Transparency of Classification Systems for Clinical Decision Support

Antoine Richard, Brice Mayag, François Talbot, Alexis Tsoukiàs, Yves Meinard

Transparency of classification systems for clinical decision support

Antoine Richard[1,3][0000−0001−8677−8910], Brice Mayag[1], François Talbot[2], Alexis Tsoukias[1][0000−0001−5772−3988], and Yves Meinard[1][0000−0001−5928−8959]

[1] Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, UMR 7243, 75016 Paris, France
[2] DSII Bron, Hospices Civils de Lyon, 61 Boulevard Pinel, 69672 Bron, France
[3] GIE Hopsis Lyon, 3 Quai de Célestins, 69229 Lyon, France
antoine.richard0@dauphine.eu

Abstract. In collaboration with the Civil Hospitals of Lyon, we aim to develop a "transparent" classification system for medical purposes. To do so, we need clear definitions and operational criteria to determine what is a "transparent" classification system in our context. However, the term "transparency" is often left undefined in the literature, and there is a lack of operational criteria allowing to check whether a given algorithm deserves to be called "transparent" or not. Therefore, in this paper, we propose a definition of "transparency" for classification systems in medical contexts. We also propose several operational criteria to evaluate whether a classification system can be considered "transparent". We apply these operational criteria to evaluate the "transparency" of several well-known classification systems.

Keywords: Explainable AI · Transparency of Algorithms · Health Information Systems · Multi-label Classification

1 Introduction

In collaboration with the Civil Hospitals of Lyon (HCL), in France, we aimed to develop and to propose decision support systems corresponding to the clinicians' needs. In 2018, the HCL received more than one million patients for medical consultations. Therefore, the decision has been made to build a decision support system focused on supporting physicians during their medical consultations. After some observations and analyses of medical consultations in the endocrinology department of the HCL [31], we drew two conclusions: physicians mainly need data on patients to reach diagnoses, and getting these data from their information system is quite time-consuming for physicians during consultations. To reduce physicians' workload, we decided to support them by using a classification system learning which data on patients physicians need in which circumstance. By doing this, we should be able to anticipate and provide the data that physicians will need at the beginning of their future consultations.

This can be formalized as a multi-label classification problem, as presented in Table 1 with fictitious data.

In this paper, a "classification system" refers to the combination of a "learning algorithm" and the "type of classifier" produced by this learning algorithm. For example, a classification system based on decision trees can use a learning system such as C4.5 [30], the type of classifier produced by this learning system being a decision-tree. This distinction is necessary because a learning algorithm and a classifier produced by this learning algorithm are not used in the same way and do not perform the same functions.

X: data known on patient				Y: data on patient needed by physician					
Sex	Age	BMI	Disease	HbA1c	Blood Sugar	HDL	LDL	Creatinine	Microalbumin
♀	42	34.23	DT2	1	1	0	0	0	0
♂	52	27.15	HChol	0	0	1	1	0	0
♂	24	21.12	DT1	1	1	0	0	1	1
♀	67	26.22	HChol	0	0	1	1	0	0

Table 1. Example of multi-label dataset based on our practical case

However, in the case of clinical decision support systems (CDSSs), a well-known problem is the lack of acceptability of support systems by clinicians [5, 19]. More than being performant, a CDSS has first to be accepted by clinicians, and "transparent" support systems are arguably more accepted by clinicians [22, 33]. Mainly because "transparency" allows clinicians to better understand the proposals of CDSSs and minimize the risk of misinterpretation. Following these results, we posit that the "transparency" of support systems is a way to improve the "acceptability" of CDSSs by clinicians.

In the literature, one can find several definition of the concept of "transparency": *"giving explanations of results"* [9, 10, 15, 20, 26, 28, 33, 36], *"having a reasoning process comprehensible and interpretable by users"* [1, 11, 12, 24, 27, 34], *"being able to trace-back all data used in the process"* [2–4, 16, 40], but also *"being able to take into account feedbacks of users"* [7, 40]. Individually, each of the above definitions highlights an aspect of the concept of "transparency" of classification systems, but do not capture all aspects of "transparent" classification systems in our context. In addition, definitions are abstract descriptions of concepts and there is a lack of operational criteria, in the sense of concrete properties one can verify in practice, to determine whether a given algorithm deserves to be called "transparent" or not.

