UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Semantic match: Debugging feature attribution methods in XAI for healthcare

Cina, G.; Röber, T.E.; Goedhart, R.; Birbil, S.I.

Publication date
2023

Document Version
Final published version

Published in
Proceedings of Machine Learning Research

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
Cina, G., Röber, T. E., Goedhart, R., & Birbil, S. I. (2023). Semantic match: Debugging feature attribution methods in XAI for healthcare. *Proceedings of Machine Learning Research*, *209*, 182-191. https://proceedings.mlr.press/v209/cina23a.html

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:16 Sept 2026

Semantic match: Debugging feature attribution methods in XAI for healthcare

Giovanni Cinà G.CINA@AMSTERDAMUMC.NL
Department of Medical Informatics, Amsterdam University Medical Center, The Netherlands
Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
Pacmed, The Netherlands

Tabea E. Röber T.E.ROBER@UVA.NL
Rob Goedhart R.GOEDHART2@UVA.NL
Ş. İlker Birbil S.I.BIRBIL@UVA.NL
Department of Business Analytics, Amsterdam Business School, University of Amsterdam, The Netherlands

Abstract

The recent spike in certified Artificial Intelligence tools for healthcare has renewed the debate around adoption of this technology. One thread of such debate concerns Explainable AI and its promise to render AI devices more transparent and trustworthy. A few voices active in the medical AI space have expressed concerns on the reliability of Explainable AI techniques and especially feature attribution methods, questioning their use and inclusion in guidelines and standards. We characterize the problem as a lack of semantic match between explanations and human understanding. To understand when feature importance can be used reliably, we introduce a distinction between feature importance of low- and high-level features. We argue that for data types where low-level features come endowed with a clear semantics, such as tabular data like Electronic Health Records, semantic match can be obtained, and thus feature attribution methods can still be employed in a meaningful and useful way. For high-level features, we sketch a procedure to test whether semantic match has been achieved.

Data and Code Availability No data was used for the paper.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Along with the blooming of Artificial Intelligence (AI) and the accompanying increase in model complexity, there has been a surge of interest in explainable AI (XAI), namely AI that allows humans to understand its inner workings (*e.g.*, Doshi-Velez and Kim, 2017; Linardatos et al., 2020; Gilpin et al., 2018; Biran and Cotton, 2017; Doran et al., 2017). This interest is particularly keen in safety-critical domains such as healthcare, where it is perceived that XAI can engender trust, help monitoring bias, and facilitate AI development (*e.g.*, Doshi-Velez and Kim, 2017; Lipton, 2018). XAI has already shown to improve clinicians' ability to diagnose and assess prognoses of diseases as well as assist with planning and resource allocation. For example, Letham et al. (2015) developed a stroke prediction model that matches the performance of the most accurate Machine Learning (ML) algorithms, while remaining as interpretable as conventional scoring methods used in clinical practice.

There is a wide variety of techniques for XAI, and many categorizations have been proposed in the literature. Techniques can roughly be grouped into local vs. global, and model-specific vs. model-agnostic approaches. Local methods aim to explain model outputs for individual samples, while global methods focus on making models more explainable at an aggregate level. Model-specific methods are tailored to explain a specific type of model, while model-agnostic methods can be applied to a range of different models. Many of the well-known techniques yield post-hoc explanations, meaning that they generate explanations for already trained, so-called 'black-box', models. Alternatively, there exist approaches that are considered inherently explainable, also known as white-box (or glass-box) models, such as decision trees and linear regression models. A detailed taxonomy is beyond the scope of this paper; for an extensive overview we refer

© 2023 G. Cinà, T.E. Röber, R. Goedhart & ¸: Birbil.

the reader to Carvalho et al. (2019), Molnar (2022), and Ras et al. (2022).

A group of XAI techniques that has enjoyed substantial fame is the set of feature attribution methods, namely techniques that assign to each feature a measure of how much it contributes to the calculation of the outcome according to the model. Popular techniques produce such explanations in a local fashion, and famous ones include SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), saliency maps (Selvaraju et al., 2017), and integrated gradients (Sundararajan et al., 2017). To give a sense of the success of these techniques, it suffices to say that some of them are now integrated as default explainability tools in widespread cloud machine learning services, while in areas such as natural language processing, researchers are starting to use such methods as the gold standard against which they judge the quality of other explanations (Mohankumar et al., 2020).

