Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2022) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

XAI & I: Self-explanatory AI facilitating mutual understanding between AI and human experts

Jacques A. Grange[a*], Henrijs Princis[b], Theodor R. W. Kozlowski[b], Aissa Amadou-Dioffo[b], Jing Wu[b], Yulia A. Hicks[c] and Mark K. Johansen[a]

[a]Cardiff University, School of Psychology, Cardiff CF10 3AT, United Kingdom
[b]Cardiff University, School of Computer Science and Informatics, Cardiff CF24 4AG, United Kingdom
[c]Cardiff University, School of Engineering, Cardiff CF24 3AA, United Kingdom

Abstract

Traditionally, explainable artificial intelligence seeks to provide explanation and interpretability of high-performing black-box models such as deep neural networks. Interpretation of such models remains difficult, because of their high complexity. An alternative method is to instead force a deep-neural network to use human-intelligible features as the basis for its decisions. We tested this approach using the natural category domain of rock types. We compared the performance of a black-box implementation of transfer-learning using Resnet50 to that of a network first trained to predict expert-identified features and then forced to use these features to categorise rock images. The performance of this feature-constrained network was virtually identical to that of the unconstrained network. Further, a partially constrained network forced to condense down to a small number of features that was not trained with expert features did not result in these abstracted features being intelligible; nevertheless, an affine transformation of these features could be found that aligned well with expert-intelligible features. These findings show that making an AI intrinsically intelligible need not be at the cost of performance.

Keywords: Self-explanatory AI; Deep neural networks; Transfer learning; XAI; Category learning

[*] Corresponding author. Tel.: +44-2920-87-0077.
E-mail address: grangeja@cardiff.ac.uk

1. Introduction

1.1. Explainable AI vs. intrinsically intelligible AI

Impressive strides have been made by Artificial Intelligence (AI) systems in matching—and sometimes surpassing—human capability. Foremost among them are systems that classify images into categories (such as animals, cars, clouds, etc.), recognise natural speech and diagnose medical conditions. The complexity and distributed nature of these networks means that their decisions are often intrinsically inscrutable to humans, even experts. This weakens the basis of their accountability, reducing trust, and diminishing their adoption, especially if applications involve the AI making decisions about health, safety and risk. Thus, understandable reasons for an AI's decisions are crucial to adopting AIs, in part because of legal requirements governing their use in high-stakes domains.

One important and extensively studied approach to aid the understanding of the reasons underlying an AI's decisions is to focus on finding explanations for those decisions. This is the aim of Explainable AI (XAI) research, which seeks to make the AI interpretable, explainable and transparent [1]. XAI attempts to achieve this by using approaches like the construction of surrogate models that approximate the black-box AI but are more interpretable, saliency methods that indicate preferential selective attention to some input features as more important than others, local perturbation methods which evaluate changes in the output as a result of changes in the input, etc. Rudin [2] has argued that most of the researchers using these methods assume a trade-off between performance and the intelligibility of the explanation for the performance. However, the author argues that using models that are intrinsically interpretable, rather than secondary models that are an approximation, circumvents black-box interpretability problems "when the data are structured, with a good representation in terms of naturally meaningful features" (Rudin, 2019, page 2). Similarly, Elton [3] has called for self-explanatory AI as an alternative to XAI and Shen et al. [4] successfully trained a network to classify lung nodule malignancy from CT scans while also predicting expert-intelligible features.

1.2. Outline of our methodological approach to constructing inherently understandable AIs

Motivated by the desire for inherently understandable AI, our methodological approach has been to constrain a deep neural network (DNN) to predict expert feature values and then train a single-layer classifier to use these predicted values as a basis for categorization. As such, our proposed architecture forces the network to be inherently understandable by experts because it provides them, for any given category instance, with both an estimation of expert feature values and a categorization judgment. This constrained network was evaluated in contrast to a baseline, unconstrained network that made use of transfer learning with Resnet50 [5] and used Resnet50's penultimate layer to directly predict categories. In between these two approaches, we assessed a partially constrained network which was forced to condense down to a small number of features that were nonetheless not forced to reflect expert-identified features. We then compared the performance of all three network architectures to establish that intrinsic AI intelligibility can be attained without substantive loss of performance compared to the baseline, black-box model.