The main objective of this paper is to propose a definition of transparency, and a set of operational criteria, applicable to classification systems in a medical context. These operational criteria should allow us to determine which classification system is "transparent" for users in our use case. Let us specify that, in this paper, the term "users" refers to physicians.

In section 2 we detail the definition and operational criteria we propose to evaluate the transparency of classification systems. In section 3, based to our definition of transparency, we explain why we choose a version of the naive bayes algorithm to handle our practical case. We briefly conclude in section 4, with a discussion on the use of an evaluation of "transparency" for practical use cases.

2 Definition of a "transparent" classification system

Even though the concept of algorithm "transparency" is as old as recommendation systems, the emergence and the ubiquity of "black-box" learning algorithms nowadays, such as neural networks, put "transparency" of algorithms back in the limelight [14]. As detailed in section 1, numerous definitions have been given to the concept of "transparency" of classification systems, and there is a lack of operational criteria to determine whether a given algorithm deserves to be called "transparent" or not.

In this paper, we propose the definition below, based on definitions of "transparency" in the literature. Let us recall that our aim here is to propose a definition, and operational criteria, of what we called a "transparent" classification system in a medical context with a user-centered point-of-view.

Definition 1. *A classification system is considered to be "transparent" if, and only if:*

- *the classification system is* ***understandable***
- *the type of classifier and learning system used are* ***interpretable***
- *results produced are* ***traceable***
- *classifiers used are* ***revisable***

2.1 Understandability of the classification system

Although transparency is often defined as *"giving explanations of results"*, several authors have highlighted that these explanations must be "understandable", or "comprehensible", by users [12, 26, 33]. As proposed by Montavon [28], the fact that something is "understandable" by users can be defined as its belonging to a domain that human beings can make sense of.

However, we need an operational criterion to be sure that users can make sense of what we will provide them. In our case, users being physicians, we can consider that users can make sense of anything they have studied during their medical training. Therefore, we define as "understandable" anything based on notions/concepts included in the school curriculum of all potential users. Based on this operational criterion, we propose the definition below of what we call an "understandable" classification systems.

Definition 2. *A classification system is considered to be understandable by users if, and only if, each of its aspects is based on notions/concepts included in the school curriculum of all potential users.*

Let us consider a classification system based on a set C of notions/concepts, and a set S of notions/concepts included in the school curriculum of all potential users, such than $S \cap C$ can be empty. Defined like this, the "understandability" of a classification system is a continuum extending from $S \cap C = \emptyset$ to $S \cap C = C$.

2.2 Interpretability of classifiers and learning system

According to Spagnolli [34], the aim of being "transparent" is to ensure that users are in a position to make informed decisions, without bias, based on the results of the system. A classification system only "understandable" does not prevent misinterpretations of its results or misinformed decisions by users. Therefore, to be considered "transparent" a classification system must also be "interpretable" by users. The criterion of "interpretability" is even more important when applied to sensitive issues like those involved in medical matters. But what could be operational criteria to establish whether a classification system is "interpretable" or not by users?

Let us look at the standard example of a classification system dedicated to picture classification [17]. In practice, the user will use the classifier produced by the learning algorithm and not directly the learning algorithm. Therefore, if the user gives a picture of an animal to the classifier and the classifier says "it's a human", then the user can legitimately ask "Why did you give me this result?" [33]. Here, we have two possibilities: the classifier provides a good classification and the user wants to better understand the reasons underlying this classification, or the classifier provides a wrong classification and the user wants to understand why the classifier didn't provide the right classification.

In the first case, the user can expect "understandable" explanations on the reasoning process that conducted to a specific result. Depending on the classifier used, explanations can take different forms such as "because it has clothes, hair and no claws" or "because the picture is similar to these others pictures of humans". In addition, to prevent misinterpretations, the user can also legitimately wonder "To what extent can I trust this classification?" and expect the classifier to give the risk of error of this result.

In the second case, the user needs to have access to an understandable version of the general process of the classifier and not only the reasoning process that conducts to the classification. This allows the user to understand under which conditions the classifier can produce wrong classifications. In addition, the user can legitimately wonder "To what extent can I trust this classifier in general?". To answer this question, the classifier must be able to provide general performances rates such as its error rate, its precision, its sensitivity and its specificity.

