Classifying malignant brain tumours from ^{1}H-MRS data using Breadth Ensemble Learning

Albert Vilamala, Lluís A. Belanche and Alfredo Vellido
Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3. 08034, Barcelona - Spain
Email: {avilamala, belanche, avellido}@lsi.upc.edu

***Abstract*—In neuro oncology, the accurate diagnostic identification and characterization of tumours is paramount for determining their prognosis and the adequate course of treatment. This is usually a difficult problem *per se*, due to the localization of the tumour in an extremely sensitive and difficult to reach organ such as the brain. The clinical analysis of brain tumours often requires the use of non-invasive measurement methods, the most common of which resort to imaging techniques. The discrimination between high-grade malignant tumours of different origin but similar characteristics, such as glioblastomas and metastases, is a particularly difficult problem in this context. This is because imaging techniques are often not sensitive enough and their spectroscopic signal is overall too similar. In spite of this, machine learning techniques, coupled with robust feature selection procedures, have recently made substantial inroads into the problem. In this study, magnetic resonance spectroscopy data from an international, multi-centre database were used to discriminate between these two types of malignant brain tumours using *ensemble learning* techniques, with a focus on the definition of a feature selection method specifically designed for ensembles. This method, Breadth Ensemble Learning, takes advantage of the fact that many of the frequencies of the available spectra convey no relevant information for the discrimination of the tumours. The potential of the proposed method is supported by some of the best results reported to date for this problem.**

I. Introduction

In oncology, the early diagnosis of a tumour and its characterization are crucial for the provision of the most accurate prognosis and the adequate treatment for the patient. Neuro oncology faces an added difficulty: the risks of undergoing surgery in such a sensitive organ as the brain. The need to limit these risks has fuelled, over the last two decades, a considerable amount of research in alternative non-invasive measurement techniques based on Nuclear Magnetic Resonance (NMR) in modality variants such as Magnetic Resonance Imaging (MRI) and Magnetic Resonance Spectroscopy (MRS), which could still provide accurate diagnosis from indirect information.

Nevertheless, the interpretation of the NMR outcome is often less than obvious in this context. MRI can be ambiguous and imprecise in some cases (such as in the differentiation of the high-grade tumours that are the subject of this study), while MRS lacks spatial global information that MRI provides and not all radiologists are trained to make sense of it. In recent years, both modalities have been merged in the form of MRSI [1] to overcome each other's limitations. In any case, computer-based methods for automated pattern recognition, often stemming from the fields of computational intelligence and machine learning can be used to ease the interpretation of the NMR outcome and thus assist experts [2], [3].

In this paper, we investigate the problem of discriminating between high-grade malignant tumours of different origin, but similar biochemical signature, namely *glioblastomas* and *metastases* (glioblastomas have their origin in the brain, while metastases can have their origin elsewhere). Distinguishing between them is paramount, given the fact that each of these pathologies requires a completely different treatment. Previous studies on this topic [4], [5], [6], based on Single-Voxel Proton MRS (SV-^{1}H-MRS), stress the complexity of the task at hand. These studies have resorted to diverse approaches, including subjective and automated feature selection, feature transformation and extraction, and mostly, simple linear classifiers.

We hypothesize that this problem can not be solved adequately by a single classifier due to the complexity of the available spectra, the intraclass variability and the high interclass similarity. Thus, minor differences between instances might be the key to a successful classification.

We propose the use of *ensemble methods* for the problem of discriminating glioblastomas from metastases. As in other ensemble methods [7], we aim at locally specializing the base classifiers of the ensemble in different subsets of instances by using different subspaces of features as predictors.

In this study, MR spectra from an international, multi-centre database were used to build a specific method, namely Breadth Ensemble Learning, able to differentiate among these two types of malignant brain tumours. It makes the most of the fact that many of the frequencies within the spectra contain no relevant information for the current task by using a feature selection strategy specifically designed for ensembles dealing with these data. The suitability of the method is supported by some of the best results reported to date for this problem.

The rest of the paper is structured as follows: The available MRS data are first introduced. This is followed by a summary description of the proposed method. In Section IV, this method is evaluated and its results compared with other ensemble and non-ensemble methods. Then, the generated solution is interpreted and its results compared to those in previous

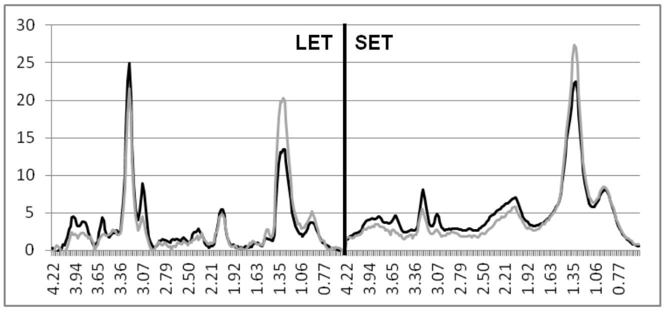

Fig. 1. Mean spectra as a function of frequency (in ppm) of glioblastomas (grey) and metastases (black) from the INTERPRET database, for both long and short echo times (LET and SET, respectively).

investigations. The study wraps up with some conclusions and an outline of future work.

