# Specifics of Medical Data Mining for Diagnosis Aid: A Survey

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

[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**

Data mining continues to play an important role in medicine; specifically, for the development of diagnosis aid models used in expert and intelligent systems. Although we can find abundant research on this topic, clinicians remain reluctant to use decision support tools. Social pressure explains partly this lukewarm position, but concerns about reliability and credibility are also put forward. To address this reticence, we emphasize the importance of the collaboration between both data miners and clinicians. This survey lays the foundation for such an interaction, by focusing on the specifics of diagnosis aid, and the related data modeling goals. On this regard, we propose an overview on the requirements expected by the clinicians, who are both the experts and the final users. Indeed, we believe that the interaction with clinicians should take place from the very first steps of the process and throughout the development of the predictive models, thus not only at the final validation stage. Actually, against a current research approach quite blindly driven by data, we advocate the need for a new expert-aware approach. This survey paper provides guidelines to contribute to the design of daily helpful diagnosis aid systems.

*Keywords:* Data Mining; Medicine; Diagnosis Aid; Explainable Artificial Intelligence

## 1. Introduction

As one of the trendiest research topics of our century, Data Mining (DM) makes key contributions to the scientific and technological advance in a considerable number of fields (Gupta, 2014; PhridviRaj and GuruRao, 2014). Coined during the nineties, the discipline is subject to a tough competition for the development of algorithms always more powerful, which aim at processing data

*Corresponding author. University of Mons, Department of Mathematics and Operations Research, 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)

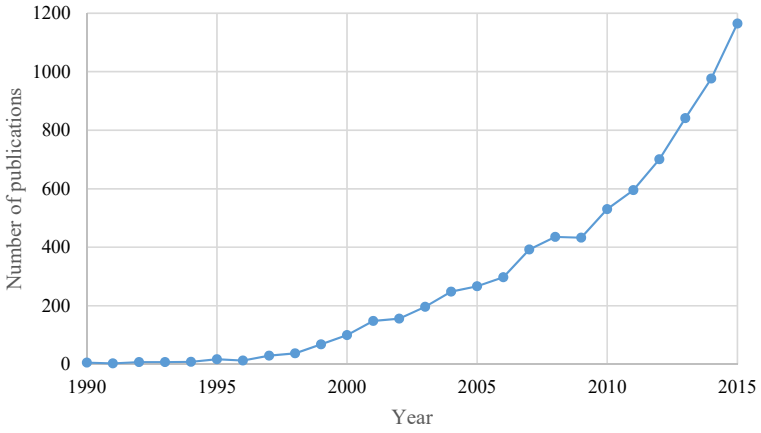


Figure 1: Evolution of the annual number of publications related to *medical data mining* in the Scopus database (Scopus) on a quarter of a century, from 1990 to 2015

to infer some knowledge in the form of patterns and/or relationships (Bellazzi and Zupan, 2008). The associated techniques are derived from the fields of both statistics and Machine Learning (ML), the latter which aims at developing computational methods able to extract generalizations from a set of data (Giudici, 2005).

Medical applications feature among the concerns of the DM community, with a significant increase in research interest over the last years (see Figure 1). This interaction comes in different disciplines (Bellazzi et al., 2011): at the cellular and molecular level (*bioinformatics*); at the tissue and organ level (*imaging informatics*); at the single patient level (*clinical informatics*); at the population and society level (*public health informatics*).

For half a century now, diagnosis prediction has been a very active research area of clinical informatics (Wagholikar et al., 2012). In this regard, with the advent of DM, research has progressively shifted away from the statistical approach long considered as a standard practice. Actually, under a hypothetico-deductive process, statistical analyses are driven to check a hypothesis stated beforehand and data samples are collected for this special purpose (Yoo et al., 2012). This statistical approach is surely adapted to raise differences between pathological and control groups, but not to set an individual assessment, i.e. a clinical examination per subject. In contrast, enriched by ML techniques, DM inductively processes a voluminous amount of data, to both extract knowledge and develop predictive models able to help in diagnosing pathologies (Vieira et al., 2017; Yoo et al., 2012; Bellazzi and Zupan, 2008). In such a process, statistics may find its place in feature engineer-

ing, before the stage of model building which is mainly based on ML methods of classification or regression (Esfandiari et al., 2014).

In that respect, it is through data mining that recent works were devoted to the early detection of cancer, e.g. see Lyu and Haque (2018); Aličković and Subasi (2017); Cichosz et al. (2016); Nahar et al. (2016); Esfandiari et al. (2014); Krishnaiah et al. (2013);Parvin et al. (2013); Gupta et al. (2011). Other pathologies, such as cardiac and pulmonary diseases, diabetes, hypertension, meningitis form besides a significant part of the research for more precise diagnoses (Esfandiari et al., 2014). Several psychiatric disorders, such as Attention Deficit Hyperactivity Disorder (ADHD) (Itani et al., 2018a; Abraham et al., 2017; Milham et al., 2012), Alzheimer (Papakostas et al., 2015), autism (Kosmicki et al., 2015), schizophrenia, depression and Parkinson (Woo et al., 2017) are also the object of extensive investigation.

As probably perceived by most of researchers, and certainly by the authors of the present paper, diagnostic decision support systems that have been proposed so far are not unanimously approved by clinicians (Wagholikar et al., 2012). Such systems, and the underlying predictive models, are notably found as being far from the field reality. It is thus most likely that data miners are not enough attentive to the specifics of medical diagnostic decision support. In particular, though the DM community was sensitized about the distinctive nature of medical applications (Cios and Moore, 2002), the predictive performance remains practically the lonely parameter within the scope of data miners, which encourages competition. This trend has been accentuated with the greater availability of open medical databases, shared by different medical and research centers worldwide (Di Martino et al., 2017; Woo et al., 2017; Di Martino et al., 2014; Esfandiari et al., 2014; Mennes et al., 2013; Ihle et al., 2012; Kerr et al., 2012; Milham et al., 2012; Poline et al., 2012). Some of these datasets were launched at the occasion of official contests, e.g. the ADHD-200 collection (Milham et al., 2012). In focusing almost exclusively on performance, these research works (1) miss challenges of better perceiving and understanding the issues proper to the medical field, (2) are exposed to the risk of yielding inconsistent models, since notably, recent studies showed that there may be no logic behind the predictions of accurate models (Ribeiro et al., 2016).

It is our strong conviction that the clinicians have to be involved in the whole development process of their decision support systems. Indeed, they bring expertise and knowledge to contribute to intelligent and expert systems. That is why, in the present paper, we will shed light upon the specifics of medical data mining for diagnosis aid and raise the related data modeling goals. For such a purpose, we will address the following questions.

(1) How can decision support models be more attractive to clinicians? What are the expressed requirements in this regard?

(2) What are the objectives corresponding to such requirements in terms of mathematical modeling?

(3) In what way medical data, particularly in this era of open medical data proliferation, makes data mining more challenging?

(4) To what extent are the current data mining techniques able to satisfy the clinicians' needs and to handle the particular nature of medical data simultaneously?

In answering these questions, we are led to describe a comprehensive *expert-aware approach* which stands out from the existing literature, through three main contributions exposed below.

- Because of the limited effectiveness of some models, Karpatne et al. (2017) push for a *theory-guided data science*. Such DM models are grounded in theoretical bases, in the domains of Physics and Chemistry mainly. In the context of medical diagnosis, we can adopt a similar approach, not guided by theory, but rather by the experts' domain knowledge. Our paper lays the bases for such an approach, in building a kind of bridge between both the medical and data mining domains.

- We not only express that the issue of diagnosis aid is of a particular nature, we also propose the translation of the associated specifics into modeling goals. Indeed, most of the papers that have interest on the specifics of the medical domain have a wide scope, and are thus not specifically focused on diagnosis, but also on prognosis and monitoring notably, which involves that modeling is not discussed with enough depth (Bellazzi and Zupan, 2008; Cios and Moore, 2002; Lavrač, 1999). Besides, we bring a more recent point of view compared to the papers that specifically addressed aided medical diagnosis (Wagholikar et al., 2012; Kononenko, 2001).

- We do not provide an overview of DM techniques and the related works; this was widely proposed in previous surveys (Kalantari et al. 2018; Kourou et al. 2015; Esfandiari et al. 2014; Wagholikar et al. 2012; Yoo et al. 2012). We rather question the existing DM techniques, given the modeling goals raised following the understanding of the problem and data. This allows us to raise some solid future research directions.

| PREDICTED AS ▷ | N | P |
|---|---|---|
| Negative (N) | TN | FP |
| Positive (P) | FN | TP |

Figure 2: Confusion matrix

The paper is organized as follows. In section 2, we expose the materials we considered to structure and make our survey. The results are presented in section 3 and discussed in section 4. Finally, we conclude this report in section 5.

## 2. Materials

### *2.1. Terminology*

Medical diagnosis is the result of a challenging task which consists of collecting and conciliating different information (Donner-Banzhoff et al., 2017; Hommersom and Lucas, 2016; Miller, 2016). The latter include the *symptoms* (subjective data) and the *signs* (objective data) of the trouble provided by clinical examinations and laboratory tests. In quest of explanations for these symptoms and signs, the clinicians come to the conclusion of the existence/absence of a trouble, i.e. the *diagnosis*.

A *test* is one among other elements that motivates a diagnosis (Gordis, 2014; Cios and Moore, 2002). The predictions of a clinical test are of several types. A patient with (respectively without) the disease $D$ predicted as such is designated as *true positive* (resp. *true negative*). In case of wrong predictions, the patients are *false positives* or *false negatives* respectively. Let $TP$ (resp. $TN$) denote the number of True Positives (resp. True Negatives) and $FP$ (resp. $FN$) the number of False Positives (resp. False Negatives); these quantities are usually exposed in a matrix of confusion (see Figure 2) (Witten et al., 2005). Different scalar metrics are computed from $TP$, $TN$, $FP$ and $FN$ to assess the performance of clinical tests; they are exposed in Table 1 (Lalkhen and McCluskey, 2008; Akobeng, 2007a,b). Let us note that *positive* and *negative predictive values* depend on the prevalence of the disease (Akobeng, 2007a): they are easily deduced from the knowledge of *sensitivity* and *specificity*, which are free from such an influence.