Based on all the above aspects, we are now able to propose the following definition of the "interpretability" of the type of classifier used in the classification system.

Definition 3. *A type of classifier is considered to be "interpretable" by users if, and only if, it is able to provide to users:*

- *understandable explanations of results, including :*
  - *the reasoning process that conducts to results*
  - *the risk of error of results*
- *an understandable version of its general process*
- *its global error, precision, sensitivity and specificity rates*

Nevertheless, although the classifier can answer the question "Why this result?", it will not be able to answer if the user asks, still to prevent a potential misinterpretation, "How the process of classification have been built? Where does it come from?". Only the learning algorithm used by the classification system can be able to bring elements of a response to users because the function of the learning system is to build classifiers, whereas the function of classifiers is to classify.

Therefore, a "transparent" classification system must be based on a type of classifier "interpretable", as defined in Definition 3, but it must also use an "interpretable" learning algorithm, still to ensure that users are in a position to make informed decisions. A first way to establish whether a learning algorithm is "interpretable" could be to evaluate if users can easily reproduce the process of the algorithm. However, evaluating "interpretability" in this way would be tedious for users. We have then to establish operational criteria of learning algorithms that can contribute to its "interpretability" by users.

First, the more linear it is, the more reproducible it is by users. However, linearity alone is not enough to allow "interpretability". For example, this is the case if the various steps of the algorithm fail to be understandable by users or if branching and ending conditions are not understandable by users. Accordingly, we proposed the following definition of the "interpretability" of a learning algorithm.

Definition 4. *A learning algorithm is considered to be "interpretable" by users if, and only if it has:*

- *a process as linear as possible*
- *understandable steps*
- *understandable branching and ending conditions*

The use of concept such as "possibility" of the algorithm implies that we cannot tell that a learning algorithm is absolutely "interpretable". By corollary, the assessment algorithm's "interpretability" is quite subjective and dependent on what we consider as "possible" in terms of linearity for an learning algorithm.

2.3 Traceability of results

Another aspect we have to take into account is the capacity to traceback data used to produce a specific classification. As introduced by Hedbom [18], a user has the right to know which of her/his personal data are used in a classification system, but also how and why. This is all the more true in medical contexts, where the data used are sensitive.

The "understandability" and "interpretability" criteria alone are not enough to ensure the ability to traceback the operations and data used to produce a given result. For example, let us suppose we have a perfectly understandable and interpretable classification system, if this system does some operations randomly, it becomes difficult to traceback operations made from a given result.

By contrast, if a classification system is totally "understandable" and "interpretable", the determinism of classifiers and the learning system is a necessary and sufficient condition to allow "traceability". We can then propose the following definition of the traceability of results.

Definition 5. *The results of a classification system are considered to be "traceable" if, and only if, the learning system and the type of classifier used have a non-stochastic process.*

2.4 Revisability of classifiers

Lastly, the concept of "transparency" can be associated with the possibility for users to make feedbacks to the classification system to improve future results [40]. When a classification system allows users to make feedbacks that are taken into account, this classification system appears less as a "black-box" system to users.

For example, in the medical context, Caruana et al. [7] have reported that physicians had a better appreciation of a rule-based classifier than of a neural network, in the case of predicting pneumonia risk and hospital readmission. This is despite the fact that neural network had better results than the rule-based classifier. According to the authors, the possibility to modify directly wrong rules of the classifier played a crucial role in the preference of physicians.

However, not all classifiers can be directly modified by users. Another way to take account of users' feedbacks is to use continuous learning algorithms (or online learning). The majority of learning algorithms are offline algorithms, but all can be modified, more or less easily, to become online learning algorithms. In that case, the classifier is considered to be partly "revisable". We then obtain the following definition of "revisability" of the type of classifier used by a classification system.

Definition 6. *A type of classifier used by a classification system is considered to be "revisable" by users if, and only if, users can directly modify the classifier's process or, at least, the learning algorithm can easily become an online learning algorithm.*

3 Evaluation of different classification systems

In this section, we use the operational criteria we have established in section 2 to evaluate the degree of "transparency" of several well-known classification systems. With this evaluation, we aim to determine whether one of these classification systems can be used in our use case, from a "transparency" point of view.

