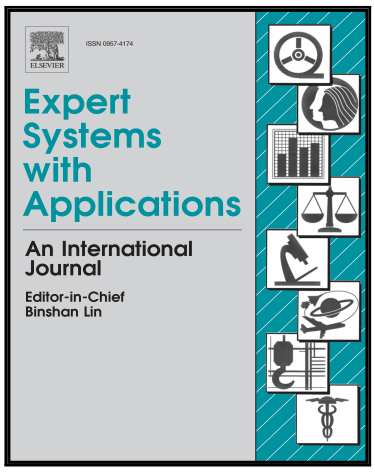

A Multi-level Classification Framework for Multi-site Medical Data: Application to the ADHD-200 Collection

Sarah Itani, Fabian Lecron, Philippe Fortemps

PII: S0957-4174(17)30592-4
DOI: 10.1016/j.eswa.2017.08.044
Reference: ESWA 11513

To appear in: *Expert Systems With Applications*

Received date: 2 May 2017
Revised date: 19 July 2017
Accepted date: 26 August 2017

Please cite this article as: Sarah Itani, Fabian Lecron, Philippe Fortemps, A Multi-level Classification Framework for Multi-site Medical Data: Application to the ADHD-200 Collection, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.08.044

Highlights

- A classification approach to face the heterogeneity of multisite medical databases
- A promising learning scheme to develop consistent aid in diagnosis models
- A case study on Attention Deficit Hyperactivity Disorder

A Multi-level Classification Framework for Multi-site Medical Data: Application to the ADHD-200 Collection

Sarah Itani[a,b], Fabian Lecron[c,*], Philippe Fortemps[b]

[a] *Fund for Scientific Research - FNRS (F.R.S.-FNRS), Brussels, Belgium*
[b] *Faculty of Engineering, University of Mons, Department of Mathematics and Operations Research, Mons, Belgium*
[c] *Faculty of Engineering, University of Mons, Department of Engineering Innovation Management, Mons, Belgium*

Abstract

Recently, the culture of sharing medical data has emerged impressively, reducing significantly the barrier to the development of medical research accordingly. As open-access large datasets result from this significant initiative, data mining techniques can be considered for the development of interpretable expert systems to help in diagnosis. However, the collaborative effort of information gathering yields heterogeneous databases because of technical and geographical factors. Indeed, on the one hand, the harmonization of protocols for data collection is still missing. On the other hand, cultural and social factors impact locally both the epidemiology and etiology of a given disease. Ignoring these factors could weaken the credibility of studies based on multi-site data. Thereby, our work tackles the development of computer-aided diagnosis systems relying on heterogeneous data. For such a purpose, we propose a multi-level approach (inspired by multi-level statistical modeling) based on decision trees (in the sense of machine learning). This framework is applied on the public ADHD-200 collection for the study of Attention Deficit Hyperactivity Disorder (ADHD).

Keywords: Clinical Decision Support Systems, Decision Trees, Multi-level Approach, Attention Deficit Hyperactivity Disorder (ADHD)

Dear Editor and Reviewers,

First of all, we would like to thank you for the time you took on reviewing our work. Your insightful and constructive comments helped us improving the manuscript and correcting some mistakes.

You will find below a point-to-point list of all the responses to the comments

*Corresponding author. University of Mons, Department of Engineering Innovation Management, Rue de Houdain, 9, 7000 Mons, Belgium.
Email addresses: sarah.itani@umons.ac.be (Sarah Itani), fabian.lecron@umons.ac.be (Fabian Lecron), philippe.fortemps@umons.ac.be (Philippe Fortemps)

raised by the reviewers. Note that the major modifications in the paper are in red, in order to facilitate the track of changes.

Kind regards, The authors.

Reviewer #1

Comment #1

The presented ideas are not clear. The paper is too long. For this reason, it is hard to read and to understand what has been done and why.

Response

Thank you for your comments that lead us to improve the general structure of the paper.

Comment #2

The contribution (page 3) is not precise.

Response

We reformulated our contribution in page 3.

Our work addresses the problem of the heterogeneity of multi-site medical data. Actually, as mentioned in our paper, this issue was already addressed in the recent work of Abraham et al. (2017). Their pioneering work adapted cross validation techniques to develop classifiers less sensitive to the heterogeneity of multi-site medical data. By contrast, our study addresses this issue in adjusting the learning processes, while the work of Abraham et al. (2017) proposes to adjust the validation processes. Our proposal consists of solving the classification problems in a hierarchical way, to check if a subject is affected by a trouble/disease, and then, in case of pathology, to detect its type. If this methodology was already used in previous works, our proposal differs as it consists of a hybrid hierarchical classification structure that we call a multi-level approach, introduced in section 3.1. Finally, our study brings another way of approaching our case study: the ADHD-200 collection. For such a purpose, we used decision trees which are interpretable classifiers, so adapted to a medical context. The classifiers are based on features that, though unusual to the neuroscience sphere, appeared to be interesting, in terms of interpretation and prediction accuracies.

Comment #3

The results are unreadable.

Response

We have taken note of your comment and hope that the corrected version of the paper will meet your expectations. For the sake of clarity, we added a section "Results reporting" to expose the different analyses that we lead on levels I and II.

Comment #4

In my opinion the paper cannot be published in this form and it needs a lot of improvements. For instance, the descriptions of datasets used in the experiments should be precise presented in tabular form, including the number of features, samples, classes, the class labels, etc. How many datasets are used? Are they available online? Why old-fashioned C4.5 is used for classification? It should be justified.

Response

The datasets of our study are extracted from the ADHD-200 collection, which is available online at http://fcon_1000.projects.nitrc.org/indi/adhd200/. We added the reference to this website.

Follow up your comments, we reorganized section 3.4 and presented the data of our study in two steps: (1) we presented the features that we computed based on fMRI signals and mentioned the phenotypic features that we used in our study (age, gender, handedness, IQ). Altogether, per subject, 584 features were computed; (2) we explicitly mentioned the subsets on which our study is based on levels I and II (KKI, NYU, OHSU, PU). In a tabular form, we give a summary of these datasets, regarding the number of samples, the distributions of instances according to class labels.

As for the use of C4.5, this algorithm is used for classification because of its readability and interpretability. The compactness of the model involves a quite rapid decision inference. The shape of decision tree is even such that no computerization is required to infer a decision. Subjects are classified just by reading the model. All these qualities are advantageous in view of the medical context of the classification. We added these justifications regarding the use of C4.5 in the paper (section 3.2).

Comment #5

Comparative analysis should include alternative approaches.

Response

Alternative approaches are exposed in section 2 regarding the classification features (connectivity, graph theory, correlation,...) and methodologies (e.g. SVM, ANN,...). In the last section of the paper, we propose, following the advice of reviewer II, a comparative analysis completed by recent results (Riaz et al., 2016), for the dissociation TD-ADHD, based on SVM. We also kept the comparison between our hybrid hierarchical classification and the hierarchical classification proposed by Colby et al. (2012).

Comment #6

There is a typo in the paper title in EES system.

Response

We corrected the title in EES system. Thank you.

Reviewer #2

Comment #1

Overall, I think the paper is a good effort and it has a remarkable work of preprocessing the data and to conciliate heterogeneous data (at the level 2). The experimental results show interpretable models that can be applied to obtain new knowledge from the ADHD-200 collection. However, the novelty of the proposal is limited and the paper has some relevant lacks.

Response

We warmly thank the reviewer for this constructive remark and the following comments that allowed us to improve the paper.

Comment #2

The methodology proposed may not be considered as innovative because it is an application of hierarchical classification to heterogeneous data sets. Furthermore, in the results page of the ADHD-200 global competition (a competition based on the ADHD-200 collection of data sets, available at http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html) this methodology was mentioned in the section "Prediction Accuracy by Chance" : "The first is a hierarchical two-class classifier where teams first decide who is TDC and who is ADHD. Next they decide in the ADHD class who is ADHD-1 or ADHD-3".

Response

We totally agree with the reviewer on the fact that the application of hierarchical classification has already been considered in previous works, notably those of the ADHD-200 contest; we recognized this fact in page 7 of the first paper version. Actually, up to now, such a cascade configuration was proposed in two distinct classification strategies.

1. A first approach consists of developing two classifiers per site, the first one to dissociate TD from ADHD cases, and the second one, to dissociate ADHD-I from ADHD-C cases. In proceeding so, the related studies emit implicitly the hypothesis that the diagnosis of ADHD should be localized and that there are no common behaviors in children affected by ADHD whatever their socio-cultural situation. So the total number of classifiers is (2×number of sites).
2. A second approach consists of developing two classifiers across all sites, so two universal classifiers are available for diagnosis, aiming at dissociating TD and ADHD subjects first, and then ADHD-I and ADHD-C subjects. In proceeding so, the related studies emit implicitly the hypothesis that cultural and social factors do not influence the diagnosis of ADHD, its etiology and epidemiology accordingly. So the total number of classifiers is 2.

In a technical point of view: (1) the first approach is weakened by the fact that, when it comes to dissociate ADHD-I from ADHD-C, the classifiers are learned on reduced training sets (including only pathological patients per site), (2) the second approach may be criticized in the consistency of its results, as it does not apply adjustments (e.g. a reduction of inter-site variability) on the training sets.

Our work proposes a third alternative as a hierarchical classification in a hybrid mode, inspired of both classification strategies mentioned below.