Despite the enthusiasm and a growing community of researchers devoting energy to XAI, there is currently no consensus on the reliability of XAI techniques, and several researchers have cast serious doubts on whether XAI solutions should be incorporated into guidelines and standards, or even deployed at all (*e.g.*, Lipton, 2018; McCoy et al., 2021; Ghassemi et al., 2021; Neely et al., 2022).

Undoubtedly, there is an inherent tension between the desire for machines performing better than humans, and the requirement for machines to provide human-understandable explanations. Together with their super-human capacities – such as the ability to juggle dozens or hundreds of factors – the sub-symbolic character of statistical learning techniques contributes to rendering many ML models opaque to humans. Simply put: humans cannot 'read off' what a neural network has learned just by looking at the matrix of weights.

The internal or 'latent' representation of a machine, namely the way in which the machine encoded the patterns found in the data, is what we would like to explain to the human, together with the way in which this internal representation interacts with a single data point to generate an output. The solution presented by feature attribution methods is also sub-symbolic, in the sense that an explanation also takes the form of a vector or matrix of values, and here lies the source of the problem. As different scholars pointed out (Ghassemi et al., 2021; Rudin, 2019), the assignment of meaning to such explanations can be tricky, sometimes lulling the humans into a false sense of understanding while the explanations are in fact flawed or misleading. This issue is particularly sensitive when such explainability techniques are used in high-stakes environments such as healthcare.

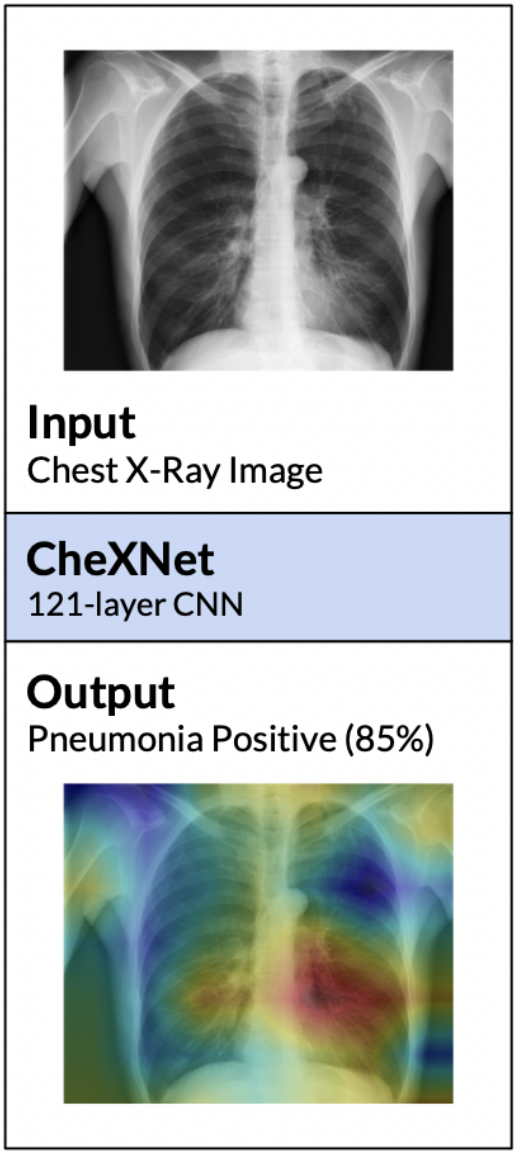


Figure 1: An example of visual explanation by heatmap for a medical image. Image courtesy of Rajpurkar et al. (2017).

Does this mean that feature attribution methods are altogether unreliable? In this article we argue that existing criticism on the viability of post-hoc local

Figure 2: While the explanation of the first classification seems intuitive, this impression is put into question when a similar explanation is offered for an absurd classification. Figure borrowed from Rudin (2019) with author's permission.

explainability methods throws away the baby with the bathwater by generalizing a problem that is specific to unstructured data such as images. We characterize the issue with feature attribution methods as a lack of semantic match between explanations and human understanding. To understand when semantic match can be obtained reliably, we introduce a distinction between feature importance of low- and high-level features. We argue that in the case of data types for which low-level features come endowed with clear semantics, such as tabular data like Electronic Health Records (EHRs), semantic match is enabled and thus feature attribution methods can still be employed in a meaningful and useful way. As for high-level features, we present a conceptual procedure to verify whether semantic match is achieved, paving the way for future operationalization of this test.