We also evaluate the performances of these algorithms on datasets similar to our use case, to evaluate the cost of using a "transparent" alogrithm in terms of performances.

3.1 "Transparency" evaluation

Our evaluation of "transparency" has been made on six different classification systems. The BPMLL algorithm (based on artificial neural networks) [42], the MLkNN algorithm (based on k-Nearest Neighbors) [41], the Naive Bayes algorithm (producing probability-based classifiers) [23], the C4.5 algorithm (producing decision-tree classifiers) [30], the RIPPER algorithm (producing rule-based classifiers) [8] and the SMO algorithm (producing SVM classifiers) [29, 25].

Fig. 1 displays a summary of the following evaluation of our different classification systems. Due to their similarities in terms of "transparency", C4.5 and RIPPER algorithms have been considered as the same entity.

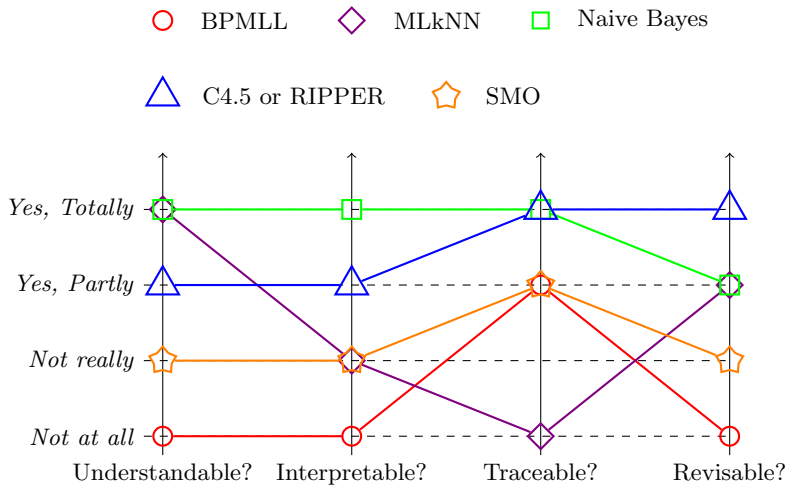

Fig. 1. Graphical representation of the potential "transparency" of different classification systems according to our operational criteria.

Let us start with the evaluation of a classification system based on the BPMLL algorithm [42] (red circles in Fig. 1). The BPMLL algorithm is based on a neural network and neural networks are based on notions/concepts that are not included in the school curriculum of users such as back-propagation and activation functions. Therefore, the steps of the BPMLL algorithm, as well as its branching/ending conditions, cannot be considered to be "understandable" by users. In addition, the learning process of neural networks is not what might be called a linear process. Accordingly, we cannot consider this classification system to be "understandable" and "interpretable" by users. However, neural networks

are generally determinist but, due to their low "understandability", they can only be considered to be partly "traceable". Finally, concerning the "revisability" of such a classification system, users cannot directly modify a wrong part of the classifier process and neural networks are not really adapted to continuous learning due to the vanishing gradient problem [21].

The ML-KNN algorithm [41] (violet diamonds in Fig. 1) is considered to be fully "understandable" because it is based on notions like distances and probabilities. Classifiers produced by the ML-KNN algorithm can produce explanations such as "x is similar to this other example". However, due to nested loops and advanced use of probabilities, the learning algorithm does not fit our criteria of "interpretable". In addition, the k-Nearest Neighbors algorithm [13], used by ML-KNN, is generally not determinist which makes the classification system not "traceable". Nevertheless, although classifiers produced by the ML-KNN algorithm cannot be directly modified by users, ML-KNN can easily be modified to become online learning. Consequently, it is partly "revisable".

The Naive Bayes algorithm [23] (green squares in Fig. 1) is considered to be fully "understandable" because, in our context, probabilities and the Bayes theorem are included in the school curriculum of all potential users. The Naive Bayes algorithm is also quite linear and all its steps, as well as its branching/ending conditions, are "understandable". Accordingly, the Naive Bayes algorithm is considered to be fully "interpretable" by users. In addition, the Naive Bayes algorithm is fully determinist, so considered to be fully "traceable". Lastly, users cannot easily modify the classifier, because its a set of probabilities, but the Naive Bayes algorithm can update these probabilities with users' feedbacks, becoming an online learning algorithm. The Naive Bayes algorithm is then considered to be partly "revisable".