II. Materials

The data used in this study come from two different sources. A total of 78 glioblastomas and 31 metastases were gathered from the international, multi-centre INTERPRET European project database [8]. An independent hold-out set consisting of 30 glioblastomas and 10 metastases was also used to evaluate the proposed method. These were gathered from the *eTumour* and *HealthAgents* research projects [9].

In more detail, each of the spectra in these datasets consists of proton resonance signals at a finite number of equally spaced frequencies. Many brain metabolites are known to generate signals at specific frequencies in the MR signal [10]. A total of 195 of these frequencies, validated by experts as corresponding to the most relevant frequency interval in the spectrum, were used in the classification experiments. Each tumour type is expected to have a characteristic spectral signature (Fig. 1). A label indicating the type of tumour was also available in the datasets. In the case of the hold-out set, labels were only used *a posteriori*.

The time of echo is a parameter of signal acquisition. It determines the presence and visibility of different metabolites within the spectrum. According to [11], data acquired at short echo time – SET ($20 - 40ms$) yield a better resolution for certain metabolites (e.g. lipids, myo-inositol, glutamine and glutamate) at the expense of overlapping some resonances (e.g. glutamine/glutamate and N-acetylaspartate), which may make the interpretation of the spectra difficult. Instead, data retrieved at long echo time – LET ($135 - 144ms$) yield less baseline distortion but also less clearly resolved metabolites.

Several studies [12], [13] have shown the differential advantage for classification of using LET and SET in combination. In this study, we employ both LET and SET by direct concatenation of the spectra.

The available data show a number of challenges that hinder the classification task. Among them, their high dimensionality compared to the low number of instances available; the non-informativeness of many features; the high correlation among features given the spectral shape of the data; and, above all, the class-unbalanced nature of the analysed data, which contain many more glioblastomas than metastases.

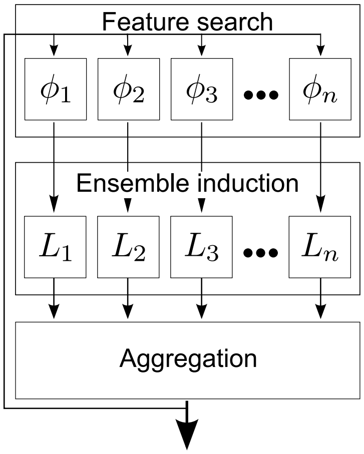

Fig. 2. The system is composed of 3 main modules: the *feature search* (composed of different subsets of features ϕ_i), the *ensemble induction* (made of several base classifiers L_i) and the *aggregation strategy*.

III. Methods

The core of any ensemble method is the block containing several base classifiers. Its purpose is having *different* classifiers which yield diverse predictions for a given data instance. The final ensemble prediction is calculated by appropriately combining the predictions of the base classifiers. Combining the predictions is far from being a trivial matter, giving rise to an active field of research within the ensemble methods community [14].

Besides, the task of designing diverse classifiers in order to predict differently can be accomplished using a variety of techniques. Among the most widely used are those which attempt to train every classifier using unequal sets of instances; training classifiers which are distinct in nature (e.g. using different classifiers or setting their parameters differently); using a unequal subset of data features for every classifier; or a combination thereof.

The coverage of the whole space of instances should be ensured by stimulating the classifiers to disagree between them, aiming towards a local specialization of each classifier. This fact might even lead individual components to lose predictive ability if the whole ensemble benefits from it. Individual inaccuracies are corrected by integrating the individual outcomes in a global prediction.

A. Structure overview

The system introduced in this section (Fig. 2), even if being composed of three basic modules (feature search, ensemble induction and aggregation strategy), works in a wrapper-like fashion, meaning that every module is not sufficient by itself, but completely dependent on the others. Thus, every decision made during the construction of the system is subject to the performance of all components.

The workflow starts in the feature search module by selecting a different subset of features associated to each base classifier. In the ensemble induction module, each classifier estimates the class membership for every instance according to the information that it possesses. Then, every output from the classifiers is aggregated for the whole ensemble to provide a single estimation. Finally, the ensemble performance is assessed and this value is used to provide feedback to the feature search module again. Notice that at each iteration, updated information regarding the performance of the ensemble is used to properly refine the selection of features.

We must stress at this point the importance of the proposed procedure followed to build the ensemble. According to most of the existing literature, ensembles are built by initially training the first base classifier with the target of improving its individual performance. Once this has been achieved, the procedure goes on constructing the rest of the classifiers either independently w.r.t. the previous ones (as in Bagging [15]), or driving the training towards those cases where the firsts classifiers were weak at predicting (as in Boosting [16]).

Contrarily, our method seeks to improve the ensemble performance by iteratively adding or removing one feature to/from each classifier in such a way that leads to the greatest improvement of the overall ensemble performance. For this reason we have named our algorithm *Breadth Ensemble Learning* (BEL).

B. Feature selection

The process of selecting the proper subset of features for each of the base classifiers is crucial in our method because all the mandatory diversity introduced in the ensemble depends only on the decisions made at this phase. A sequential selection search properly adapted to deal with the proposed ensemble method was chosen.