The criticism on local feature attribution methods

In this section we expand on the problem that feature attribution methods are confronted with. Applied to images, explanations generated by feature attribution methods present themselves as heatmaps or colored overlays, indicating the contribution of specific pixels to the prediction of the model on the input at hand. Intuitively, highlighted regions comprise pixels which were considered 'important' by the model. *Prima facie*, one might be led to believe that this allows humans to check that the model is paying attention to the right elements of the image, therefore increasing our trust in the specific prediction and in the model more generally.

In Rajpurkar et al. (2017), the authors developed an algorithm to predict the probability of pneumonia from chest X-rays with the performance exceeding those of practising radiologists. To interpret the outcome, the authors use a heatmap to visualize the areas of the image that are most indicative for the prediction using class activation mappings. Figure 1, taken from said paper, shows an example of an X-ray image for which the algorithm predicts pneumonia with a probability of 85%. The heatmap shows which area of the image contributes most to the prediction of the model.

However, when a certain area of an image is highlighted we simply do not know if what we recognize, say the shape of a kidney or the beak of a bird, is the same as what the AI recognizes. As several researchers have pointed out, *e.g.*, Ghassemi et al. (2021) and Lipton (2018) among others, what look like plausible explanations at first may turn out to be ungrounded or spurious explanations when subjected to closer scrutiny. Rudin (2019) exemplifies this mishap with the example in Figure 2. A classification model was trained to classify images into some classes, one of which is 'Siberian husky' and another one is 'transverse flute'. The evidence for the animal being a Siberian husky seems intuitive and would probably let us believe our model looks at the right places (the snout and paw of the dog) to make the classification. However, the evidence for the animal being a transverse flute looks almost identical. This shows that heatmaps display what parts the model is looking at, which does not necessarily translate to what we look at on a conceptual level. This opens the door for different kinds of biases, both on the side of the user,

who can project their beliefs erroneously onto the machine, and on the side of the XAI, which might provide feature attribution information that is inconsistent or misleading.

The root of the problem, we maintain, lies in the inability of humans to attribute meaning to a sub-symbolic encoding of information. What is needed is a systematic way to translate sub-symbolic representations to human-understandable ones, in a way that respects how we assign meaning. This is represented schematically in Figure 3. We call such a commuting diagram a **semantic match**. An ideal explanation would provide content of the bottom-left node and a suitable translation in order to obtain a semantic match: the explanation of a certain sub-symbolic representation encoded by the machine should have (i) a clearly defined meaning and (ii) an unambiguous way to translate to human terms with the same (or very similar) meaning. Semantic match, which is a crude simplification of complex cognitive and linguistic phenomena, offers a handy conceptual tool to debug explanations.

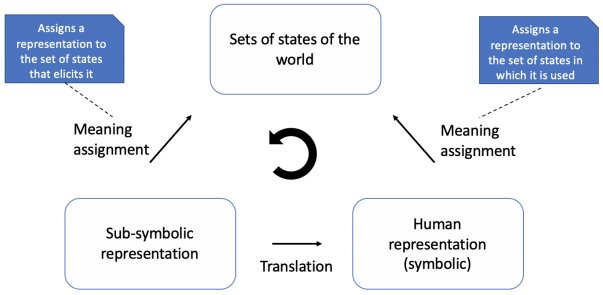


Figure 3: The diagram representing a semantic match.