When several tests are required to check the presence of a medical condition, these tests may be assessed globally in terms of *net sensitivity* and *net specificity*. The values of these indicators depend on the way in which the tests were administered, i.e. sequentially or simultaneously (Gordis, 2014). Figures 3 and 4 present the mechanisms of sequential and parallel testing. For illustration

| Metric | Definition | Formula |
|---|---|---|
| Accuracy ($A$) | Rate of successful predictions | $A = \frac{TP+TN}{TP+FP+TN+FN}$ |
| Sensitivity or true positive rate ($tp$) | ▷ Ability to detect patients with a given disease.<br>▷ Probability that a patient with disease tests positive. | $tp = \frac{TP}{TP+FN}$ |
| Specificity or true negative rate ($tn$) | ▷ Ability to detect patients without a given disease.<br>▷ Probability that a patient without disease tests negative. | $tn = \frac{TN}{TN+FP}$ |
| Positive Predictive Value ($PPV$) | Chance that a patient, predicted as having a given disease, is truly so. | $PPV = \frac{TP}{TP+FP}$ |
| Negative Predictive Value ($NPV$) | Chance that a patient, predicted as free from a given disease, is truly so. | $NPV = \frac{TN}{TN+FN}$ |

Table 1: Performance metrics of screening tests

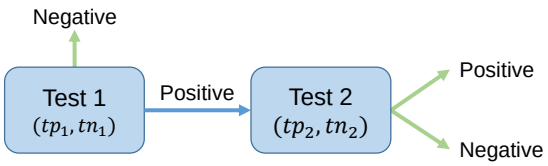


Figure 3: Sequential testing

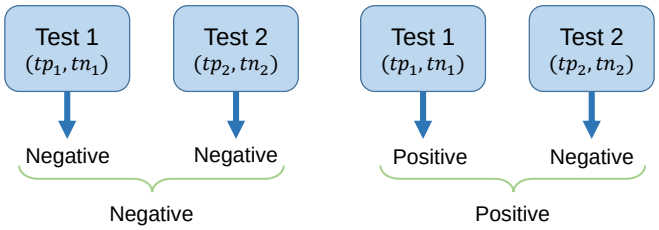


Figure 4: Parallel testing

purposes, the example presents the case of two tests; the associated reasoning may be generalized to situations involving more tests. In case of sequential testing, a patient is submitted to another round of examination if he/she tested positive, in order to settle definitely his/her medical condition. If the patient tests positive following a second round of examination, the subject is diagnosed with the disease in question. Thus, if one of both tests presents a negative result, the patient is considered as disease-free. The associated net sensitivity and specificity are expressed as:

$$tp = tp_1 \cdot tp_2 \quad \text{and} \quad tn = tn_1 + tn_2 - tn_1 \cdot tn_2.$$

In contrast, in case of parallel testing, a patient is considered as negative once all tests confirm this condition simultaneously. In this case, the associated net specificity and sensitivity are given by:

$$tn = tn_1 \cdot tn_2 \quad \text{and} \quad tp = tp_1 + tp_2 - tp_1 \cdot tp_2.$$

In the same way that a clinician can ask for the opinion of an expert confrere, he/she can resort to *models* for diagnosis aid. The only difference between both scenarios rests on the external nature of the diagnostic support, either human or computerized. The data of one or several test(s) are potential inputs for diagnosis aid models. It should be noted that *non-interpreted* outcomes of testing (e.g. a cholesterol level, a scan) constitute the model inputs, and not the value of the test(s), i.e. positive or negative. Actually, it is the role of the predictive model to determine a patient's medical condition in output.

In light of the foregoing, in the present survey, what we refer to as a *model* is different from a *test*, the latter being a potential input of the former. A model provides a recommendation of diagnosis; a test provides a result that allows, among other potential information, to make a diagnosis.

### *2.2. The knowledge discovery process*

The extraction of knowledge for the purpose of diagnosis aid fits into a Knowledge Discovery Process (KDP). Since its pioneer formalization by Fayyad et al. (1996), alternative models were proposed, either academically- or industrially-minded (Kurgan and Musilek, 2006). In particular, the KDP was adapted for medical applications and illustrated for the issue of diagnosis aid by Cios et al. (2007, 2000). The associated steps are summarized below.

**Understanding of the problem** The process is initiated by the problem statement, the definition of the objectives, and the sufficient appropriation of a domain-specific vocabulary. Obvi-

ously, interactions with domain experts are essential. At this level, the choice of data mining techniques is partially foreseen given the expressed requirements.

**Understanding of the data** This step consists of collecting and exploring data, i.e. observing and analyzing the information.

**Preparation of the data** The creation of target datasets (Fayyad et al., 1996) involves notably noise removal as well as checking the completeness and consistency of data. Then, data are processed through engineering, selection and possible reduction of pertinent features.

**Data mining** This process receives the prepared datasets, and extracts knowledge, i.e. patterns, relationships (Bellazzi and Zupan, 2008).

**Evaluation of the discovered knowledge** The results are closely considered: they are expected to bring new and interesting elements, to be understood and to make sense. Here, domain experts have to play an important role in their ability to interpret and assess the results.

**Use of the discovered knowledge** It can lead to action taking, decision making or systems deployment (Fayyad et al., 1996).

The KDP is not strictly a one-way process as it is not excluded to reconsider the work of previous stages: this allows to reinforce the consistency of the results (Cios et al., 2007). For example, the final evaluation may ask for refining the results. Or to better understand the data, a re-understanding of the problem may strengthen the domain-specific knowledge.

### *2.3. Acceptance criteria*

One difficulty related to medical DM is that it may target different publics with the resulting necessity to address different expectations.

Actually, a DM approach may be requested in the medical field by researchers and specialists in order to study a given pathology through the identification of explanatory factors. In that case, the extracted knowledge is validated if it carries a certain level of credibility, measured by means of criteria related to statistical power notably. If endorsed by the scientific community, such results may be taken into consideration (directly or indirectly) by clinicians faced with a diagnosis task.

As suggested in section 2.2, the extracted knowledge may also be deployed in the form of a computerized diagnosis aid. Despite they are the lonely users of such technologies, the clinicians are influenced in their expectations, e.g. by the patients who place a lot of hope in a fair diagnosis.

Different models were developed in an effort to explain how a clinician may accept a technology and integrate it to his/her working practices (Andargoli et al., 2017; Ketikidis et al., 2012; Holden and Karsh, 2010; Yarbrough and Smith, 2007). The most popular is the Technology Acceptance Model (TAM), introduced by Davis et al. (1989) and revised by Venkatesh and Davis (2000) (TAM2). Enjoyed for its concise structure, the model depicts the psychological process which, influenced by material and social factors, leads to the intention of using a computerized application in different contexts (Yarbrough and Smith, 2007).

Venkatesh and Davis (2000) report that the acceptance of technology is acquired in practice once its *usefulness* and *ease of use* are both perceived by the user. Moreover, the ease of use is one of the factors influencing the user's perception of the usefulness of the application. The perception of usefulness rests also on social factors: the *subjective norm*, i.e. the user's (professional or private) surroundings' opinion regarding his/her decision to adopt (or not) the application, and the *image*, i.e. the social status the application provides to the user (Mun et al., 2006; Chismar and Wiley-Patton, 2002).

The subjective norm impacts directly the intention of use. This influence is exerted on the clinician by his/her patients but also by the professional environment. Indeed, the physician is sensitive to the opinion of confreres, particularly of reference people in the domain, even though this opinion may be contrary to the physician's beliefs (Mun et al., 2006). As for the influence of the patients, the study of Shaffer et al. (2013) shows they often tend to demonize computerized diagnostic support. Conversely, non computer-assisted practices are perceived as a pledge of professionalism; may the clinician resort to the opinion of an expert confrere is even perceived as an intelligent act. Yet in both last cases, the clinicians might base their decision on elements provided in the literature and extracted from a DM approach. Thus, the involvement of computing in the diagnostic process, if only to have an advice, would in itself lead the physician's image to take a hit towards colleagues and/or patients (Mun et al., 2006).

In the present survey, we will highlight the specifics of DM to develop diagnostic decision support models which meet the requirements of the clinicians. We will deal with how to make computerized diagnosis aid fulfill criteria of *output quality* and *results demonstrability* advocated by TAM. Nevertheless, it must be recognized that adopting a suitable approach of modeling does not guarantee exclusively the acceptance of the models since some related factors (e.g. subjective norm) do not fall within DM concerns.

| | | Knowledge Discovery Process | Nature of Medical Data | Overview of DM Techniques | Performance Evaluation | Specifics of Medical DM |
|---|---|---|---|---|---|---|
| Lavrač (1999) | Selected techniques for data mining in medicine | | ✓ | ✓ | ✓ | ✓ |
| Kononenko (2001) | Machine learning for medical diagnosis: history, state of the art and perspective | | | ✓ | | ✓ |
| Cios and Moore (2002) | The uniqueness of medical data mining | ✓ | ✓ | | ✓ | ✓ |
| Bellazzi and Zupan (2008) | Predictive data mining in clinical medicine: current issues and guidelines | ✓ | ✓ | ✓ | | ✓ |
| Harrison (2008) | Introduction to the mining of clinical data | | ✓ | | | |
| Iavindrasana et al. (2009) | Clinical data mining: a review | ✓ | | ✓ | ✓ | |
| Pandey and Mishra (2009) | Knowledge and intelligent computing system in medicine | | | ✓ | | |
| Yardimci (2009) | Soft computing in medicine | | | ✓ | | |
| Wagholikar et al. (2012) | Modeling paradigms for medical diagnostic decision support: a survey and future direction | | | ✓ | | |
| Yoo et al. (2012) | Data mining in healthcare and biomedicine: a survey of the literature | | ✓ | ✓ | | |
| Esfandiari et al. (2014) | Knowledge discovery in medicine: current issue and future trend | ✓ | ✓ | ✓ | ✓ | |
| Chen and Fawcett (2016) | Using data mining strategies in clinical decision making: a literature review | ✓ | | ✓ | | |
| Patel and Patel (2016) | Survey of data mining techniques used in healthcare domain | ✓ | | ✓ | | |
| Miotto et al. (2017) | Deep learning for healthcare: review, opportunities and challenges | | ✓ | ✓ | | |

Table 2: Baseline reports on medical DM, presented by covered topics

### *2.4. Previous reports on medical data mining*

Our survey is based on fourteen narrative reports that address medical DM exhaustively; some focus specifically on the issue of diagnosis aid (Wagholikar et al., 2012; Kononenko, 2001). These reports are listed in Table 2, according to the themes to which they refer (directly or indirectly). It must be emphasized these reports do not necessarily address the themes with the same depth. Moreover, the distribution in themes is only indicative and it makes no necessarily representation of the papers' structure.

Some reports depict the KDP; different standards are addressed. It is generally highlighted that DM is included within (and is thus distinct from) the KDP. The **nature of medical data** and the related terminologies is a frequently exposed subject. Most of studies propose an **overview on DM techniques** which were considered in medical applications and/or provide examples of works having used such techniques. Emphasis is placed on subsets of DM techniques (Miotto et al., 2017; Pandey and Mishra, 2009; Yardimci, 2009) or specific methods that are pertinent/often requested in the medical domain (Bellazzi and Zupan, 2008; Kononenko, 2001; Lavrač, 1999). Other papers give a general topology of DM techniques, fully explored and discussed (Esfandiari et al., 2014; Iavindrasana et al., 2009). Furthermore, as mentioned before, the **performance evaluation** is a crucial aspect of DM: different indicators may achieve such an assessment (Esfandiari et al., 2014; Iavindrasana et al., 2009; Cios and Moore, 2002; Lavrač, 1999). The triplet accuracy/specificity/sensitivity is the most popular set of indicators in the medical domain. Finally, some papers give distinctive elements of medical decision making and/or applications, i.e. what we designate as the **specifics of medical data mining**. The study of Kononenko (2001) gives some indications on specific requirements that models for diagnosis aid should meet to be efficient, while the studies of Bellazzi and Zupan (2008); Cios and Moore (2002); Lavrač (1999) focus their consideration on medicine in general, e.g. on diagnosis, prognosis or monitoring.