The C4.5 and RIPPER algorithms are considered to be partly "understandable" because, even though decision trees or rulesets are notions fully "understandable" by users, these two learning algorithms are based on the notion of Shannon's entropy [32], a notion that is not included into the school curriculum of all potential users. With the same logic, even though decision trees or rulesets are fully "interpretable" classifiers, these learning algorithms are quite linear but their steps and branching/ending are not "understandable" by users because based on Shannon's entropy. The only difference between C4.5 and RIPPER could be on the linearity of their learning algorithm, because RIPPER may be considered to be less linear than C4.5, so less "interpretable". Accordingly, C4.5 and RIPPER are considered to be partly "interpretable" by users. In addition, the C4.5 and RIPPER algorithms are determinists, so fully traceable, and they are considered to be fully "revisable", because users can modify directly classifiers such as decision trees or rulesets.

Lastly, concerning the SMO algorithm, it is mainly based on mathematical notions, such as a combination of functions, that are not necessarily included in the school curriculum of all potential users. The SMO algorithm is not considered to be really "understandable" and "interpretable" by users. The SMO algorithm is determinist but, due to its low "interpertability" it could be more diffcult to

traceback its results. It is then considered to be partly "traceable". In addition, the SMO algorithm can become online [35], but not as easily as ML-kNN or Naive Bayes algorithms (for example), it is not considered to be really "revisable".

Consequently, if we start from the classification system with less operational criteria of "transparency" checked, to the classification system with a majority of operational criteria checked, we obtain: BPMLL, SMO, MLkNN, RIPPER, C4.5 and Naive Bayes. Accordingly, a classification system based on the Naive Bayes algorithm can be considered as the best alternative, from a "transparency" perspective, to treat our medical use case.

3.2 Naive Bayes algorithm for multi-label classification

As developed in section 3.1, the Naive Bayes algorithm can be considered to be "transparent" according to our operational criteria. A common way to apply a one-label classification system to a multi-label classification problem, like in our case, is to use the meta-learning algorithm RAkEL [37]. However, the use of RAkEL, which is stochastic and combine several classifiers, makes classification systems less "interpretable" and "traceable". We proposed then a version of the Naive Bayes algorithm, developed in Algorithm 1, to treat directly multi-label classification problems staying as "transparent" as possible.

Algorithm 1: A Naive Bayes algorithm for multi-label classification

Data: a learning dataset $\mathcal{I}$, a set of variables $\mathcal{X}$ and a set of labels $\mathcal{Y}$
Result: sets of approximated probabilities $P_{\mathcal{Y}}$ and $P_{\mathcal{X}|\mathcal{Y}}$

`// Computing subsets of numerical variables`
1 **foreach** *variable* $X \in \mathcal{X}$ **do**
2 Discretize domain of X according to its values in $\mathcal{I}$

`// Counting occurences of` $\mathcal{Y}$ `and` $\mathcal{X} \cap \mathcal{Y}$
3 **foreach** *instance* $I \in \mathcal{I}$ **do**
4 **foreach** *label* $Y \in \mathcal{Y}$ **do**
5 $y_I \leftarrow$ value of Y for instance I
6 Increment by one the number of occurences of $Y = y_I$
7 **foreach** *variable* $X \in \mathcal{X}$ **do**
8 $t_X^I \leftarrow$ the subset of X corresponding to its value in instance I
9 Increment by one the number of occurences of $Y = y_I \cap X = t_X^I$

10 Compute probabilities $P_{\mathcal{Y}}$ and $P_{\mathcal{X}|\mathcal{Y}}$ from computed number of occurences
11 **return** $P_{\mathcal{Y}}$ *and* $P_{\mathcal{X}|\mathcal{Y}}$

To treat numerical variables, the first step of our algorithm is to discretize these numerical variables into several subsets (Algorithm 1, line 2). Discretizing numerical variables allows us to treat them as nominal variables. For each instance of the learning dataset, we get the subset corresponding to the value of

each variable for the instance (Algorithm 1, line 8). Then, our algorithm counts occurences of each value of label and variables, and computes their frequency of occurence.

