Therefore, for a classical, non-incremental approach to ML, it is necessary to wait long enough to collect all the required data. In contrast, an incremental learning approach may allow us to build an initial classifier with a smaller number of samples and update it incrementally when new data are collected. In this study, an incremental learning algorithm for Gaussian Discriminant Analysis (iGDA) based on the Graybill and Deal weighted combination of estimators is introduced. Each time a new set of data becomes available, a new estimation is carried out and a combination with a previous estimation is performed. iGDA does not require access to the previously used data and is able to include new classes that were not in the original analysis, thus allowing the customization of the models to the distribution of data at a particular clinical center. An evaluation using five benchmark databases has been used to evaluate the behaviour of the iGDA algorithm in terms of stability-plasticity, class inclusion and order effect. Finally, the iGDA algorithm has been applied to automatic brain tumour classification with magnetic resonance spectroscopy, and compared with two state-of-the-art incremental algorithms. The empirical results obtained show the ability of the algorithm to learn in an incremental fashion, improving the performance of the models when new information is available, and converging in the course of time. Furthermore, the algorithm shows a negligible instance and concept order effect, avoiding the bias that such effects could introduce.

Key words: machine learning, incremental learning, Graybill-Deal estimator, automatic brain tumour diagnosis, magnetic resonance

* Corresponding author: Biomedical Mining Group. IBIME, ITACA. Camino de Vera s/n, Building 8G, 1st floor. 46022, Valencia, SPAIN. Tel.(+34)963877000 Ext. 75278. Fax:(+34)963877279.

Email address: `vesaltor@upv.es` (Salvador Tortajada).

URL: `http://www.ibime.upv.es` (Salvador Tortajada).

1 Introduction

During the last decade, three European projects (INTERPRET (2000-2002) [1,2], eTUMOUR (2004-2009) [3], and HEALTHAGENTS (2005-2008) [4]) have endeavoured to develop a non-invasive diagnostic tool using machine learning (ML) techniques applied to proton magnetic resonance spectroscopy (^{1}H MRS) data from brain tumours. A major aim was to minimize the need for an invasive histological diagnosis of a brain tumour biopsy as is currently required for the diagnosis and management of brain tumours. Non-invasive brain tumour diagnosis using ^{1}H MRS has shown considerable promise in aiding patient management but is not in widespread clinical use due mainly to the difficulties of data interpretation. Machine learning (ML) has been successfully applied to this problem providing automated analysis of ^{1}H MRS [2,5,6]. However, the development of robust brain tumour classifiers requires a large number of cases to be acquired for each tumour type and at present the approach has only been used for a few common tumours. Cases are accrued from a large number of hospitals over many years and data transferred to a centralised database. This approach has several disadvantages, ethical approval and patient consent needs to be obtained to send and store data. In order to expand the applicability of ML techniques to MRS of a wider range of tumours, more cases need to be collected over a more prolonged period of time and the logistics of using a centralised database to provide this have so far proved insurmountable. Distributed databases where the data is held at the data collecting hospitals have major advantages [4] and such a system in which classifiers can be trained without moving the data from the hospital at which it was collected would provide a practical solution. The ability to retrain the classifiers as new data accumulates is also an important requirement and to meet these needs, an incremental learning algorithm is required.

Until now, the different Clinical Decision Support Systems (CDSS) developed for automatic brain tumour diagnosis have only used non-incremental classification models [2–4]. Non-incremental classifiers entail an implicit assumption that learning stops when the training set has been processed. Hence, the performance of a non-incremental automatic classifier strongly depends on the availability of a representative training set for each class. However, the gathering of these data is often expensive and time-consuming, and a strategy to wait long enough as to gather enough data all in one set may be undesirable and/or impractical. Furthermore, there are situations where the access to previous data may be forbidden. There are also types of data sources where the underlying distributions may evolve over time rather than be stationary. In particular, this happens in the *concept drift* [7–9], where the target distribution $p(c|x)$ may change in the course of time, and in the *covariate shift* [7,10], where the data distribution $p(x)$ changes continously. Under these circumstances an incremental learning algorithm might be a practical and more effective solution.

The easiest way to take advantage from new observations is to build a new model from scratch using all the historic data. But this solution may be more expensive than modifying an already trained system, or even impractical if older training set data is not readily accessible. Typically, an alternative has been to keep a relevant subset of the previous data available. This approach was used in the partial memory learning [9] and in the so-called boundary methods, or maximum margin methods [11,12]. In this paper, it is assumed that previous data are not accessible at all. In the last two decades, various approaches have been developed for providing learners with incremental learning ability. A number of incremental techniques were designed for decision trees [13–15]. Other incremental decision trees techniques have been applied for data streaming [16]. Incremental learning has also been used for connectionist models based on structural adaptation [17–20] or on weight adaptation [21,22]. There are some approaches to incremental principal component analysis [23,24] that update the projection matrix incrementally. Moreover, incremental algorithms for Fisher's Linear Discriminant Analysis have also been developed in the last decade [25–27].

Following the definitions of Langley [28] and Giraud-Carrier [29], an **incremental learning algorithm** is a learning algorithm that produces a sequence of classifiers $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n$ for any given training set of samples $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ available at different moments $t_1, t_2, \ldots, t_n$, such that $\mathcal{H}_{i+1}$ is determined by $\mathcal{H}_i$ and $\mathcal{S}_{i+1}$. The main characteristics of an incremental learning algorithm are: a) it should be able to learn additional information from new data without completely forgetting its previous knowledge; b) it should not require access to previous data; c) since each $\mathcal{H}_i$ can be viewed as the best approximation of the target application, the performance should improve over time.

This definition is related to a general problem for classification models called the *stability-plasticity dilemma* [30]. This dilemma reveals that some information may be lost when new information is learned (*gradual forgetting*) and highlights the difference between stable classifiers and plastic classifiers. To summarize, the problem is how to design a learning system that is sensitive to new input without being radically disrupted by such input.

In addition, Polikar et al. stated in [17] that an incremental learning algorithm should be able to admit new classes when they are introduced with the new data. This means that a new target concept appears over time while the rest of the target concepts remain stable.

Another issue to be considered is the problem of order effects in incremental learning, which has been addressed by several authors [28,31,32]. An incremental learning algorithm suffers from an *order effect* when different ordered sequences of the same instances lead to different models. In this sense, the selection of the final models may be biased due to the ordering of the introduced inputs. An order effect is *benign* if this effect produces classifiers of nearly equal scores on some metric; otherwise it is *malignant*. The order effect can appear at three levels: attribute level, instance level, and concept or class level.

In this work, an ML-based method is proposed to continously adapt an automatic brain tumour diagnosis model to reflect the most recent information included in newly acquired cases. Therefore, an incremental learning algorithm based on a weighted combination of Gaussian parameter estimation is presented for automatic brain tumour diagnosis. Our method relies on the Graybill and Deal combination of unbiased estimators [33,34] originally developed for the estimation of a common mean when several sets of data come from different measurement methods or different laboratories. The Graybill-Deal estimator is known to be unbiased for the mean [34,35] and it has been applied to discriminant analysis to develop a straightforward method for updating the parameters of each class when new observations arrive. Although the Gaussian assumption restricts this method to datasets with continuous variables, the proposed algorithm is able to learn from new information when it arrives, adjusting the parameters of the model to incorporate new classes in the discriminant space when needed, and showing a benign order effect at instance and concept level. Some benchmark experiments have been carried out to show these issues and, finally, the incremental algorithm has been applied for brain tumour diagnosis.

2 Methods

The formal purpose of classification is to assign instances to one class among $|\mathcal{C}|$ possible classes based on a set of features obtained from each observation. A decision rule δ is a function that maps an object $\mathbf{x} \in \mathbb{R}^d$ into a class $c \in \mathcal{C}$. An error is incurred if the decision rule assigns the instance to a wrong class. The final objective is to minimize the error for discriminating among different classes. In discriminant analysis, each class is represented by a function $g_i(\mathbf{x}), i = 1, \ldots, |\mathcal{C}|$. A classifier $\delta(\mathbf{x})$ assigns the class c_j if $g_j(\mathbf{x}) > g_i(\mathbf{x}), \forall j \neq i$. When a 0-1 loss function is used, finding the class that maximizes the log-likelihood of the posterior probability $p(c|\mathbf{x})$ is equivalent. Using Bayes' rule, assuming that the density functions follow a multivariate normal, $p(\mathbf{x}|c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, and taking into account that the prior probabilities are parameters to be estimated, then the expression can be evaluated using

$$g_c(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{W}_c\mathbf{x} + \mathbf{w}_c^{\mathrm{T}}\mathbf{x} + w_{c0} \ , \tag{1}$$

where $\mathbf{W}_c = -\frac{1}{2}\mathbf{\Sigma}_c^{-1}$, $\mathbf{w}_c = \mathbf{\Sigma}_c^{-1}\boldsymbol{\mu}_c$ and $w_{c0} = \log \pi_c - \frac{1}{2}\log|\mathbf{\Sigma}_c| - \frac{1}{2}\boldsymbol{\mu}_c^{\mathrm{T}}\mathbf{\Sigma}_c^{-1}\boldsymbol{\mu}_c$. The mean, $\boldsymbol{\mu}_c$, the covariance matrix, $\mathbf{\Sigma}_c$, and the prior probabilities, π_c, of the class c are the parameters that can be estimated by the maximum-likelihood method using a set of labeled samples [36].