1.3. Choice of natural category domain: subtypes of rocks

A dominant focus of psychological studies of human categorisation is on assessing the representation of the categories in terms of similarity and in terms of rules in the mind. There are many mathematical models of human categorisation based on types of representation, including prototypes[6], exemplars [7]–[9] and various kinds of rules and mixed representations [10], [11]. An underlying assumption of exemplar and prototype models is that human categorisation is fundamentally based on similarity. As such, the research domain has dominantly focused on assessing artificially constructed, simple categories which facilitate the systematic assessment of the relevant similarities. A key problem has been extending this to real-world categories because the similarity characteristics of those categories are so complex, they are very difficult and expansive to assess. However, recent advances have shown that DNNs show considerable promise in terms of being able to predict human similarity judgments for real-world categories [12], [13]. While these findings are important for Psychology, they also show that DNNs have or at least can have shared properties with humans in terms of an underlying representational basis in similarity. This suggests that mutual intelligibility between humans and AIs is not only possible but will also facilitate interaction between them and enhance the performance of both. Historically, computer-based intelligence has been largely based on raw computational power. In contrast, human beings have fairly limited and highly constrained computational capacity but nonetheless have sophisticated and culturally transmissible concepts and concept-based reasoning. It is these kinds of sophisticated concepts that AI has struggled to approximate. However, DNNs are now having some success in terms of learning and approximating these natural concepts [5]. Nevertheless, the basis for this success is not necessarily clear. Similarly, humans tend to have very little insight about the basis for their categorisation, though domain experts sometimes have and can articulate more explicit categorisation processes. Ultimately, a mutual understanding between AI and experts must be based on shared concepts.

The domain choice for the present research has been motivated by the need for a set of real-world categories associated with well-specified features that are intelligible to experts and have, along with their similarity properties, been previously assessed. Nosofsky and colleagues' evaluation of human rock categorisation [13]–[15] have the above properties. Their primary motivation was in terms of characterising the psychological assessment of these categories as founded in similarity in naïve humans, rather than the development of an expert AI. A key cost of characterising similarity-based categories in human has been the need for extensive evaluation of similarity relationships between category instances, to then be able to develop similarity-based representations for these categories that approximate those representations that humans have. For example, pairwise similarity judgments can be used as input to multi-dimensional scaling (MSD), resulting in a hopefully low-dimensional similarity space. However, positioning additional category instances in that space requires new similarity judgments, at some considerable cost. As such, a key success of Sanders and Nosofsky [13] is to circumvent this cost by training a DNN to predict the coordinates of new instances in the MDS space they originally derived from naïve participants' similarity ratings [14]. In addition, they demonstrated that highlighting of expert-identified features facilitated natural category learning in naïve participants [15]. Ultimately, while their intent was to characterise psychological aspects of categorisation, mainly in naïve participants, their assessment of expert-identified features and the similarity properties of category instances and categories in this reasonably constrained topic area is useful for our purposes; this motivates our initial choice of this domain as a place to explore the development of an AI based on a DNN that is nonetheless inherently understandable by human experts.

2. Constraining a network to make it inherently intelligible by experts

2.1. Selected data set

Nosofsky, Sanders and colleagues used 30 rock categories organized in three super-ordinate categories (igneous, sedimentary and metamorphic). Each of the 30 categories were represented by images of 12 instances [16] later complemented with an additional 4 instances per category [13]. From their extensive evaluation, we have taken their set of 480 images, and used expert-identified features as well as the inferred MDS feature dimension values based on naïve participants' similarity ratings.