Let Θ be the full set of features and let n denote the number of base classifiers (which is constant). We denote by $L_i(\phi)$ the i-th base classifier developed using the feature subset ϕ. The ensemble at time (iteration) t can then be expressed as $\mathcal{L}(t) = \{L_1(\phi_1(t)), \ldots, L_n(\phi_n(t))\}$, where $\phi_i(t) \subseteq \Theta$.

To form the next ensemble, $\mathcal{L}(t+1)$, from $\mathcal{L}(t)$, we proceed as follows. For the i-th base classifier, three possibilities are considered: add the best feature to $\phi_i(t)$, remove the worst feature from $\phi_i(t)$, or leave $\phi_i(t)$ unchanged. The choice that leads to the highest *overall ensemble* performance will be selected. The best feature $B_i(t+1)$ for L_i is the feature that, when added to $\phi_i(t)$, leads to the best ensemble performance:

$$B_i(t+1) = \underset{\theta \in \Theta \setminus \phi_i(t)}{\arg\max}\ P(\{L_1(\phi_1(t)), \ldots, L_i(\phi_i(t) \cup \{\theta\}), \ldots, L_n(\phi_n(t))\})$$

where P is the ensemble performance measure, described in Section III-D. Conversely, the worst feature $W_i(t+1)$ for L_i is the feature that, when removed from $\phi_i(t)$, leads to the best ensemble performance:

$$W_i(t+1) = \underset{\theta \in \phi_i(t)}{\arg\max}\ P(\{L_1(\phi_1(t)), \ldots, L_i(\phi_i(t) \setminus \{\theta\}), \ldots, L_n(\phi_n(t))\})$$

Then $\phi_i(t+1)$ is set to either $\phi_i(t) \cup \{B_i(t+1)\}$, $\phi_i(t) \setminus \{W_i(t+1)\}$ or $\phi_i(t)$, depending on which choice leads to the best performance when $L_i(\phi_i(t+1))$ is used. This updating process is repeated for all the base classifiers to form $\mathcal{L}(t+1)$.

Because we allow each classifier to add, remove or keep a feature in every iteration, it might be the case that different subsets have different number of features. This is not prevented, since there is no reason to enforce feature subsets of the same size.

C. Base classifiers

A sensitive decision when designing an ensemble is the selection of the base classifiers. One of the requirements we ask to our classifiers is the possibility to provide soft decisions to their predictions in the form of posterior probabilities. By providing those probabilities, each classifier is able to give not only a crisp qualitative measure about which class a given instance belongs to, but also a quantitative value regarding the degree of membership of the instance w.r.t. the different classes. This will be useful in the aggregation phase, since the contribution of each classifier will be automatically weighted according to the confidence expressed by the posterior probability.

According to this requirement, Bayesian classifiers were chosen, which provide probabilities in a natural manner. Specifically, we employ Naive Bayes (NB), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Previous studies dealing with the same kind of data [17] concluded that linear methods were better suited to the classification task at hand than nonlinear ones, specially emphasizing the results obtained by LDA. Therefore, our priority will be using this last classifier.

Another classifier known for its versatility and which has been shown to achieve high performance in many domains, namely Support Vector Machines (SVM), is also considered. More precisely, we use a linear least squares SVM (LS-SVM) with the add-on proposed by Platt [18], whose output is able to approximate posterior probabilities by fitting the distances between support vectors and instances to a sigmoid function previously trained using cross-validation.

Furthermore, we want to explore the use of tree classifiers, since they are the most commonly employed in ensemble architectures. This is because of their relatively fast computation and their ability to provide high variance among the different trees within the ensemble, which might end up generating the desired diversity property. For instance, both Bagging and Boosting, as well as Random Forest, were originally designed to use trees as their component classifiers. In our study we use the CART [19] classifier.

D. Aggregation strategy

A strategy to combine the results outputted by the base classifiers must be defined. Among the wide range of techniques to compute a final ensemble decision, we have selected a simple well-known arithmetic strategy [14]. Because the outputs of the base classifiers provide a class membership probability, the confidence that every classifier gives to each prediction is already implicitly taken into consideration by the value provided. Thus, a simple measure as the average among all the predictions $\rho(t) = \frac{1}{n}\sum_{i=1}^{n}\psi_i(t)$, where ρ is the ensemble prediction and ψ_i the prediction of the i-th base classifier at time t, will suffice for our purpose. Moreover, it has the nice property of being a soft output, meaning that the prediction of an unknown sample can be interpreted as the class membership probability, or *confidence*, that the whole ensemble assigns to this prediction.

The performance measure P (e.g. Area Under the ROC Curve), used in Section III-B, which evaluates how well the ensemble is doing depending on the available data, is calculated at this point and its value fed again to the system in order to aid the feature search.

We are aware of the importance of the aggregation strategy in the final ensemble performance, because the same outputs from the base classifiers might lead to completely different ensemble final predictions depending on the aggregation technique. However, we must keep in mind that averaging is used over the whole ensemble construction, which will be optimized for that measure. Comparing the performance of the breadth ensemble learning algorithm as a function of the aggregation strategy is out of the scope of the current study.