This decision rule divides the sample space into $|\mathcal{C}|$ decision regions. The points $\mathbf{x}$ of the sample space which satisfy that $g_j(\mathbf{x}) = g_i(\mathbf{x}), i \neq j$ make the decision boundary. These discriminant functions describe quadratic decision boundaries except when the covariance matrices of all the classes are identical. If a common covariance matrix is used the quadratic terms of Equation (1) cancel giving rise to a linear boundary. Linear and quadratic versions are both available in the proposed incremental algorithm.

2.1 Graybill-Deal combination of estimators

Given k sets with N_i instances $x_1, \ldots, x_{N_i}$ in each set, for $i = 1, \ldots, k$, it is possible to estimate the common mean of the population using a weighted mean, where the weights w_i depend on the number of instances and the population variance, provided that all the variances are known. When the true variance is not known, the sample variance is used instead. In this case, the weighted mean is computed by

$$\bar{X}_{GD} = \sum_{i=1}^{k} \hat{w}_i \bar{X}_i \ , \tag{2}$$

where $\bar{X}_i$ is the mean value of the i-th set. The weights are calculated using the sample variance as

$$\hat{w}_i = \frac{N_i/S_i^2}{\sum_{j=1}^{k} N_j/S_j^2} \ , \tag{3}$$

where S_i^2 is the sample estimate variance of the corresponding set. Notice that the estimation given by (3) gives higher weight to those sets with larger number of instances N_i and smaller variance S_i^2. Graybill and Deal [33] demonstrated that the estimation of the mean μ using $\bar{X}_{GD}$ is unbiased, that is $E[\bar{X}_{GD}] = \mu$.

It is trivial to extend this result to the combination of the mean for multivariate samples. However, the need of a combined estimator for the covariance matrix of the samples presents a harder challenge. In the next subsection, a solution is proposed as part of the developed incremental algorithm.

2.2 Incremental Gaussian Discriminant Analysis based on Graybill-Deal estimation of weights

Let the training dataset be obtained in different samples, $\mathcal{S}_i$, that are available in times $t_i, i = 1, \ldots, T$. The Graybill-Deal estimation for Gaussian discriminant analysis begins with the adjustment of the parameters for each class based on maximum log-likelihood using the first set of samples. Hence, the prior probabilities for each class, $\pi_c^{(1)}$; the mean vector, $\boldsymbol{\mu}_c^{(1)}$; and the covariance matrix, $\boldsymbol{\Sigma}_c^{(1)}$ are estimated following the usual maximum log-likelihood estimation [36].

A model $\mathcal{H}_i$ is composed of a mean $\boldsymbol{\mu}_c^{(i)}$, a covariance matrix $\boldsymbol{\Sigma}_c^{(i)}$, a prior probability $\pi_c^{(i)}$, and the number of instances $N_c^{(i)}$ of each class. The first iteration of the algorithm estimates the parameters of the first model. When a new dataset $\mathcal{S}_i$ is available in time t_i, $i > 1$, the first step is to carry out a new parameter estimation from $\mathcal{S}_i$, where $\hat{\boldsymbol{\mu}}_c$, $\hat{\boldsymbol{\Sigma}}_c$, π_c, and N_c are calculated for each class. Then, the prior probabilities of the new model $\mathcal{H}_i$ are updated based on the number of samples per class:

$$N_c^{(i)} = N_c^{(i-1)} + N_c \ , \tag{4}$$

$$N^{(i)} = N^{(i-1)} + \sum_c N_c \ , \tag{5}$$

$$\pi_c^{(i)} = \frac{N_c^{(i)}}{N^{(i)}} \ , \tag{6}$$

where N_c is the number of instances of $\mathcal{S}_i$ and the class c. The new mean and a weighted covariance matrix are calculated as

$$\boldsymbol{\mu}_c^{(i)} = w_c^{(i)} \boldsymbol{\mu}_c^{(i-1)} + (1 - w_c^{(i)}) \hat{\boldsymbol{\mu}}_c \ , \tag{7}$$

$$\boldsymbol{\Sigma}_c^{(i)} = w_c^{(i)} \boldsymbol{\Sigma}_c^{(i-1)} + (1 - w_c^{(i)}) \hat{\boldsymbol{\Sigma}}_c \ , \tag{8}$$

where $w_c^{(i)}$ and $(1 - w_c^{(i)})$ are the weights for updating the parameters of class c. The weights are calculated using the Graybill-Deal combination of estimators (3) and they are subject to $0 \leq w_c^{(i)} \leq 1$ and $\sum_i w_c^{(i)} = 1$. We propose to adapt the variance to multivariate distributions by means of the sum of the variances $S_i^2 = tr(\boldsymbol{\Sigma}_i)$, which is equivalent to the sum of the eigenvalues of the covariance matrix.

After estimating and incrementally updating the parameters, the common covariance matrix can be used to obtain a linear discriminant instead of a quadratic discriminant as previously explained. The pseudocode for the iGDA algorithm is given in the Appendix.

One interesting property is that the algorithm allows the possibility of introducing new classes if required. Therefore, if a new set of samples includes data from a new class, an estimation of the additional parameters is carried out. The prior probabilities are updated according to the new data set and the parameters of the new class are retained within the model, thus modifying the final decision boundaries and the regions described for each class. This is due to the generative model approach followed by the algorithm.

2.3 Comparison with other algorithms

Learn^{++} is a well-known incremental learning algorithm proposed by Polikar in [17]. Learn^{++} is inspired by the AdaBoost algorithm [37], which was developed to improve the classification performance of weak learners [1] . Schapire [38] showed that a *weak learner* can be transformed into a *strong learner* using a *boosting* procedure.

Learn^{++} uses the concept of boosting to incrementally improve the performance of the classification. In contrast with AdaBoost, Learn^{++} does not extract the subsets from the same training set but from the successive observations available throughout time. Learn^{++} uses a weak learner to generate multiple hypotheses from different subsets of data. Therefore, each hypothesis learns only a portion of the input space. The weak learner is based on a perceptron, thus each hypothesis defines a linear hyperplane as a decision boundary. When the algorithm learns with a new set of samples, it generates a new set of hypotheses. The outputs of all the hypotheses are combined using a weighted majority voting. Therefore, Learn^{++} does not require access to previously used data during the incremental learning and it does not forget previously acquired knowledge.

Another well-known incremental learning algorithm is the incremental Linear Discriminant Analysis (iLDA) proposed by Pang et al. [25]. iLDA uses a constructive method for deriving an updated discriminant eigenspace for classification. A typical Linear Discriminant Analysis (LDA) seeks directions in the D-dimensional space that are efficient for discrimination, projecting the

[1] A *weak learner* is a learning algorithm that performs slightly better than random guessing.

observations to a P-dimensional space where $P < D$. To obtain the projection matrix $\mathbf{W}$, the ratio between the *between-class* scatter matrix $\boldsymbol{S}_b$ and the *within-class* scatter matrix $\boldsymbol{S}_w$ must be maximized. Once the observations are projected, different ML techniques can be used for classification purposes [39].

The iLDA method aims to obtain a new discriminant eigenspace model Φ by combining two discriminant eigenspace models Ω_t and Ω_{t+1} from different samples $\mathcal{S}_t$ and $\mathcal{S}_{t+1}$ acquired at time t and $t+1$ respectively. This new model, Φ, updates the sample mean, the $\boldsymbol{S}_w$ matrix and the $\boldsymbol{S}_b$ matrix and results in a new projection matrix $\mathbf{W}$. Once the data are projected in the new discriminant eigenspace, a nearest neighbour algorithm is used for classification purposes. For technical details see [25]. iLDA does not require access to previously seen data and it can also include new classes if needed.

Finally, a naive incremental Gaussian model is used as a baseline for comparison with the above methods. This model updates its parameters from scratch. That is, the previous data and the current data are used to train a new model using quadratic discriminant analysis [36].

3 Benchmark experiments

The behaviour of the iGDA algorithm has been tested on several databases with a threefold purpose: 1) to show that the developed algorithm is able to incrementally learn and adapt the parameters of the classifier, improving its performance without incurring in catastrophic forgetting; 2) to show how the iGDA algorithm is able to introduce new concepts or classes into its knowledge representation; 3) to analyze whether the order in which the instances are introduced into the analysis have a crucial influence in the final hypothesis, that is, if the algorithm is order dependent or not. The selected datasets have only real attributes since the iGDA is restricted to that set of numbers. In order to avoid possible bias, every experiment was evaluated following a K random sampling train-test strategy, where $K = 100$.