To discretize numerical variables, we first decided to use the fuzzy c-means clustering algorithm [6]. The fuzzy c-means allows to determine an "interpretable" set of subsets T_X of a variable X based on the distribution of observed values in this variable domain. Therefore, the subset t corresponding to a new value $x \in X$ is the subset $t \in T_X$ with the highest membership degree $\mu_t(x)$ (Equation 1).

$$t_X \leftarrow \underset{t \in T_X}{\arg\max}\ \mu_t(x) \tag{1}$$

However, we see here that the use of the fuzzy c-means algorithm requires introducing new concepts such as fuzzy sets, membership functions and membership degrees [39]. These concepts are not included into the school curriculum of users, reducing the "transparency" of the classification systems.

Therefore, we propose to use another discretizing method, more "transparent". This method, inspired by histograms, consists in splitting the variable domain into n subsets of equal size. Therefore, the subset t corresponding to a new value $x \in X$ is the subset $t \in T_X$ such as $min(t) \leq x < max(t)$. This method was preferred due to its simplicity and its potential better "transparency".

3.3 The search for a right balance between performances and transparency

Now that we have evaluated the "transparency" of several classifier systems, and we have identified the Naive Bayes algorithm as the most "transparent" alternative in our context, a question still remains: Does "transparency" have a cost in terms of performances?

To answer this question we evaluated classifiers presented at the beginning of this section on performance criteria for different well-known multi-label datasets and a dataset named *consultations* corresponding to our use case. Table 1 is an example based on this dataset. Currently, our dataset contains 50 instances with 4 features (patients' age, sex, BMI and disease) and 18 labels corresponding to data potentially needed by endocrinologists during consultations.

Our aim in this sub-section is to determine if the use of our version of the Naive Bayes algorithm offers suitable performances in our use case. If this is not the case, we won't have the choice but to envisage using a less "transparent" algorithm if it offers better performances.

These evaluations were made by using the Java library Mulan [38], which allowed to use several learning systems and cross-validation metrics. The program to reproduce these evaluations can be found on the GitLab of the LAMSADE[4].

Fig. 2 shows the distribution of macro-averaged F-measures of classifier systems computed for different multi-label datasets. The F-measure is a harmonic mean of the precision and the recall of evaluated classification systems. These

[4] https://git.lamsade.fr/a_richard/transparent-performances

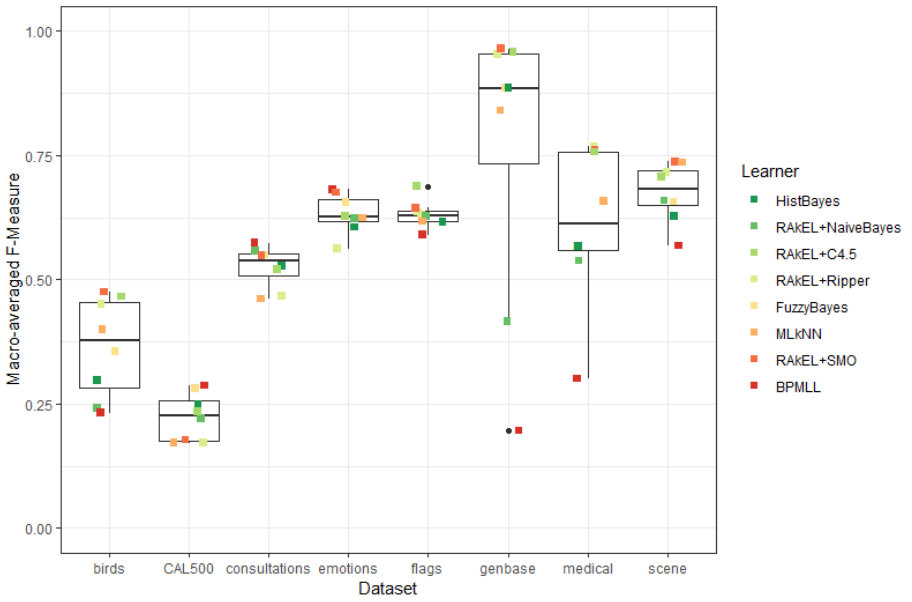

Fig. 2. Distribution of macro-averaged F-measures of several multi-label classification systems for different datasets. Results obtained by cross-validation.

results have been obtained by cross-validation. Classification systems have been ordered by their degree of "transparency" according to the definition developed in section 2. Green for the most "transparent", red for the less "transparent". Although a macro-averaged F-measure alone does not allow a precise evaluation, it allows us to have an overview of classification systems' performances.

