Confirmation trees: A simple strategy for producing hybrid intelligence

Frederik Andersen, Pantelis P. Analytis and Diana Verdes
Department of Business and Management, University of Southern Denmark

Kristian P. Lorenzen
Department of Electrical and Computer Engineering, Aarhus University

Julian Berger and Ralf H.J.M. Kurvers
Center for Adaptive Rationality, Max Planck Institute for Human Development

Abstract

Artificial agents now perform on par with or better than experts on several challenging decision-making tasks. People, however, remain reluctant to allow algorithms to make decisions on their behalf and legal constraints may prevent it altogether. How can we harness artificial intelligence, while maintaining trustworthiness and accountability? We propose *confirmation trees*, a decision-tree strategy for hybrid intelligence that can improve accuracy while maintaining human control. First, decisions are elicited from a human expert and an artificial agent. If they agree, that decision is adopted. If they disagree, a second human expert is consulted to break the tie. Hence, a human expert always approves the final decision. Our approach outperforms human experts or algorithms alone at diagnosing malign skin lesions. Crucially, it performs better than a strong human baseline, using substantially fewer human ratings. Our results show the potential of this approach for medical diagnostics and beyond.

Keywords: collective intelligence, hybrid intelligence, majority voting, decision trees, augmented intelligence, neural networks.

Introduction

Artificial intelligence (AI) and deep neural networks have entered our lives for good. Artificial agents can now accurately identify and categorize objects in images, produce images from text prompts, and even write code and identify bugs. One of the domains where deep neural networks hold the greatest promise to improve or even disrupt the current working routines is in diagnostic medicine. In several diagnostic tasks, deep neural networks have reached a level of performance comparable to or even better than that of seasoned experts (Topol, 2019; Marchetti et al., 2020; Haggenmüller et al., 2021; Hannun et al., 2019). At the same time, professionals and lay people are known to be reluctant to adopt decision-making algorithms in their daily routines (Dawes, Faust, & Meehl, 1989; Dietvorst, Simmons, & Massey, 2015), and there is currently a legal gap in attributing legal and moral responsibility to algorithms (Santoni de Sio & Mecacci, 2021; Grote & Berens, 2020). This creates a difficult puzzle: How can we harness the diagnostic capacities of deep neural networks, while keeping the final responsibility with human decision makers?

A promising way forward is to design decision-making processes that combine human decision makers and artificial agents. In the same way that combining the predictions—or abilities—of different human decision makers or models can boost decision accuracy (R. H. Kurvers et al., 2016; He, Analytis, & Bhatia, 2022; R. Kurvers et al., 2023), combining human decision makers and artificial agents (Grosz, 1996; Kamar, 2016) can produce hybrid intelligence, which can outperform both humans and artificial agents (Patel et al., 2019). A hybrid approach maintains human agency and control, rather than replacing humans with algorithms, and can thus help to overcome people's reluctance to rely on algorithms as well as potential legal constraints.

Here, we present *hybrid confirmation trees*, a simple sequential decision-making strategy that does exactly that (Figure 1). First, a prediction is elicited from a human expert and an artificial agent. If they agree, that decision is adopted. In cases of disagreement, a second human expert is consulted to break the tie and make the final decision. To ensure maximum independence between predictions, none of the decision agents (human or AI) have access to the predictions of the other. We show that hybrid confirmation trees have the potential to improve overall diagnostic performance in terms of the achieved true positive and false positive rates while substantially decreasing the overall decision-making cost compared to human confirmation trees. Importantly, this approach always has at least one human decision maker onboard approving a decision. Similar strategies have been documented in the context of hypothesis testing (Bruner & Austin, 1986) and been proposed as plausible and effective strategies in multi-attribute choice (2006). Confirmation trees are also a manifestation of majority decision-making (Hastie & Kameda, 2005), with the difference that when the first two agents agree, the third agent is redundant and thus, as in other frugal decision-making algorithms, pruned (Luan, Schooler, & Gigerenzer, 2011).