E. Initial conditions

Selecting the initial most appropriate features is not a crucial matter in our method, given the possibility to remove a poor feature at any step of the algorithm. However, it is advisable to start the search from an already advantageous status, guiding it towards a promising path.

Starting the algorithm with empty feature subsets would lead the system to propose adding the same feature to each one of the feature subsets, hence generating no diversity at all.

The random selection of features would also be a poor alternative in our domain, since the vast majority of the available features are known to contain little relevant information.

The current implementation relies on the Relieved [20] algorithm, which is a version of Relief-F that samples all the instances exactly once in the task of choosing the initial most informative features. More precisely, Relieved outputs a list of features ranked according to their discrimination power for the current task. Then, the n best positioned features are kept and distributed such that every feature subset at $t = 0$ is composed of only one feature, that is, $\phi_i(0)$ contains the i-th feature from the list provided by Relieved.

IV. Results

The INTERPRET dataset, after standardization, was used in the training phase for feature selection and model fitting using a leave-one-out cross-validation procedure. This was driven by iteratively maximizing the *Area Under the ROC Curve (AUC)* as a measure of performance [21]. The process was stopped whenever it was not possible to improve the performance measure in the next iteration. At the end of the execution, the selection of features that achieved this maximum performance were kept as solution.

Subsequently, the system was retrained using the whole INTERPRET dataset with the best subsets of features selected in the previous step. The system was tested against the standardized hold-out set to gauge the real system performance.

AUC was used as loss function due to its independence from the classification threshold parameter.

The suitability of our method was assessed by calculating not only the AUC, but also the *Area Under the Convex Hull of the ROC Curve (AUH)* metric [22]. The AUC might underestimate the quality of the prediction in small datasets while AUH might slightly overestimate it. Thus, both metrics were calculated.

Furthermore, and in order to allow comparison with previous studies, the *accuracy (ACC)*, the *F-measure (F)*, and *balanced error rate (BER)* were also calculated. We took advantage of the posterior probabilities provided by our system which, by performing a ROC analysis on the validation set, allowed us to select the most suitable threshold where to split the data for classification purposes. This threshold was used in the hold-out set to calculate these measures of performance.

The only hyperparameter that our method requires is the number of base classifiers n. In this study an educated choice for n was made, setting the value to 50. The evaluation of the performance depending on the number of classifiers n is out of the scope of this study. The test bench included five different base classifiers (NB, LDA, QDA, LS-SVM and CART).

Notice that our method is completely deterministic for ensembles using NB, LDA, QDA and LS-SVM as base classifiers for a given dataset. We have removed all sources of randomness by using leave-one-out as the cross-validation strategy every time a resampling was needed and none of these algorithms generates a source of variation. For this reason, only a point performance value is provided.

However, due to the computational cost of the experiments with CART, 10-times 10-fold cross-validation was carried out instead for the ensembles composed of these base classifiers.

According to the results shown in Table I for 50 classifiers, the use of LDA as base classifiers seems the most adequate choice, since the system built with them is able to achieve an AUC of 0.88 and an AUH of 0.91 for the hold-out set, clearly outperforming its competitors. Furthermore, the rest of the calculated metrics support the previous statement, as they reach their best values for LDA.

Second best is the system consisting of LS-SVM, achieving a remarkable AUC of 0.84 and an AUH of 0.88.

For a variety of reasons, the ensembles composed by either NB, QDA or CART achieve poor performance, with values under 0.65 as AUC in all cases. QDA might fail due to the fact that the system is overfitting the training data by trying to fit quadratic functions. Also, CART and NB are unable to model the structure of the data. The former may select a wrong strategy to achieve its purpose and the latter might be too simplistic.

TABLE I
BREADTH ENSEMBLE LEARNING PERFORMANCE USING DIFFERENT BASE CLASSIFIERS

	# classifiers	AUC	AUH	ACC	F	BER
NB	1	0.59	0.68	0.80	0.80	0.40
	50	0.61	0.74	0.85	0.87	0.33
LDA	1	0.79	0.83	0.82	0.86	0.35
	50	0.88	0.91	0.87	0.88	0.22
QDA	1	0.58	0.68	0.77	0.79	0.37
	50	0.61	0.72	0.77	0.81	0.47
LS-SVM	1	0.68	0.76	0.80	0.86	0.35
	50	0.84	0.88	0.82	0.88	0.22
CART	1	0.58 ± 0.07	0.58 ± 0.07	0.75 ± 0.00	0.78 ± 0.05	0.46 ± 0.10
	50	0.65 ± 0.06	0.74 ± 0.04	0.81 ± 0.02	0.83 ± 0.02	0.37 ± 0.03

A. Setting the baseline

In order to validate the necessity of using an ensemble architecture instead of a single classifier, results in the previous section shall be compared with the results obtained when only one base classifier (n=1) is used.

For these tests, the same procedure previously explained was used to set up the initial conditions for the single classifier.