3.1 Stability/Plasticity dilemma

3.1.1 Vehicle Silhouette Database.

The vehicle silhouette database has been extracted from the UCI Machine Learning Repository [40]. The purpose of this database is to classify a given silhouette into one of four different types of vehicle using a set of 18 features. The database consisted of 846 instances. It was divided into a training parti-

Dataset	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$	$\mathcal{H}_6$	$\mathcal{H}_7$
$\mathcal{S}_1$	99.93	97.42	95.48	94.22	93.22	92.57	92.11
$\mathcal{S}_2$	–	97.43	95.04	93.52	92.90	92.26	91.86
$\mathcal{S}_3$	–	–	95.31	94.16	93.18	92.53	91.90
$\mathcal{S}_4$	–	–	–	94.24	93.38	92.64	92.08
$\mathcal{S}_5$	–	–	–	–	92.68	92.06	91.54
$\mathcal{S}_6$	–	–	–	–	–	92.27	91.79
$\mathcal{S}_7$	–	–	–	–	–	–	91.73
TEST	62.00	79.08	81.53	82.82	83.62	84.09	84.54
CI ($\alpha = 1\%$)	±1.44	±0.71	±0.62	±0.65	±0.65	±0.63	±0.59

Table 1
Training and test accuracy for the Vehicle Silhouette Database using a quadratic iGDA. The rows indicate the different datasets $\mathcal{S}_1, \ldots, \mathcal{S}_7$ and the columns show the hypothesis or models $\mathcal{H}_j$ built from a previous model $\mathcal{H}_{j-1}$ and the new dataset $\mathcal{S}_j$, except $\mathcal{H}_1$ which is built from $\mathcal{S}_1$ only. Each column shows the average performance (%) on the current and the previous training datasets for the current model. The last rows (TEST, CI) indicate the evolution of the average accuracy of the models in the course of time evaluated with an independent test set and the confidence interval ($\alpha = 1\%$).

tion (630 instances) and a test partition (216 instances). The training partition was split again into 7 training sets $\mathcal{S}_1, \ldots, \mathcal{S}_7$ of 90 instances with a similar prevalence to the original database for each class. Table 1 shows that there is a gradual loss of information relating to the previous training datasets when new observations are introduced using the quadratic iGDA. However, the overall performance increases from 62% to 84%. The linear iGDA showed an increase from 73% to 78%, also with a gradual forgetting when new information was added (Table not shown). These results are comparable to the performance of a completely new quadratic classifier trained with the entire training dataset (85%) and to a linear classifier (80%).

3.1.2 Wisconsin Breast Cancer Database.

The Wisconsin Breast Cancer Database from the UCI Machine Learning Repository consists of 569 instances with 30 variables from a digitalized image of a fine needle aspirate (FNA) of a breast mass. The objective in this problem is to classify the instances into a malignant (37.3%) or a benign (62.7%) breast tumour. The database was divided into a test partition (169 instances) and a training partition (400 instances) that were also split into five different sets of 80 instances $\mathcal{S}_1, \ldots, \mathcal{S}_5$. Each partition had the same prevalence for each class as the whole database. The results of the quadratic classifier are

Dataset	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$
$\mathcal{S}_1$	69.11	99.14	98.16	97.77	97.44
$\mathcal{S}_2$	–	99.16	98.32	97.70	97.31
$\mathcal{S}_3$	–	–	98.24	97.64	97.25
$\mathcal{S}_4$	–	–	–	97.55	97.17
$\mathcal{S}_5$	–	–	–	–	97.39
TEST	52.21	94.12	94.95	95.20	95.34
CI ($\alpha = 1\%$)	±3.56	±0.48	±0.46	±0.43	±0.42

Table 2
Training and test accuracy (%) for the Wisconsin Breast Cancer Database using a quadratic iGDA.

shown in Table 2. The linear iGDA also showed an improvement on accuracy: from 91.14% to 94.38% for the independent test set. As shown in the previous experiment, there is generally an improvement in overall classification as the new data are used for incremental learning, but a gradual forgetting is observed with respect to the previous datasets. The poor performance of the first classifier in the quadratic iGDA may be due to the low number of instances in the first dataset $\mathcal{S}_1$ and it is known that quadratic discriminant classification rules generally require larger samples than those based on linear discriminant analysis [41].

3.2 Introduction of new classes

3.2.1 Concentric Circle Database.

The concentric circle database is a synthetic set of five classes each belonging to a concentric ring of data. This database is used to test the ability of the incremental algorithm to introduce new classes. The data is bidimensional with a uniform distribution inside each ring (see Figure 1, left). The database was split into 6 different sets: $\mathcal{S}_1$ and $\mathcal{S}_2$ included 50 instances from each of classes 1, 2, and 3; $\mathcal{S}_3$ and $\mathcal{S}_4$ included 50 instances from classes 1 to 3 and 100 instances from class 4; finally, $\mathcal{S}_5$ and $\mathcal{S}_6$ contained 100 instances from classes 1 to 4 and 200 instances from class 5. Therefore, equal prior probabilities were kept for the number of instances of each class. An independent test set was generated with 10 000 instances from each class. In order to simulate the general behaviour of the algorithm in a real scenario, the test set included all the five classes. Since the database describes quadratic boundaries, only quadratic iGDA was employed (see results in Table 3).

Dataset	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$	$\mathcal{H}_6$
$\mathcal{S}_1$	88.89	90.53	90.82	90.73	91.16	91.60
$\mathcal{S}_2$	–	89.97	90.70	90.45	90.64	91.11
$\mathcal{S}_3$	–	–	69.58	87.50	88.93	89.12
$\mathcal{S}_4$	–	–	–	87.44	88.97	89.15
$\mathcal{S}_5$	–	–	–	–	62.40	84.76
$\mathcal{S}_6$	–	–	–	–	–	84.99
TEST	50.78	52.93	60.52	68.28	71.73	83.50
CI ($\alpha = 1\%$)	±0.60	±0.48	±0.61	±0.78	±0.62	±1.01

Table 3
Training and test accuracy (%) for the Concentric Circle Database using a quadratic iGDA.

As demonstrated in Table 3, iGDA has the ability to include new classes with an increase in overall classification performance for the test set as soon as data from new classes appear in the new datasets.

3.2.2 Image Segmentation Database.

The Image Segmentation database from the UCI Machine Learning Repository consists of 2 310 instances with 18 attributes for segmenting the images from 7 outdoor images. The seven classes are: brickface, sky, foliage, cement, window, path, and grass. The database was split into three training subsets $\mathcal{S}_1$ (including classes brickface, sky and foliage), $\mathcal{S}_2$ (including all the classes except path and grass), $\mathcal{S}_3$ (including all the classes), and one test partition (231 instances) were all the classes were represented. The prior probabilities of all classes were made equal as for the previous experiment. The results for the linear version of the iGDA algorithm are shown in Table 4 and are comparable to that in Muhlbaier et al. [18], where the best improvement went from a 42.2% to a 91.0% after the third dataset. Although there was an improvement for the quadratic version, the results obtained were poor: from 22.2% to 58.8%.

3.3 Order effects

3.3.1 Instance level order effects.

A synthetic dataset with two categories drawn from different multivariate normal distributions (shown right in Figure 1) has been used to analyze the

Dataset	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$
$\mathcal{S}_1$	98.45	99.69	99.78
$\mathcal{S}_2$	–	88.25	87.45
$\mathcal{S}_3$	–	–	94.63
TEST	42.14	64.30	91.42
CI ($\alpha = 1\%$)	±0.15	±0.41	±0.43

Table 4
Training and test accuracy (%) for the Image Segmentation Database using a linear iGDA.

instance level order effects. A training set of 400 instances and a test set of 4 000 instances were drawn from the distributions with equal prior probabilities for each category. The training set was split into 20 different training samples with 20 instances in each sample. The samples were used for incremental learning to build consecutive models as explained before. To evaluate the order effects, the instances were permuted in 100 experiments and the mean accuracies and the decision boundaries of the models of each iteration in the experiments were compared.

The Vehicle Silhouette database was also used to reinforce the analysis. The same configuration as in Section 3.1 was prepared, but the instances were permuted 100 times to test the effect of the instance order. Figure 2 shows the convergence in accuracy for these two experiments, whereas Figure 3 shows the iterative convergence of the decision boundary for the two-dimensional synthetic dataset.

3.3.2 Concept level order effects.

The Concentric Circle database was used to analyze the effect of the concept order on the iGDA. The database was divided into six different samples as in Section 3.2.1. To avoid the problem of imbalanced classes [42], the prior probabilities were forced to be equal. With this set-up of samples and classes and considering that there are five possible categories, the possible combinations for introducing different categories in each sample are 20. Therefore, 100 repetitions of 20 different combinations of samples were analysed. Figure 4 depicts the convergence of the incremental algorithm. The results show a *benign* concept lever order effect when the prior probabilities of the categories are equal.