- In a first level, we dissociate TD from ADHD subjects per site, so there are as many classification models as there are sites. In proceeding so, we respect the scientifically recognized fact that the etiology of a trouble is influenced by local (social and cultural) factors.

- In a second level of classification, we dissociated ADHD-I from ADHD-C subjects, in developing a universal classifier across all sites. We thus emit the hypothesis that, at this level of classification, the prediction is free from local factors. This hypothesis appears reasonable as at this level, we have to predict the way in which the trouble is expressed, and we consider this expression as free from social, cultural or other local considerations.

So under our proposal, the total number of classifiers is (number of sites + 1). On the second level, in a technical point of view, we adjust the data and reduce the inter-site variability in applying data transformations.

We enhanced this contribution in the introduction section of the corrected version of the paper.

Comment #3

In the section 2.2, the authors affirm that learn a model for each one of the data sets D1,..., Dk, ignores possible global characteristics of the whole collection D, but at the first level of their proposal they learn a model for each data set.

Response

We understand the perplexity which comes from this affirmation given our own classification practice. In that assertion, we wanted to enhance the limitations of a classification framework exclusively based on local models. As mentioned beforehand (see answer to comment 2), our classification approach is hybrid.

Comment #4

Another discrepancy found is that the authors in section 3.1 propose to group data sets in entities but in the experimental results of the first level, in section 4.3, they do not group any data set.

Response

Through this paper, our aim was to propose a very general proposition; the ADHD-200 collection is just a case study, for which data groupings in entities were not required.

Grouping data sets in entities may be interesting in situations where multi-site medical collections would include different subsets for the same geographical entity. For example, if the data collection includes subsets collected in different sites (research centers/hospitals/universities) in New-York (e.g. NY1, NY2, NY3,...), these subsets should be merged in a same data set since they are related to patients from the same socio-cultural entity, which is consistent with our idea of developing a classification model per entity, at the first classification level (TD vs ADHD). In the case of the ADHD-200 collection, only one New-York site provided data : the New York University Medical Center. There were no other NY sites involved in the data collection process. In the same way, only Peking University provided data. So in both cases, the geographical locations were represented by subjects collected by a unique site: no grouping was required.

In the corrected version of the paper, we specify that for the ADHD-200 collection, there exists no reason for merging different sites in entities. A socio-cultural analysis showed us that each site has to be treated as its own entity.

Comment #5

Another question is why only two data sets (PU and NYU) are utilized, using only 481 instances out of 776. Since the authors affirm the open data initiative provides a greater number of available data, one may ask why the authors have not used all the available data. It would be also interesting use the ADHD-200 global competition result for comparisons, not only to show the predictive accuracy in two individual test sets, but for the predictive accuracy across all the data sets too.

Response

The objective of the paper is to propose a multi-level approach able to cope with open multi-site medical data. But the paper does not aim at competing with other works on prediction accuracies. Indeed, the comparisons with other studies was provided to validate our approach in showing that our prediction rates are close to those of the literature. In other terms, the ADHD-200 collection constitutes a case study, but we were not looking at relaunching the challenge through the paper. We wanted also to enhance the importance of interpretation and extraction of knowledge. To that goal, decision trees were presented and commented, which may constitute a lot of information in addition to those of the prediction rates. Thus, for the sake of clarity, we presented the results necessary to illustrate our proposal. Even so, reviewer #1 reported that the paper is too long in its current state.

Moreover, our proposal is surely founded on the idea of maximizing the use of the data, but their complete use is not always guaranteed. For example, in the ADHD-200 collection, the data sets of The University of Pittsburgh and

Washington University in St. Louis provide only control groups (TD) whose use may not be considered; these data sets were thus discarded. At the first level of classification, we selected NYU and PU data sets among the other available data sets (The Kennedy Krieger Institute, NeuroImage and Oregon Health and Science University). Indeed, at this classification level, let us remind that our approach consists of developing a model per site to take into consideration the specificities of every geographical entity. So, opposing NYU and PU data sets was sufficient to illustrate and advocate our approach since both localizations are associated to geographically opposed socio-cultural factors. Incidentally, the resulting decision trees appeared to be different; this observation supports therefore the interest of the approach. Moreover, NYU and PU data are the largest data sets with the less amount of missing values so their use allowed to focus only on the multi-level approach.

By contrast, at the second level of classification, we crossed all the remaining available ADHD instances (from NYU, PU, KKI & OHSU sets), except for NeuroImage data because of missing IQ values. We thus decided to drop the NeuroImage instances as they represent only 3% of the training set. Finally, prediction accuracies were provided for NYU and PU sites on the whole classification problem, over both levels.

Comment #6

Also, I encourage the authors to explain explicitly the validation method used. It seems that the validation method used is hold out validation but I think it is not explicitly declared in the paper. It is clear that cross validation is used to obtain the "m" parameter. However I do not see so clear whether a train and test approach is used to compute the accuracy presented in the Tables 3,4,6 and 7. If I misinterpreted their words and cross validation is used to obtain the accuracy of the model, the results in Table 7 might not be valid since cross validation uses much more instances to build the model than the train and test procedure and the ADHD-200 competition results were based on a holdout data set.

Response

The cross validation is indeed used to obtain the parameter m. We confirm the validation method used is holdout validation; all the accuracies presented in the Tables were acquired under the train and test approach. We are explicit regarding this fact in the corrected version of the paper. Moreover, for the sake of clarity, we presented the results regarding the tuning of parameter m with 10-fold CV.

Comment #7

It strikes me that the default value is used for the parameter c (pruning confidence) and a wrapper search is made for the parameter m (minimum number of instances). This is only worth adjusting if you are going to turn off post-pruning, since the parameter m is used for pre-pruning. Moreover, it is usually

better to prune the full tree, as J.R. Quinlan stated :"Growing and pruning trees is slower but more reliable"[1].

Response

Actually, Quinlan designates by "prepruning" an operation by which the development of additional splits is conditioned by additional error rates that might occur on the training set [1]. On this matter, Quinlan said: "*The typical approach is to look at the best way of splitting a subset and to assess the split from the point of view of statistical significance, information gain, error reduction, or whatever.*"[1].

By contrast, parameter m of C4.5 is not used in the aim of reducing errors but to control overfitting; the use of m does not necessarily exempt from post-pruning. A description of the parameter is provided in [1]: "*Near-trivial tests in which almost all the training cases have the same outcome can lead to odd trees with little predictive power. C4.5 requires that any test used in the tree must have at least two outcomes with a minimum number of cases (or, to be more precise, the sum of the weights of the cases for at least two of the subsets Ti must attain some minimum). The default minimum is 2, but can be changed by this option; a higher value may be a good idea for tasks where there is a lot of noisy data.*" In view of that, we somewhat imprecisely qualified m as a prepruning parameter (we corrected the description of parameter m in the last version of the paper).

To check this point, as an example, we did the test on the NYU set PHEN-$V_{15,27,40,70,87,88}$ and acquired the following accuracies :

$m = 9$	Training (CV)	Test	$m = 1$	Training (CV)	Test
Unpruned	73,3%	65,9%	Unpruned	70,0%	56,1%
Pruned	73,8%	68,3%	Pruned	70,0%	58,5%

We notice that, whatever the value of m, pruning has a better effect than unpruning on accuracies.

Comment #8

It would be most appropriate to consider the use of some methodology for handling unbalanced data for the girls problem in the Pekin data set [2], instead of manually adjust the decision tree.

Response

Actually, Peking data set is particular, as it includes a kind of double unbalance. Indeed, the data set includes approximately 26% of girls. Among them, only 13% are affected by ADHD.

We drew some inspiration from [2] to try to handle this problem and considered the SMOTE technique (implementation provided by WEKA).

First, we considered girls as a minority class in the dataset and produced additional synthetic data. In this way, we certainly restored the balance between

boys and girls, but this did not remove the unbalanced representation of TD and ADHD girls. So, the problem of losing a discussion on girls appeared once again, and the manual adjustment of the decision tree remained necessary.

We tried then to apply SMOTE twice: first to solve the unbalance representation of TD and ADHD in girls, and second, in balancing the proportion of boys and girls. In this way, the decision trees that we acquired were difficult to interpret and sometimes too long, which lead us to believe that data were overfitted, probably because of an excessive generation of synthetic data. This intuition was confirmed with low accuracies on the test set (less than 70%) and a large gap with training accuracies.

Comment #9

Using some updated performances for this problem, such as those found in [2] or [3], will improve the paper.

Response

Thank you for the proposal. We compared our results to those of [2] as it provides results regarding both NYU and PU sites. This comparison was only possible at level I, since the work of [2] solves the problem TD vs ADHD.

Comment #10

Minor comments:

- There is a typo in the first footnote number at page 20.
- Section 3.3., third line: "The parameter enables", I think it might be "The parameter m enables"
- The end of section 4.2: "Actually, we can that"
- For the References section: it would be useful and desirable to include DOIs.

Response

Thank you for these comments. We corrected the paper.