## 3. Results

The development of diagnostic decision support models is enrolled in a KDP (see section 2.2). A first step in the process is the definition of the problem and its scopes following concertation with domain experts (Cios et al., 2007). The scopes are then translated into *DM goals* that we propose and discuss in the first part of the present section. Then, we focus on understanding medical data and their specifics, which allows in turn to reinforce our understanding of the problem, with the challenges that it brings.

| | Requirements | DM goals |
|---|---|---|
| R1 | Ability to validate the model against the expert personal domain knowledge, through scientific arguments which support the model | Global interpretability |
| R2 | Ability to validate a decision on a per-patient basis | Local interpretability |
| R3 | A reliable predictive model | Performance & Robustness |
| R4 | A cautious predictive model: errors have not the same impact in both senses | Appropriate balance between specificity and sensibility |
| R5 | Ability to estimate the possibility of disease in a subject | A probabilistic output |

Table 3: Requirements and translation into DM goals

### *3.1. The problem of diagnosis aid*

It is estimated that one out of ten medical diagnoses is mistaken (Graber, 2013). Alternatively, two clinicians may come to different diagnoses with regard to a given patient. Setting up a diagnostic decision support model is thus surely a challenging objective (Vieira et al., 2017). Moreover, it is complex to reproduce the intuition of the physician acquired after several years of expertise (Hommersom and Lucas, 2016). But DM has a strong side: it can process a large amount of data where humans fail due to the limits of their mental processes (Donner-Banzhoff et al., 2017).

But what are the requirements of clinicians towards models for diagnosis aid? We propose here to state general needs, gathering some elements found in the literature (Ribeiro et al., 2016; Bellazzi and Zupan, 2008; Mencar et al., 2007; Kononenko, 2001; Lavrač, 1999) and those suggested by clinical collaborators. The expectations and associated DM goals are presented in Table 3; these will be discussed and justified in the following sections. The list may need to be extended by additional aspects depending on each specific context of diagnosis.

#### *3.1.1. Interpretability*

The clinicians' foremost requirement is to be able to justify and validate the prediction of a diagnosis aid model. To this end, the recommandation must be based on an *interpretable* model (Mencar et al., 2007; Kononenko, 2001; Lavrač, 1999). Efforts have been recently deployed to define this concept (Doshi-Velez and Kim, 2017; Lakkaraju et al., 2016; Lipton, 2016). According to Doshi-Velez and Kim (2017), interpretability is "*the ability to explain or to present in understandable terms to humans*". Based on this definition, we worked on defining the properties of interpretability regarding the problem of diagnosis aid.

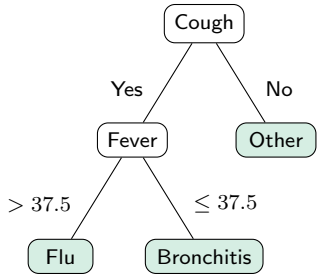


Figure 5: An example of decision tree (taken from Bock et al. 2013)

*Transparency (readability, human readability).* Through a code or a language, transparent models are able to show how their output (e.g. the predicted diagnosis) is deduced from the inputs (e.g. the patient's data) (Bellazzi and Zupan, 2008). This is the property of white boxes in contrast to black boxes which are completely opaque (Baesens et al., 2009). For example, decision trees are white boxes as they provide all the progression that leads to make a final decision, under a top-bottom reading approach (see Figure 5). Artificial neural networks are black boxes: no link can be made between the inputs and the output of the models (Bellazzi and Zupan, 2008). To sum up, the content of transparent models is accessible, a feature which several papers also express as *(human-)readability* (Hajiloo et al., 2013; Kukenys et al., 2011; Anglade et al., 2009; Corney, 2002).

*Intelligibility.* Caruana et al. (2015); Lou et al. (2013, 2012) introduced intelligibility as the ability to demonstrate clearly the influence of each model input in the final decision. In particular, they advocate as intelligible Generalized Additive Models (GAM), given by:

$$g(y) = f_1(x_1) + \ldots + f_n(x_n)$$

if $x_i$ denotes the model inputs for $i = 1, \ldots, n$, $f_i$ the shape function, $g$ the link function, $y$ the target variable. Here, the learning phase has to predict the shape functions; the contribution of every input $x_i$ to the final result is represented by its image through $f_i(x_i)$. But basically, GAMs can not model potential interactions between inputs. Generalized Additive Models plus Interactions ($GA^2Ms$) address this inconvenient (Lou et al., 2013):

$$g(y) = \sum_i f_i(x_i) + \sum_{i,j} f_{i,j}(x_i.x_j).$$

The underlying algorithm deduces the most likely existing interactions. Taking into consideration some selected pairwise interactions improve the performances and do not impair the intelligibility of the models. As a matter of fact, these interactions can be viewed on heatmaps (Lou et al., 2013). GA$^2$Ms were considered for clinical prediction with success, e.g. for the prediction of pneumonia risk (Caruana et al., 2015). Naturally, GAMs are not the lonely existing intelligible models: as suggested by Lipton (2016), decision trees comply with intelligibility as each model node with its incident branches can be translated literally, in *plain text description*. If we look again at Figure 5, the first node coupled with its right incident branch can be read as: "patients who do not have a cough".

*Simplicity (compactness).* According to Occam's razor, simplicity must be found in any process. Enhanced as a feature of interpretability (Lipton, 2016; Guillaume, 2001; Revelle and Rocklin, 1979), simplicity is acquired through a reduced number of variables (Kononenko, 2001). This results in a faster decision inference (Lipton, 2016) and in a certain convenience for both the clinician and the subject. Indeed, a sophisticated diagnosis may require significant amount of data, whereas the collection of medical data may be expensive, time consuming, and uncomfortable for the patient (Kononenko, 2001). Among the simplest models, 1R (Holte, 1993) is well-known: it considers a single attribute to make predictions. To go further, TAR2 (Menzies and Hu, 2006) proposes compact theories in two components: the minimal sets of constraints that change the sample distribution to focus on favorable or unfavorable situations respectively, i.e. health and illness in our context. It may provide simpler theories than decision trees.

*Comprehensibility.* The ability to apprehend a global and synthesized vision of the model constitutes an important factor of interpretability. Obviously, it is likely that a non-simple model is not comprehensible; however, simplicity is not a guarantee of comprehensibility (Domingos, 1999). If simplicity is easily measurable in numerical terms (e.g. the number of levels in a decision tree), comprehensibility is much more complex to assess because it may depend on how a clinician, personally, perceives the model. For example, as diagnosis aid tools, some clinicians may prefer decision trees while others would rather prefer a list of rules. Moreover, regardless the clinicians' preference related to the nature of a model, between a short one and a long one, having both the same predictive accuracy, the long model may be found unanimously as more interesting, e.g. if the latter involves fewer attributes.

*Meaningfulness.* The model expressed in non domain-specific terms may not be clear to experts. In other terms, any part, any parameter or feature of the model must make sense, i.e. refer to the usual vocabulary/knowledge of the experts (Mencar et al., 2007). This quality influences necessarily the nature of the model and the pertinent features on which diagnosis aid should be built, but also the type of operations applicable to the features.

A model may not be declared as interpretable in absolute terms (Lipton, 2016; Ribeiro et al., 2016). For example, the interpretability of linear models decreases as the number of input variables increases because of loss in simplicity (Mencar et al., 2007). Such an observation is also true concerning decision trees, as the number of nodes increases (Otte, 2013). Besides, performing a principal component analysis on features leads to new ones that, though constituting a combination of the original features, are less perceptible; whatever the white box models in which such new features are introduced, meaningfulness is practically lost.

In Table 3, we presented *global* and *local* interpretability in answer to requirements R1 and R2 respectively. Indeed, both notions are slightly different (Ribeiro et al., 2016). On the one hand, global interpretability is a prerequisite to the global validation of a model as a whole, i.e. from the inputs to their individual involvement and interaction in the final decision. Translating this concept in the context of diagnosis aid, this means the clinician can raise explanations for diagnosis recommendation, and for a range of patient profiles. If this synthesized view of the model behavior is pertinent, the model is globally validated by the user. On the other hand, local interpretability is a prerequisite to the local validation of a model. In this context, this means that, for a given patient, the clinician is provided with explanations for the diagnosis recommendation; on this basis, he/she has the ability to trust (or not) the prediction of a model.

Naturally, global interpretability involves local interpretability: if a model may be interpreted by its user whatever the values of the inputs, interpretation is acquired for any combination of input values. Conversely, a model may be locally, though not globally interpretable: this is the case if a property of interpretability is missing at a global level. For example, a lost of global simplicity may affect the ability to understand the model in its whole, while the chain of reasoning which leads to a given decision may be simple.

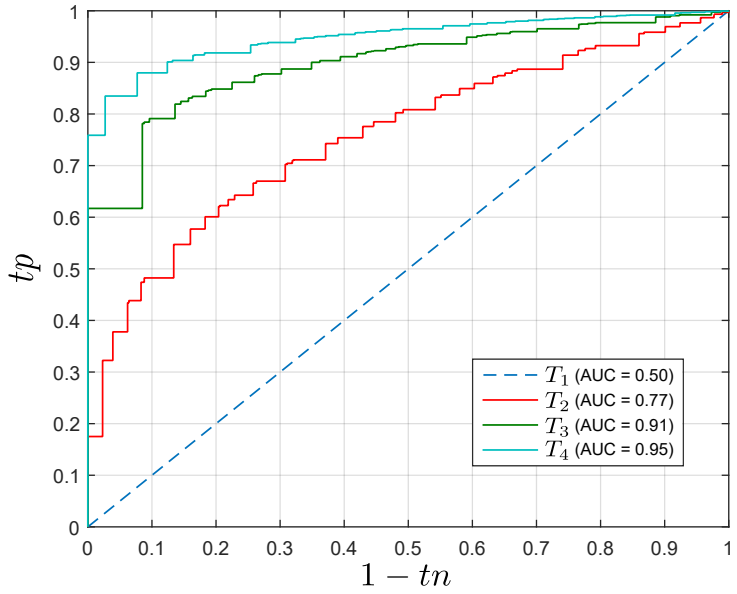


Figure 6: ROC graph: comparison of four clinical tests

*3.1.2. Performance & Robustness*

Reliability is a requirement that models for diagnosis aid should meet. In DM terms, it means that they must be performant and robust.

To measure performance, many are the evaluation metrics in DM: they are derived from different disciplines (Sokolova and Lapalme, 2009; Witten et al., 2005; Klösgen and Zytkow, 2002; Baldi et al., 2000). It is understood that the assessment of models for diagnosis aid should be preferably achieved with the same indicators that the clinicians usually use for the assessment of their own clinical tests and to which they are therefore familiar, i.e. accuracy, specificity and sensitivity (see section 2.1 and Table 1). Furthermore, a graphic indicator is also used to assess the relevance of clinical tests, and may thus be considered to check the performance of predictive models.

Let us consider a clinical test with a continuous outcome $s$. Given a threshold $\alpha$, we suppose that a patient is predicted as disease-free if $s < \alpha$ and pathological otherwise. Depending on the value of $\alpha$, the true and false positive fractions, i.e. the sensitivity ($tp$) and the complementary of the specificity (1-$tn$) respectively, are variables. Indeed, the higher the threshold $\alpha$, the higher the fraction of false negatives is, and the less sensitive the test is. The couples of values (1-$tn$, $tp$) corresponding to different levels of threshold $\alpha$ are the points of a Receiver Operating Characteristic (ROC) curve (Lasko et al., 2005). The Area Under the Curve (AUC) may be interpreted as the average sensitivity of the test. Figure 6 shows the ROC curves of four Tests ($T$). Obviously,

$T_4$ presents the best behavior since it notably keeps the highest levels of sensitivity among the other classifiers, in the range of tested values for $\alpha$. With an AUC of 0.5, $T_1$ is such a random test. Actually, the ROC curve is convenient since it allows to compare different tests directly and visually. Many techniques exist to select the value of $\alpha$ that achieves interesting performances (see Kumar and Indrayan 2011).