2.2. Baseline: transfer learning in rock classification with a feature-unconstrained network

Resnet50 was selected based on its recent successes in basic categorisation performance on real-work categories [5]. Resnet50 has been shown to classify images into 1000 real-world categories such as clouds, trees, cats, planes, etc. As such, Resnet is a general classifier whose accuracy closely approximates, and in some cases, supersedes that of humans. Our approach is to extract features from Resnet50 as a general classifier and use these features in the rock-classification domain. Starting with a fully trained version of Resnet50, we discarded the last, classification layer and used the penultimate layer of 2048 abstracted features as an input to a new classification system for 30 rock categories with various learned feature layers inserted. The new system (Figure 1, top path) consisted of a 50% random drop-out layer (there to avoid overfitting during learning), a fully connected 256-node layer, followed by a ReLU function, and a fully connected 30-node layer followed by a Softmax function that ultimately classifies images into 30 rock categories.

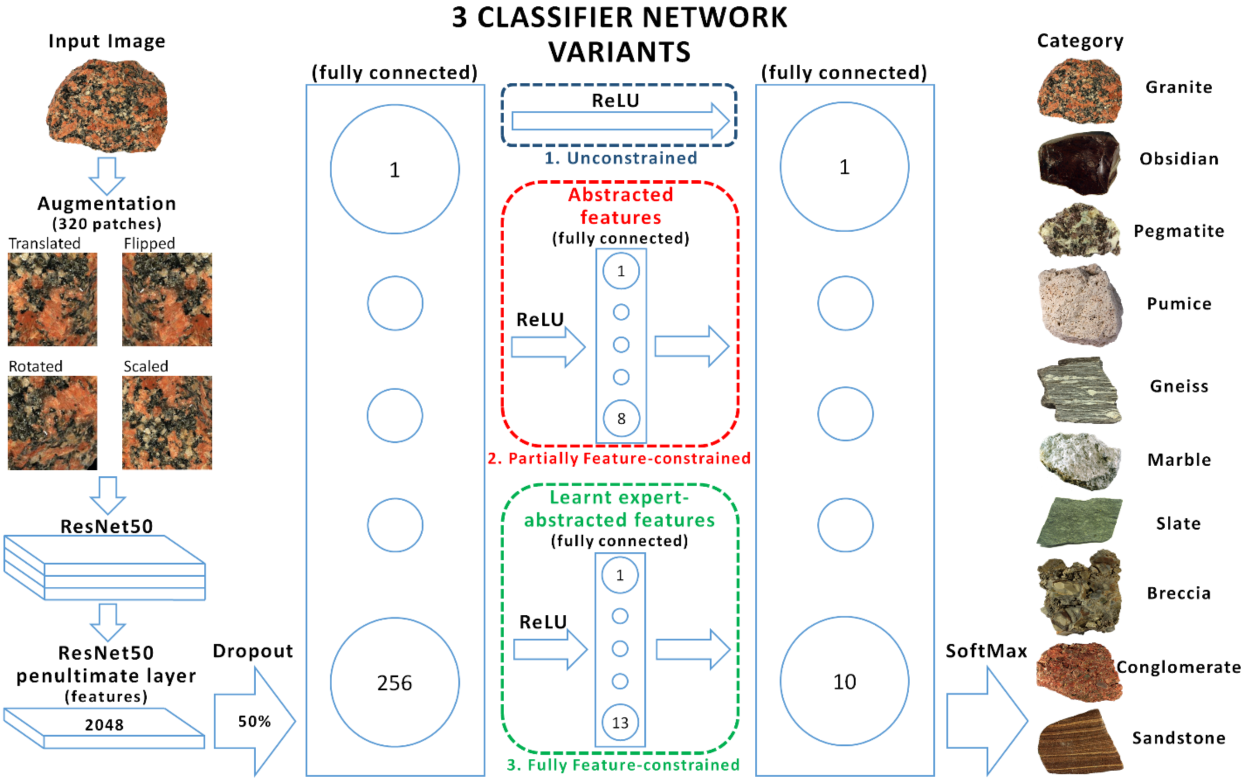

Fig. 1. Illustration of the three classifier network variants making use of transfer learning with Renset50. Rock images are first augmented before being fed through Resnet50, the penultimate layer of Resnet50 is taken through a 50% drop-out layer, a fully connected 256-node layer, then one of three paths before a 10-node category activation layer is passed through a SoftMax function: (1) an unconstrained variant path, where the 256-node layer fully connects to the 10-node layer through a ReLU function, (2) a partially feature-constrained variant path, where the 256-node layer activations are condensed, via a ReLU function, down to an 8-node, 'Abstracted Features' layer and (3) a fully feature-constrained variant path, where the abstracted feature layer is forced to be made of transfer-learned expert-abstracted features.

Augmentation of the set of 480 rock images was achieved by generating 320 random patches of 224 x 224 pixels (with translated, flipped, rotated and scaled), all patches covering as much of the rock as possible. Initial simulation runs showed that 200 learning epochs were sufficient for performance to asymptote without overfitting. The rock classification system was trained on the augmented set for 13 training images over 200 epochs and subsequently validated on the augmented set for the remaining 3 images in each category. Note Resnet50 itself was fixed and not further trained. Only the new classification layers were trained. Classification performance of the network, both training and validation, was measured by patch-voting for the various categories based on averaged classification probabilities across all 320 patches for a given image.