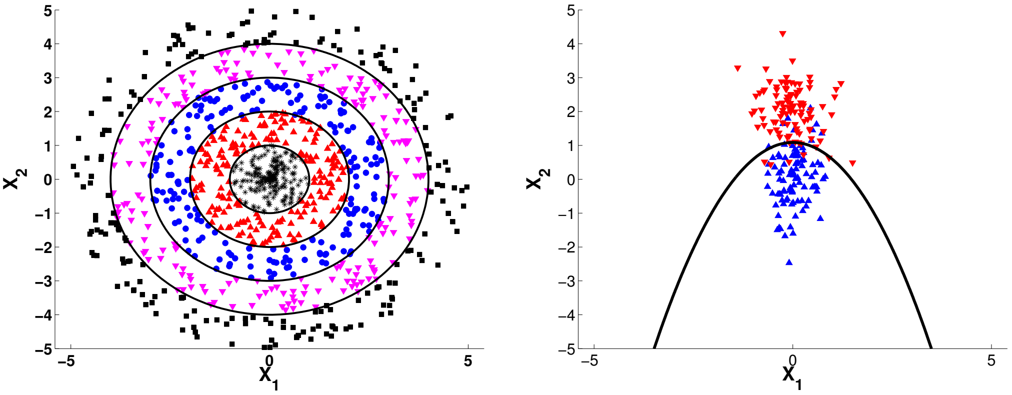


Fig. 1. The Concentric Circle dataset is shown on the left. Five classes are drawn following a uniform distribution in their corresponding ring. Assuming Gaussian distributions the decision boundaries can be obtained. In addition, the two-dimensional synthetic dataset is shown on the right. The class c_1 follows $p(c_1|\boldsymbol{x}) \sim \mathcal{N}\left(\binom{0}{0}, \begin{pmatrix}1/8 & 0\\ 0 & 1/2\end{pmatrix}\right)$, and class c_2 follows a distribution $p(c_2|\boldsymbol{x}) \sim \mathcal{N}\left(\binom{0}{2}, \begin{pmatrix}1/4 & 0\\ 0 & 1/2\end{pmatrix}\right)$. The decision boundary is a parabolic curve.

4 Experimental design for brain tumour diagnosis

So far, the behaviour of the iGDA algorithm has been studied using different benchmark datasets with a focus on various properties. In this section, the iGDA algorithm is applied to a real biomedical problem of high medical relevance: automatic brain tumour classification with ^1H MRS. The current gold standard classification of a brain tumour is a histopathological analysis of biopsy; but this is an invasive surgical procedure with potential adverse consequences for the patient. An alternative is a diagnosis based on ^{1}H MRS, which is a non-invasive technique that provides biochemical information on tissue *in vivo*. The database used for our evaluation contains single voxel proton magnetic resonance spectra (SV ^{1}H MRS) acquired at 1.5T from brain tumours at nine European and one Argentinian hospitals. Data used in this work was gathered during three European projects: INTERPRET, eTUMOUR, and HEALTHAGENTS. An acquisition protocol was defined in INTERPRET to provide maximum compatibility of the spectra obtained using different MRS systems at the different participant hospitals [43,44]. This acquisition protocol was extended to the data acquisition procedure in eTUMOUR and HEALTHAGENTS. The spectra were acquired with MR scanners of several manufacturers: Siemens, General Electric and Philips. The acquisition protocols included Point Resolved Spectroscopy (PRESS) and Stimulated Echo Acquisition Mode (STEAM) sequences [45] with a range in the Time of Repetition (TR, between 1600 and 2020 ms), the Time of Echo (TE, 20 or 30-32 ms), the spectral width (1000-200 Hz), and the number of data-points (512, 1024 or 2048) [2]. Each spectrum was semi-automatically pre-processed

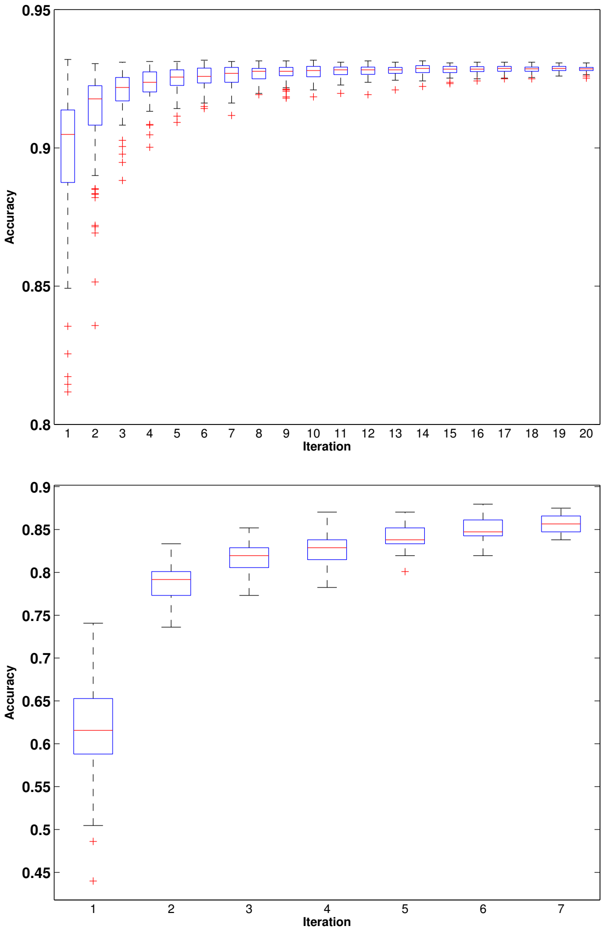
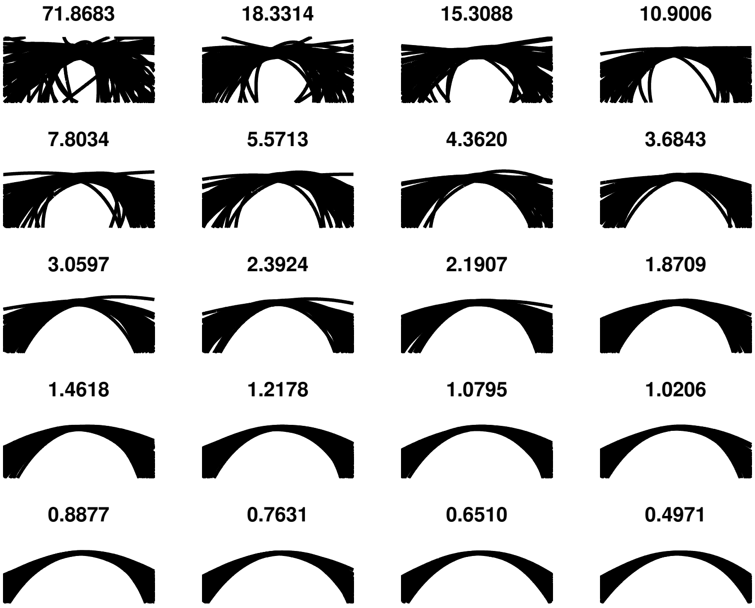


Fig. 2. Boxplots of the accuracy of the models trained with different permutations of the instances. The X-axis shows the iterations of the incremental models. The top Figure shows the results for the two-dimensional synthetic database with 20 iterations. The bottom Figure shows the results for the Vehicle Silhouette database. The convergence of the accuracy proves that the instance order has a *benign* effect on the final models for both datasets.

in order to suppress the water peak, perform a phase correction, suppress the base line, normalize the spectrum area and correct the frequency shift as described in [6].

Spectral patterns contain resonance peaks related to the concentration of different metabolites in the tissue analyzed which are useful for tumour classification purposes [5,6]. Based on a biochemical prior knowledge, a total number of 15 features were obtained from the integration of the signal under a spectral region associated with each metabolite of interest (see Figure 5). Signal quality and the diagnosis associated with each spectrum was validated by the INTERPRET Clinical Data Validation Committee [2], the eTUMOUR Clinical Validation Committee, and expert spectroscopists. In INTERPRET and eTUMOUR the class of each case was determined by a panel of histopathologists, while in HEALTHAGENTS the class was established by the original


71.8683 18.3314 15.3088 10.9006

7.8034 5.5713 4.3620 3.6843

3.0597 2.3924 2.1907 1.8709

1.4618 1.2178 1.0795 1.0206

0.8877 0.7631 0.6510 0.4971

Fig. 3. Convergence of the decision boundaries of each model in 20 iterations for the two-dimensional synthetic database. The variance of the different parameters of the decision boundaries are also shown at the top of each iteration. The iterations are shown left-to-right, top-to-bottom. It can be seen that the first models present arbitrary decision boundaries because their parameters are adjusted from the first sample only. When further samples are used for learning, the decision boundaries and their parameters begin to converge until the final iteration.