Indeed, just as we are unable to understand a latent representation, we are unable to relate to a heatmap unless it comes paired with a well-defined meaning assignment and translation, as illustrated in the diagram. Overlaying the heatmap to an image encourages us to use our visual intuition as translation, but alas this last step is an ill-advised one, since it gives us the illusion of a semantic match while in fact the explanations do not conform to expectations. Figure 4 exemplifies a failed semantic match. In this scenario a heatmap is generated as an explanation for the behavior of the model on an image. From the image, it may appear that the machine 'recognizes' a certain feature (a dog's head), and therefore classifies the image as a specific class ('dog'). However, another spurious input generates the same heatmap, and hence the translation is invalid and unreliable. To replicate the uncanny observation of Figure 2, one can construct another scenario where states of the world are pairs of input images and predictions. These are cases of semantic mismatch: the states of the world in which the heatmap is produced do not correspond to the states where one would plausibly use the concept of dog's head to classify a dog image.

This criticism seems to undercut the utility of such local feature attribution methods: if they are potentially misleading and bring no clear added value, should they be used at all?

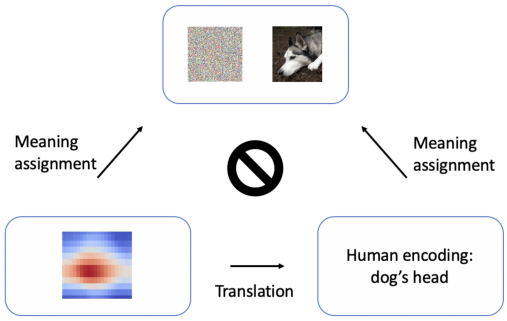


Figure 4: A scenario depicting a semantic mismatch. The images eliciting a certain explanation do not comply with our recollection of the explanation.

Distinguishing low- and high-level features

In order to understand the utility of such methods, one needs to step back and consider what are the characteristics of the examples that give rise to the problem, and assess whether they generalize. Not unsurprisingly, most of the debate revolves around images, given the astounding success of deep learning in the realm of computer vision. Images are a prime example of what is sometimes called 'unstructured data', namely data that does not come equipped with additional structure and is akin to raw sensory data. Note that this is a misnomer, since such data typically is endowed with a rich structure (in the mathematical sense) deriving from the notion of distance defined between pixels or between parts of an ordered sequence. We will nonetheless adopt this terminology to facilitate readability, since it is ubiquitous.

One of the defining characteristics of images – and unstructured data more generally – is that a single feature has no intrinsic meaning: a certain color value of a pixel means nothing by itself. Only a pattern of values for a group of features can be attributed meaning, *e.g.*, a cloud of pixels with the shape and color of a dog's head. In other words, we recognize properties or entities in an image by matching activation patterns with our visual intuitions. Following a widespread habit in the ML community, we will refer to patterns in groups of features as 'high-level features', while single features, namely the entries of the vector representing an input, will be dubbed 'low-level features' by contrast. With this terminology, we can succinctly state that in images only high-level features can be attributed meaning while low-level features cannot.

On the contrary, in structured data each feature is usually conferred specific meaning; *e.g.*, in EHR data a certain value might contain the information pertaining to the blood pressure of a patient at a certain time. When such low-level features are not defined with a specific protocol – such as the ones for measurements of vital signs in clinical settings – they refer to standard concepts in natural language. We can, therefore, easily interpret what these low-level features mean regardless of the values of the other features. For instance, a value of 180 in the feature corresponding to systolic blood pressure gives us a piece of information that we can understand and process, even without knowing other features of a patient.

Surely, there are also high-level features in the case of structured data. Continuing with the example of EHRs, if low-level features comprise eGFR and serum creatinine, a high-level feature could be a pattern of values eGFR and serum creatinine indicating stages of kidney damage. In short, also in tabular data any concept emerging from the combination of low-level features qualifies as a high-level feature. Hence, the crucial difference between structured and unstructured data is that the former has clear meaning for the low-level features, while on the high-level features the two data types behave similarly.