Table I summarizes the different measures of performance obtained depending on the employed base classifier. CART seems to be the worst performing one. The reason for this low performance may be the way that features are selected: it might pick irrelevant features to split the data.

QDA achieves, again, poor performance in the hold-out set. The low number of available observations makes the computation of the covariance matrices an unreliable process.

NB follows it in low rank performance. Its simplicity when modelling the underlying data distribution might be the probable reason for its failure.

LS-SVM, known to work reasonably well in many problems, also present an acceptable performance in the domain of this study, outperforming the previously mentioned techniques, and thus reinforcing the hypothesis that linear models are suitable to approach our problem.

LDA is the classifier that performs best for our tests in terms of AUC, AUH and accuracy. This result correlates with previous studies [23] which prove the suitability of this technique to deal with the discrimination between glioblastomas and metastases.

Nevertheless, the performance of all single classifiers was improved when an ensemble strategy was applied. This gain supports the comparative advantage of using an ensemble setting. The most remarkable improvement was achieved precisely by LS-SVM and LDA. LDA, for instance, increased its performance from an AUC of 0.79 in a non-ensemble architecture to a 0.88 using the ensemble of 50 base classifiers.

B. Comparing against classical ensemble methods

Our method, specifically designed to deal with the discrimination of gioblastomas from metastases using ^{1}H-MRS data, has been compared to three of the most commonly used general-purpose ensemble methods, namely Random Forest [7], Bagging [15] and Boosting (Adaboost.M1 [16]).

The parameters have been optimized according to the advice provided by the authors. The best settings were actively sought to achieve the highest average AUC in a battery of 100 executions for each setting:

- For Random Forest, 500 trees were grown, each one of them randomly picking 20 features (the square root of the total number of features) per node.
- In the case of Bagging, 100 trees were grown, whose nodes were split after accumulating 20 instances and increasing its fit by 0.5.
- Boosting was tuned with the same values as Bagging but the parameter controlling the fit increment was set to 0.4.

The system was constructed using the INTERPRET dataset and the performance was assessed on the hold-out. For those algorithms not providing a posterior probability, the quotient between the number of trees voting the positive class over the total number of votes was used as posterior probability.

As seen in Table II, the majority of ensemble methods evaluated in our testing environment achieve quite low performance results. When no feature selection was applied beforehand, Random Forest was only able to reach an AUC of 0.67 ± 0.01 in the hold-out set. The method seems to fail in fitting the data, possibly due to the selection of the wrong features leading to a poor performance in test.

Similarly, Bagging obtained an AUC of 0.69 ± 0.04. As with Random Forest, it seems that the algorithm is not capable of modelling the data.

Finally, Adaboost.M1 reached a slightly higher AUC value of 0.71 ± 0.02. This algorithm achieves the best performance within the general-purpose ensemble methods assessed in this study. However, these results are significantly poorer than those obtained by our technique.

The inability of these classical ensemble methods to model the data might be attributed to the lack of a wise feature selection. Therefore, we have performed another set of experiments where the most informative features have been selected prior to the ensemble execution in order to reduce the disadvantage of these methods against our BEL strategy, which does embed feature selection.

TABLE II
PERFORMANCE OF DIFFERENT ENSEMBLE METHODS ON ^{1}H-MRS DATA

Feature Selection	Ensemble Technique	AUC	AUH	ACC	F	BER
None	Random Forest (CART)	0.67 ± 0.01	0.77 ± 0.01	0.77 ± 0.02	0.86 ± 0.01	0.44 ± 0.07
	Bagging (CART)	0.69 ± 0.04	0.72 ± 0.04	0.75 ± 0.05	0.83 ± 0.04	0.35 ± 0.04
	Boosting (Adaboost.M1, CART)	0.71 ± 0.02	0.78 ± 0.02	0.74 ± 0.04	0.83 ± 0.03	0.36 ± 0.03
Filter (Relieved)$_{(m=14)}$	Random Forest (CART)	0.59 ± 0.02	0.68 ± 0.02	0.75 ± 0.00	0.86 ± 0.00	0.50 ± 0.00
	Bagging (CART)	0.62 ± 0.03	0.70 ± 0.03	0.73 ± 0.03	0.83 ± 0.02	0.39 ± 0.03
	Boosting (Adaboost.M1, CART)	0.62 ± 0.04	0.68 ± 0.03	0.76 ± 0.03	0.85 ± 0.02	0.40 ± 0.05
Filter (Random Forest)$_{(m=23)}$	Random Forest (CART)	0.67 ± 0.01	0.74 ± 0.01	0.78 ± 0.02	0.86 ± 0.01	0.35 ± 0.02
	Bagging (CART)	0.71 ± 0.04	0.73 ± 0.04	0.75 ± 0.06	0.83 ± 0.05	0.34 ± 0.04
	Boosting (Adaboost.M1, CART)	0.72 ± 0.02	0.77 ± 0.02	0.72 ± 0.04	0.81 ± 0.03	0.38 ± 0.03
Embedded	Breadth Ensemble Learning (LDA)	0.88	0.91	0.87	0.88	0.22

More precisely, Random Forest itself was used to select the best features. That is, we launched this method using the parameters already described, and sorted the features according to the average Gini index [19] in a sequence of 100 runs. A second strategy consisting in applying Relieved [20] was also applied to sort the features.