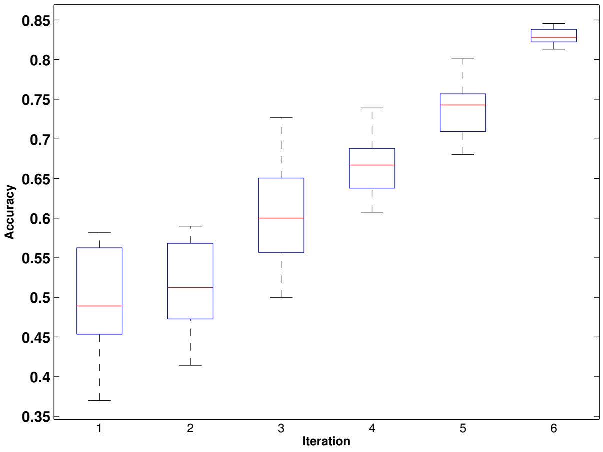


Fig. 4. Convergence of the median accuracies of each combination of samples for the Concentric Circle database. The X-axis shows 6 iterations of the incremental models, each one corresponding to a sample $\mathcal{S}_i$. The convergence proves that the concept order has a slight effect on the accuracy of the models.

histopathologist.

Three types of brain tumour classes were taken into account in the experiments: aggressive brain tumours (AGG), including Glioblastomas and Metastases; low-grade glial tumours (LGG), including grade II Astrocytoma, Oligodendroglioma and Oligoastrocytoma; and Meningioma (MEN). The prevalence of the brain tumour classes considered in this study is shown in Table 5.

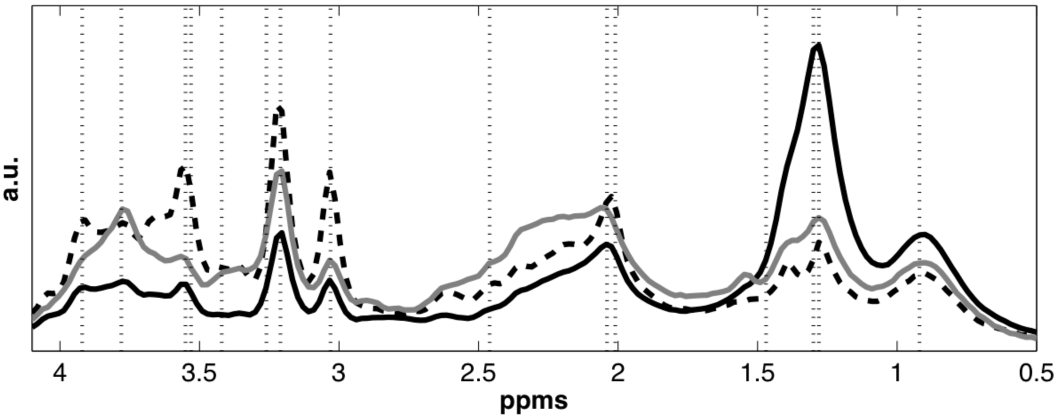


Fig. 5. The features selected for classification are the peak integration of the metabolites observed in the brain (vertical dotted lines): Creatine (3.93 ppm and 3.02 ppm), Choline (3.21 ppm), N-Acetyl Aspartate (2.01 ppm), Myo-Inositol (3.26 ppm and 3.53 ppm), Glycine (3.55 ppm), Taurine (3.26 ppm), Glutamate/Glutamine (2.04 and 2.46 ppm), Alanine (1.47 and 3.78 ppm), Lactate (1.31 ppm), and Lipids (1.29 and 0.92 ppm). The peak integration computes the value of the area under the peaks considering an interval of 0.15 ppm from the assumed peak centre. The mean spectrum of each class of brain tumour is shown: aggressive (solid black line), low grade glial (dashed line), and meningioma (solid grey line).

A gaussian assumption is made since all the variables are continuous. Furthermore, both quadratic and linear classifiers have previously been shown to be powerful enough to achieve good results in automatic brain tumour classification [2,6]. Although there may be more sophisticated feature selection techniques for this problem [46,6], the use of peak integration is a good trade-off between complexity and performance, and it is independent of the different incremental data subsets. Finally, the evaluation method is based on K-random sampling train-test where $K = 100$ because the iterative incremental procedure makes the use of cross-validation or bootstrapping difficult. From the K repetitions the mean accuracy is shown and the standard deviation is used to estimate the confidence interval.

In these experiments, three specific desired features of a clinical decision support system (CDSS) based on machine learning techniques were analyzed: 1) the convergence of the classifiers in terms of stability/plasticity; 2) the effect of including new classes; 3) the customization of the classifiers in relation to the distributions of data in different hospitals.

4.1 Convergence of the iGDA

Following the methodology applied in Section 3.1, we tried to show how the iGDA algorithm is able to learn brain tumour discrimination with MRS in an incremental fashion from different subsets of training data. This was evaluated using the whole brain tumour database to show how the iGDA performance improved in the course of time when new observations were used to update the classifier. The whole database (see Table 5) was randomly split into a training partition (300 samples, 39.5%) and a test partition (460 samples, 60.5%). The decision of using only 39.5% of data for training is justified by the need of simulating a real scenario where the number of instances might be small. Although more incremental iterations could have been performed at the cost of having fewer instances for testing, the selected samples are enough to demonstrate the convergence of the algorithm and reduction in the standard errors of the results. The training partition was split into ten subsets of 30 samples. The whole test partition was used as an independent test set for each new updated classifier. The performance of the classifiers was measured in terms of the accuracy. The linear and quadratic versions of iGDA and the results were also compared to the performance of the other incremental algorithms.

4.2 Inclusion of new classes

In this second experiment, centers in Table 5 were used in order to address the inclussion of new classes. Each center initially contained only two classes (LGG, MEN). An initial classifier was trained from the first group of hospitals (CEN_0). Subsequently, using data from the rest of hospitals, the remaining class (AGG) was included in the following subsets and each generation of the classifier was evaluated with an independent test set. When introducing new classes, a problem of imbalanced classes may appear [42], resulting in a classifier with null sensitivity for the new class. In order to detect such a bias, a geometric mean of sensitivities was used to evaluate the classifiers in these experiments ($\mathrm{G} = \sqrt[|\mathcal{C}|]{\prod_{i=1}^{|\mathcal{C}|} \mathrm{sen}_i}$, where sen_i is the sensitivity of class i). The general behaviour of the G measure is high when all the sensitivities are high and in equilibrium.

4.3 Customization to different centers

The third experiment simulates the customization of the classifier for a hospital by adapting a general model into the specific distribution of one hospital. Data

Center	Classes			Total
	AGG	LGG	MEN	
CEN_0	111	44	29	184
CEN_1	108	48	34	190
CEN_2	114	44	33	191
CEN_3	120	26	49	195
TOTAL	453	162	145	760

Table 5
The different centers and the number of samples per class. AGG: aggressive, LGG: low-grade glial, and MEN: Meningioma.

from three hospitals (CEN_0) were used to train an initial classifier. Three other groups from two hospitals (CEN_1, CEN_2, and CEN_3) were made for testing the iGDA. These groups were chosen to balance the number of samples in each center. In addition, all the centers were grouped together in order to obtain a general behaviour of the convergence of the algorithm to compare with. Table 5 shows the prevalence of each class in the dataset according to the four data groups used. Each center was divided into a test set and four subsets with 20 random samples in each one. Once the initial classifier was trained, it was used to automatically classify data from the test set of the other centers. Then, the first sample $\mathcal{S}_1$ of CEN_1 was used to update the classifier with the iGDA algorithm. The same process was performed with the first sample $\mathcal{S}_1$ of the other two centers, thus obtaining a total of three new incrementally updated classifiers. After incremental updating of the classifier of each center, a new evaluation was carried out using the independent test set of the corresponding center.

5 Results in brain tumour classification with MRS

5.1 Convergence of the iGDA

The comparison with the Learn^{++} and the iLDA algorithms shows that the accuracies of all these methods converge asymptotically (see Figure 6). This result suggest that the iGDA algorithm works properly as an incremental learning algorithm.

Generally speaking, the linear version of the iGDA algorithm performs better than the Learn^{++} and the iLDA algorithms. However, the quadratic version of iGDA needs three incremental updates to reach a comparable accuracy with

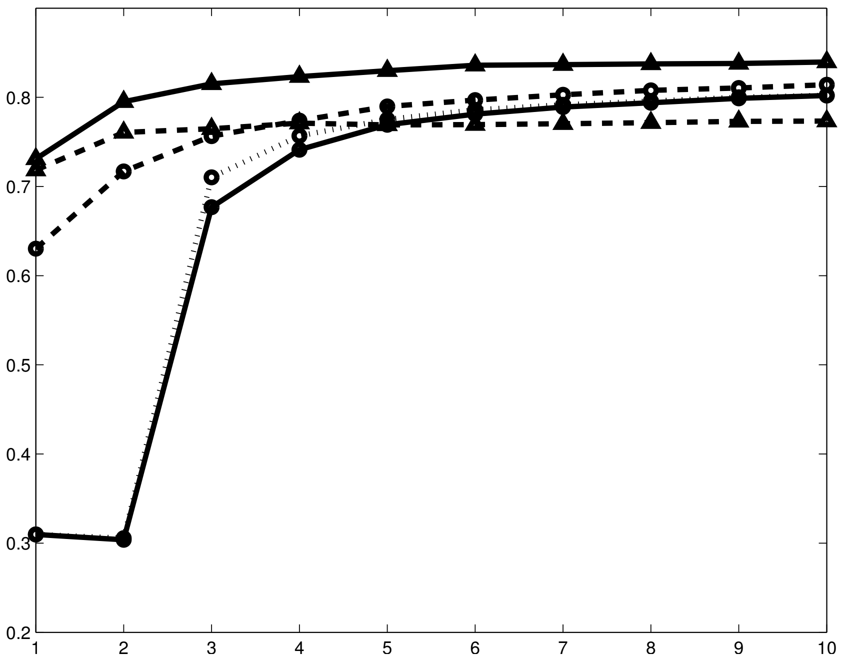


Fig. 6. Comparison of the evolution of the accuracies of the linear iGDA (solid line and triangles), the quadratic iGDA (solid line and circles), iLDA (dashed line and triangles), and Learn^{++} (dashed line and circles) incremental learning algorithms. Also, a naive Gaussian classification model updated from scratch is compared (dotted line and circles). The first iteration is the performance of the initial classifier. From the second bin on, the incremental algorithm is executed. The experiment was repeated 100 times. The plots represent the mean value of all the experiments. The x-axis shows the different moments of time, t_i, of new observed data. The y-axis shows the accuracy. The iGDA using Graybill-Deal weight estimation shows a very good performance and it converges asymptotically.