Training and validation accuracy were computed as the average accuracy for runs employing twelve different splits of the images in each category into 13 training and 3 validation instances. This was repeated 12 times to assess the repeatability/variability of performance. Grand means are reported to reflect accuracy. The performance achieved with this transfer learning approach on the 30 rock categories was a validation accuracy of 61% following a training accuracy of 100%. Because our focus was on developing a network intrinsically understandable by experts, we needed performance that was sufficiently good to be worth explaining. Nosofsky and colleagues selected a subset of categories from the initial 30 categories to do a psychological assessment of enhanced learning based on explicitly identified, highlighted visual features in the images [15]. We selected 10 of these categories because of their strong grounding in visually identifiable features: granite, obsidian, pegmatite, pumice, gneiss, marble, slate, breccia, conglomerate and sandstone. Transfer learning performance on these 10 categories was a validation accuracy of 86.8% (SD 0.8%) subsequent to a training accuracy of 100%.

2.3. Partially feature-constrained network

In contrast with the unconstrained network, a partially constrained network was forced to condense down to a small number of features (Figure 1, middle path); however, the nature of these features was unconstrained. The number of features was set to 8 to facilitate comparison to the 8 feature dimensions identified by Nosofsky and colleagues from multi-dimensional scaling solutions based on similarity judgments for pairs of rock images made by naïve/non-expert participants [15].

The performance of the partially constrained network was a validation accuracy of 87.1% (SD 0.8%), subsequent to a training accuracy of 100%. A key finding is that even though the nature of the 8 feature dimensions was not constrained, an optimised affine transformation of these dimensions strongly correlated with the MDS dimensions identified by Sanders and Nosofsky [13] (Figure 2), suggesting that the network's categorisation behaviour is grounded in a similarity-based representation, much like humans. This bodes well for the intelligibility of a fully constrained network specifically trained to predict expert-identified features which then form the basis for classification, thus making the basis for classification inherently intelligible.

2.4. Fully feature-constrained network, using expert features

A fully constrained network was forced to use expert-identified feature values for the rock images by a phased process of transfer learning (Figure 1, bottom path). Using the penultimate layer of Resnet50, the network was first trained to predict feature values. In a second phase, a classifier network was trained to use these predicted features to classify the images.