Comments bibliography

[1] Quinlan, J. R. (1993). C4. 5: programs for machine learning. Elsevier.
[2] Riaz A., Alonso E., Slabaugh G. (2016) Phenotypic Integrated Framework for Classification of ADHD Using fMRI. In: Campilho A., Karray F. (eds) Image Analysis and Recognition. ICIAR 2016. Lecture Notes in Computer Science, vol 9730. Springer, Cham
[3] Guo X., An X., Kuang D., Zhao Y., He L. (2014) ADHD-200 Classification Based on Social Network Method. In: Huang DS., Han K., Gromiha M. (eds) Intelligent Computing in Bioinformatics. ICIC 2014. Lecture Notes in Computer Science, vol 8590. Springer, Cham

1. Introduction

During the last decade, the sharing of large-scale medical data appears to be a growing trend encouraged by the scientific community. Several medical databases were launched publicly to address different health concerns (Esfandiari et al., 2014; Di Martino et al., 2014; Ihle et al., 2012; Kerr et al., 2012; Milham et al., 2012; Church, 2005; Hunter et al., 2005; Mueller et al., 2005). Some initiatives aim not only at sharing databases, but also software tools to manage information at best (Milham, 2012). Such a culture of data sharing is valuable to the research sphere (Ross, 2016; Mennes et al., 2013; Ross & Krumholz, 2013; Piwowar et al., 2008). Indeed, access to data is henceforth facilitated, especially to researchers not having the medical information on hand usually. Among others, specialists of computer science and mathematics can contribute their expertise at the technological level notably. The interaction of medicine with other disciplines is made possible therefore. Moreover, with the sharing of large-scale data, a same issue is approached in different ways worldwide, so research is enriched and accelerated. At last, a common framework of open-access data encourages local and international research centers to make their own databases available online. This virtuous circle multiplies the mass of available information, and the quality of studies is improved accordingly.

In particular, opening such access to data allows to focus on the explanation of some diseases/troubles through the detection of physiological foundations and/or typical symptoms. The advent of data mining allows to meet these needs, and to develop expert systems for the purpose of aid in diagnosis (Esfandiari et al., 2014; Parvathi & Rautaray, 2014). Such a decision support system should preferably be *interpretable*, i.e. able to show how a diagnosis is acquired (Lavrač, 1999), and *readable*, through assessment criteria making sense (Wagholikar et al., 2012)

The efforts deployed to gather data collaboratively are undeniably outstanding, nevertheless, they yield databases whose use is challenging. Indeed, the harmonization of data collection protocols is still missing (Milham et al., 2012). Across sites, data acquisition differs in terms of equipment calibration, experimental conditions, and sampling methodologies (Abraham et al., 2017). There exist also variations in the strategies that are used to process medical images (Abraham et al., 2017). Besides, the influence of cultural and social factors on both the epidemiology and etiology of diseases is established since the past century (Trostle, 2005; Link & Phelan, 1995; Landy, 1977). These factors of disparities are added to the natural heterogeneity of medical data, caused by the different natures of information available on a patient (interviews, phenotype, scans) and their interpretation by physicians (Wasan et al., 2006; Cios & Moore, 2002). Hence, multi-site medical data remain in reality a patchwork of subsets that cannot be merged into a single dataset without any adaptation. As inconsistency may arise from the use of heterogeneous medical data, studies generally focus on subsets withdrawn from these large databases (Abraham et al., 2017). The collective attempts to share large open-access data are therefore partially rewarded: admittedly, the access to medical data is easier, but the open-access

databases remain not fully exploited.

In our study, we attempt to tackle the development of interpretable and readable diagnosis support models able to cope with a medical multi-site database. We propose an application to the ADHD-200 collection (Milham et al., 2012), an example of stable and recent release of multi-site medical database launched for the study of Attention Deficit Hyperactivity Disorder (ADHD). The main contributions of our work are exposed below.

- After the study of Abraham et al. (2017), we propose another way to deal with the issue of heterogeneous multi-site medical databases. The work of Abraham et al. (2017) proposed two cross-validation strategies that enable to acquire models less sensitive to the heterogeneity of multi-site medical databases. In our work, instead of questioning the validation phase to address this issue, we propose a rethinking of the learning process to develop models able to help in diagnosing a disease/trouble, under a novel *hierarchical* spirit.

- In previous works, to predict a diagnosis, hierarchical systems were set up to practice classification in two steps: (1) to dissociate healthy and pathological cases and (2) to detect the type of pathology. The methodologies lied on the development of either two classifiers per dataset (site) or two classifiers across the whole database. The novelty of our proposal is related to the hybrid nature of our hierarchical classifier where *intra-site* and *inter-site variabilities* play both a role at different levels to deliver a final robust diagnosis. To the best of our knowledge, such a procedure is innovative and allows to take into consideration geographical parameters in contrast to the work of Abraham et al. 2017. We will name our proposition as a *multi-level approach* since it is inspired by the theory of multi-level analysis.

- The work presents new results as concerns the ADHD-200 collection. In particular, we do not consider domain-specific features to solve the classification task, but rather features having demonstrated success in other domains. The results show that these *meaningful features* provide interesting interpretations for helping in diagnosing ADHD.

In section 2, we will expose our case study, before developing the materials and methods of this work in section 3. The results of our study will be presented and discussed in section 4. Finally, we will conclude this paper in section 5.

2. A case study: Attention Deficit Hyperactivity Disorder

Approximately five to seven percent of children and teenagers are likely to be confronted one day with Attention Deficit Hyperactivity Disorder (ADHD)[1].

[1] http://adhd-institute.com/burden-of-adhd/epidemiology/

Also affecting adults, this mental trouble is characterized by inattention, and/or hyperactive-impulsive attitudes. Generally, people with ADHD have to deal with a reduced self-control impairing notably their ability to express serenely their feelings, to the detriment of their social and professional daily life.

Today, the subject's environment (mainly parents and teachers) constitutes almost the only source of information that practitioners have to make a diagnosis, unmistakably subjective. Research is still ongoing to better understand the physiological bases of the trouble.

In 2011, researchers from various fields of expertise were challenged to propose an objective assessment of ADHD in the context of the ADHD-200 contest (Milham et al., 2012). As a working basis, a multi-site medical database, called the ADHD-200 collection, was released online[2]. The database includes clinical (phenotypic) and neuroimaging data (resting-state functional and structural magnetic resonance images) on altogether 947 patients. Typically Developing (TD) and ADHD affected patients are included in the database. ADHD cases are expressed in three types: Inattentive (ADHD-I), Hyperactive-Impulsive (ADHD-HI) and a Combination of both types (ADHD-C). Eight sites contributed to the collection of data: Peking University (PU), Kennedy Krieger Institute (KKI), NeuroImage (NI), New-York University (NYU), Oregon Health & Sciences University (OHSU), University of Pittsburgh (Pitt.U), Washington University in St. Louis (WU) and Brown University[3]. Tables 1 and 2 present the subjects' distribution according to multiple criteria as regards both training and test sets (see also Milham et al. 2012). Let us note that the ADHD-HI type (Hyperactive-Impulsive) was not predicted as the associated population is very low in the training set.

The ADHD-200 collection can be qualified as a heterogeneous medical dataset as it includes instances of various geographical origins whereas social factors influence the local prevalence of ADHD (Russell et al., 2014); the subsets were not collected according to a common protocol (Bellec et al., 2017); the gender representativeness as well as the healthy and pathological cases proportions can significantly vary according to sites.

Actually, the inter-site variability of the ADHD-200 collection has been largely raised as a complex aspect to deal with (Bellec et al., 2017). To manage it, two main strategies were envisaged by previous works:

- learn a single model on the whole data collection D, ignoring local specificities of the underlying datasets $D_1,\ldots, D_k$, and identifying the latter potentially by a nominal variable, so hoping that the learning process will be sensitive to this information, e.g. in Deshpande et al. 2015; Brown et al. 2012; Dai et al. 2012; Eloyan et al. 2012; Sidhu et al. 2012;
- learn a model on each one of the datasets $D_1,\ldots, D_k$, so ignoring possible global characteristics of the whole collection D, e.g. in Colby et al. 2012.

[2]The dataset is available at http://fcon_1000.projects.nitrc.org/indi/adhd200/.

[3]Brown University data were discarded in this work as the results of diagnosis are not supplied for this site.

Table 1: Summary of the ADHD-200 training data

Site	Age	Gender		TD	ADHD	Total
		F	**M**			
PU	8-17	52	142	116	78	194
KKI	8-13	37	46	61	22	83
NI	11-22	17	31	23	25	48
NYU	7-18	77	145	99	123	222
OHSU	7-12	36	43	42	37	79
Pitt.U	10-20	43	46	89	0	89
WU	7-22	28	33	61	0	61
Total		290	486	491	285	776

Table 2: Summary of the ADHD-200 test set

Site	Age	Gender		TD	ADHD	Total
		F	**M**			
PU	8-15	19	32	27	24	51
KKI	8-12	1	10	8	3	11
NI	13-26	13	12	14	11	25
NYU	7-17	13	28	12	29	41
OHSU	7-12	17	17	28	6	34
Pitt.U	14-17	2	7	5	4	9
WU	-	-	-	-	-	-
Total		65	106	94	77	171

In our sense, the first approach simplifies drastically the modeling of discrepancies. The second approach intercepts better the specificity of each group but does not allow any cross of information on a more global level.