Robustness regards the predictive behavior of the model under changes impacting its inputs or parameters. Conceptually, this involves that the model output should not be different for patients having similar profiles. Moreover, a model should not be significantly influenced by weak parameter shifts. Thus, taking inspiration from Odenbaugh (2011), we can define robustness from two perspectives.

- *Robustness against model tuning*: let $M = \{M_1, M_2, M_3, \dots M_n\}$ denote a set of models sharing the same settings except for a certain parameter. For a given instance $x_i$, a robust prediction $y_i$ with regards to $M$ is such that $M_j(x_i) = y_i$, $\forall M_j \in M$.
- *Robustness against input variability*: let $X = \left\{x^{(1)}, x^{(2)}, x^{(3)}, \dots x^{(n)}\right\}$ denote a set of similar instances, i.e. close in the sense of a distance metric or sharing the same feature values, except for one. A robust prediction $y$ with regards to $X$ is such that $M(x^{(i)}) = y$, $\forall x^{(i)} \in X$.

Generally, robustness is assessed on a range of values which correspond to low deviations of a reference situation.

### *3.1.3. Appropriate balance between specificity and sensitivity*

Depending on the way by which they are committed, some errors of prediction may not be tolerated. The analysis of sensitivity and specificity allows to measure such a risk.

In some cases like the diagnosis of psychiatric disorders in children, should predictive models make errors, it must be ensured that these wrong predictions have the lowest impact on them. Indeed, a model having a low ability to detect healthy subjects (i.e. low specificity) is far from being cautious, since they are exposed to the risk of being prescribed an unnecessary medication (with risks for still growing children). The opposite situation is less damaging: if a model has a low ability to detect pathology (i.e. low sensitivity), the detection of the trouble may be just delayed in time in the child's development. Thus, for a given accuracy, models may not be necessarily cautious: it depends on the balance between sensitivity and specificity. To illustrate this point, let us consider the confusion matrices of two models of Attention Deficit Hyperactivity Disorder (ADHD) diagnosis having the same accuracy: see Tables 4 and 5. In both cases, the accuracy is of 68%, but model B

| **Predicted as >** | ADHD | TD |
|---|---|---|
| ADHD | 48 | 2 |
| TD | 30 | 20 |

Table 4: Example A

| **Predicted as >** | ADHD | TD |
|---|---|---|
| ADHD | 20 | 30 |
| TD | 2 | 48 |

Table 5: Example B

(Table 5) is more cautious than model A (Table 5): indeed, the specificity of B is of 96%, while the specificity of model A is poor with 40% of Typically Developing (TD) children detected only.

This being said, high sensitivity is important for the diagnosis of severe diseases (e.g. cancer) (Lalkhen and McCluskey, 2008). Yet simultaneous testing (see section 2.1) reduces the number of false negatives, so it increases the net sensitivity. Indeed, under such a modality of testing, a patient is considered positive once he/she was predicted as so by at least one test (Gordis, 2014). In the same way, data miners can develop a set of several models, based on the (non-interpreted) outcomes of parallel testing and come to the existence of disease in case at least one of these models predicts positivity.

*3.1.4. A probabilistic output*

As discussed before, the imprecision of diagnostic tests involves the impossibility to predict with certainty a subject's condition. In reality, the result of a diagnosis is more nuanced than that. That is why Bellazzi and Zupan (2008) suggest that predictive DM models should provide an output probability for diagnosis with confidence intervals, which is rarely completed.

In clinical practice, the knowledge of specificity and sensitivity is sufficient to compute the probability of contracting a disease for a given patient. To this end, two indicators are computed: the likelihood ratios for positive and negative tests (Akobeng, 2007b; Deeks and Altman, 2004).

- *The likelihood ratio for a positive test* ($LR+$) expresses how many more times a patient with a given disease is likely to test positive than a patient without the disease:

$$LR+ = \frac{\text{Sensitivity}}{1 - \text{Specificity}} = \frac{tp}{1 - tn}.$$

  A test such that $LR+ > 1$ involves that a subject with disease is more likely to test positive than a subject without the same disease.

- *The likelihood ratio for a negative test* ($LR-$) expresses how many more times a patient with

a given disease is likely to test negative than a patient without the disease:

$$LR- = \frac{1 - \text{Sensitivity}}{\text{Specificity}} = \frac{1 - tp}{tn}.$$

An interesting test is such that $LR- < 1$: it is naturally expected that a patient with disease is less likely to test negative than a subject without the disease. In general, the inverse of $LR-$ is more intuitive to interpret: it expresses how many more times a patient without disease is likely to test negative than a patient with the disease.

Before medical examination, a clinician can estimate the *pre-test probability* $P_0$ that a patient is affected by a disease given his/her medical history and the exhibited symptoms. Once the patient has been tested, it is possible to deduce the probability a posteriori that he/she has the disease, i.e. the *post-test probability* $P_1$, given the likelihood ratio $LR$ of the test. Indeed, the Bayes theorem expresses that (Akobeng, 2007b):

$$\text{Post-test odds} = LR \times \text{Pre-test odds} \quad \Leftrightarrow \quad \frac{P_1}{1 - P_1} = LR \times \frac{P_0}{1 - P_0}$$

$P_1$ is deduced using the likelihood ratio ($LR+$ or $LR-$) corresponding to the test result ($+$ or $-$).

With the use of a Fagan's nomogram, the post-test probability is deduced easily, in a graphical form (see Figure 7). By extending the segment joining the pre-test probability (left axis) and the likelihood ratio (middle axis), the intersection with the right axis gives the post-test probability. Figure 7 shows an example with a pre-test probability of 45%. The clinical test is characterized by $tp = 93\%$ and $tn = 65\%$. If the subject tests positive, the probability of disease increases to 68.5%; in case of negative result, the probability decreases to 8.8%. Note that a test with a unit likelihood ratio (dotted line) gives no additional information to a clinician. The reasoning may be transposed to predictive models for diagnosis aid as they are also characterized by sensitivity and specificity values.

### *3.2. Medical Data*

In contrast to the problem of diagnosis aid, the nature of medical data is a subject widely covered by reviews on medical DM. Table 6 presents the features of these data, and the reports that cover

[1] Drawn with the function `nomogrammer` of the `ggplot2` package in R software (Adam M. Chekroud, 2016; R Core Team, 2013)

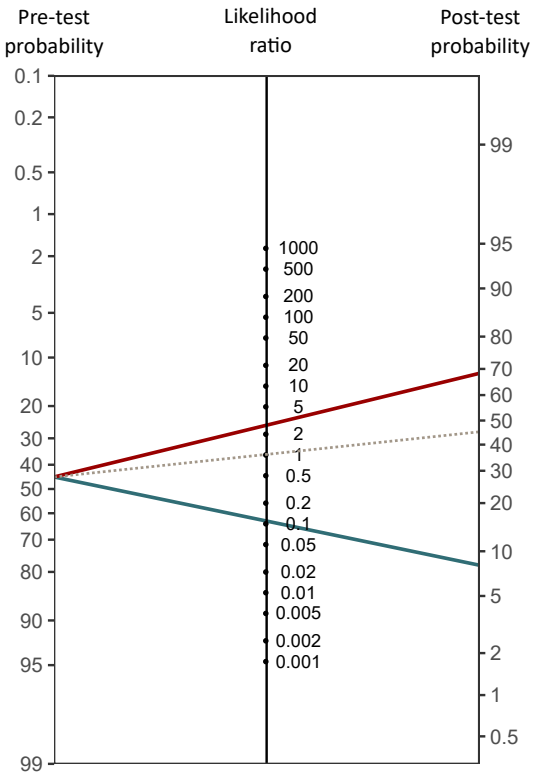


Figure 7: Fagan's nomogram[1]

them partly or completely. In particular, Cios and Moore (2002) and Harrison (2008) gave overviews which practically corroborate one another, except for the concepts encompassed by the *heterogeneity* feature of medical data. Indeed, Cios and Moore (2002) consider that heterogeneity is explained by the varied nature, volume and imprecision of medical data, their non-single interpretation by physicians, the non-existence of a single disease terminology as well as the poor compatibility with the logic of mathematics. On the other hand, Harrison (2008) explains the heterogeneity of medical data by their various nature and by the existence of different methodologies to acquire a same measurement value. In the rest of the present survey, we associate the heterogeneity of data with the features preventing their uniform processing whether between records or attributes. This brings us to expand the description of heterogeneity provided by Harrison (2008) with additional features. In the second part of this section, we outline the factors that increased this natural heterogeneity of medical data in open databases released by the culture of data sharing.

### *3.2.1. Characteristics of medical data*

By their particular nature, medical data require specific handling, with thus additional (implicit) needs to consider for the development of models for diagnosis aid. In the following paragraphs, we

| | Lavrač (1999) | Cios and Moore (2002) | Bellazzi and Zupan (2008) | Harrison (2008) | Yoo et al. (2012) | Esfandiari et al. (2014) | Miotto et al. (2017) |
|---|---|---|---|---|---|---|---|
| Different data types & measurement methods | | ✓ | ✓ | ✓ | | ✓ | |
| High dimensionality | | ✓ | | ✓ | | ✓ | ✓ |
| Imprecision | | ✓ | | ✓ | | | ✓ |
| Incompleteness | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Inconsistency | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ |
| Non-single disease terminology | | ✓ | | ✓ | | | |
| Non-single interpretation | | ✓ | | ✓ | | | |
| Poor compatibility with the logic of mathematics | | ✓ | | ✓ | | ✓ | |
| Temporal components | | | | ✓ | | | ✓ |
| Sensitive property & use issues | | ✓ | | ✓ | ✓ | ✓ | |

Table 6: Medical data features covered by different reviews

summarize these features; some are illustrated in the specific context of diagnosis.

*Heterogeneity.* Several sources of heterogeneity exist in medical data.

- Medical data are of different types such as biological data, images, signals, discrete values, interviews.
- Several scales exist to assess some discrete indicators (Harrison, 2008). For example, Stanford-binet (Bain and Allin, 2005), Raven's matrices (Raven, 2000), Wechsler (Kaufman et al., 2015; Wechsler, 2014, 2003) are different tests used to measure the intellectual quotient.
- A form of heterogeneity may characterize diseases; types of cancer are incidentally reviewed from time to time (Miotto et al., 2017; Kourou et al., 2015; Allison and Sledge, 2014).

*High dimensionality.* For a given patient, many types of data may be available. Moreover, imaging and signal acquisition technologies generate voluminous files and raw data (Cios and Moore, 2002). For example, at a given instant, a brain MRI consists of multiple 2D images acquired as slices of the whole brain volume, during a complete rotation of the MRI equipment. In consequence, to assess the brain activity over a certain time period, several brain volumes, i.e. sets of 2D slices, are acquired per patient (Bates, 2011; Reimer et al., 2010).

*Imprecision.* Test results, observations and diagnoses are subject to imprecision (Harrison, 2008; Cios and Moore, 2002). Actually, a diagnosis is established from a set of tests which do not achieve perfection; they are also characterized by specificity and sensitivity values (Maxim et al., 2014; Cios and Moore, 2002).