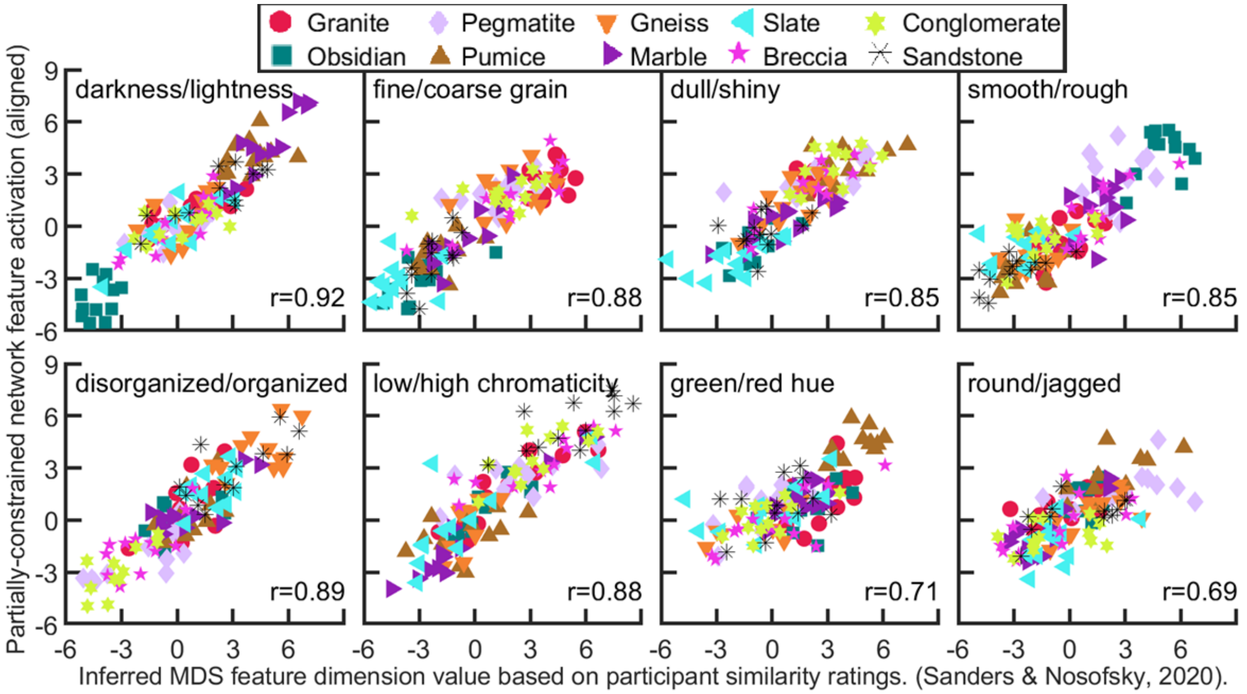

Fig. 2. For the partially feature-constrained network, classifier abstracted features are found to be an affine transform of the Sanders and Nosofsky (2020) inferred MDS feature dimension values based on naïve participants' similarity ratings. Each panel specifies the MDS feature dimension and the Pearson's r correlation coefficient.