the other algorithms. This behaviour may be explained by the low number of samples of the less prevalent classes in each subset. Nevertheless, there is asymptotic convergence of all methods: the data fits to the Gaussian model assumed by the iGDA, which describes linear or quadratic boundaries, as well as to the model assumed by the Learn^{++} algorithm, which divides the sample space using multiple hyperplanes.

The significance of differences ($\alpha = 5\%$) among algorithms was evaluated with a multiple comparison test using a Friedman's nonparametric two-way analysis of variance test with Tukey's honestly significant difference criterion from the first to the last iteration. The linear iGDA always displays a significant difference with respect to the other algorithms except with the iLDA in the first iteration. From iteration 8 to 10 the differences among the algorithms are all significant ($p < 0.01$).

5.2 Inclusion of new classes

The mean accuracy of the results obtained when a new class appears inside the new observed samples improve from 0.29 to 0.78 in 10 incremental iterations. Since the convergence is asymptotic, the first two iterations show the biggest improvement: from 0.29 to 0.45 and to 0.57. Thereafter, the improvement is slower. The geometric mean of sensitivities (G) improves from 0 to 0.76. Our results show that the first classifier is unable to correctly classify any sample belonging to the new class and thus $G = 0$. But, after further learning from two additional samples that include cases of the new class, the subsequent classifiers converge, obtaining not only a good accuracy but also a good G without forgetting to classify the initial classes. Our results show that the iGDA is able to introduce the new class into its knowledge base.

5.3 Customization to different centers

The third experiment tried to simulate a practical environment where a trained classifier is used for classification with data coming from different populations of patients and/or different acquisition machines. The results in Figure 7 show how the initial classifier exhibits a performance that clearly needs improvement. Therefore, when the classifier is updated with the new observations, the performance increases significantly with a small additional set of samples. In every new center, the accuracy of the incremental classifier improves in the course of new observations being used to incrementally train the classifier. Each observation included 20 new samples. The centers were joined in a unique set to compare the evolution of each center with the evolution of all the centers and show that the accuracy tends to converge asymptotically.

In general, the sensitivities for the first classification model in the centres CEN_1, CEN_2, and CEN_3 are between 0.71 to 0.76 for AGG, 0.85 to 0.86 for LGG, and 0.29 to 0.58 for MEN. After four incremental iterations the sensitivities vary from 0.79 to 0.83 for AGG, 0.74 to 0.84 for LGG, and 0.51 to 0.73 for MEN. Therefore, the incremental algorithm seems to be prone to balance the sensitivities of the different tumour types, increasing the sensitivities of the AGG and MEN tumour types while slightly decreasing the sensitivity of the LGG tumour types.

Again, a multiple comparison test ($\alpha = 5\%$) was carried out. Initially, only CEN_2 and CEN_3 showed significant differences but by iteration 5, only CEN_3 showed significant differences against the other centers ($p < 0.01$).

The same multiple comparison test ($\alpha = 5\%$) was used to analyze the statistical differences in the incremental models developed in the iterations of each center. These tests showed that the models of CEN_1 and CEN_3 had significant differences among iterations, except for the results of iteration 4 and 5. With respect to the models of center CEN_2 there were significant differences between iteration 1 and 2 and between iterations 2 and 4, and iterations 3 and 5.

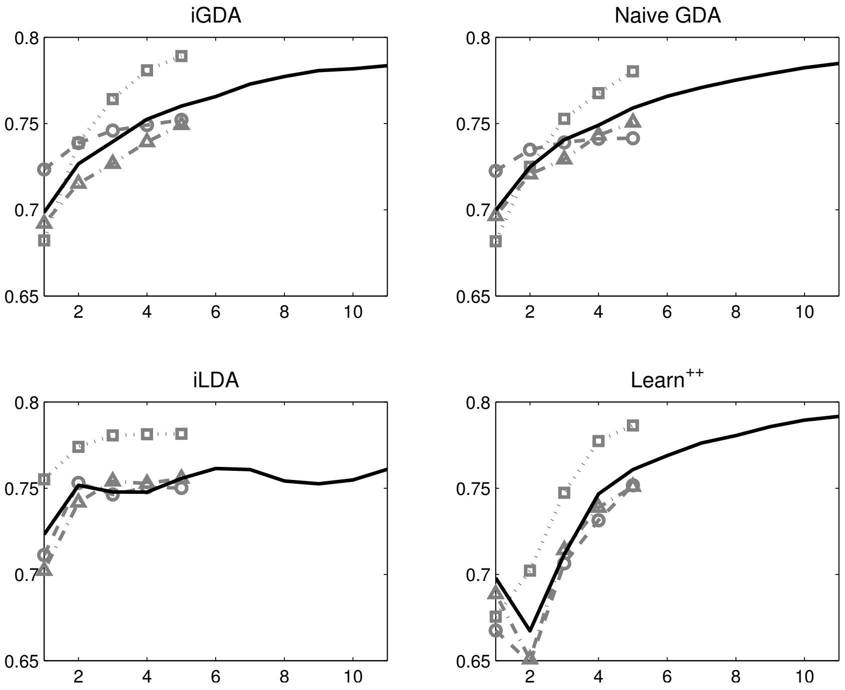


Fig. 7. Comparison of the evolution of the mean accuracies of the different incremental learning algorithms trained with data from center CEN_0 and tested with data of new centers: CEN_1 (grey dash dotted line with triangles), CEN_2 (grey dashed line with circles), CEN_3 (grey dotted line with squares), and the evolution of the convergence for the union of centers CEN_1 to CEN_3 (black).

6 Discussion and conclusions

6.1 Technical aspects of the iGDA

The iGDA algorithm is presented as a new incremental algorithm for Gaussian discriminant analysis based on a weighted combination of different parameter estimations. It obeys the definition about the incremental learning algorithm given by several authors [28,29,17]. iGDA does not use any previous original datasets, but updates its knowledge by means of the information of the newly observed data and its already acquired knowledge. Therefore, it can be used when dealing with problems where past information is inaccessible or where there are problems gathering an appropriate dataset in a reasonable time. In such situations, this incremental learning algorithm can avoid the waiting time by using a small amount of information to build an initial simpler model and then update the model incrementally, and allow for additional classes, as new information arrives.

The implementation of the algorithm is straightforward and the models can be estimated in polynomial time. The complexity of the algorithm is $O\Big(cd^2(N+d)\Big)$, where c is the number of classes, d is the number of variables, and N is the number of instances.

Figure 7 shows that the evolution of the updated classifiers in centers CEN_1 and CEN_2 is comparable to the evolution of the classifiers from all centers taken together. However, the evolution using the dataset from center CEN_3 shows the highest improvement. This may be explained by the prevalence of the different brain tumour types in center CEN_3, which has an influence on the prior probabilities of the models. Hence, while the updated classifiers of CEN_1 and CEN_2 are improving the knowledge concerning the conditional distributions $p(x|c)$, the updated classifiers of CEN_3 are reinforcing the knowledge of the conditional distributions as well as the prior probabilities π_c. The final accuracy reached is similar to the median accuracy rate achieved in [6] for quadratic and linear discriminant analysis. In our results, the iGDA is comparable with the baseline model and the other incremental algorithms.

Since the experiments were repeated 100 times to avoid any possible bias, the results show a general behaviour of the iGDA algorithm. However, when the convergence to a minimum error has been achieved, there may be situations where addition of a new biased dataset results in a model with a slightly poorer performance to that previously, but without statistical significance. Thus, when a convergence has been reached small oscillations in the accuracy of the models may be observed, similar to other iterative procedures.

An interesting feature of iGDA is that it does not have a *malignant* order effect [28], nor at instance level or at concept level. This means that the order of the instances may give rise to slightly different models, but with similar discrimination accuracies. Our results show that the decision boundaries of the models are also similar regardless of the order in which the instances appear, or even the order in which the classes are introduced into the analysis.