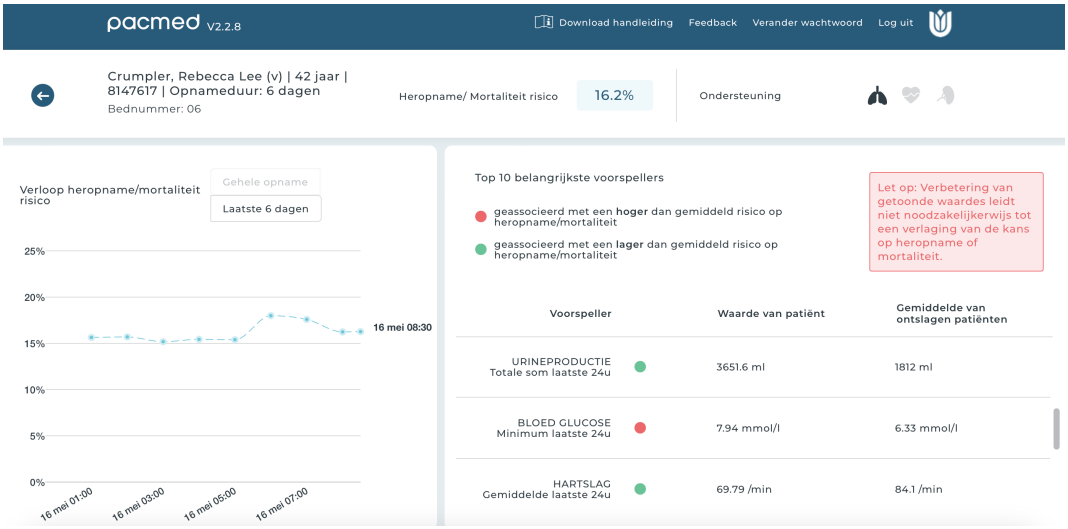
We maintain that the usage of a post-hoc feature attribution method hinges on the application of the corresponding semantic match diagram. Without a rigorously defined meaning and translation, a heatmap remains a matrix of values without rhyme or reason. We cannot extract information from such a matrix just as we cannot fathom what is encoded in the latent space of a neural network by eyeballing the value of a point in said space. So far the attribution of meaning to a heatmap-explanation has been carried out in an intuitive way, which, as previous scholars argued, is prone to error and a false source of confidence.

Saving feature importance for low-level features

The distinction between low-and high-level features helps untangle the cases in which semantic match works out-of-the box from those in which it fails. A post-hoc local feature attribution method can be used appropriately on low-level features when they have a predefined translation, as is the case for medical tabular data. For example, suppose a risk prediction model is trained on EHR data. When a patient is presented to the model, the model might provide a risk score associated with feature importance values for the patient's lab values. Suppose systolic blood pressure is marked as the most important factor increasing the risk of a specific patient. In this case, there is no ambiguity: such level of importance is attributed to systolic blood pressure and nothing else, and any user with sufficient training knows what systolic blood pressure is. Note that this has nothing to do with how the importance is calculated (*e.g.*, if it takes into account feature interactions) or if it is a sensible level of importance; all we are stating is that the user can unambiguously understand what the importance is attributed to. In other words, the importance is attributed to something which semantically matches our concept of 'systolic blood pressure'.

The user of such a model can then engage with the feature importance and assess whether it is sensible, while addressing questions like "Is this level of risk reasonable given a high importance of systolic blood pressure?" Such a question may be answered in the positive, if the clinician believes the value of systolic blood pressure is concerning, or in the negative, if for example, the clinician knows that the value of systolic blood pressure is a byproduct of medications they can control. An example of such an approach is displayed in Figure 5, where the top features contributing to clinical risk are displayed along with a color code indicating if they are risk increasing or decreasing.

Crucially, semantic match allows the user of an explanation to engage with it, and to decide whether it is agreeable or not. In contrast, when the same risk prediction is trained on image data, this reasoning falls apart. What feature attribution can highlight are just



(*a*) The user-interface of Pacmed Critical for a single patient (in Dutch).

(*b*) A close-up of the right side of the interface, displaying feature attribution for low-level features (in Dutch).

Figure 5: An example of the front-end of Pacmed Critical, a medical device currently used in the Amsterdam University Medical Center for the prediction of adverse outcomes after ICU discharge. Courtesy of Pacmed; for background on the prediction model see (Thoral et al., 2021; de Hond et al., 2022). The displayed patient data is synthetic.

high-level features, *e.g.*, the portion of an image with a kidney shape. But lacking a semantic match, clinicians cannot trust whether what they see in a heatmap matches the machine's internal representation.