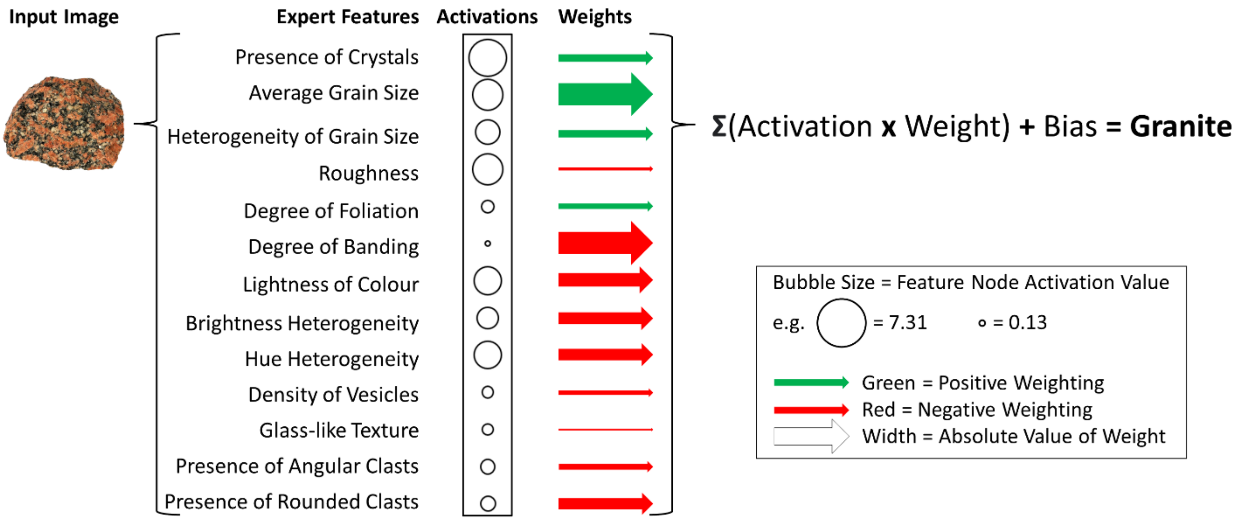

Fig. 3. Example categorisation of one instance of the granite images. Activations of the Resnet50 transfer-learned expert-abstracted features is represented by the surface area of a bubble. The magnitude of the weights applied to these feature activations are represented by the width of the arrows, green arrows indicating positive weights and red arrows negative weights.