One limitation of the iGDA algorithm is that it assumes that the data will follow a Gaussian distribution. This assumption may be useful for real number variables, even when they do not follow a Gaussian distribution, but this approach is useless for discrete distributions, such as Bernoulli or multinomial distributions. Nevertheless, the extension of these concepts may be of interest to other distributions, including discrete ones. The unimodal Gaussian assumption also restricts the type of decision boundaries to linear or quadratic boundaries.

Another feature of the iGDA is its ability to include new classes. However, this ability may lead to an imbalanced class problem [42] if the new class is underrepresented compared to the previous classes. This may be also related to the outvoting problem that occurs in incremental learners based on voting schemes such as the Learn^{++} [17,18]. Furthermore, the behaviour of the weights in multivariate distributions and the combination of the covariance matrices using the Graybill-Deal estimation must still be theoretically studied and is the focus of future work.

6.2 Potential clinical interest of iGDA for brain tumour diagnosis

Primary brain tumours are proportionately less frequent than other cancers, but they are devastating diseases with high mortality. An accurate initial diagnosis of brain tumours has important consequences for therapeutic decisions and prognosis. Compared to most other tumours, obtaining brain tumour tissue for diagnostic purposes is relatively difficult even when using the advance technique of stereotactic biopsy [47]. The clinical DSS that are based on ML techniques and ^{1}H-MRS have shown a promising results for non-invasive brain tumour diagnosis. However, the development of robust classifiers requires acquisition of a large number of cases. Furthermore, in multi-center projects it is usually assumed that the data have similar distributions, however in practice we may expect some differences in data distributions or class assignments. A straightforward application of the incremental method presented here is its ability to customize an already trained classifier to the specific distribution of a particular hospital. In other words, if a hospital has a limited number of samples for a particular class, a classifier trained with data from other hospitals can be used as an initial model and then adapted to the distribution of the patient population or the hospital scanner performance. Thus a classifier can be developed that has a customization to the hospital, but without the need for an unachievable acquisition of local data. The development of new models in the course of time as new data is acquired is related to the concepts of temporal and external validation reported by Altman et al. in [48]. Based on the results, our incremental algorithm could enhance the performance of such models when evaluated with subsequent patients coming from new hospitals.

In the framework of a clinical DSS the iGDA algorithm that has been developed may take advantage of the availability of new information to adapt the knowledge of the current system to the evolution of the data domain and also to extend the lifecycle of the system in a real clinical environment. Assuming that new information is ready for supervised classification at different times, the iGDA algorithm can learn from such new data without access to the previously seen data, even when a new class arises.

The ability to customize a model to a specific clinical centre could be used to improve the behaviour of a state-of-the-art CDSS for aiding brain tumour diagnosis. Further work will include the integration of the incremental algorithm developed in this work into a generic and dynamic **DSS** for clinical environments such as the aforementioned CDSSs and CURIAM [49]. The CURIAM Brain Tumour version [50] offers orientation on brain tumour diagnosis and is currently being tested in a clinical setting at several hospitals in Europe. The incremental learning method shown here may also complement to an audit model of brain tumour classifiers [51] and help provide dynamic optimisation of a CDSS.

Acknowledgements

This work has been partially funded by the Spanish Instituto de Salud Carlos III (ISCiii) through the RETICS Combiomed (RD07/0067/2001). The authors thank the Programa Torres Quevedo from Ministerio de Educación y Ciencia, co-founded by the European Social Fund (PTQ05-02-03386 and PTQ08-01-06802). We thank eTUMOUR, HEALTHAGENTS and INTERPRET partners for providing data, in particular W. Gajewicz (MUL), J. Calvar (FLENI), A. Heerschap (RUNMC), J. Capellades (IDI-Badalona), C. Majós (IDI-Bellvitge), and W. Semmier (DKFZ-Heidelberg). CIBER-BBN is an initiative funded by the VI National R&D&I Plan 2008-2011, CIBER Actions are financed by the Instituto de Salud Carlos III with assistance from the European Regional Development Fund.

Appendix A.iGDA algorithm pseudocode

Algorithm 1 Incremental Gaussian Discriminant Analysis

```
Input: $\mathcal{S}_{i+1} = \{(\mathbf{x}_n, c_n)_{n=1}^N\}; \mathcal{H}_i$
Output: $\mathcal{H}_{i+1}$
Require: $\forall c \in \mathcal{S}_i, N_c > 1$
  for all $c \in \mathcal{S}$ do
    $\pi_c \leftarrow N_c/N$
    $\boldsymbol{\mu}_c \leftarrow \frac{1}{N_c} \sum \boldsymbol{x}_n$
    $\boldsymbol{\Sigma}_c \leftarrow \frac{1}{N_c} \sum (\boldsymbol{x}_n - \boldsymbol{\mu}_c)^{\mathrm{T}} (\boldsymbol{x}_n - \boldsymbol{\mu}_c)$
  end for
  if $\mathcal{H}_{i-1} \neq \emptyset$ then
    for all $c \in \mathcal{S}$ do
      $\omega_{i+1}, \omega_i \leftarrow \textbf{Graybill-Deal}(N_i, \Sigma_i, N_{i+1}, \Sigma_{i+1})$
      $\pi_c \leftarrow \frac{N_c^{i+1} + N_c^i}{N^{i+1} + N^i}$
      $\boldsymbol{\mu}_c \leftarrow \omega_{i+1} \boldsymbol{\mu}_c^{i+1} + \omega_i \boldsymbol{\mu}_c^i$
      $\boldsymbol{\Sigma}_c \leftarrow \omega_{i+1} \boldsymbol{\Sigma}_c^{i+1} + \omega_i \boldsymbol{\Sigma}_c^i$
    end for
  end if
  if linear then
    for all $c \in \mathcal{S}$ do
      $\boldsymbol{\Sigma}_c \leftarrow \boldsymbol{\Sigma}$
    end for
  end if
  for all $c \in \mathcal{S}$ do
    $\boldsymbol{W}_c \leftarrow -(1/2) \boldsymbol{\Sigma}_c^{-1}$
    $\boldsymbol{w}_c \leftarrow \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c$
    $w_{c0} \leftarrow \log \pi_c - (1/2) \log |\boldsymbol{\Sigma}| - (1/2) \boldsymbol{\mu}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c$
  end for
  return $\mathcal{H}_{i+1}$
```

Algorithm 2 Graybill-Deal computation of weights

```
Input: $N_1, N_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$
Output: $\omega_1, \omega_2$
  $S_1 = trace(\boldsymbol{\Sigma}_1)$
  $S_2 = trace(\boldsymbol{\Sigma}_2)$
  $\omega_1 = \frac{N_1/S_1}{\sum_{i=1}^{2} N_i/S_i}$
  $\omega_2 = 1 - \omega_1$
  return $\omega_1, \omega_2$
```

References

[1] INTERPRET Consortium. INTERPRET. Web site, 1999-2001. IST-1999-10310, EC, http://gabrmn.uab.es/interpret/.

[2] Anne R Tate, Joshua Underwood, Dionisio M Acosta, Margarida Julià-Sapé, Carles Majós, À Moreno-Torres, Franklyn A Howe, Marinette van der Graaf, Virginie Lefournier, Mary M Murphy, Alison Loosemore, Christophe Ladroue, Pieter Wesseling, Jean Luc Bosson, Miquel E Cabañas, Arjan W Simonetti, Witold Gajewicz, Jorge Calvar, Antoni Capdevila, Peter R Wilkins, B Anthony Bell, Chantal Rémy, Arend Heerschap, Des Watson, John R Griffiths, and Carles Arús. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. *NMR in Biomedicine*, 19(4):411–434, 2006.

[3] eTUMOUR Consortium. eTumour: Web accessible MR Decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data. Web site. FP6-2002-LIFESCIHEALTH 503094, VI framework programme, EC, http://www.etumour.net.

[4] Horacio González-Vélez, Mariola Mier, Margarida Julià-Sapé, Theodoros N Arvanitis, Juan Miguel García-Gómez, Montserrat Robles, Paul H Lewis, Srinandan Dasmahapatra, David Dupplaw, Andrew C Peet, Carles Arús, Bernardo Celda, Sabine Van Huffel, and Magí Lluch i Ariet. HealthAgents: Distributed multi-agent brain tumor diagnosis and prognosis. *Applied Intelligence*, 30(3):191–202, June 2009.

[5] Juan Miguel García-Gómez, Salvador Tortajada, César Vidal, Margarida Julià-Sapé, Jan Luts, Àngel Moreno-Torres, Sabine Van Huffel, Carles Arús, and Montserrat Robles. The effect of combining two echo times in automatic brain tumor classification by MRS. *NMR in Biomedicine*, 21(10):1112–1125, Sep 2008.