*Incompleteness.* It is common to raise missing information in medical datasets. The problem of missing values can occur because of economical, ethical, medical or technical reasons (Yoo et al., 2012; Cios and Moore, 2002).

- Economical reasons: generally, medical datasets are not collected for research purposes. Thus, the idea of collecting unnecessary data (such as images and signals) from a medical point of view is mostly given up to save money and time.
- Ethical reasons: patients may have not given their consent for the use of their data and/or to undergo an exam on a voluntary basis. Questions of privacy remain sensitive at this level.
- Medical reasons: the medical status of a patient may forbid the acquisition of some data. This status is defined by factors such as age, family and/or patient medical history. Thereby, all kinds of information may not necessarily be collected regarding a single patient.
- Technical reasons: the entirety of medical information is not necessarily featured in computerized systems. Indeed, some data remain on paper, which involves missing values in information systems. This is the case in hospitals that have not been pushed to a full computerization of their processes.

Incompleteness may also occur in case of medical datasets including representatives of a single (healthy/pathological) population. In such a case, there is no possibility to assess the features of a population against another.

*Inconsistency.* The quality of medical data is not guaranteed. Indeed, imaging and signal acquisition technologies are subject to noise, which may result in inconsistent data (Cios and Moore, 2002). Though preprocessing pipelines exist, it is not ensured that noise has been well removed. The full elimination of disturbance sources may even sometimes be impossible, notably because some of these disruptions could be ignored by the data miner.

Sometimes, inconsistency arises form the data incompleteness. For example, it happens that patients, though presenting similar features, have not the same diagnosis. This is explained by the fact that clinicians can take into consideration other facts (e.g. patient medical history) that are not necessarily consigned in the database used by the data miner.

*Non-single interpretation.* The interpretation of medical data by physicians is subjective. Incidentally, research works were dedicated to the best representation of the agreement between physicians on diagnoses (McHugh, 2012; Pies, 2007), e.g. the Kappa statistics (McHugh, 2012; Cohen, 1960):

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

with $Pr(a)$, the agreement rate among physicians i.e. the *observed agreement*; $Pr(e)$, the probability of a random agreement i.e. the *expected agreement* (Viera et al., 2005). For example, if clinicians agreed on 55 diagnoses among 100, then $Pr(a) = 0.55$. $Pr(e)$ is computed as:

$$Pr(e) = \sum_{i=1}^{m} \prod_{j=1}^{n} p_{ij}$$

if $p_{ij}$ is the probability that clinician $j$ gives the diagnosis $i$, $m$ being the number of possible diagnoses, and $n$, the number of clinicians. Actually, this probability takes into consideration the prevalence of the disease within the population soliciting the clinician in question.

Table 7 shows Kappa coefficients for psychiatric disorders in children (Freedman et al., 2013). Higher $\kappa$ are observed in diseases such as breast cancer, with a rate of 89% (Pies, 2007).

*Poor compatibility with the logic of mathematics.* The medical context involves special requirements that the logic of mathematics can not completely meet (Cios and Moore, 2002; Moore and Hutchins, 1980). Medicine is a discipline opened to nuances in interpretations, such that the expected responses should not necessarily be definitive. The problem of canonical forms is also an illustration of this partial incompatibility: on the opposite of medicine, mathematics are based on canonical forms through the ability to simplify expressions in reduced forms (Cios and Moore, 2002).

| Mental Disorder | $\kappa$ (%) |
|---|---|
| Autism Spectrum Disorder | 69 |
| Attention Deficit Hyperactivity Disorder (ADHD) | 61 |
| Bipolar I Disorder | 52 |
| Avoidant/Restrictive Food Intake Disorder | 48 |
| Conduct Disorder | 46 |
| Posttraumatic Stress Disorder | 34 |
| Major Depressive Disorder | 28 |
| Mixed Anxiety-Depressive Disorder | 5 |
| Nonsuicidal Self-Injury | -3 |

Table 7: Kappa coefficients for psychiatric disorders (adapted from Freedman et al. 2013)

*Temporal components.* Medical data may include temporal relations, between symptoms for example. Yet the order in which symptoms appear is far from insignificant in the process of diagnosis (Combi et al., 2010). To illustrate this fact, Wagholikar et al. (2012) mentions the example of traumatic brain injury and fainting: the sense of the causal relationship may have a specific impact on the final diagnosis.

*Sensitive property and use issues.* Modalities of data storage, analysis methods and property are the subject of extensive debates in ethics committees to ensure respect for the patients' privacy and the careful use of the medical data to avoid any form of abuse (Yoo et al., 2012; Cios and Moore, 2002).

#### *3.2.2. Characteristics of open medical data*

With the culture of open data sharing, voluminous databases are available for research; they are released to address a large range of diseases (see Esfandiari et al. 2014 and Woo et al. 2017) and collected from various research/hospital centers worldwide. Open medical data are thus characterized by an accentuated heterogeneity. Indeed, once medical databases are related to volunteers assessed in different parts of the world, cultural and social factors, if ignored, may lead to bias in research studies. Yet such factors are most likely involved in both trouble causes and prevalence (Al-Kuraya et al., 2005; Trostle, 2005; Link and Phelan, 1995; Landy, 1977). What is more, according to sites, the nature and calibration of equipment, experimental conditions and recruitment procedures differ, which increases inter-site variability (Abraham et al., 2017).

## 4. Discussion

Our analysis shows that modeling the process of diagnostic decision is an art in its own right. Indeed, it involves to satisfy the clinicians' requirements in this respect (see section 3.1) without

ignoring the particular nature of medical data (see section 3.2) to achieve a final useful, pertinent and consistent modeling. In the present section, we discuss the potential of DM to meet such an objective. This allows us, in a second stage, to raise the limitations and strengths of the advocated research approach.

### *4.1. On the potential of data mining to meet the expressed needs*

One of the major challenges to face with the problem of diagnosis aid is to succeed in targeting model **interpretability**, without loss of performance. On that regard, decision trees are valued for the interpretability integrated within their proper structure; it is in this sense that decision trees, among others, are qualified as *ante-hoc* methods (Holzinger et al., 2017). Besides, under an appropriate tuning of the related parameters, a good level of performance and robustness may be expected. Such benefits make decision trees very pertinent in the medical domain (Podgorelec et al., 2002). However, concerning performance, we can do much more with neural networks and the recent advent of deep learning (Miotto et al., 2017). Of course, opting for such techniques means shifting away from interpretability.

Thus, we must recognize that the most accurate methods of machine learning remain those which are the most sophisticated, and the less interpretable accordingly. It is here legitimate to ask how achieving a fair compromise between performance and interpretability, despite the obvious hurdle of complexity. In response to this questioning, the work of Ribeiro et al. (2016) advocates LIME – Local Interpretable Model-agnostic Explanations – which may be qualified as a *post-hoc* method (Holzinger et al., 2017). Indeed, in a limited neighbor of an instance $x$, the method approximates a classifier (e.g. a neural network), a posteriori, by an interpretable model as concise and faithful as possible. The latter may be a linear model whose weights show, by their intensity, the features that have a meaningful impact on the prediction. In the same spirit, the technique SP-LIME (Submodular Pick Local Interpretable Model-agnostic Explanations) tries to interpret a model in its integrity (Ribeiro et al., 2016). SP-LIME selects some instances from the initial training set in a way that the set of explanations provided by LIME for each selected instance gives a global insight on the model.

A priori, both LIME and SP-LIME techniques are well suited to meet the requirements expressed beforehand for diagnosis aid since they allow to validate predictive models locally and globally, while giving the opportunity to resort to performant models. Incidentally, LIME was applied successfully on medical data to extract explanations from the predictions of a random forest (Katuwal and Chen, 2016). But the method remains still to be explored on technical and computational planes (Ribeiro

et al., 2016). In point of fact, depending on the nature of the intermediate classifier, the notion of model complexity is not obvious to quantify. Moreover, the work of Ribeiro et al. (2016) has not experienced decision trees as intermediate explainers. Yet this may be one avenue worth exploring since decision trees, by their intuitive shape, are very attractive for medical applications (Podgorelec et al., 2002). By contrast, the interest of using SP-LIME is less well understood: after all, it relies on training different interpretable models to explain globally a single classifier. In comparison, training a single interpretable classifier seems equally relevant while consuming less time and resources.

**Performance**, measured notably by the accuracy of the models, is surely a factor influencing the practitioners' confidence on diagnosis aid. However, **robustness** constitutes also a key element to ensure a full reliability. For example, our previous DM-based work (Lecron et al., 2013) proposed a robust methodology for 3D spine reconstruction for the diagnosis and treatment of scoliosis. Actually, to perform a reconstruction, the physician has to manually set landmarks on radiographs, so it must be insured that an imprecise localization of such landmarks has the least impact on the quality of the results. In particular, reference methods appeared to be highly sensitive to the position of the initial landmarks, with a drastic increase in errors as these markers are located far from their appropriate position. On the opposite, our method presents a very slow increase in errors on a large range of deviation. This robustness, acquired on any input of the algorithm, is crucial to get a final reliable reconstruction on which diagnosis is based.

Dealing with the heterogeneity of **open multi-site medical databases** is another challenging aspect of the development of models for diagnosis aid. The recent work of Abraham et al. (2017) proposes two cross-validation methodologies to develop classifiers less sensitive to the heterogeneity of multi-site medical data. To constitute a validation set, *intra-site cross-validation* randomly withdraws instances from each site subset in proportion to their representation in the whole training set, while *inter-site cross-validation* excludes a site completely. Such techniques make classifiers less sensitive to the problem of technical heterogeneities. But let us assume for a moment that these technical disparities do not exist: can we still consider a multi-site database as a homogeneous entity? The answer is likely to be negative. As stated before, socio-cultural factors influence the epidemiology and etiology of troubles (Al-Kuraya et al., 2005; Trostle, 2005; Link and Phelan, 1995; Landy, 1977). Regrouping site subsets, based on similar socio-cultural conditions, may thus lead to more consistent diagnoses. Moreover, troubles may exist in different conditions, e.g. psychological disorders are often qualified as spectrum disorders (Maser and Patterson, 2002). So another question to solve is whether socio-cultural factors influence the existence of the trouble or the way in which

a trouble is expressed. To sum up, as we showed in a recent work (Itani et al., 2018a), research should not necessarily be focused on a universal diagnosis for a given condition.

Researchers are frequently faced with unbalanced datasets in pathological/healthy subjects, or even with datasets including only representatives of a single class. A large number of studies aimed at circumventing this issue, e.g. Majid et al. (2014); Parvin et al. (2013); Li et al. (2010); Cohen et al. (2006), to bring back the learning process to the classical multi-class scheme. But in this respect, a recent option deserves some interest: the process of One-Class Classification (OCC) (Krawczyk et al., 2015; Khan and Madden, 2014; Pimentel et al., 2014; Chandola et al., 2009; Khan and Madden, 2009; Mazhelis, 2006; Hodge and Austin, 2004; Markou and Singh, 2003a,b). As coined by Moya and Hush (1996); Moya et al. (1993), OCC aims to find a description of a target set of examples, and to detect objects that fit into the description. This classification mode only assumes the availability of a single class examples, considered as positive instances. Negative instances may exist, and play the role of counter-examples; they are often scarce. The OCC task actually helps to identify *anomalies*, *outliers* or *novelties* when faced with the study of a particular phenomenon.