From neuroimaging data, typical information such as correlation measures between brain regions of interest, graph theory metrics, morphometric features are computed and used for learning (Deshpande et al., 2015; Guo et al., 2014; Eloyan et al., 2012; Colby et al., 2012; Dai et al., 2012; Fair et al., 2012; Sidhu et al., 2012; Smith et al., 2011; Rubinov & Sporns, 2010; Telesford et al., 2010; Bullmore & Sporns, 2009; Marrelec et al., 2006). As a large number of features is raised accordingly, successful tools such as Principal Component Analysis (PCA), Support Vector Machine with Recursive Features Elimination (SVM-RFE, see Guyon et al. 2002), as well as other variants of SVM (Aytug, 2015) are considered for dimensionality reduction. Even if some of these features could be easily interpreted, they are either introduced in a process of dimensionality reduction that could lead to loose the real sense of the information, or drown in a large mass of data, from which they hardly override. Furthermore, the classifiers that are considered in this context do not contribute to increase the interpretability of the features. Indeed, countless studies have privileged Support Vector Machine (SVM) as a predictor of ADHD (Riaz et al., 2016; Strigo et al., 2013; Chang et al., 2012; Colby et al., 2012; Dai et al., 2012; Fair et al., 2012; Sidhu et al., 2012; Wee et al., 2012; Mueller et al., 2011; Anuradha et al., 2010). In Deshpande et al. 2015, accuracies accounting for 90% in ADHD and healthy subjects dissociation, and nearly 95% between ADHD subtypes could be reached thanks to Artificial Neural Networks (ANN). Such classifiers achieve high prediction accuracies, but they lack of interpretability and readability (Wagholikar et al., 2012). Indeed, such models, through equations (SVM) and networks of weighted sums (ANN), can be assimilated to black boxes, i.e. the relation between the model inputs (patient's features) and output (patient's diagnosis) is difficult to establish clearly.

3. Materials and methods

The medical context involves special requirements (Cios & Moore, 2002; Moore & Hutchins, 1980). Therefore, we propose a multi-level approach that is consistent with the medical way of thinking and which addresses simultaneously the issue of heterogeneous multi-site medical data. We will justify and develop this approach in section 3.1. Afterwards, we will advocate the use of decision trees as white-box models in section 3.2. We will expose how to tune and validate such models in an appropriate way in section 3.3. Finally, we will present the data of our case study in section 3.4, and expose the modalities of results reporting in section 3.5.

3.1. Multi-level approach

Numerous disorders exist in several types. Although basically suffering from the same disorder, patients may exhibit different types of this disorder. That is the case, for example, for allergies where patients suffering from a

common hypersensitivity to environmental factors can express it in different forms such as asthma, rhinitis, conjunctivitis (Tanno et al., 2014). In the same way, psychological disorders are often characterized by a spectrum of conditions such that these conditions appear as different ways in which a same trouble is expressed (Maser & Akiskal, 2002). For example, as previously stated, ADHD exist in three types: inattentive and/or hyperactive-impulsive.

In each case, the diagnosis question could be solved in a hierarchical way:

1. first, in checking the absence or presence of the disorder;
2. then, in detecting the disorder type in case of a positive diagnosis.

Such a hierarchical structure is interesting as it allows to consider different assessment criteria in two steps. Indeed, a medical dataset can contain voluminous information on a subject (Cios & Moore, 2002), though the information is not necessarily required in its integrity to solve each of the problems consisting of either detecting the disorder or defining the disorder type. Thus, to simplify the learning process, a feature extraction process can help to raise the most interesting information at each step of the diagnosis. From a medical point of view, this hierarchical structure could potentially lead to technical simplifications, e.g. if expensive and sophisticated information (as scans) are required on the second stage of a diagnosis only.

Obviously, this cascade configuration appears as justified; besides, it has been considered in previous works, e.g. in Colby et al. (2012). However, we propose to adapt this approach in the case of multi-site data. Indeed, the latter involve geographical and technical variations that we could hardly ignore. As mentioned beforehand, geographically distinct sites may be associated to variations in the epidemiology and etiology of a trouble. Moreover, the interpretation of medical data may be different according to physicians, such a difference being perhaps strengthened by this same factor of localization. Besides, technical disparities include variations impacting data acquisition and processing strategies.

Let us suppose a multi-site database $D_1,\ldots,D_k$, each D_i representing a *site* subset. Let us assume, first, that there are no technical heterogeneities, i.e. data were collected according to a same core protocol. It is consistent to group subsets that would have been collected in sites belonging to a same cultural and social entity. In proceeding so, we obtain datasets called $E_1,\ldots,E_j$, each E_i representing an *entity* subset with $j \leq k$. Using these data grouped by entity, we propose a multi-level approach to solve the hierarchical classification problem exposed in the preceding paragraphs.

- In a first level, a classifier should be trained to detect a potential anomaly in a patient. This classifier is developed by homogeneous entity, i.e. based on a training set E_i related to a same cultural and social identity i.
- In a second level, a classifier should be trained to predict a disorder type. At this stage, if a model is developed by entity, there could be a few number of available instances because the corresponding training set consists of pathological instances only, i.e. $E_i^{Disorder} = E_i \setminus E_i^{Healthy}$. That is why, it is better to use the union of sets $\cup_{i=1}^{j} E_i^{Disorder}$.

This approach is inspired by the multi-level analysis theory that has addressed different issues in varying methods (Segenreich et al., 2015; Lecron et al., 2012; Timmerman, 2006). For example, Jansen et al. (2005) introduced the Multi-level Simultaneous Component Analysis (MSCA). The main idea of such a method consists of relying on PCA by decomposing information of instances into two contributions. The first one concerns the evolution of every individual's properties in time and is, consequently, expressed in a *within-individual space*. The second component is expressed in a common space to all instances, called *between-individual space* and allows to raise a global tendency. We transpose this principle in our work, not in the context of an analysis, but to develop a classification scheme. In a first level, we develop a model per entity to detect a potential trouble. Then, in a second level, to detect the trouble type, we cross all the instances regardless of the entity to which they belong. In proceeding so, we make the hypothesis that the classification of a disorder type should be non cultural or social dependent. This may be a realistic hypothesis as, at this level, the question to solve is to detect how a trouble is expressed.

The problem of technical heterogeneities is solved on a case-by-case basis. For example, if signals were collected in the same experimental conditions, but were amplified with different gains, then the problem is solved in applying a gain factor to one of both signals groups to make the comparison possible. It must be emphasized that the full elimination of variations may sometimes be impossible, notably because some of these variations could be ignored by the user of the medical database. That is why it is preferable to cross-validate the classifiers trained on such data. Indeed, cross-validation techniques randomly withdraw instances from the training set as validation sets (see section 3.3), which allows to integrate the robustness to variations as part of the assessment (Abraham et al., 2017).

3.2. Decision trees as interpretable models

As mentioned in introduction of this paper, an aid in diagnosis system should be able to provide physicians with explanations on how a decision is taken (Stoean & Stoean, 2013; Wagholikar et al., 2012; Lavrač, 1999). Different models, such as trees and rules, can fit within this requirement, as they are so called white-box models. In our work, we focused on decision trees.

A decision tree encode a reasoning under the intuitive shape of trees, which makes it a suitable candidate in our goal of developing practical models, i.e. interpretable and readable. Furthermore, in a medical context, a decision tree can be used in situations of extreme emergence, as the time for decision inference is low. Moreover, a decision tree does not require any computerization, as it may be simply read (Lavrač, 1999).

Underlying on a recursive and greedy algorithm (Bishop, 2007; Quinlan, 1993, 1986), the learning of a decision tree process unfolds in accordance with the logic of dividing and conquering. The model is developed in dividing gradually all training instances contained in the root node, to establish the most coherent possible groupings based on common or at least, comparable characteristics. So the learning process is run on a set of training features which is not necessarily

needed in its entirety; only the most substantive features are selected to develop a decision tree. The model enhances directly a subset of interesting features for diagnosis, which is not the case of other categories of classifiers.

3.3. Validation procedures

Given a training set, decision trees learning may be adjusted by two parameters: the minimal number of training instances required by leaf m and the confidence level c about pruning (Quinlan, 1993). The parameter m constitutes a stop criterion of the learning process; it also tunes the granularity of the model. Thus, increasing m (by default set at 2) may provide more overarching models and makes them less sensitive to noise. As for parameter c, it is associated to a post-pruning process; it is lead to achieve a fair compromise between the length of decision trees and their predictive error rates.

In our work, we maintained the default value of c (25%). As regards the parameter m, it has been adjusted by a 10-fold cross-validation technique (10-fold CV). We admitted this parameter varied between 5 and 20, the latter representing approximately 10% of the size of the considered training datasets; the classifier whose parameter m is associated with the highest cross-validation prediction rate was held as relevant (see Algorithm 1).

The classifiers were then assessed by a holdout validation procedure, with respect to the test sets provided by the ADHD-200 collection. We reported performance indicators (Witten & Frank, 2005; Klösgen & Zytkow, 2002) that are well suited to the medical context of this work (Cios & Moore, 2002); they are defined hereafter. Let us note TP (respectively TN), the number of true positives (resp. negatives), that is, the number of patients rightly predicted by the model as pathological (resp. healthy) and FP (resp. FN), the number of false positives (resp. negatives), that is, the number of patients wrongly predicted as pathological (resp. healthy).