[6] Juan Miguel García-Gómez, Jan Luts, Margarida Julià-Sapé, Patrick Krooshof, Salvador Tortajada, Javier Vicente, Willem Melssen, Elies Fuster-Garcia, Iván Olier, Geert Postma, Daniel Monleón, Àngel Moreno-Torres, Jesús Pujol, Ana-Paula Candiota, M Carmen Martínez-Bisbal, Johan Suykens, Lutgarde Buydens, Bernardo Celda, Sabine Van Huffel, Carles Arús, and Montserrat Robles. Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 22(1):5–18, 2009.

[7] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence, editors. *Dataset Shift in Machine Learning.* The MIT Press, 2009.

[8] Nick W Street and YongSeog Kim. A streaming ensemble algorithm (SEA) for large-scale classification. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 377–382. ACM, 2001.

[9] Marcus A Maloof and Ryszard S Michalski. Incremental learning with partial instance memory. *Artificial Intelligence*, 154(1-2):95–126, 2004.

[10] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000.

[11] José-Luis Sancho, William E Pierson, Batu Ulug, Aníbal R Figueiras-Vidal, and Stanley C Ahalt. Class separability estimation and incremental learning using boundary methods. *Neurocomputing*, 35(1-4):3–26, 2000.

[12] Alistair Shilton, Marimuthu Palaniswami, Daniel Ralph, and Ah Chung Tsoi. Incremental training of support vector machines. *IEEE Transactions on Neural Networks*, 16:114–131, 2005.

[13] Jeffrey C Schlimmer and Douglas H Fisher. A case study of incremental concept induction. In *5th National Conference on Artificial Intelligence*, pages 496–501, 1986.

[14] Paul E Utgoff. Incremental induction of decision trees. *Machine Learning*, 4:161–186, 1989.

[15] Zhi-Hua Zhou and Zhao-Qian Chen. Hybrid decision tree. *Knowledge-Based Systems*, 15(8):515 – 528, 2002.

[16] J. Gama and P. Medas. Learning decision trees from dynamic data streams. *Journal of Universal Conputer Science*, 11(8):1353–1366, 2005.

[17] Robi Polikar, Lalita Udpa, Satish S Udpa, and Vasant Honavar. Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews*, 31(4):497–508, 2001.

[18] Michael Muhlbaier, Apostolos Topalis, and Robi Polikar. Learn++.NC: Combining Ensemble of Classifiers Combined with Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes. *IEEE Transactions on Neural Networks*, 20(1):152–168, 2009.

[19] Gail A Carpenter, Stephen Grossberg, Natalya Markuzon, John H Reynolds, and David B Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3:698–713, Sep 1992.

[20] Koichiro Yamauchi, Takayuki Oohira, and Takashi Omori. Fast incremental learning methods inspired by biological learning behavior. *Artificial Life and Robotics*, 9(3):128–134, 2005.

[21] Sheng Wan and Larry E Banta. Parameter Incremental Learning Algorithm for Neural Networks. *IEEE Transactions on Neural Networks*, 17(6):1424–1438, 2006.

[22] LiMin Fu. Incremental knowledge acquisition in supervised learning networks. *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, 26(6):801–809, November 1996.

[23] S Chandrasekaran, B S Manjunath, Y F Wang, J Winkeler, and H Zhang. An eigenspace update algorithm for image analysis. *Graphical Models Image Processing*, 59(5):321–332, 1997.

[24] Peter Hall, David Marshall, and Ralph Martin. Merging and splitting eigenspace models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1042–1049, 2000.

[25] Shaoning Pang, Seiichi Ozawa, and Nikola Kasabov. Incremental linear discriminant analysis for classification of data streams. *IEEE Transactions on System, Man and Cybernetics*, 35(5):905–914, 2005.

[26] Tae-Kyun Kim, Shu-Fai Wong, Björn Stenger, Josef Kittler, and Roberto Cipolla. Incremental linear discriminant analysis using sufficient spanning set approximations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7.

[27] Zhibin Huang, Kai Ding, Lianwen Jin, and Xue Gao. Writer Adaptive Online Handwriting Recognition Using Incremental Linear Discriminant Analysis. pages 91–95, 2009.

[28] Pat Langley. Order effects in Incremental Learning. In P Reimann and H Spada, editors, *Learning in humans and machines: Towards an interdisciplinary learning science*, pages 1–17. Elsevier, Oxford, 1995.

[29] Christophe Giraud-Carrier. A note on the utility of incremental learning. *AI Communications*, 13(4):215–223, 2000.

[30] Stephen Grossberg. Nonlinear neural networks: principles, mechanisms and architectures. *Neural Networks*, 1(1):17–61, 1998.

[31] Antoine Cornuéjols. Getting order indepence in Incremental Learning. In *AAAI Spring Symposium on Training Issues in Incremental Learning*, pages 43–54, 1993.

[32] Nicola Di Mauro, Floriana Esposito, Stefano Ferilli, and Teresa M A Basile. Avoiding Order Effects in Incremental Learning. In *In S. Bandini and S. Manzoni (Eds.), Advances in Artificial Intelligence (AI*IA05) LNCS*, pages 110–121. Springer-Verlag, 2005.

[33] Franklin A Graybill and R B Deal. Combining unbiased estimators. *Biometrics*, 15:543–550, 1959.

[34] Nien F Zhang. The uncertainty associated with the weighted mean of measurement data. *Metrologia*, 43:195–204, 2006.

[35] Nabendu Pal, Jyh-Jiuan Lin, Ching-Hui Chang, and Somesh Kumar. A revisit to the common mean problem: Comparing the maximum likelihood estimator with the Graybill-Deal estimator. *Computational Statistics & Data Analysis*, 51(12):5673–5681, August 2007.

[36] Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification.* Wiley-Interscience, New York, NY, 2001.

[37] Yoav Freund and Robert E Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *Computer Systems Science*, 57(1):119–139, 1997.

[38] Robert E Schapire. The Strength of Weak Learnability. *Machine Learning*, 5:197–227, 1990.

[39] Christopher M Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[40] A Asuncion and D J Newman. UCI Machine Learning Repository, 2007.

[41] Jerome H Friedman. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84(405):165–175, Mar 1989.

[42] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: a systematic study. *Intelligent Data Analysis Journal*, 6(5):429–449, Nov 2002.

[43] Margarida Julià-Sapé, Dionisio M Acosta, Mariola Mier, Carles Arús, and Des Watson. A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 19(1):22–33, Feb 2006.

[44] Marinette van der Graaf, Margarida Julià-Sapé, Franklyn A Howe, Anne Ziegler, Carles Majós, Àngel Moreno-Torres, Mark Rijpkema, Dionisio M Acosta, Kirstie S Opstad, Yvonne M van der Meulen, Carles Arús, and Arend Heerschap. MRS quality assessment in a multicentre study on MRS-based classification of brain tumours. *NMR Biomed*, 21(2):148–158, 2008.

[45] Luigi Landini, Vincenzo Positano, and Maria Filomena Santarelli, editors. *Advanced Image Processing in Magnetic Resonance Imaging.* CRC Press, 2005.

[46] Jan Luts, Jean-Baptiste Poullet, Juan M García-Gómez, Arend Heerschap, Montserrat Robles, Johan AK Suykens, and Sabine Van Huffel. The effect of feature extraction for brain tumour classification based on short echo time 1h mr spectra. *Magnetic Resonance in Medicine*, 60(2):288–298, 2008.

[47] Ruben Dammers, Joost W Schouten, Iain K Haitsma, Arnauld JPE Vincent, Johan M Kros, and Clemens MF Dirven. Towards improving the safety and diagnostic yield of stereotactic biopsy in a single centre. *Acta Neurochir (Wien)*, 152(11):1915–21, 2010.

[48] Douglas G Altman, Yvonne Vergouwe, Patrick Royston, and Karel G Moons. Prognosis and prognostic research: validating a prognostic model. *BMJ*, 338(b605):1432–1435, 2009.

[49] Carlos Sáez, Juan Miguel García-Gómez, Javier Vicente, Salvador Tortajada, Miguel Esparza, Alfredo T Navarro, Elies Fuster-Garcia, Montserrat Robles, Luis Martí-Bonmatí, and Carles Arús. A generic Decision Support System featuring an assembled view of predictive models for Magnetic Resonance and clinical data. In *ESMRMB 2008: 25th Annual Scientific Meeting*, October 2008.

[50] Carlos Sáez, Juan Miguel García-Gómez, Javier Vicente, Salvador Tortajada, Elies Fuster-Garcia, Miguel Esparza, Alfredo T. Navarro, and Montserrat Robles. Curiam BT 1.0, decision support system for brain tumour diagnosis. In *ESMRMB 2009: 26th Annual Scientific Meeting*. Springer, october 2009.

[51] Javier Vicente, Juan Miguel García-Gómez, Salvador Tortajada, Alfredo T Navarro, Franklyn A Howe, Andrew C Peet, Margarida Julià-Sapé, Bernardo Celda, Pieter Wesseling, Magí Lluch-Ariet, and Montserrat Robles. Ranking of Brain Tumour Classifiers Using a Bayesian Approach. In *Bio-inspired systems: Computational and Ambient Intelligence (Part I)*, volume 5517 of *Lecture Notes in Computer Science*, pages 33–50. Springer, 2009.