In previous works, OCC was considered for the detection of suspicious masses in mammographies (Cichosz et al., 2016; Hoffmann, 2007) and scoliosis, by the classification of Magnetic Resonance Images (MRI) (Karpate et al., 2015). Heart diseases were also addressed by Cabral and de Oliveira (2014). López-de Ipiña et al. (2016) showed that OCC outperforms multi-class classification for the diagnosis of Alzheimer's disease. But OCC is an area of research that still needs to be further developed, particularly as regards medical applications. An important question arises with respect to the performance of OCC models depending on whether they target pathological or control cases. The recent work of Retico et al. (2016), dedicated to the study of autism, shows that the model of pathology rules out well negative instances. The same would not apply as regards models developed based on the control group, which appeared heterogeneous. Concerning interpretability, techniques such as LIME are still to be adapted to OCC.

An **output probability** does not appear indispensable for clinical diagnostic decision models; so crisp classification is not necessarily to question. Indeed, as we exposed in section 3.1, sensitivity and specificity values are sufficient to deduce the probability that a patient has a given disease. The usefulness of a probabilistic output might be rather of a technical order, to tune the balance between sensitivity and specificity, in setting the appropriate threshold beyond which the existence of the trouble may be confirmed. In practical terms, a model with a probabilistic output would be of some interest to clinicians if they accept to drop their traditional Bayesian methodology to trust

the probabilities suggested by predictive models. But to this end, we still need to find a place to devote to the pre-test probability since it is not a proper input of predictive DM models. This could be the case if the data on which the clinicians estimate this pre-test probability play a role in the prediction process, which is not insured as this information (e.g. the medical history) is not always available to the data miner.

Finally, we have to acknowledge the fundamental role of the couple sensitivity-specificity to assess both the performance and the prudent nature of predictive models. The ideal to strive is acquiring both high specificity and sensitivity. Generally, the one is acquired at the expense of the other: there must be therefore an aspiration to achieve a fair compromise between both (see section 3.1.3). Technically, this may involve to play on the balance between pathological and control instances in the initial training set, to tune appropriately modeling parameters, or to develop several predictive models intended to be handled sequentially or simultaneously.

### *4.2. On the limitations and strengths of the proposed research methodology*

The current research in the domain of DM-based diagnosis aid mainly relies on the choice or development of a predictive algorithm in quest of the best level of accuracy. To strengthen the value of the results, comparisons with other techniques or previous works for the same problem are often proposed, e.g. Riaz et al. (2016); Chen et al. (2013); Yeh et al. (2011); Aslandogan and Mahajani (2004); Dreiseitl et al. (2001). The culture of open medical data sharing has intensified this competition spirit, rather than encouraging the development of models having a concrete applicability. Yet as we showed, the latter objective could only be achieved through a more complete research approach, targeting performance among other context-oriented objectives (see Table 3). If this knowledge-guided data mining methodology presents several strengths, it has also limitations.

*Limitations.*

- Under the current state of the art, it is concretely hard to find a DM algorithm which dominates all others on the set of modeling goals defined in Table 3.
- The success of our research methodology is conditioned by non-technical variables, namely :
  (1) the clinicians' availability in time to interact with the data miners on the very first steps to understand the problem and data, but also throughout the whole development of a given diagnostic decision support system ;

(2) the ability for each of both actors to make their domain-specific vocabulary accessible to the other.

*Strengths.*

- Though theoretical at the moment, the methodology shows the way towards the development of predictive diagnosis aid models at the core of systems which are expected to integrate clinicians' expertise, and improve their agreement on a given diagnosis. As such, it is important to consider the advocated approach, in conducting the evolution of the current DM modeling techniques towards the expected requirements.

- As regards its underlying DM goals, the methodology promotes the development of scientific knowledge in the medical domain. Indeed, if DM is widely requested for diagnosis aid, it is above all in the hope of raising some physiological bases, notably about pathologies that still remain unknown.

- Through criteria such as simplicity and intelligibility, the methodology promotes models which may induce technological progress in terms of data acquisition devices. On this regard, the issues raised by the models may be highly interesting, e.g. do we still need to resort to an uncomfortable brain MRI exam while the predictive model of a given mental pathology is based on the activity and/or structure of a limited number of brain zones, i.e. a limited part of the information provided by MRI? Such a result would be well worth to explore in order to develop less energy-consuming, less expensive and why not, more user-friendly equipment.

## 5. Conclusion and future work

An expert and intelligent system has to be designed based on the context of the application to which it is devoted. It is from this perspective that we proposed the present survey, intended for researchers specialized in Data Mining (DM) and working on predictive models at the core of diagnostic decision support systems. It seems that, far from being trivial, the implementation of models for diagnosis aid needs to rely on a specific approach of data modeling and processing, based on the clinicians' requirements. In many respects, such a method differs from the common practice which promotes continuous improvements of the predictive performances, without being enough challenged by the issue of knowledge extraction required for a better detection and understanding of diseases.

The main conclusion of our paper is the need for the permanent medical expert's presence throughout the development of diagnostic decision support systems. It is in this sense that we advocate an *expert-aware approach*. Indeed, at the early stages of such a project, the clinician (i.e. the medical expert) helps the data miner (i.e. the DM expert) to understand the problem and the related data; in contrast, the data miner allows the medical expert to perceive what DM may realistically provide to build an expert system. On this regard, the survey proposes a list of requirements translated into DM goals. This list is naturally not fixed; it may depend on the pathology which is subject to diagnosis, that is why interaction is so important. During the development of the predictive model, the continuous interaction between both domain experts leads to the engineering and the selection of a set of attributes which are simultaneously efficient, pertinent and meaningful. Afterwards, the interaction allows to guide the selection of an appropriate model, in terms of performance and interpretability, among other context-oriented goals that we exposed in the survey. At the last stages, the medical expert has an important role to play in the implementation of the expert and intelligent system, in order to take into consideration other practical constraints such as intuitive ergonomics and timeliness.

Thus, with this survey, we shed light on the importance of the interaction between both data miners and clinicians for the development of meaningful models for diagnosis aid. At no moment one of both actors should prevail over the other: a kind of balance has to be ensured, at each stage, to practice a form of data mining driven by knowledge, while leaving some place for discovery and innovation. Within this consideration, we advocate an approach which supposes that the data miner does not work for the clinicians, but closely with them.

Compared to previous papers, our survey makes new contributions as regards the issue of DM-based diagnosis aid, in presenting a complete report of (1) the related modeling aspects, (2) the constraints related to medical data processing. Actually, we proposed an overview of the specifics of medical data, in comparing several surveys of the literature (Miotto et al., 2017; Esfandiari et al., 2014; Yoo et al., 2012; Harrison, 2008; Bellazzi and Zupan, 2008; Cios and Moore, 2002; Lavrač, 1999) and revised the notion of data heterogeneity, given notably the current context of open medical data sharing. Finally, we developed a critical perspective on the current modeling techniques as regards their ability to meet the formulated modeling requirements, in thus complementing previous papers that widely surveyed DM methods (Miotto et al., 2017; Esfandiari et al., 2014; Wagholikar et al., 2012; Yoo et al., 2012).

This paper naturally opens the door to promising research perspectives. As a matter of fact, the

advocated DM research approach raises new issues to reach simultaneously the proposed modeling goals.

*Hybridization.* – In the current state of knowledge, it is practically excluded to find or to develop a single algorithm which may process patient's data to predict a diagnosis through a model reaching the proposed ambitious goals, i.e. interpretability, performance, robustness, caution, and nuanced outputs. There is rather a need for a hybrid strategy, in which both tasks of feature engineering and model building are likely to be considered separately. In such a hybrid process, it remains essential that the predictive model presents the qualities mentioned above in order to justify its results. In contrast, the inputs of the latter model may be derived from the initial training set through a mechanism which does not necessarily need to be justified, provided that it leads to features making sense for the clinicians (as required by *meaningfulness*). On this regard, feature learning may present specific architectures, like those derived from deep learning, e.g. autoencoders and convolutional neural networks (Ching et al., 2018; Ravı et al., 2017). But these algorithms require important quantity of data, which is not necessarily available in the context of medical applications: this constitutes an important obstacle to overcome. Another challenge for data miners is to succeed in reaching high levels of performance through hybrid strategies, i.e. while using the features derived from deep architectures as the inputs of predictive models of another nature, ideally white boxes. Indeed, if such a hybridization recently provided interesting results for diseases such as breast cancer (Selvathi and AarthyPoornila, 2017), it resulted in less relevant levels of accuracy as regards mental pathologies such as ADHD and autism (Sen et al., 2018).

*Iterative training process.* – Methods such as LIME (Ribeiro et al., 2016) are grafted onto classifiers, typically black boxes, in order to interpret predictions. Though giving the advantage of combining performance and interpretability, such components extend the decision chain so it becomes legitimate to question the related robustness. Another drawback of these methods is the lack of global vision of the model behavior: it is surely possible to interpret a prediction related to a given instance, but it is conversely harder to raise the different combinations of factors to explain a given prediction. That is why the design of models such as decision trees or GAMs (Generalized Additive Models), seems more suitable as they present an interpretable structure by nature. However, these methods need still to be improved to reach higher levels of accuracy. To achieve this, it can be interesting to consider an iterative training process, in which the model is refined thanks to the intervention of a clinician who may interpret it to adjust the training set content, e.g. in selecting features belonging

to a coherent structural or functional ensemble according to the medical knowledge.

*Profile-based features.* – It would be interesting to predict a diagnosis through the way in which a patient is related to one or several reference profile(s). A classical approach would rely on principal component analysis, which is a well-known method, but whose results are less easy to interpret in terms of reference profiles. For such a purpose, we may rather consider Non-negative Matrix Factorization (NMF) which proceeds to interpretable dimensionality reductions in a wide range of applications. Indeed, NMF consists of a decomposition $X \approx WH$ with $X \geq 0, W \geq 0, H \geq 0$ (Gillis, 2014). If $n$ and $r$ denote respectively the number of columns in $X$ and $W$ ($r \ll n$), NMF is actually achieved by optimizing an error measure, for a given factorization rank $r$. The columns of $W$ are seen as basis elements whose positive linear combination may reconstitute any instance of the original dataset. These basis elements are interpreted according to the context in which they are raised. For example, in text mining, the columns of $W$ are interpreted as topics, i.e. a set of words. Given that any instance, i.e. a document in this context, is expressed as a non-negative linear combination of these topics, the corresponding weights show to which extent the document is involved in each topic. The work of Anderson et al. (2014) takes inspiration from this idea for diagnosis prediction of ADHD, considering a topic as a set of elements about a patient (e.g. phenotype, neuroimaging features); a decision tree was then trained on the resulting NMF weights. Actually, we think the interpretability of the NMF in this context may be improved in decomposing data of the same nature, thus considering that the columns of $W$ correspond rather to homogeneous reference patients' profiles. Besides, some questions still remain to investigate, in particular :

(1) the choice of values for parameter $r$;
(2) the influence of certain variables on the decomposition, e.g. the patients' gender as well as the impact of a (un)supervised decomposition;
(3) the way in which the constraint of having positive data in $X$ may be circumvented, notably by means of an alternative matrix representation for both positive and negative data.