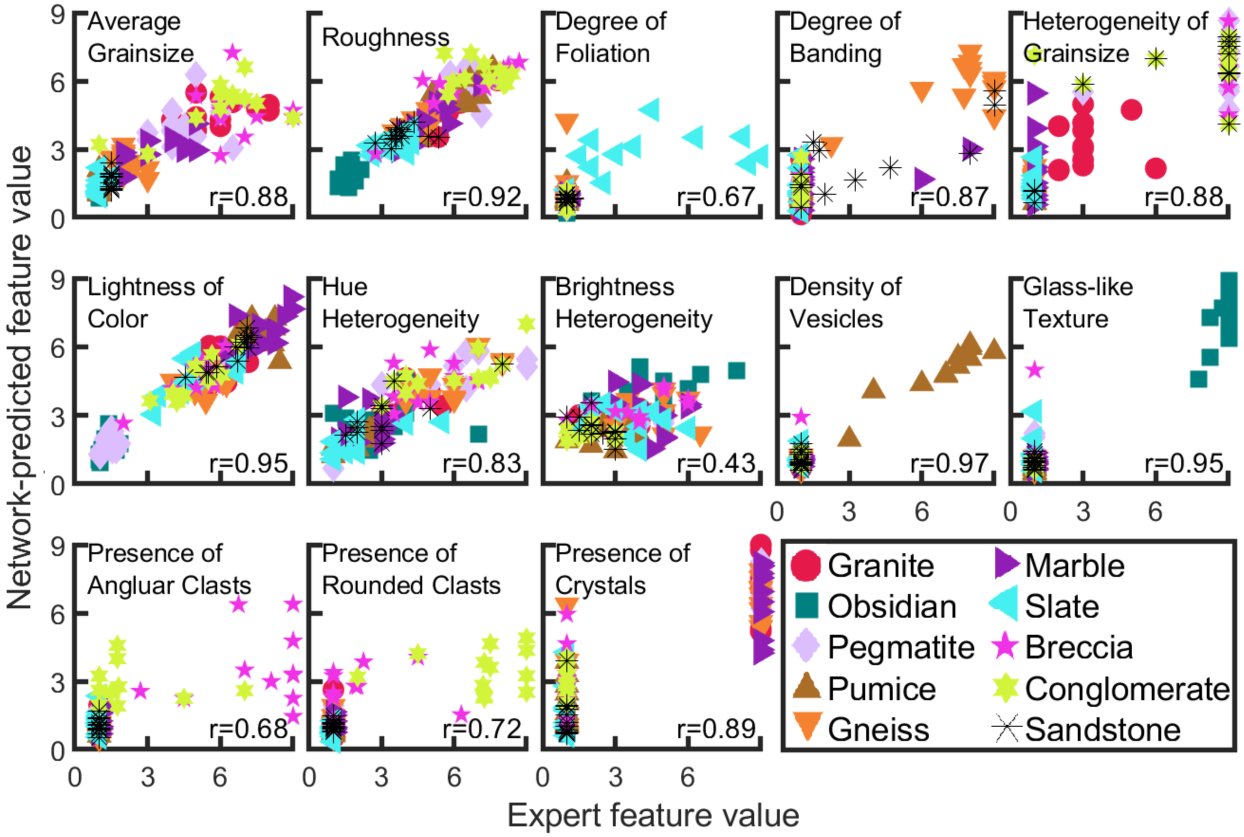

Fig. 4. For the expert-feature fully constrained network, the classifier's transfer-learned feature values are found to correlate well with expert-rated feature values. Each panel specifies the results for the expert feature and the Pearson's r correlation coefficient.

The expert features (presence of crystals, average grain size, heterogeneity of grain size, roughness, degree of foliation, degree of banding, lightness of colour, brightness heterogeneity, hue heterogeneity, density of vesicles, glass-like texture, presence of angular clasts and presence of rounded clasts) were rated for each image by the authors, informed by feature definitions by geologists. These identified features closely matched the expert-identified features from Nosofsky, Sanders and colleagues [16].

The key finding is that the network's validation performance was only 1.7% short, when forced to use expert-identified features (mean 85.1%, SD 0.7%), of the unconstrained network's performance (mean 86.8%, SD 0.8%). An illustration of how network-predicted feature values are combined together to categorise a rock image is provided in Figure 3. The magnitude of node activation is represented by the size of the feature circles and the magnitude of the weight between the features and the correct category is represented by the thickness of the arrows, with green and red arrows representing positive and negative weights, respectively. Expert feature values and network-predicted feature values were strongly correlated across all validation images, as illustrated in Figure 4.

3. Discussion

In an effort to demonstrate that inherently human-intelligible models are a viable alternative to black-box models, we constrained a DNN to first predict expert features in rock images before forcing the network to use only these predicted features as a basis for categorising the images. This fully feature-constrained network performed almost as well as transfer learning using Resnet50 to directly predict categories. A partially feature-constrained network was also constrained to condense down to a small number of abstracted features, but these features were unconstrained in that they were not trained to predict expert features. As such, these features were not intelligible by humans, but their feature space was found to be an affine transform of an MDS space from Nosofsky and Sanders [13], and thus could be made intelligible when rotated to align with human-relevant features. Our findings demonstrate that intelligibility of a DNN model is not necessarily achieved at any performance cost, consistent with the arguments presented by Rudin [2] and contrary to the broadly held view that such a trade-off between performance and intelligibility is necessary.

A limitation of this research is that it is focused solely on the use of features mutually intelligible by humans and AI and that it did not really consider in much detail how these features are combined together into category representations. Nosofsky, Sanders and McDaniel [14] have compellingly argued for an exemplar-based model for representing these rock category sub-types rather than prototype representation. This suggests that our use of a single-layer classifier that approximates a prototype representation is at best simplistic. It is also worth noting that Nosofsky and colleagues [17] clearly indicated that the superordinate rock categories (igneous, sedimentary and metamorphic) violate the family-resemblance principle that these categories can be coherently represented by best instances, i.e., prototypes. Ultimately, this implies that a fully intelligible AI may need to include explicit information about similarity to particular category instances in addition to the use of more fully articulated expert-specified rules based on intelligible features.

Success in achieving an AI which provides intrinsically explainable classification of rocks can be readily extended to other domains such as medicine or security. The assignment of instances (medical images or X-ray images of luggage) to categories (a disease or the presence of explosives) is a pervasive practical problem, involving discriminations of large numbers of images, often larger than human experts can reliably review. At minimum, an expert AI could provide a triage of what human experts should look at more closely and why. Ultimately, intrinsically intelligible AI needs to share a common conceptual framework with its users for mutual benefit and trust.