- *Accuracy* corresponds to the rate of successful predictions:

$$A = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{\text{Number of instances}}$$

- *Specificity* corresponds to the ability to detect healthy patients, i.e. the true negative rate:

$$tn = \frac{TN}{TN + FP} = \frac{TN}{\text{Number of healthy individuals}}$$

- *Sensitivity* corresponds to the ability to detect pathological patients, i.e. the true positive rate:

$$tp = \frac{TP}{TP + FN} = \frac{TP}{\text{Number of pathological individuals}}$$

Algorithm 1 Training a classifier

```
1: procedure TRAINCLASSIFIER(validationMethod)
2:     for m ← 5, 20 do
3:         classifier ← executeTraining(m, validationMethod)
4:         if classifier.accuracy > bestAccuracy then
5:             bestAccuracy ← classifier.accuracy
6:             bestModel ← classifier
7:         end if
8:     end for
9:     return bestModel
10: end procedure
```

3.4. Data

3.4.1. Learning features

From the ADHD-200 collection features, age, gender, handedness and intellectual quotient were used for training as phenotypic data. In the paper, they are denoted as AGE, GEN, HAND, IQ respectively. We use the notation PHEN to designate the set of four clinical data, i.e. PHEN = {AGE, GEN, HAND, IQ}.

As for brain images, we focused exclusively on resting-state functional Magnetic Resonances Images (fMRI) rather than structural ones since recent studies showed functional brain activity involvement as significant in neurological phenomena (Abraham et al., 2017; Sidhu et al., 2012; Purdon et al., 2011). Brain images were preprocessed under the initiative of the Neuro Bureau according to the called *The Athena* pipeline (The Neuro Bureau, 2011). Such a major work lead to the extraction of BOLD (Blood Oxygenation Level Dependent) signals, measuring the functional brain activity. The associated signals reflect variations in oxyhemoglobin and deoxyhemoglobin concentrations in blood caused by the neuronal activity (Aguirre et al., 1998; Logothetis & Wandell, 2004). Timecourse values of the signals are extracted for each cerebral Region of Interest (ROI). Brains are parceled according to a standard: Automated Anatomical Labeling (AAL) atlas was considered in this work (116 ROIs), which involves a set of 116 time series available for each subject (Tzourio-Mazoyer et al., 2002). For the sake of simplicity, the cerebral zones are numbered from 1 to 116. The Atlas matches these numbers with the names of the cerebral zones. Two successive numbers (odd and even, as, for example, 3 and 4) indicate regions having the same spatial location within the left and right hemispheres of the brain.

From fMRI signals, we computed statistical features on the one hand, and frequency features by applying a Discrete Fourier Transform on the second hand. These features are successfully used in several domains, e.g. in music signals classification and speech recognition (Lambrou et al., 1998; Le et al., 2011; Tzanetakis & Cook, 2002) (see Table 3). When these features are computed with regard to a precise brain region, they are indexed by a number, representing the code of the cerebral zone to which the attribute refers, i.e. V_{15} indicates the signal variance of the cerebral zone 15, $V_{15,27,40}$ represents a set of three values

concerning the signal variance of zones 15, 27, 40.

Table 3: Features computed on fMRI signals

Feature	Key	Meaning	Interpretation
Variance	V	Energy of fMRI signal	Always positive. The higher the variance is, the more energetic the signal is.
Skewness	S	Symmetry of fMRI intensity distribution	If positive (negative) skewness, fMRI distribution concentrated on the left (right) of the mean signal value.
Kurtosis	K	Aspect of fMRI intensity distribution	If kurtosis superior (inferior) to three, fMRI intensity distribution more (less) shaped than a Gaussian one.
Frequency	F	Frequency associated to the line of maximal amplitude of DFT spectrum.	The higher this frequency is, the more dynamic the fMRI signal is.
Centroid	C	Center of mass of DFT spectrum.	Gives an idea on the global dynamism of the fMRI signal.

Thus, a set of 116 features by modality (variance, skewness, kurtosis, frequency, spectral centroid) was computed, which means that, for each patient, biomarkers accounted for 580 features altogether, in addition to the four clinical attributes (age, gender, handedness and IQ). A feature extraction was clearly required to avoid the curse of dimensionality (Witten & Frank, 2005). This was achieved thanks to a correlation-based feature subset selection (Hall, 1999), an efficient heuristic used to detect a combination of attributes that weakly correlates (to avoid overlapping data) but are highly correlated with the prediction variable.

3.4.2. Selected datasets

As mentioned in section 3.1, in multi-site datasets, it is consistent to group subsets that would have been collected in sites belonging to a same cultural and social entity. However, in the case of the ADHD-200 collection, a preliminary socio-cultural analysis showed us that each site has to be treated as its own entity.

The multi-level approach presented in section 3.1 was applied on some datasets of the ADHD-200 collection. Let us note that the datasets of the University of Pittsburgh (Pitt.U) and of Washington University in St. Louis (WU) include only Typically Developing (TD) subjects (see Table 1). They were thus discarded from our study. The NeuroImage (NI) dataset was also discarded because IQ values are missing.

For the detection of a possible trouble (first level), we illustrated our approach with Peking University (PU) and New-York University (NYU) datasets which are associated to geographically opposed socio-cultural factors. For the identification of the trouble type (second level), we crossed the information of both sites

in addition to those of Kennedy Krieger Institute (KKI) and Oregon Health & Sciences University (OHSU) to enrich the resulting training dataset. Some instances were discarded because of missing values regarding phenotypic data and/or brain images. Tables 4 and 5 present the number of instances that were used in this study on levels I and II respectively. As mentioned before, we computed 584 features for each dataset.

Table 4: Summary of the dataset used in level I

Sites	Set	Total	TD	ADHD
NYU	Training	210	93	117
	Test	41	12	29
PU	Training	193	116	77
	Test	51	27	24

Table 5: Summary of the dataset used in level II

Sites	Set	Total	Inattentive (ADHD-I)	Combined (ADHD-C)
PU/KKI/NYU/OHSU	Training	245	137	108
	Test	60	38	22

Before learning on the second level, it was necessary to consider how to conciliate sources of heterogeneity impairing data that were to cross in this level. In particular, we realized on one hand that discrepancies affected variance evolving in different ranges of values according to sites since measured signals were probably amplified with different factors. To inhibit this effect, the measures of variances were reported to the average variances measured by site and by cerebral zone. On the other hand, frequencies associated with the maximal amplitude of the Fourier spectra were reported to the maximal frequency of these spectra to homogenize data whose sampling times differ. Finally, it shall be noted that New-York site measured handedness according to the Edinburgh method in continuous values comprised between -1 and 1 (Oldfield, 1971), while other sites provided labels (0 - left-hander, 1 - right-hander, 2 - ambidextrous). Thus, on the second level, New-York handedness values were adjusted in nominal variables in accordance with Edinburgh scale.

3.5. Results reporting

In the following section, we present the results and the discussion regarding our multi-level approach, based on selected datasets from the ADHD-200 collection. For each site of level I and for level II, our work was based on four variants of

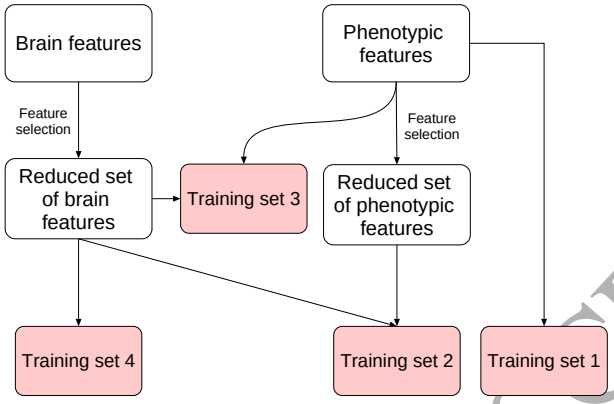

Figure 1: Training sets used to develop classification models

training sets, as shown by Figure 1. A first model was learned on phenotypic features exclusively (Training set 1). Then, we learned a model constituted of reduced phenotypic and brain features (Training set 2). As there are only four phenotypic features, we also trained models on training sets constituted by these phenotypic features and the reduced set of brain features (Training set 3). Finally, we considered a training set including only reduced brain features (Training set 4). Training sets 1 and 4 were considered to check if either phenotypic or brain features may alone provide a reliable classifier and a diagnosis accordingly.

The content of these training sets are explicitly given in section 4. Besides, the prediction rates we report are related to the holdout validation. The tables present also the value of parameter m (selected by 10-fold CV).

4. Results & Discussion

In this section, we aim to enhance the relevance of the multi-level approach as a classification scheme implemented to deal with multi-site medical datasets and to address the problem of aid in diagnosis of troubles existing in different types. As previously stated, this approach is evaluated in the context of Attention Deficit Hyperactivity Disorder. The results of the learning process are exposed in sections 4.1 and 4.2. For the development of decision trees, we considered the implementation of algorithm C4.5 proposed by Weka software[4]. Note that the numbers given in the nodes of the illustrated decision trees extracted from this software (Figures 2-7) relate respectively to the total number of training cases covered by the associated branching n_t and the number of instances wrongly

[4]Available at http://www.cs.waikato.ac.nz/ml/weka/.

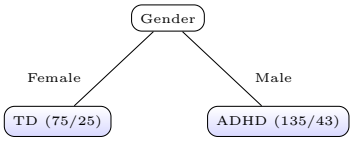

Figure 2: New-York phenotype-based diagnosis

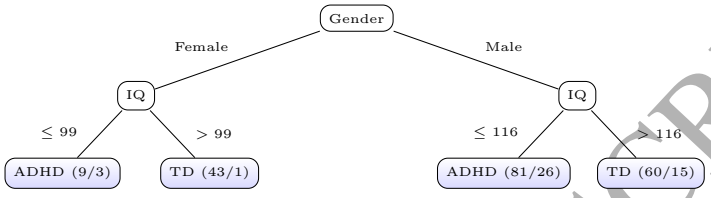

Figure 3: Peking phenotype-based diagnosis

classified e_t. This is noted as n_t/e_t. Finally, a synthesis of the accuracies on both levels of classification is provided in section 4.3.