The detection of reference profiles may be integrated to a training process, by the detection of central profiles around which some intervals define the extent of a given healthy or pathological condition. One of our recent works implements such an idea, in proposing a decision tree for one-class classification (Itani et al., 2018b).

*Coalition-based prediction.* – Given some factors such as the health care policy in a country, the local clinical practices and the equipment availability, patients do not follow the same assessment stages

for a diagnosis. This results in the availability of different sources of data from patient to patient, for the diagnosis of a given pathology. Data mining should cope with this situation when it comes to develop diagnosis aid models. One way to handle this issue is to develop predictive models which suit their local contexts. But, one challenging research opportunity is to aspire to more universal solutions. Decision making models, like MR-Sort (Leroy et al., 2011), constitute a good source of inspiration on this regard: they rely on a classification of instances based on sufficient subsets of the criteria (called coalitions). Currently, MR-Sort addresses quite small problems (up to 10 criteria and 5 classes) (Sobrie et al., 2013). Thus, transposing the related idea to machine learning approaches is certainly an interesting research work. In the case of decision trees, this would involve to question the availability of a given information for a patient, and guide the following questioning in function of this availability. This would necessarily involve some changes in the underlying greedy algorithm, by including a degree of dependence between certain levels. Indeed, if the result of a specific exam is available, the next question should be based on the results of this exam. Overall, depending on each decision chain of the resulting tree, the required information would not be the same. Of course, having such a universal model does not exempt from revising some of its aspects depending on the country and the population, i.e. mainly the thresholds on which the subdivisions are based and that allow finally to define the borders between healthy and pathological populations. This seems obvious, but quite challenging to achieve technically.

Therefore, in spite of the progress made in data science, more work is required, especially to address the issue of aided diagnosis. On this regard, it is through a DM research focused on specific needs that we can gain both the clinicians' trust and help towards predictive models – and more generally systems – having a practical applicability in their everyday working life.

## 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 and Thierry Pham Hoang (Faculty of Psychology and Education, University of Mons, Belgium) for their advice and interest in this work. Finally, the authors would like to thank the anonymous reviewers for their insightful comments and suggestions to improve the paper.

## 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.

Adam M. Chekroud, 2016. nomogrammer: Fagans nomogram using ggplot2. https://achekroud.github.io/nomogrammer_vignette.html. [Online; accessed 19-11-2017].

Akobeng, A.K., 2007a. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta paediatrica 96, 338–341.

Akobeng, A.K., 2007b. Understanding diagnostic tests 2: likelihood ratios, pre-and post-test probabilities and their use in clinical practice. Acta paediatrica 96, 487–491.

Al-Kuraya, K., Schraml, P., Sheikh, S., Amr, S., Torhorst, J., Tapia, C., Novotny, H., Spichtin, H., Maurer, R., Mirlacher, M., et al., 2005. Predominance of high-grade pathway in breast cancer development of Middle East women. Modern Pathology 18, 891–897.

Aličković, E., Subasi, A., 2017. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications 28, 753–763.

Allison, K.H., Sledge, G.W., 2014. Heterogeneity and cancer. Oncology 28, 772–772.

Andargoli, A.E., Scheepers, H., Rajendran, D., Sohal, A., 2017. Health information systems evaluation frameworks: A systematic review. International journal of medical informatics 97, 195–209.

Anderson, A., Douglas, P.K., Kerr, W.T., Haynes, V.S., Yuille, A.L., Xie, J., Wu, Y.N., Brown, J.A., Cohen, M.S., 2014. Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD. NeuroImage 102, 207–219.

Anglade, A., Ramirez, R., Dixon, S., et al., 2009. Genre classification using harmony rules induced from automatic chord transcriptions, in: 10th International Society for Music Information Retrieval Conference (ISMIR), pp. 669–674.

Aslandogan, Y.A., Mahajani, G.A., 2004. Evidence combination in medical data mining, in: ITCC 2004 - International Conference on Information Technology: Coding and Computing, pp. 465–469.

Baesens, B., Mues, C., Martens, D., Vanthienen, J., 2009. 50 years of data mining and OR: upcoming trends and challenges. Journal of the Operational Research Society 60, S16–S23.

Bain, S.K., Allin, J.D., 2005. Book review: Stanford-binet intelligence scales. Journal of Psychoeducational Assessment 23, 87–95.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.

Bates, M.J., 2011. Understanding information retrieval systems: management, types, and standards. Auerbach Publications.

Bellazzi, R., Ferrazzi, F., Sacchi, L., 2011. Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1, 416–430.

Bellazzi, R., Zupan, B., 2008. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics 77, 81–97.

Bock, H.H., Lenski, W., Richter, M.M., 2013. Information Systems and Data Analysis: Prospects – Foundations – Applications. Springer Science & Business Media.

Cabral, G.G., de Oliveira, A.L.I., 2014. One-class classification for heart disease diagnosis, in: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. pp. 2551–2556.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N., 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 1721–1730.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR) 41, 15:1–15:58.

Chen, H.L., Huang, C.C., Yu, X.G., Xu, X., Sun, X., Wang, G., Wang, S.J., 2013. An efficient diagnosis system for detection of Parkinsons disease using fuzzy k-nearest neighbor approach. Expert systems with applications 40, 263–271.

Chen, L.Y.A., Fawcett, T.N., 2016. Using data mining strategies in clinical decision making: A literature review. CIN: Computers, Informatics, Nursing 34, 448–454.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al., 2018. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface 15, 20170387.

Chismar, W.G., Wiley-Patton, S., 2002. Test of the technology acceptance model for the internet in pediatrics, in: AMIA Symposium, American Medical Informatics Association. pp. 155–159.

Cichosz, P., Jagodziński, D., Matysiewicz, M., Neumann, Ł., Nowak, R.M., Okuniewski, R., Oleszkiewicz, W., 2016. Novelty detection for breast cancer image classification, in: Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016, International Society for Optics and Photonics. pp. 1003135–1003135.

Cios, K.J., Moore, G.W., 2002. Uniqueness of medical data mining. Artificial intelligence in medicine 26, 1–24.

Cios, K.J., Pedrycz, W., Swiniarski, R.W., Kurgan, L., 2007. Data Mining, A Knowledge Discovery Approach. Springer US.

Cios, K.J., Teresinska, A., Konieczna, S., Potocka, J., Sharma, S., 2000. A knowledge discovery approach to diagnosing myocardial perfusion. IEEE Engineering in Medicine and Biology Magazine 19, 17–25.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Geissbuhler, A., 2006. Learning from imbalanced data in surveillance of nosocomial infection. Artificial intelligence in medicine 37, 7–18.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20, 37–46.

Combi, C., Keravnou-Papailiou, E., Shahar, Y., 2010. Temporal information systems in medicine. Springer Science & Business Media.

Corney, D., 2002. Food bytes: intelligent systems in the food industry. British Food Journal 104, 787–805.

Davis, F.D., Bagozzi, R.P., Warshaw, P.R., 1989. User acceptance of computer technology: a comparison of two theoretical models. Management science 35, 982–1003.

Deeks, J.J., Altman, D.G., 2004. Diagnostic tests 4: likelihood ratios. Bmj 329, 168–169.

Di Martino, A., OConnor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., et al., 2017. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. Scientific Data 4, 170010.

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 large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19, 659.

Domingos, P., 1999. The role of occam's razor in knowledge discovery. Data mining and knowledge discovery 3, 409–425.

Donner-Banzhoff, N., Seidel, J., Sikeler, A.M., Bösner, S., Vogelmeier, M., Westram, A., Feufel, M., Gaissmaier, W., Wegwarth, O., Gigerenzer, G., 2017. The phenomenology of the diagnostic process: A primary care–based survey. Medical Decision Making 37, 27–34.

Doshi-Velez, F., Kim, B., 2017. A roadmap for a rigorous science of interpretability, in: arXiv preprint arXiv:1702.08608.

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., Binder, M., 2001. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics 34, 28–36.

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.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39, 27–34.

Freedman, R., Lewis, D.A., Michels, R., Pine, D.S., Schultz, S.K., Tamminga, C.A., Gabbard, G.O., Gau, S.S.F., Javitt, D.C., Oquendo, M.A., et al., 2013. The initial field trials of dsm-5: New blooms and old thorns. American Journal of Psychiatry 170, 1–5.

Gillis, N., 2014. The why and how of nonnegative matrix factorization. Regularization, Optimization, Kernels, and Support Vector Machines 12.

Giudici, P., 2005. Applied data mining: statistical methods for business and industry. John Wiley & Sons.

Gordis, L., 2014. Epidemiology (Fifth edition.). Elsevier Saunders.

Graber, M.L., 2013. The incidence of diagnostic error in medicine. BMJ Quality & Safety 22, ii21–ii27.

Guillaume, S., 2001. Designing fuzzy inference systems from data: An interpretability-oriented review. IEEE Transactions on fuzzy systems 9, 426–443.

Gupta, G., 2014. Introduction to data mining with case studies. PHI Learning Pvt. Ltd.

Gupta, S., Kumar, D., Sharma, A., 2011. Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian Journal of Computer Science and Engineering (IJCSE) 2, 188–195.

Hajiloo, M., Sapkota, Y., Mackey, J.R., Robson, P., Greiner, R., Damaraju, S., 2013. ETHNO-PRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. BMC bioinformatics 14, 61.

Harrison, J.H., 2008. Introduction to the mining of clinical data. Clinics in laboratory medicine 28, 1–7.

Hodge, V.J., Austin, J., 2004. A survey of outlier detection methodologies. Artificial intelligence review 22, 85–126.

Hoffmann, H., 2007. Kernel PCA for novelty detection. Pattern Recognition 40, 863–874.

Holden, R.J., Karsh, B.T., 2010. The technology acceptance model: its past and its future in health care. Journal of biomedical informatics 43, 159–172.

Holte, R.C., 1993. Very simple classification rules perform well on most commonly used datasets. Machine learning 11, 63–90.

Holzinger, A., Biemann, C., Pattichis, C., Kell, D., 2017. What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 .

Hommersom, A., Lucas, P.J., 2016. Foundations of Biomedical Knowledge Representation: Methods and Applications. volume 9521. Springer.

Iavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R., Geissbuhler, A., et al., 2009. Clinical data mining: a review. Yearb Med Inform 2009, 121–133.

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.

López-de Ipiña, K., Faundez-Zanuy, M., Solé-Casals, J., Zelarin, F., Calvo, P., 2016. Multi-class versus one-class classifier in spontaneous speech analysis oriented to alzheimer disease diagnosis, in: Recent Advances in Nonlinear Speech Processing. Springer, pp. 63–72.

Itani, S., Lecron, F., Fortemps, P., 2018a. A multi-level classification framework for multi-site medical data: Application to the ADHD-200 collection. Expert Systems with Applications 91, 36 – 45.

Itani, S., Lecron, F., Fortemps, P., 2018b. A one-class decision tree based on kernel density estimation. arXiv preprint arXiv:1805.05021 .

Kalantari, A., Kamsin, A., Shamshirband, S., Gani, A., Alinejad-Rokny, H., Chronopoulos, A.T., 2018. Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. Neurocomputing 276, 2–22.

Karpate, Y., Commowick, O., Barillot, C., 2015. Probabilistic one class learning for automatic detection of multiple sclerosis lesions, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 486–489.