Finally, each ensemble method evaluated in this section was run 100 times using a subset of the best m features returned by our feature selection strategies, were m was set according to the *elbow criterion* [24].

In light of the results shown in Table II, applying Relieved in our domain as the feature selection strategy is not a good choice, since all the models perform worse than using no feature selection. This might be due to the inability of Relieved to take into account the redundancies between nearby features.

On the other hand, the use of Random Forest as a feature selection might slightly improve the performance of some models, but not much difference is appreciated.

C. Comparing to previous studies

In the tests using the hold-out set, an AUC of 0.88 and a quite coherent AUH of 0.91 were obtained. The overall accuracy was 87.5%. These results compare favourably with those recently reported in [6], where an AUC of 0.86 and an AUH of 0.91 were achieved using the same data, and therefore rank with the best obtained to date, according to the authors' knowledge.

D. Calculating diversity

In this section we use three different measures of diversity to assess the level of discrepancy among the different classifiers conforming the ensemble.

The first measure is the Disagreement Measure [14], which computes the average number of instances that are classified differently for every pair of base learners.

The second measure is the Q-statistic [14], a pairways measure that evaluates the dependency between two pairs of classifiers. Its values range between -1 and 1, being 0 whenever the two classifiers are independent.

Finally, we have also computed the Entropy measure [14], as an unpaired measure to assess the overall ensemble diversity.

TABLE III
DIVERSITY ON BREADTH ENSEMBLE LEARNING

Dataset	Disagreement	Q-statistic	Entropy
INTERPRET	0.32	0.47	0.47
Hold-out	0.31	0.46	0.47

The scores obtained by these measures of diversity on the INTERPRET and Hold-out set, shown in Table III, confirm the existence of a certain degree of diversity in our algorithm, even though each measure rates it differently. They demonstrate the variability on measuring diversity and the difficulty of a useful empirical evaluation [14].

E. Using a synthetic dataset

Apart from evaluating the performance of the BEL technique in the domain of brain tumour diagnosis, we have also performed several tests on a set of synthetic data, specifically generated to have subgroups of highly correlated features, as in our domain of application, without the sample size limitation.

A set of $1,000$ training instances and $100,000$ testing samples were generated. Each instance was composed of 500 features of which only the firsts 50 were informative for the binary classification task at hand and the remaining 450 were noise. The 50 valuable features consist of 5 independent groups of 10 highly correlated features. A thorough explanation about the generating process can be found in [25].

The construction of the BEL system, made of LDA base classifiers, has been carried out by a 10-times 10-fold cross-validation strategy on the training instances and has been evaluated over the whole testing set.

The results obtained can be seen in Table IV, which show the suitability of our method for this kind of data. Despite the already good performance of BEL using a single base classifier, when using a larger number of classifiers its classification ability is even better.

Focusing on the percentage of features picked by each model (Fig. 3), the system mainly selects the firsts 50 features as expected, even if selecting some uninformative features due to the very unfavourable ratio between number of observations and number of features. Notice that when using only 1 base

TABLE IV
PERFORMANCE OF BREADTH ENSEMBLE LEARNING ON SYNTHETIC DATA

	# classifiers	AUC	AUH	ACC	F	BER
Breadth Ensemble Learning (LDA)	1	0.993 ± 0.001	0.993 ± 0.001	0.952 ± 0.003	0.950 ± 0.004	0.051 ± 0.004
	50	0.995 ± 0.002	0.995 ± 0.002	0.960 ± 0.007	0.958 ± 0.007	0.042 ± 0.007

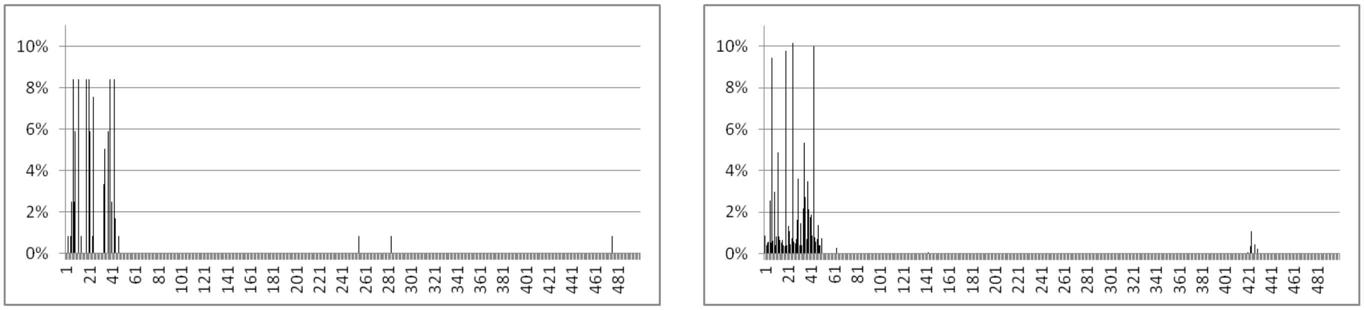

Fig. 3. Percentage of features selected after 10 runs of Breadth Ensemble Learning for classifying the synthetic dataset. The figure on the left corresponds to the ensemble using only 1 base classifier and the figure on the right to the ensemble made of 50 classifiers.

classifier, many correlated features that are partly represented by their neighbours are not selected, while when using 50 learners, all the features are selected in more or less proportion, a fact that contributes to its slight improvement.