Debugging feature importance for high-level features

Does this mean that we have to give up feature attribution for high-level features? Not in our view. In fact, what we need is a procedure to test whether semantic match is present or if it is violated. In what follows we present such procedure at a conceptual level.

Suppose a ML model f has been trained on labeled data of the shape (x, y), where x represent an input vector and y an output. We denote a local feature attribution method with M and say that $M(f, x, y) = e$ is the explanation furnished by the model f on such pair of input and output. We are interested in testing whether we have semantic match with the explanation e, or in other words, if what we 'see' in the explanation is indeed what the explanation captures.

At an intuitive level, what we want to ascertain is that an explanation matches our translation of it. This is encoded in the commutation of the semantic match diagram, namely that all the data points giving rise to a certain explanation are also complying to our translation hypothesis, which we will indicate with θ, and vice versa. We are thus looking at answering two questions:

- Is every input-output pair that generates an explanation similar to e also an example in which θ is the case?
- Is every input-output pair in which θ is satisfied going to generate an explanation that is similar to e?

To exemplify the procedure, suppose one has developed an algorithm to classify pictures of animals. Presented with a picture classified as a dog and an explanation e, one formulates the translation hypothesis θ that the explanation highlights the tail of the animal. To answer the aforementioned questions, one would need to take images that generate explanations similar to e and check whether the explanations of those samples highlight tails. For the second question, one would need to take images of animals with tails and consider how similar their explanations are to e.

Note that in this procedure one may also select data points whose label is different from y. This may not be an issue, since the high-level feature used for the prediction of x as y may in theory also be used on another data point and another label (and hence be highlighted in the explanation). In the animal classification example, one may collect images of a panther in which a tail is displayed, and rule that the explanations correctly highlights tails in those images.

Discussion

In this article we have reviewed the reliability problem of feature attribution methods and have proposed to diagnose the issue by means of the semantic match diagram. We have argued that without clear meaning and translation, semantic match cannot be obtained for high-level feature importance. A corollary of this statement is that current methods for feature attribution may not be appropriate for unstructured data unless semantic match is verified. While a full-fledged test for semantic match remains to be developed, we have moved the first steps by sketching a procedure on how to measure such quantity.

Even without such a test, structured data may still benefit from feature attribution, since for this data type low-level features have an in-built semantic match with human concepts. This allows humans to engage with explanations and exercise that all-important oversight that is required to spot failure modes of ML applications. Recent fairness concerns in the realms of human resources, healthcare, and law enforcement underscore the need for continued human control over semi-automated decisions.

When it comes to the limit of this analysis, it is important to remark that, even in the presence of semantic match, explanations can still fail to deliver on their promise. If an explanation is not faithful to the model, that is, it does not consistently represent the machine's behavior, the user may not leverage the explanation to agree or disagree with the machine's output; recent work on this point shows a concerning divergence in explanations (Neely et al., 2022). It should also be added that not all data types neatly fall into the categories of structured and unstructured data. Textual data, for example, contains both tokens that have intrinsic meaning and tokens that have only contextual meaning, and therefore sits somewhat in the middle. Time series data is in a similar spot, exhibiting sequences whose single values may have defined meaning but whose evolution over time is

harder to grasp. In these cases explanatory methods should be employed with caution and awareness of potential semantic mismatch. In the same vein, it should be recognized that, before the advent of neural networks that could process raw data, there was a long tradition of image processing by hand-crafting meaningful image features (see for example Street et al. (1993)). In essence, such approaches build features with an intrinsic semantic match by encoding expert knowledge with feature engineering, turning an unstructured dataset into a structured one. These approaches may be rediscovered as ways to attribute meaning to explanations in consultation with domain experts. Finally, semantic match is a user-dependent concept: while users with the relevant background may correctly interpret an explanation, others may not, and in healthcare settings it is crucial to clearly identify user groups and provide proper training.

Beside expanding and refining the procedure we introduced in the previous sections, future work aiming to obviate problems deriving from failed semantic match could direct attention to generating explanations that comply with human categories by design, possibly even with special categories employed by the targeted user. One direction of future research explores the possibility to capture information on machine behavior in a symbolic manner by means of hybrid models (Sarker et al., 2021). Another option might contemplate combining visual explanation with image segmentation, to fix the semantics of entities used in the explanation. Finally, more elaborate explanations may require access to ontologies regulating the relationship between entities (as in *e.g.*, Lecue and Wu (2018); Liartis et al. (2021)), which should themselves match ontologies used – more or less implicitly – by humans.