Karpatne, A., Atluri, G., Faghmous, J., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering 29, 2318–2331.

Katuwal, G.J., Chen, R., 2016. Machine learning model interpretability for precision medicine. arXiv preprint arXiv:1610.09045 .

Kaufman, A.S., Raiford, S.E., Coalson, D.L., 2015. Intelligent testing with the WISC-V. John Wiley & Sons.

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.

Ketikidis, P., Dimitrovski, T., Lazuras, L., Bath, P.A., 2012. Acceptance of health information technology in health professionals: An application of the revised technology acceptance model. Health informatics journal 18, 124–134.

Khan, S.S., Madden, M.G., 2009. A survey of recent trends in one class classification, in: Artificial Intelligence and Cognitive Science: 20th Irish Conference (AICS 2009), Springer. pp. 188–197.

Khan, S.S., Madden, M.G., 2014. One-class classification: taxonomy of study and review of techniques. The Knowledge Engineering Review 29, 345–374.

Klösgen, W., Zytkow, J.M. (Eds.), 2002. Handbook of Data Mining and Knowledge Discovery. Oxford University Press, Inc., New York, NY, USA.

Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine 23, 89–109.

Kosmicki, J., Sochat, V., Duda, M., Wall, D., 2015. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Translational psychiatry 5, e514.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal 13, 8–17.

Krawczyk, B., Woźniak, M., Herrera, F., 2015. On the usefulness of one-class classifier ensembles for decomposition of multi-class problems. Pattern Recognition 48, 3969–3982.

Krishnaiah, V., Narsimha, D.G., Chandra, D.N.S., 2013. Diagnosis of lung cancer prediction system using data mining classification techniques. International Journal of Computer Science and Information Technologies 4, 39–45.

Kukenys, I., Browne, W., Zhang, M., 2011. Transparent, online image pattern classification using a learning classifier system. Applications of Evolutionary Computation , 183–193.

Kumar, R., Indrayan, A., 2011. Receiver operating characteristic (ROC) curve for medical researchers. Indian pediatrics 48, 277–287.

Kurgan, L.A., Musilek, P., 2006. A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review 21, 1–24.

Lakkaraju, H., Bach, S.H., Leskovec, J., 2016. Interpretable decision sets: A joint framework for description and prediction, in: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 1675–1684.

Lalkhen, A.G., McCluskey, A., 2008. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain 8, 221–223.

Landy, D., 1977. Culture, disease and healing: Studies in medical anthropology.

Lasko, T.A., Bhagwat, J.G., Zou, K.H., Ohno-Machado, L., 2005. The use of receiver operating characteristic curves in biomedical informatics. Journal of biomedical informatics 38, 404–415.

Lavrač, N., 1999. Selected techniques for data mining in medicine. Artificial intelligence in medicine 16, 3–23.

Lecron, F., Boisvert, J., Mahmoudi, S., Labelle, H., Benjelloun, M., 2013. Three-dimensional spine model reconstruction using one-class svm regularization. IEEE Transactions on Biomedical Engineering 60, 3256–3264.

Leroy, A., Mousseau, V., Pirlot, M., 2011. Learning the parameters of a multiple criteria sorting method, in: International Conference on Algorithmic DecisionTheory, Springer. pp. 219–233.

Li, D.C., Liu, C.W., Hu, S.C., 2010. A learning method for the class imbalance problem with medical data sets. Computers in biology and medicine 40, 509–518.

Link, B.G., Phelan, J., 1995. Social conditions as fundamental causes of disease. Journal of health and social behavior , 80–94.

Lipton, Z.C., 2016. The mythos of model interpretability. arXiv preprint arXiv:1606.03490 .

Lou, Y., Caruana, R., Gehrke, J., 2012. Intelligible models for classification and regression, in: 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 150–158.

Lou, Y., Caruana, R., Gehrke, J., Hooker, G., 2013. Accurate intelligible models with pairwise interactions, in: 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 623–631.

Lyu, B., Haque, A., 2018. Deep learning based tumor type classification using gene expression data. bioRxiv , 364323.

Majid, A., Ali, S., Iqbal, M., Kausar, N., 2014. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. Computer methods and programs in biomedicine 113, 792–808.

Markou, M., Singh, S., 2003a. Novelty detection: a review - part 1: statistical approaches. Signal processing 83, 2481–2497.

Markou, M., Singh, S., 2003b. Novelty detection: a review - part 2: neural network based approaches. Signal processing 83, 2499–2521.

Maser, J.D., Patterson, T., 2002. Spectrum and nosology: implications for DSM-V. Psychiatric Clinics of North America 25, 855–885.

Maxim, L.D., Niebo, R., Utell, M.J., 2014. Screening tests: a review with examples. Inhalation toxicology 26, 811–828.

Mazhelis, O., 2006. One-class classifiers: a review and analysis of suitability in the context of mobile-masquerader detection. South African Computer Journal 2006, 29–48.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia medica 22, 276–282.

Mencar, C., Castellano, G., Fanelli, A.M., 2007. On the role of interpretability in fuzzy data mining. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15, 521–537.

Mennes, M., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2013. Making data sharing work: the FCP/INDI experience. Neuroimage 82, 683–691.

Menzies, T., Hu, Y., 2006. Just enough learning (of association rules): the tar2 treatment learner. Artificial Intelligence Review 25, 211–229.

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.

Miller, R.A., 2016. Diagnostic decision support systems, in: Clinical decision support systems. Springer, pp. 181–208.

Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T., 2017. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics , bbx044.

Moore, G.W., Hutchins, G.M., 1980. Effort and demand logic in medical decision making. Metamedicine 1, 277–303.

Moya, M.M., Hush, D.R., 1996. Network constraints and multi-objective optimization for one-class classification. Neural Networks 9, 463–474.

Moya, M.M., Koch, M.W., Hostetler, L.D., 1993. One-class classifier networks for target recognition applications. Technical Report. Sandia National Labs., Albuquerque, NM (United States).

Mun, Y.Y., Jackson, J.D., Park, J.S., Probst, J.C., 2006. Understanding information technology acceptance by individual professionals: Toward an integrative view. Information & Management 43, 350–363.

Nahar, J., Ali, A.S., Imam, T., Tickle, K., Chen, P., 2016. Brain cancer diagnosis-association rule based computational intelligence approach, in: 2016 IEEE International Conference on Computer and Information Technology (CIT), IEEE. pp. 89–95.

Odenbaugh, J., 2011. True lies: realism, robustness, and models. Philosophy of Science 78, 1177–1188.

Otte, C., 2013. Safe and interpretable machine learning: a methodological review, in: Computational Intelligence in Intelligent Data Analysis. Springer, pp. 111–122.

Pandey, B., Mishra, R., 2009. Knowledge and intelligent computing system in medicine. Computers in biology and medicine 39, 215–230.

Papakostas, G.A., Savio, A., Graña, M., Kaburlasos, V.G., 2015. A lattice computing approach to Alzheimers disease computer assisted diagnosis based on MRI data. Neurocomputing 150, 37–42.

Parvin, H., Minaei-Bidgoli, B., Alinejad-Rokny, H., 2013. A new imbalanced learning and dictions tree method for breast cancer diagnosis. Journal of Bionanoscience 7, 673–678.

Patel, S., Patel, H., 2016. Survey of data mining techniques used in healthcare domain. International Journal of Information Sciences and Techniques 6, 53–60.

PhridviRaj, M., GuruRao, C., 2014. Data mining–past, present and future–a typical survey on data streams. Procedia Technology 12, 255–263.

Pies, R., 2007. How objective are psychiatric diagnoses?:(guess again). Psychiatry (Edgmont) 4, 18–22.

Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L., 2014. A review of novelty detection. Signal Processing 99, 215–249.

Podgorelec, V., Kokol, P., Stiglic, B., Rozman, I., 2002. Decision trees: an overview and their use in medicine. Journal of medical systems 26, 445–463.

Poline, J.B., Breeze, J.L., Ghosh, S., Gorgolewski, K., Halchenko, Y.O., Hanke, M., Haselgrove, C., Helmer, K.G., Keator, D.B., Marcus, D.S., et al., 2012. Data sharing in neuroimaging research. Frontiers in neuroinformatics 6.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Development Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Raven, J., 2000. The Raven's progressive matrices: change and stability over culture and time. Cognitive psychology 41, 1–48.

Ravı, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.Z., 2017. Deep learning for health informatics. IEEE journal of biomedical and health informatics 21, 4–21.

Reimer, P., Parizel, P.M., Meaney, J.F., Stichnoth, F.A., 2010. Clinical MR imaging. Springer.

Retico, A., Gori, I., Giuliano, A., Muratori, F., Calderoni, S., 2016. One-class support vector machines identify the language and default mode regions as common patterns of structural alterations in young children with autism spectrum disorders. Frontiers in neuroscience 10.

Revelle, W., Rocklin, T., 1979. Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behavioral Research 14, 403–414.

Riaz, A., Alonso, E., Slabaugh, G., 2016. Phenotypic integrated framework for classification of ADHD using fMRI, in: 13th International Conference on Image Analysis and Recognition (ICIAR), Springer. pp. 217–225.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you? Explaining the predictions of any classifier, in: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 1135–1144.

Scopus, . Scopus - Document search. https://www.scopus.com/. [Online; accessed 31-07-2017].

Selvathi, D., AarthyPoornila, A., 2017. Performance analysis of various classifiers on deep learning network for breast cancer detection, in: Signal Processing and Communication (ICSPC), 2017 International Conference on, IEEE. pp. 359–363.

Sen, B., Borle, N.C., Greiner, R., Brown, M.R., 2018. A general prediction model for the detection of ADHD and Autism using structural and functional MRI. PloS one 13, e0194856.

Shaffer, V.A., Probst, C.A., Merkle, E.C., Arkes, H.R., Medow, M.A., 2013. Why do patients derogate physicians who use a computer-based diagnostic support system? Medical Decision Making 33, 108–118.

Sobrie, O., Mousseau, V., Pirlot, M., 2013. Learning the parameters of a multiple criteria sorting method from large sets of assignment examples, in: 77th meeting of the EWG on MCDA, Rouen, France.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45, 427–437.

Trostle, J.A., 2005. Epidemiology and culture. volume 13. Cambridge University Press.

Venkatesh, V., Davis, F.D., 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management science 46, 186–204.

Vieira, S., Pinaya, W.H., Mechelli, A., 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neuroscience & Biobehavioral Reviews 74, 58 – 75.

Viera, A.J., Garrett, J.M., et al., 2005. Understanding interobserver agreement: the kappa statistic. Fam Med 37, 360–363.

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.

Wechsler, D., 2003. WISC-IV Wechsler Intelligence Scale for Children: Technical and Interpretative: Manual. Pearson.

Wechsler, D., 2014. Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV).

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2005. Data Mining: Practical machine learning tools and techniques.

Woo, C.W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017. Building better biomarkers: brain models in translational neuroimaging. Nature neuroscience 20, 365–377.

Yarbrough, A.K., Smith, T.B., 2007. Technology acceptance among physicians: a new take on TAM. Medical Care Research and Review 64, 650–672.

Yardimci, A., 2009. Soft computing in medicine. Applied Soft Computing 9, 1029–1043.

Yeh, D.Y., Cheng, C.H., Chen, Y.W., 2011. A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications 38, 8970–8977.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F., Hua, L., 2012. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems 36, 2431–2448.