4.1. ADHD - TD dissociation

In the ADHD - TD dissociation, the learning process was first exclusively exercised out of the only clinical dataset (PHEN) to measure its impact on prediction rates. The results of the learning process exerted on the phenotypic New-York and Peking sets are exposed in Tables 6 and 7 respectively. Figures 2 and 3 present the corresponding decision trees. While observing the results, we can already notice differences in the expression of clinical factors. In both cases, the diagnosis is surely gender-based, but the intellectual quotient is particularly significant in Peking site. In this latter case, a high prediction rate was acquired by the exclusive use of the phenotype in comparison to New-York site.

We prepared the other training sets (as exposed in section 3.5), following the results of the feature extraction exposed by Table 8. As expected, among the phenotypic features, the gender was raised as relevant, in addition to IQ as regards Peking site.

On New-York site, the feature selection process practiced on the neurological attributes returned signal variances of cerebral zones labeled 15, 27, 32, 40, 70,

Table 6: Prediction rates with regard to New-York test set

Modality	m	A (%)	tn (%)	tp (%)
PHEN	17	58.5	33.3	69.0
GEN - $V_{15,27,32,40,70,87,88}$	6	61.0	75.0	55.2
PHEN - $V_{15,27,32,40,70,87,88}$	**9**	**68.3**	**75.0**	**65.5**
$V_{15,27,32,40,70,87,88}$	15	51.2	83.3	37.9

Table 7: Prediction rates with regard to Peking test set

Modality	m	A (%)	tn (%)	tp (%)
PHEN	9	82.4	85.2	79.2
GEN, IQ - $V_{32,37}$, K_{38}, $F_{16,62}$	6	78.4	92.6	62.5
PHEN - $V_{32,37}$, K_{38}, $F_{16,62}$				
* *Original*	12	76.5	81.5	70.8
* *Adjusted*	-	**82.4**	**81.5**	**83.3**
$V_{32,37}$, K_{38}, $F_{16,62}$	13	51.0	88.8	8.3

Table 8: Results of the feature extraction process

	New-York	**Peking**
Phenotypic attributes	GEN	GEN, IQ
Brain attributes	$V_{15,27,32,40,70,87,88}$	$V_{32,37}$, K_{38}, $F_{16,62}$

87, 88 as the most meaningful features. From this result, we considered three variants of training sets. The associated prediction rates with regards to the test set are exposed by Table 6.

- [GEN - $V_{15,27,32,40,70,87,88}$] is the set of features strictly suggested by the feature selection process. The resulting classifier (see Figure 4) makes a first dissociation based on gender, before developing a further discussion about the variance of fMRI signals issued by some cerebral regions.
- [PHEN - $V_{15,27,32,40,70,87,88}$] includes notably all phenotypic attributes. The learning process gives raise to a model which differs from the previous one on the last subdivisions mainly (see Figure 5).
- [$V_{15,27,32,40,70,87,88}$] includes only variance features. The associated model presents poor performances on both training and test sets, matching practically with the luck of belonging to one of both groups (ADHD and TD).

The same process was applied as regards Peking site. The feature selection process applied to neurological factors reveals that significant attributes are $V_{32,37}$, K_{38}, $F_{16,62}$. From this result, three training sets were constituted as previously and the performances of the associated models are reported in Table 7. The decision tree acquired by modality [PHEN - $V_{32,37}$, K_{38}, $F_{16,62}$] presents a low granularity but the corresponding m parameter value leads to make discussion about IQ for girls impossible. That is why, as a final proposition, we manually adjusted the decision tree, in re-establishing the subdivision based on IQ (see Figure 6). We shall note an under-representation of girls in the Peking collection,

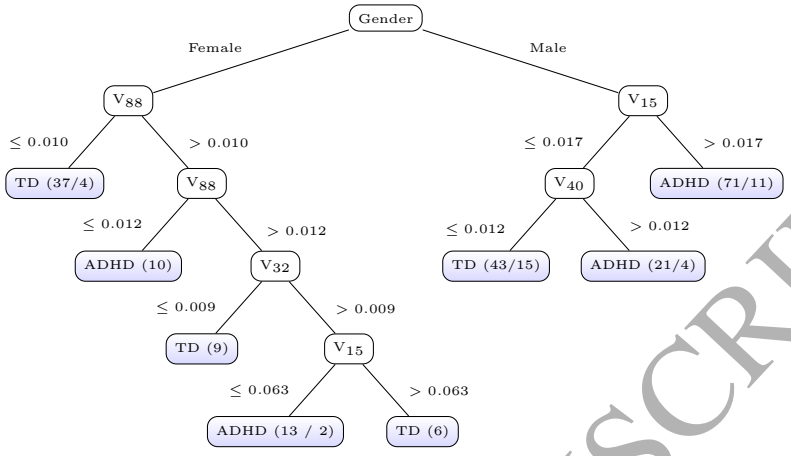

Figure 4: Model learned on the New-York set [GEN - $V_{15,27,32,40,70,87,88}$]

as well as an imbalance of ADHD and TD cases proportions within the same population. Indeed, the training set includes 52 girls and among them, only seven are affected by ADHD. In this actual state, it was practically impossible to emphasize further parameters in the goal of diagnosing girls, if it were not of focusing examination on IQ, based on a threshold of 99.

It is actually interesting to notice that subdivisions towards variance are operated on values which do not have an intrinsic meaning. Indeed, these values depend on measurement conditions. However, as variance is a measure of energy of fMRI signals, the subdivisions are based on how vigorous is the neuronal activity on a given region of interest. The role of gender seems crucial as it can lead to opposite interpretations for a same cerebral zone, e.g. in Figure 4, as for zone 15. Actually, we can notice that the trees are able to explain how the diagnosis is stated for every patient, through understandable features.

4.2. ADHD-I - ADHD-C dissociation

We first exercised learning on phenotypic data (PHEN) to notice that IQ has no more importance in the ADHD-I - ADHD-C dissociation. We noticed also a change in the impact of the gender, no more imposed at the root of the developed decision trees to make way for age, even if gender keeps a crucial role in the underlying subdivisions (see Figure 7). The resulting prediction rate on the test set is **70.0%** ($m = 17$, $tn = 76.3\%$, $tp = 59.1\%$). As for feature selection on the whole set of features, this indicates the relevance of the following attributes: AGE, V_{47}, $F_{33,46,64,65,66}$, $C_{12,63}$. Among trees acquired by variants of combinations of these attributes, the one learned on [GEN, AGE - V_{47}, $F_{33,46,64,65,66}$, $C_{12,63}$] set caught our attention as providing the highest cross-validation accuracy on the training set ($m = 7$, $A = 66.7\%$, $tn = 65.8\%$, $tp = 68.1\%$); but it has a depth of 8 instead of 3 for the previous tree. Finally, we prefer to keep the decision

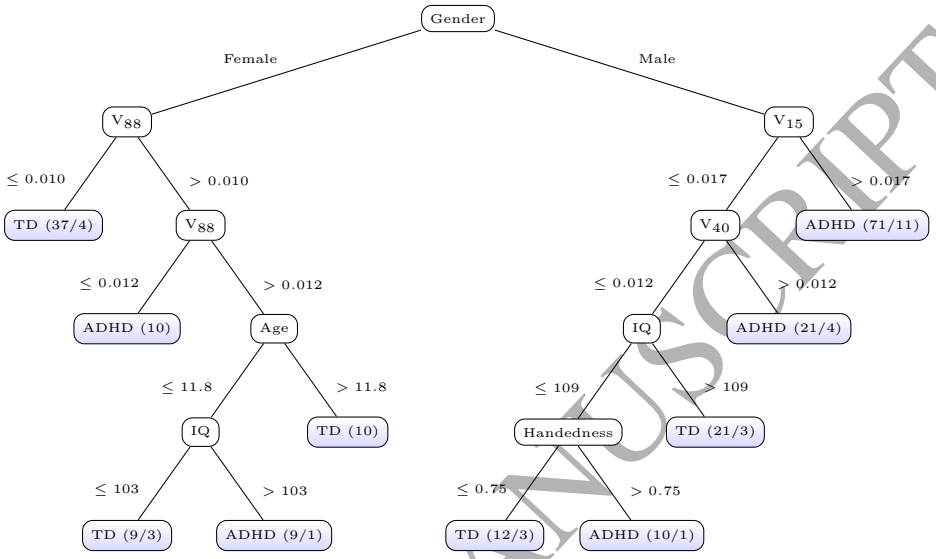

Figure 5: Model learned on the New-York set [PHEN - $V_{15,27,32,40,70,87,88}$]

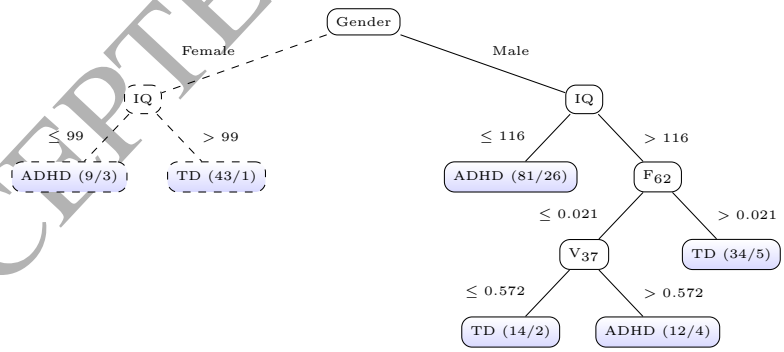

Figure 6: Model learned on the Peking set [PHEN - $V_{32,37}$, K_{38}, $F_{16,62}$], adjusted for further discussion as regards girls (dotted line)

Table 9: Synthesis of prediction rates

		New-York (%)	Peking (%)
Level I	Accuracy	68.3	82.4
	Specificity	75.0	81.5
	Sensitivity	65.5	83.3
Level II	Effective accuracy on ADHD types	66.0	60.0
Levels I - II	Global accuracy	58.0	66.7

tree based on phenotypic data since it is more compact and thus more readable without loosing on the accuracy.