We can see that the most "transparent" classification systems (greenest squares in Fig. 2) are not necessarily offering the worst performances. We can also see that, in some cases, "transparent" classification systems can offer performances close to the performances of the less "transparent" ones. In our case, represented by the *consultations* dataset, although the BPMLL algorithm offers the best F-Measure with 0.57, we can see that our version of the Naive Bayes algorithm (HistBayes) offers a quite close F-Measure with 0.53. Note that these results have to be nuanced by the small size of our dataset.

4 Discussion

As introduced in section 2, the definition and operational criteria of "transparency" we proposed are centered on our use case: classification systems in medical contexts. Because this context is sensitive, we had to establish clear operational criteria of what we called a "transparent" classification system. Based on these definitions we have been able to determine what kind of classification

system we must use in priority. Besides, we can suppose that the operational criteria we proposed can be used to evaluate the "transparency" of healthcare information systems in general. It would also be interesting to establish operational criteria of "transparent" systems in other contexts than medicine and to compare these operational criteria.

However, these definitions and operational criteria have their limitations. First, they are mainly based on our definitions of "transparency" and on our understanding of the medical context(as computer scientist and engineers). Consequently, they are not exhaustive and can be improved. And secondly, operational criteria were chosen to be easily evaluated without creating additional workload to clinicians, but it could be interesting to integrate them in the evaluation process. For example, the "understandability" of provided explanations could be evaluated directly in practice by clinicians.

Nevertheless, we claim that establishing clear operational criteria of "transparency" can be useful for decision-makers to determine which systems or algorithm is more relevant in which context. These operational criteria of "transparency" must be balanced with performance criteria. Depending on the use case, performances could be more important than "transparency". In our case, the medical context requires to be as "transparent" as possible. Fortunately, as developed in sub-section 3.3, in our case being "transparent" had not a lot of impact on performances and did not implies the use of a less "transparent" classification system with better performances.

Acknowlegment

This paper was made in collaboration with employees of the Civil Hospitals of Lyon (France). Thanks to all of them. Special thanks to Pr. Moulin and Dr. Riou for their suggestions and instructive discussions.

References

1. Abdollahi, B., Nasraoui, O.: Transparency in fair machine learning: The case of explainable recommender systems. In: Human and Machine Learning, pp. 21–35. Springer (2018). https://doi.org/10.1007/978-3-319-90403-0_2
2. Akkermans, H., Bogerd, P., Van Doremalen, J.: Travail, transparency and trust: A case study of computer-supported collaborative supply chain planning in high-tech electronics. European Journal of Operational Research **153**(2), 445–456 (2004). https://doi.org/10.1016/S0377-2217(03)00164-4
3. Amiribesheli, M., Hosseini, M., Bouchachia, H.: A principle-based transparency framework for intelligent environments. In: Proceedings of the 30th International BCS Human Computer Interaction Conference: Fusion! p. 49. BCS Learning & Development Ltd. (2016). https://doi.org/10.14236/ewic/HCI2016.68
4. Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. new media & society **20**(3), 973–989 (2018). https://doi.org/10.1177/1461444816676645
5. Berner, E.S.: Clinical decision support systems. Springer, 3rd edn. (2016)