V. DISCUSSION

The proposed method performed a parsimonious selection of subsets of features, giving raise to solutions where multiple features appeared several times in different subsets. Fig. 4 shows the relative percentage of appearances for each frequency in the MRS spectrum for the algorithm using 50 LDA.

The most frequently selected features are in consonance with those found in the literature. For example, those located at the interval between $3.30 - 3.45ppm$ might correspond to taurine and are consistent with those found by [13] regarding LET. Also in line with this study, the features located at $3.58 - 3.60ppm$ might correspond to glycine. Those located at $2.36ppm$ and $2.42ppm$ in LET might correspond to glutamine and glutamate (Glx) metabolites. Another frequently selected interval of the spectrum is located at both LET and SET, at $2.90 - 3.07ppm$, with the creatine peak at $3.03ppm$. NAA also plays a relevant role around $2.05ppm$, as mentioned by [6].

As also reported in [6], our method found more relevant features in the LET dataset than in the SET one, when spectra are used in concatenation.

It is interesting to compare this chart with the mean spectral signature for glioblastomas and metastases in Fig. 1, as it gives an indication that this central measure can be misleading in terms of inferring feature relevance. In the spectra for LET, two peaks of high amplitude correspond to choline ($3.20ppm$) and lipids/macromolecules ($1.40ppm$), which is especially clear for SET. Our method barely ever selects these frequencies despite they could be thought as relevant for the task, given the great difference in the main values between glioblastomas and metastases that they (especially the second) show.

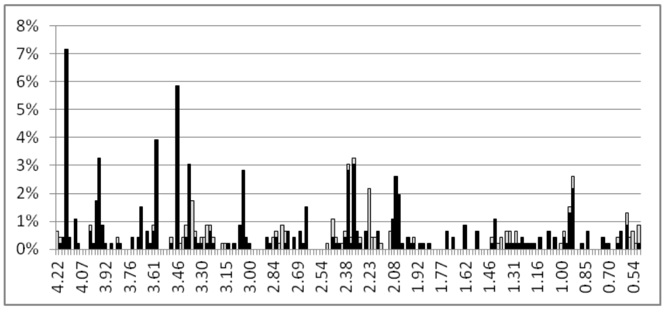

Fig. 4. Relative percentage of appearances for each feature (frequencies in ppm) from the ^{1}H-MRS spectrum using a BEL ensemble of 50 LDA. Black columns represent appearances in the Long Echo Time spectrum whilst white columns are the appearances in the Short Echo Time.

Interestingly, there are two close groups of frequencies that are highly selected, which, to date, have not been reported to contain any important information. They occur around $4.20ppm$ and $3.95ppm$ and might correspond to choline and either creatine or alanine, respectively. Further research should be done to elucidate this phenomenon.

VI. CONCLUSIONS

Glioblastomas and metastases are two types of high-grade brain tumours that are difficult to distinguish from NMR information. Their correct diagnosis is critical for the choice of adequate treatment. Individual classifiers have been shown to struggle in this task. Thus, in this study, we have investigated the alternative ensemble learning approach.

A parsimonious sequential feature selection technique was employed to feed each probabilistic classifier within the ensemble. Contrarily to many ensemble building techniques, the system was constructed in breadth, seeking to directly maximize the overall performance and not as a by-product of maximizing the performance of each of the component parts.

The classification decision was made according to the confidence expressed by each *local* classifier. Moreover, the ensemble outputs not only the crisp classification for a given instance, but also the probability that this instance belongs to a particular tumour type.

Although considering non-trivial low-relevant features as predictors might contribute to the success of ensemble methods, using too many of them can produce the opposite effect and mislead the whole system. Therefore, in our system, the feature selection strategy is tightly coupled with the way the ensemble is built.

We conjecture that the inability of general-purpose ensemble methods to cope with the problem posed in this study is precisely due to the lack of a wise feature selection strategy. Such strategy should be able to skip the high number of uninformative redundant features that the spectral data is composed of, picking only the most informative ones.

Despite most ensemble systems perform better using weak classifiers as their component parts, our system uses stable classifiers (e.g. LDA and SVM). An explanation for this is that the diversity required by the ensemble is introduced by means of a specific feature selection.

The excellent results obtained with the proposed method match the best ones reported so far in the literature [6], and reinforce the thesis that an adequate feature selection is key to solve this kind of problems. Importantly, the feature selection process also improves the interpretability of the diagnostic discrimination decision, which is a necessary requirement for the practical implementation of the method.