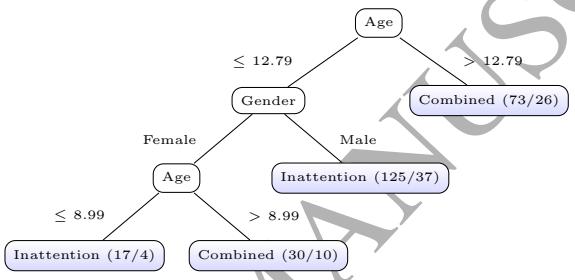

Figure 7: Second level phenotype-based diagnosis

4.3. Synthesis

Let us now aggregate the performances on both levels of classification in New-York and Peking sites (see Table 9). The models that we finally selected to compute these performances (sections 4.3 and 5) achieve a fair compromise between the model size and their performance on the training set. Among the subjects correctly predicted as affected by ADHD, we measured the proportion of subjects for whom the ADHD type was also well predicted, the latter rate which we called *effective accuracy on ADHD types* in Table 9.

5. Conclusion

Our results suggest that the multi-level approach is sufficiently relevant and efficient for the ADHD-200 collection.

On the first level of classification aiming at dissociating ADHD and TD subjects, we acquired interesting prediction rates in comparison to the recent state of the literature (Riaz et al., 2016) (see Table 10). This comforts the idea that the decision tree classifier and the chosen features are interesting for prediction.

Table 10: Comparison of performances with regard to New-York and Peking test sets on the binary problem ADHD vs TD

	Our results	Riaz et al. (2016)
NYU	**68.3%**	61.0%
PU	**82.4%**	64.7%

On the other hand, Table 11 presents the prediction rates we acquired in our study against the average results of the ADHD-200 contest[5] and those of Colby et al. (2012) on the global problem (TD vs ADHD-I vas ADHD-C). The comparison with the results of Colby et al. (2012) is quite interesting, since it advocates the advantage of a hybrid approach. As a matter of fact, they solved the problem in a hierarchical way and processed each site separately.

Table 11: Comparison of performances with regard to New-York and Peking test sets on the full problem

	Our results	ADHD-200	Colby et al. (2012)
NYU	**58.0%**	35.2%	37.0%
PU	**66.7%**	51.0%[6]	57.0%

Actually, our study shows that it is possible to aim at clarity through white boxes such as decision trees based on interpretable features without damaging the qualities of prediction. Moreover, the multi-level framework appears as promising for other multi-site data collections. Of course, this multi-step approach is questionable, notably as regards the underlying hypotheses on which it is build. Furthermore, it needs that every group contributing to data gathering delivers enough instances to develop a local model on the first level. Moreover, there should be, in each group, a representation of both ADHD and TD population. This problem is partly solved when several groups delivering data belong to the same homogeneous entity (in a cultural and social sense).

Our clinical collaborators welcomed the multi-level approach and the initiative of focusing on interpretable and readable models. Moreover, they believe promising the results achieved on the detection of ADHD and its characterization. Actually, through this work, we propose a basis for new perspectives in neuroscience, notably as regards ADHD. Indeed, the interpretability and readability of both the feature extraction and learning processes allowed to raise interesting observations. In particular, it should be noted that only a limited amount of cerebral zones were raised as pertinent among 116 zones altogether. Among

[4]This prediction rate is given on 44% of the Peking set by the ADHD-200 consortium as an estimation of the actual value.

[5]Available at http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html.

[6]This prediction rate is given on 44% of the Peking set by the ADHD-200 consortium as an estimation of the actual value.

the features considered for learning, the variances of fMRI signals were largely involved in the tree subdivisions with an interesting interpretation related to the dynamism of the neuronal activity. Furthermore, the different results acquired per site confirms the influence of cultural and social factors in the diagnosis. The importance of both gender and intellectual quotient for the Peking subjects in contrast to New-York subjects illustrates perfectly this point.

As future prospects, a particular attention could be brought on a differentiated accentuation of the true positive and negative rates. Indeed, an aid in diagnosis model characterized by a high specificity and low sensitivity is a priori preferable on a model that has an opposite tendency. Indeed, in the latter case, some patients could be more often diagnosed positively with a trouble. Yet, we may conceive the risks that such a diagnosis represents if the concerned patient, in reality not affected by the trouble does not need any therapy or, maybe more viciously, undergoes an inadequate one as affected by another trouble. At last, it can be worthwhile considering one-class classification techniques to develop models able to leave a third possibility in the first step of the diagnosis: the patient is not healthy, nor affected by the trouble to which special attention is paid, but by another unknown trouble. To this end, largest datasets are still to be gathered and shared.

6. Acknowledgments

This work is funded by the Belgian Fund for Scientific Research (F.R.S.-FNRS). We would like to thank Professors Mandy Rossignol, Laurent Lefebvre, Thierry Pham Hoang (Faculty of Psychology and Education, University of Mons, Belgium) and Doctor Isabelle Schonne (Service of Child Psychiatry, Saint-Joseph Hospital, Mons, Belgium) for their advice and interest in this work. We would like also to address our thanks to the ADHD-200 consortium for their significant initiative of data gathering and sharing.

References

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, *147*, 736–745. doi:10.1016/j.neuroimage.2016.10.045.

Aguirre, G., Zarahn, E., & D'esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *Neuroimage*, *8*, 360–369. doi:10.1006/nimg.1998.0369.

Anuradha, J., Ramachandran, V., Arulalan, K., Tripathy, B. et al. (2010). Diagnosis of ADHD using SVM algorithm. In *Proceedings of the Third Annual ACM Bangalore Conference* (p. 29). ACM. doi:10.1145/1754288.1754317.

Aytug, H. (2015). Feature selection for support vector machines using Generalized Benders Decomposition. *European Journal of Operational Research*, *244*, 210–218. doi:10.1016/j.ejor.2015.01.006.

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., & Craddock, R. C. (2017). The Neuro Bureau ADHD-200 Preprocessed Repository. *Neuroimage*, *144*, 275–286. doi:10.1016/j.neuroimage.2016.06.034.

Bishop, C. (2007). Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*, .

Brown, M. R., Sidhu, G. S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P. H., Greenshaw, A. J., & Dursun, S. M. (2012). ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Frontiers in systems neuroscience*, *6*, 69. doi:10.3389/fnsys.2012.00069.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*, 186–198. doi:10.1038/nrn2575.

Chang, C.-W., Ho, C.-C., & Chen, J.-H. (2012). ADHD classification by a texture analysis of anatomical brain MRI data. *Frontiers in systems neuroscience*, *6*, 66. doi:10.3389/fnsys.2012.00066.

Church, G. M. (2005). The personal genome project. *Molecular systems biology*, *1*. doi:10.1038/msb4100040.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine*, *26*, 1–24. doi:10.1016/S0933-3657(02)00049-0.

Colby, J. B., Rudie, J. D., Brown, J. A., Douglas, P. K., Cohen, M. S., & Shehzad, Z. (2012). Insights into multimodal imaging classification of ADHD. *Frontiers in systems neuroscience*, *6*, 59. doi:10.3389/fnsys.2012.00059.

Dai, D., Wang, J., Hua, J., & He, H. (2012). Classification of ADHD children through multimodal magnetic resonance imaging. *Frontiers in systems neuroscience*, *6*, 63. doi:10.3389/fnsys.2012.00063.

Deshpande, G., Wang, P., Rangaprakash, D., & Wilamowski, B. (2015). Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. *IEEE transactions on cybernetics*, *45*, 2668–2679. doi:10.1109/TCYB.2014.2379621.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M. et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, *19*, 659–667. doi:10.1038/mp.2013.78.

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., Barber, A. D., Joel, S., Pekar, J. J., Mostofsky, S. H. et al. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in systems neuroscience*, *6*, 61. doi:10.3389/fnsys.2012.00061.

Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, *41*, 4434–4463. doi:10.1016/j.eswa.2014.01.011.

Fair, D. A., Nigg, J. T., Iyer, S., Bathula, D., Mills, K. L., Dosenbach, N., Schlaggar, B. L., Mennes, M., Gutman, D., Bangaru, S. et al. (2012). Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data. *Frontiers in systems neuroscience*, *6*, 80. doi:10.3389/fnsys.2012.00080.