6. Cannon, R.L., Dave, J.V., Bezdek, J.C.: Efficient implementation of the fuzzy c-means clustering algorithms. IEEE transactions on pattern analysis and machine intelligence (2), 248–255 (1986). https://doi.org/10.1109/TPAMI.1986.4767778
7. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015)
8. Cohen, W.W.: Fast effective rule induction. In: Twelfth International Conference on Machine Learning. pp. 115–123. Morgan Kaufmann (1995)
9. Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction **18**(5), 455 (2008). https://doi.org/10.1007/s11257-008-9051-3
10. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE symposium on security and privacy (SP). pp. 598–617. IEEE (2016)
11. Dinka, D., Nyce, J.M., Timpka, T.: The need for transparency and rationale in automated systems. Interacting with computers **18**(5), 1070–1083 (2006)
12. Doran, D., Schulz, S., Besold, T.R.: What does explainable ai really mean? a new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (2017)
13. Dudani, S.A.: The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics (4), 325–327 (1976)
14. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., Holzinger, A.: Explainable ai: the new 42? In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 295–303. Springer (2018). https://doi.org/10.1007/978-3-319-99740-7_21
15. Göritzlehner, R., Borst, C., Ellerbroek, J., Westin, C., van Paassen, M.M., Mulder, M.: Effects of transparency on the acceptance of automated resolution advisories. In: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 2965–2970. IEEE (2014). https://doi.org/10.1109/SMC.2014.6974381
16. Groth, P.: Transparency and reliability in the data supply chain. IEEE Internet Computing **17**(2), 69–71 (2013). https://doi.org/10.1109/MIC.2013.41
17. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web **2** (2017)
18. Hedbom, H.: A survey on transparency tools for enhancing privacy. In: IFIP Summer School on the Future of Identity in the Information Society. pp. 67–82. Springer (2008). https://doi.org/10.1007/978-3-642-03315-5_5
19. Heeks, R., Mundy, D., Salazar, A.: Why health care information systems succeed or fail. Information Systems for Public Sector Management (1999)
20. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM conference on Computer supported cooperative work. pp. 241–250. ACM (2000)
21. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **6**(02), 107–116 (1998)
22. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
23. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338–345. Morgan Kaufmann, San Mateo (1995)

24. Karsenty, L., Botherel, V.: Transparency strategies to help users handle system errors. Speech Communication **45**(3), 305–324 (2005)
25. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to platt's smo algorithm for svm classifier design. Neural Computation **13**(3), 637–649 (2001). https://doi.org/10.1162/089976601300014493
26. Kim, T., Hinds, P.: Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In: ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication. pp. 80–85. IEEE (2006). https://doi.org/10.1109/ROMAN.2006.314398
27. Michener, G., Bersch, K.: Identifying transparency. Information Polity **18**(3), 233–242 (2013). https://doi.org/10.3233/IP-130299
28. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)
29. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)
30. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)
31. Richard, A., Mayag, B., Meinard, Y., Talbot, F., Tsoukiàs, A.: How AI could help physicians during their medical consultations: An analysis of physicians' decision process to develop efficient decision support systems for medical consultations. In: PFIA 2018. Nancy, France (2018)
32. Shannon, C.E.: A mathematical theory of communication. Bell system technical journal **27**(3), 379–423 (1948)
33. Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: CHI'02 extended abstracts on Human factors in computing systems. pp. 830–831. ACM (2002). https://doi.org/10.1145/506443.506619
34. Spagnolli, A., Frank, L.E., Haselager, P., Kirsh, D.: Transparency as an ethical safeguard. In: International Workshop on Symbiotic Interaction. pp. 1–6. Springer (2017). https://doi.org/10.1007/978-3-319-91593-7_1
35. Tax, D.M., Laskov, P.: Online svm learning: from classification to data description and back. In: 2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No. 03TH8718). pp. 499–508. IEEE (2003)
36. Tintarev, N., Masthoff, J.: Effective explanations of recommendations: user-centered design. In: Proceedings of the 2007 ACM conference on Recommender systems. pp. 153–156. ACM (2007). https://doi.org/10.1145/1297231.1297259
37. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multi-label classification. IEEE Transactions on Knowledge and Data Engineering **23**(7), 1079–1089 (2011). https://doi.org/10.1109/TKDE.2010.164
38. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research **12**, 2411–2414 (2011)
39. Zadeh, L.A., Klir, G.J., Yuan, B.: Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers, vol. 6. World Scientific (1996)
40. Zarsky, T.: Transparency in data mining: From theory to practice. In: Discrimination and Privacy in the Information Society, pp. 301–324. Springer (2013)
41. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern Recogn. **40**(7), 2038–2048 (2007)
42. Zhang, M., Zhou, Z.: Multi-label neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering **18**, 1338–1351 (2006). https://doi.org/10.1109/TKDE.2006.162