Acknowledgements

This work was funded by an internal grant from the Cardiff University Centre for Artificial Intelligence, Robotics and Human-Machine Systems (IROHMS) and conducted in the School of Psychology IROHMS Simulation Laboratory as part of the Human Factors Excellence (HuFEx) research group.

References

[1] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, “Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,” *Lect. Notes Comput. Sci.*, vol. 11700, p. 435, 2019.

[2] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[3] D. C. Elton, “Self-explaining ai as an alternative to interpretable ai,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12177 LNAI, pp. 95–106, 2020, doi: 10.1007/978-3-030-52152-3_10.

[4] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, “An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification,” *Expert Syst. Appl.*, vol. 128, pp. 84–95, 2019, doi: 10.1016/j.eswa.2019.01.048.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[6] J. D. Smith and J. P. Minda, “Thirty Categorization Results in Search of a Model,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 26, no. 1, pp. 3–27, 2000, doi: 10.1037/0278-7393.26.1.3.

[7] R. M. Nosofsky, “Attention, Similarity, and the Identification-Categorization Relationship,” *J. Exp. Psychol. Gen.*, vol. 115, no. 1, pp. 39–57, 1986, doi: 10.1037/0096-3445.115.1.39.

[8] D. L. Medin and M. M. Schaffer, “Context theory of classification learning,” *Psychol. Rev.*, vol. 85, no. 3, pp. 207–238, 1978, doi: 10.1037/0033-295X.85.3.207.

[9] J. K. Kruschke, “ALCOVE: An exemplar-based connectionist model of category learning,” *Psychological Review*, vol. 99, no. 1. pp. 22–44, 1992, doi: 10.1037/0033-295X.99.1.22.

[10] F. G. Ashby, L. A. Alfonso-Reese, A. U. Turken, and E. M. Waldron, “A Neuropsychological Theory of Multiple Systems in Category Learning,” *Psychol. Rev.*, vol. 105, no. 3, pp. 442–481, 1998, doi: 10.1037/0033-295X.105.3.442.

[11] M. A. Erickson and J. K. Kruschke, “Rules_Category_Learning.Pdf,” vol. 1996, no. November, pp. 1–62, 1996.

[12] J. C. Peterson, J. T. Abbott, and T. L. Griffiths, “Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations,” *Cogn. Sci.*, vol. 42, no. 8, pp. 2648–2669, 2018, doi: 10.1111/cogs.12670.

[13] C. A. Sanders and R. M. Nosofsky, “Training Deep Networks to Construct a Psychological Feature Space for a Natural-Object Category Domain,” *Comput. Brain Behav.*, vol. 3, no. 3, pp. 229–251, 2020, doi: 10.1007/s42113-020-00073-z.

[14] R. M. Nosofsky, C. A. Sanders, and M. A. McDaniel, “Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain,” *J. Exp. Psychol. Gen.*, vol. 147, no. 3, pp. 328–353, 2017, doi: 10.1037/xge0000369.

[15] T. Miyatsu, R. Gouravajhala, R. M. Nosofsky, and M. A. McDaniel, “Feature highlighting enhances learning of a complex natural-science category,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 45, no. 1, pp. 1–16, 2019, doi: 10.1037/xlm0000538.

[16] R. M. Nosofsky, C. A. Sanders, B. J. Meagher, and B. J. Douglas, “Toward the development of a feature-space representation for a complex natural category domain,” *Behav. Res. Methods*, vol. 50, no. 2, pp. 530–556, 2018, doi: 10.3758/s13428-017-0884-8.

[17] R. M. Nosofsky, C. A. Sanders, A. Gerdom, B. J. Douglas, and M. A. McDaniel, “On Learning Natural-Science Categories That Violate the Family-Resemblance Principle,” *Psychol. Sci.*, vol. 28, no. 1, pp. 104–114, 2017, doi: 10.1177/0956797616675636.