Guo, X., An, X., Kuang, D., Zhao, Y., & He, L. (2014). Adhd-200 classification based on social network method. In *International Conference on Intelligent Computing* (pp. 233–240). Springer. doi:10.1007/978-3-319-09330-7_28.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, 389–422. doi:10.1023/A:1012487302797.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Ph.D. thesis The University of Waikato.

Hunter, M., Smith, R. L., Hyslop, W., Rosso, O. A., Gerlach, R., Rostas, J., Williams, D., & Henskens, F. (2005). The Australian EEG database. *Clinical EEG and neuroscience*, *36*, 76–81. doi:10.1177/155005940503600206.

Ihle, M., Feldwisch-Drentrup, H., Teixeira, C. A., Witon, A., Schelter, B., Timmer, J., & Schulze-Bonhage, A. (2012). EPILEPSIAE–A European epilepsy database. *Computer methods and programs in biomedicine*, *106*, 127–138. doi:10.1016/j.cmpb.2010.08.011.

Jansen, J. J., Hoefsloot, H. C., van der Greef, J., Timmerman, M. E., & Smilde, A. K. (2005). Multilevel component analysis of time-resolved metabolic fingerprinting data. *Analytica chimica acta*, *530*, 173–183. doi:10.1016/j.aca.2004.09.074.

Kerr, W. T., Lau, E. P., Owens, G. E., & Trefler, A. (2012). The future of medical diagnostics: large digitized databases. *Yale J Biol Med*, *85*, 363–377.

Klösgen, W., & Zytkow, J. M. (2002). Knowledge discovery in databases: the purpose, necessity, and challenges. In *Handbook of data mining and knowledge discovery* (pp. 1–9). Oxford University Press, Inc.

Lambrou, T., Kudumakis, P., Speller, R., Sandler, M., & Linney, A. (1998). Classification of audio signals using statistical features on time and wavelet

transform domains. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on* (pp. 3621–3624). IEEE volume 6. doi:10.1109/ICASSP.1998.679665.

Landy, D. (1977). Culture, disease and healing. *Studies in medical anthropology*, (p. 467).

Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial intelligence in medicine*, *16*, 3–23. doi:10.1016/S0933-3657(98)00062-1.

Le, P. N., Ambikairajah, E., Epps, J., Sethu, V., & Choi, E. H. (2011). Investigation of spectral centroid features for cognitive load classification. *Speech Communication*, *53*, 540–551. doi:10.1016/j.specom.2011.01.005.

Lecron, F., Boisvert, J., Mahmoudi, S., Labelle, H., & Benjelloun, M. (2012). Fast 3D spine reconstruction of postoperative patients using a multilevel statistical model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 446–453). Springer. doi:10.1007/978-3-642-33418-4_55.

Link, B. G., & Phelan, J. (1995). Social conditions as fundamental causes of disease. *Journal of health and social behavior*, (pp. 80–94). doi:10.2307/2626958.

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annu. Rev. Physiol.*, *66*, 735–769. doi:10.1146/annurev.physiol.66.082602.092845.

Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., & Benali, H. (2006). Partial correlation for functional brain interactivity investigation in functional MRI. *Neuroimage*, *32*, 228–237. doi:10.1016/j.neuroimage.2005.12.057.

Maser, J. D., & Akiskal, H. S. (2002). Spectrum concepts in major mental disorders. *Psychiatric Clinics of North America*, *25*, xi – xiii. doi:10.1016/S0193-953X(02)00034-5.

Mennes, M., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2013). Making data sharing work: the FCP/INDI experience. *Neuroimage*, *82*, 683–691. doi:10.1016/j.neuroimage.2012.10.064.

Milham, M. P. (2012). Open neuroscience solutions for the connectome-wide association era. *Neuron*, *73*, 214–218. doi:10.1016/j.neuron.2011.11.004.

Milham, M. P., Fair, D., Mennes, M., Mostofsky, S. H. et al. (2012). The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, *6*, 62. doi:10.3389/fnsys.2012.00062.

Moore, G. W., & Hutchins, G. M. (1980). Effort and demand logic in medical decision making. *Metamedicine*, *1*, 277–303. doi:10.1007/BF00882620.

Mueller, A., Candrian, G., Grane, V. A., Kropotov, J. D., Ponomarev, V. A., & Baschera, G.-M. (2011). Discriminating between ADHD adults and controls using independent ERP components and a support vector machine: a validation study. *Nonlinear biomedical physics*, *5*, 1. doi:10.1186/1753-4631-5-5.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, *15*, 869–877. doi:10.1016/j.nic.2005.09.008.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, *9*, 97–113. doi:10.1016/0028-3932(71)90067-4.

Parvathi, I., & Rautaray, S. (2014). Survey on data mining techniques for the diagnosis of diseases in medical domain. *International Journal of Computer Science and Information Technologies*, *5*, 838–846.

Piwowar, H. A., Becich, M. J., Bilofsky, H., Crowley, R. S. et al. (2008). Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med*, *5*, e183. doi:10.1371/journal.pmed.0050183.

Purdon, S. E., Waldie, B., Woodward, N. D., Wilman, A. H., & Tibbo, P. G. (2011). Procedural learning in first episode schizophrenia investigated with functional magnetic resonance imaging. *Neuropsychology*, *25*, 147. doi:10.1037/a0021222.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106. doi:10.1007/BF00116251.

Quinlan, J. R. (1993). *C4. 5: Programs for Machine Learning* volume 1. Morgan Kaufmann.

Riaz, A., Alonso, E., & Slabaugh, G. (2016). Phenotypic integrated framework for classification of adhd using fmri. In *International Conference Image Analysis and Recognition* (pp. 217–225). Springer. doi:10.1007/978-3-319-41501-7_25.

Ross, J. S. (2016). Clinical research data sharing: what an open science world means for researchers involved in evidence synthesis. *Systematic Reviews*, *5*, 159. doi:10.1186/s13643-016-0334-1.

Ross, J. S., & Krumholz, H. M. (2013). Ushering in a new era of open science through data sharing: the wall must come down. *Jama*, *309*, 1355–1356. doi:10.1001/jama.2013.1299.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, *52*, 1059–1069. doi:10.1016/j.neuroimage.2009.10.003.

Russell, G., Ford, T., Rosenberg, R., & Kelly, S. (2014). The association of attention deficit hyperactivity disorder with socioeconomic disadvantage: alternative explanations and evidence. *Journal of Child Psychology and Psychiatry*, *55*, 436–445. doi:10.1111/jcpp.12170.

Segenreich, D., Paez, M. S., Regalla, M. A., Fortes, D., Faraone, S. V., Sergeant, J., & Mattos, P. (2015). Multilevel analysis of ADHD, anxiety and depression symptoms aggregation in families. *European child & adolescent psychiatry*, *24*, 525–536. doi:10.1007/s00787-014-0604-1.

Sidhu, G. S., Asgarian, N., Greiner, R., & Brown, M. R. (2012). Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. *Frontiers in systems neuroscience*, *6*, 74. doi:10.3389/fnsys.2012.00074.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., & Woolrich, M. W. (2011). Network modelling methods for FMRI. *Neuroimage*, *54*, 875–891. doi:10.1016/j.neuroimage.2010.08.063.

Stoean, R., & Stoean, C. (2013). Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. *Expert Systems with Applications*, *40*, 2677–2686. doi:10.1016/j.eswa.2012.11.007.

Strigo, I., Matthews, S., & Simmons, A. (2013). Decreased frontal regulation during pain anticipation in unmedicated subjects with major depressive disorder. *Translational psychiatry*, *3*, e239. doi:10.1038/tp.2013.15.

Tanno, L. K., Calderon, M. A., Goldberg, B. J., Akdis, C. A., Papadopoulos, N. G., & Demoly, P. (2014). Categorization of allergic disorders in the new World Health Organization International Classification of Diseases. *Clinical and translational allergy*, *4*, 42. doi:10.1186/2045-7022-4-42.

Telesford, Q. K., Morgan, A. R., Hayasaka, S., Simpson, S. L., Barret, W., Kraft, R. A., Mozolic, J. L., & Laurienti, P. J. (2010). Reproducibility of graph metrics in fMRI networks. *Frontiers in neuroinformatics*, *4*, 117. doi:10.3389/fninf.2010.00117.

The Neuro Bureau (2011). NITRC: neurobureau:AthenaPipeline - NITRC Wiki. http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline. [Online; accessed 18-06-2017].

Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, *59*, 301–320. doi:10.1348/000711005X67599.

Trostle, J. A. (2005). *Epidemiology and culture* volume 13. Cambridge University Press. doi:10.1017/CBO9780511806025.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*, 293–302. doi:10.1109/TSA.2002.800560.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*, 273–289. doi:10.1006/nimg.2001.0978.

Wagholikar, K. B., Sundararajan, V., & Deshpande, A. W. (2012). Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems*, *36*, 3029–3049. doi:10.1007/s10916-011-9780-4.

Wasan, S. K., Bhatnagar, V., & Kaur, H. (2006). The impact of data mining techniques on medical diagnostics. *Data Science Journal*, *5*, 119–126. doi:10.2481/dsj.5.119.

Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., Welsh-Bohmer, K. A., Wang, L., & Shen, D. (2012). Identification of MCI individuals using structural and functional connectivity networks. *Neuroimage*, *59*, 2045–2056. doi:10.1016/j.neuroimage.2011.10.015.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.