When it comes to medical AI, we want clinicians to be able to interact with machines in a meaningful way, namely with the right tools to adjudicate when the machine's advice is worth following. Framing the problem in terms of semantic match helps shedding light on the issue that explanations are still too ambiguous and too far from clinicians' reasoning. We should be building explanations in the clinician's language, rather than asking clinicians to rely on intuition or to learn to think like a computer scientist.

References

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: a survey. *IJCAI-17 Workshop On Explainable AI (XAI)*, 8(1):8—13, 2017. URL http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: a survey on methods and metrics. *Electronics*, 8(8):832, 2019. doi:10.3390/electronics8080832.

Anne AH de Hond, Ilse MJ Kant, Mattia Fornasa, Giovanni Cinà, Paul WG Elbers, Patrick J Thoral, M Sesmu Arbous, and Ewout W Steyerberg. Predicting readmission or death after discharge from the icu: External validation and retraining of a machine learning model. *Critical Care Medicine*, pages 10–1097, 2022. doi:10.1097/CCM.0000000000005758.

Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv*, 2017. URL https://arxiv.org/pdf/1710.00794.pdf).

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017. URL https://arxiv.org/pdf/1702.08608.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021. ISSN 2589-7500. doi:10.1016/s2589-7500(21)00208-9.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: an overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 00:80–89, 2018. doi:10.1109/dsaa.2018.00018.

Freddy Lecue and Jiewen Wu. Semantic Explanations of Predictions. *arXiv*, 2018. doi:10.48550/arxiv.1805.10587.

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015. ISSN 1932-6157. doi:10.1214/15-aoas848.

Jason Liartis, Edmund Dervakos, Orfeas Menis-Mastromichalakis, Alexandros Chortaras, and Giorgos Stamou. Semantic queries explaining opaque machine learning classifiers. In *DAO-XAI*, 2021. URL https://daoxai.inf.unibz.it/papers/DAO-XAI-2021-paper-2.pdf.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: a review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020. doi:10.3390/e23010018.

Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018. ISSN 1542-7730. doi:10.1145/3236386.3241340.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Liam G. McCoy, Connor T.A. Brenna, Stacy Chen, Karina Vold, and Sunit Das. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, 142:252–257, 2021. ISSN 0895-4356. doi:10.1016/j.jclinepi.2021.11.001.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. *arXiv preprint*, 2020. URL https://arxiv.org/pdf/2004.14243.

Christoph Molnar. *Interpretable machine learning: a guide for making black box models explainable*. Leanpub, second edition, 2022. ISBN 9798411463330. URL https://christophm.github.io/interpretable-ml-book/.

Michael Neely, Stefan F Schouten, Maurits Bleeker, and Ana Lucic. A song of (dis) agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing. *arXiv preprint*, 2022. URL https://arxiv.org/pdf/2205.04559.pdf.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.org/abs/1711.05225.

Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. Explainable deep learning: a field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397, 2022. doi:10.1613/jair.1.13200.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-Agnostic Interpretability of Machine Learning. *arXiv*, 2016. URL https://arxiv.org/pdf/1606.05386.pdf?source=post_page.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi:10.1038/s42256-019-0048-x.

Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-Symbolic Artificial Intelligence: Current Trends. *arXiv*, 2021. doi:10.48550/arxiv.2105.05330.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf.

W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE, 1993. URL https://minds.wisconsin.edu/bitstream/handle/1793/59692/TR1131.pdf?sequence=1.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. URL http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf.

Patrick J Thoral, Mattia Fornasa, Daan P de Bruin, Michele Tonutti, Hidde Hovenkamp, Ronald H Driessen, Armand RJ Girbes, Mark Hoogendoorn, and Paul WG Elbers. Explainable machine learning on amsterdamumcdb for icu discharge decision support: uniting intensivists and data scientists. *Critical Care Explorations*, 3(9), 2021. doi:10.1097/CCE.0000000000000529.