These results have been achieved using a combination of spectra acquired at different echo times. A natural extension of the current study would entail reproducing the experiments using only either LET or SET data. Future research could also extend the study of the performance of this technique with other tumour types (e.g., discriminating high grade malignant tumours from low grade gliomas) or different problems dealing with data of similar characteristics as the ones studied here.

Acknowledgment

This research was funded by Spanish MICINN R+D Plan TIN2009-13895-C02-01 project. The authors gratefully acknowledge former INTERPRET, eTUMOUR and HealthAgents European project partners. Data providers: Dr J. Capellades (Institut de Diagnòstic per la Imatge, Spain); Dr F.A. Howe and Professor J. Griffiths (St George's University of London, UK); Prof. A. Heerschap (RU); Prof. L. Stefanczyk and Dr J. Fortuniak (Uniwersytet Medycznyw Lodz, Lodz, Poland); Dr J. Calvar (Fundación para la Lucha contra las Enfermedades Neurológicas de la Infancia, Buenos Aires, Argentina). Data curators: Dr A.P. Candiota, Dr I. Olier, Ms T. Delgado, Ms J. Martín, Ms M. Camisón and Mr A. Pérez (all from GABRMN-UAB).

References

[1] J. Luts, T. Laudadio, A. Idema, A. Simonetti *et al.*, "Nosologic imaging of the brain: segmentation and classification using MRI and MRSI," *NMR in Biomedicine*, vol. 22, no. 4, pp. 374–390, 2009.

[2] P. Lisboa, A. Vellido, R. Tagliaferri, F. Napolitano, M. Ceccarelli *et al.*, "Data mining in cancer research," *IEEE Computational Intelligence Magazine*, vol. 5, no. 1, pp. 14–18, 2010.

[3] A. Vellido, E. Biganzoli, and P. Lisboa, "Machine learning in cancer research: implications for personalised medicine," in *Procs. of the 16th European Symposium on Artificial Neural Networks (ESANN 2008)*, M. Verleysen, Ed. d-side publications, 2008, pp. 55–64.

[4] K. Opstad, M. Murphy *et al.*, "Differentiation of metastases from high-grade gliomas using short echo time ^{1}H spectroscopy," *Journal of Magnetic Resonance Imaging*, vol. 20, no. 2, pp. 187–192, 2004.

[5] A. Server, R. Josefsen, B. Kulle, J. Maehlen, T. Schellhorn, O. Gadmar *et al.*, "Proton magnetic resonance spectroscopy in the distinction of high-grade cerebral gliomas from single metastatic brain tumors." *Acta Radiologica*, vol. 51, no. 3, pp. 326–328, 2010.

[6] A. Vellido, E. Romero, M. Julià-Sapé, C. Majós, A. Moreno-Torres *et al.*, "Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel ^{1}H-MRS," *NMR in Biomedicine*, In Press.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] M. Julià-Sapé, D. Acosta, M. Mier, C. Arús, and D. 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 (MAGMA)*, vol. 19, pp. 22–33, 2006.

[9] H. González-Vélez, M. Mier, M. Julià-Sapé *et al.*, "Healthagents: Distributed multi-agent brain tumor diagnosis and prognosis," *Journal of Applied Intelligence*, vol. 30, no. 3, pp. 191–202, 2009.

[10] V. Govindaraju, K. Young, and A. Maudsley, "Proton NMR chemical shifts and coupling constants for brain metabolites," *NMR in Biomedicine*, vol. 13, no. 3, pp. 129–153, 2000.

[11] C. Majós, M. Julià-Sapé, J. Alonso, M. Serrallonga, C. Aguilera, J. Acebes *et al.*, "Brain tumor classification by proton MR spectroscopy: comparison of diagnostic accuracy at short and long TE," *AJNR American Journal of Neuroradiology*, vol. 25, no. 10, pp. 1696–1704, 2004.

[12] J. García-Gómez, S. Tortajada *et al.*, "The influence of combining two echo times in automatic brain tumor classification by magnetic resonance spectroscopy," *NMR in Biomedicine*, vol. 21, pp. 1112–1125, 2008.

[13] F. F. González-Navarro, L. A. Belanche *et al.*, "Feature and model selection with discriminatory visualization for diagnostic classification of brain tumors," *Neurocomputing*, vol. 73, no. 4-6, pp. 622–632, 2010.

[14] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[15] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[16] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, 1996, pp. 148–156.

[17] A. R. Tate, J. Underwood, D. M. Acosta, M. Julià-Sapé, C. Majós *et al.*, "Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra." *NMR in Biomedicine*, vol. 19, no. 4, pp. 411–434, 2006.

[18] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, ser. Wadsworth statistics/probability series. Wadsworth International Group, 1984.

[20] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[21] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[22] C. Barber, D. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hull," *ACM Transaction on Mathematical Software*, vol. 22, no. 4, pp. 469–483, 1996.

[23] J. M. García-Gómez, J. Luts *et al.*, "Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy," *Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA)*, vol. 22, no. 1, pp. 5–18, 2009.

[24] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. The MIT Press, 2010.

[25] Y. Han and L. Yu, "A variance reduction framework for stable feature selection," in *Proceedings of the 10th International Conference on Data Mining (ICDM-10)*, 2010, pp. 206–215.