We test our approach in the context of melanoma classification. Skin cancer is one of the most common and aggressive forms of cancer, affecting more than five million people annually in the United States alone (Rogers, Weinstock, Feldman, & Coldiron, 2015). An early and correct diagnosis of melanoma is key for successful treatment and survival (Rigel & Carucci, 2000). As of 2017, algorithms based on convolutional neural networks (CNNs) achieved performance comparable to that of medical experts (Esteva et al., 2017; Tschandl et al., 2019; Hekler, Utikal, Enk, Berking, et al., 2019) for some diagnostic tasks, while the most recent studies indicate that CNNs can even outperform medical experts (Brinker, Hekler, Enk, et al., 2019; Haggenmüller et

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

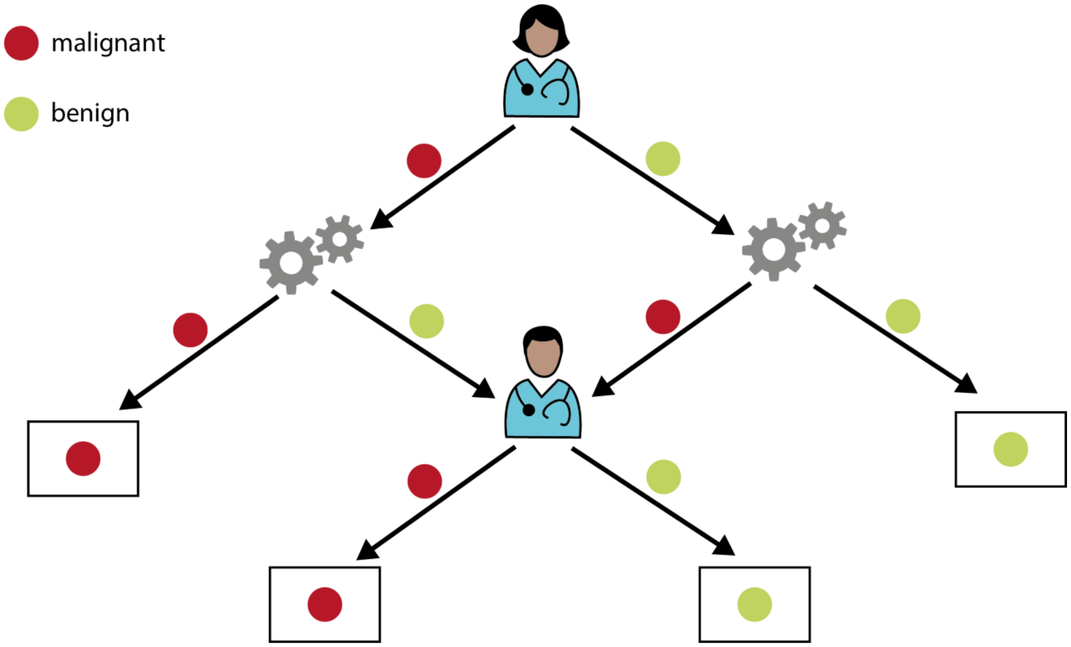

Figure 1: Visual representation of the confirmation tree decision process. The decision of a human expert is compared to the decision of an algorithm. In cases of agreement, that decision is adopted. In cases of disagreement, a second human expert breaks the tie. Boxes indicate the final decision (which is always supported by at least one human expert). Note that this is a condensed visualization of the tree structure. Both dis-confirmed branches should be fully expanded to conform with the tree structure in a graph-theoretical sense.

al., 2021). A few recent studies have examined the possibility of combining human and machine predictions, either by providing AI-generated advice to human experts (Han et al., 2020) or by training a meta-algorithm to optimally combine their predictions (Hekler, Utikal, Enk, Hauschild, et al., 2019), showing promising results. Our study contributes both to the cognitive sciences, by advancing a simple yet generic algorithm for combining the predictions of humans and artificial agents, as well as to the emerging medical literature on finding efficient decision-making processes for melanoma classification.

Methods and approach

Dataset and human experts

To assess the performance of individual medical experts, groups of medical experts, and human-algorithm hybrids, we sought a dataset where many doctors have made diagnoses using the same visual input. We identified the Melanoma Classification Benchmark (MClass) as the most suitable dataset (Brinker, Hekler, Hauschild, et al., 2019), and focused on doctors' diagnostic performance on the non-dermoscopic images in the dataset. The (openly available) data contains diagnoses rendered by 145 dermatologists who were presented with 100 images of skin lesions and asked to provide a management decision (treat/biopsy lesion or reassure the patient). The dermatologists were all practicing in hospitals at the time of data collection (42 individuals had <2 years of practical experience, 67 individuals between 2-12 years, and 36 individuals >12 years). The images were drawn from a large database of medical images to create a diverse data set. The images consisted of 80 benign (atypical nevi) and 20 malignant (melanoma) images. The ground truth of malignant lesions was verified through histopathological examination, and for benign lesions via a consensus protocol amongst experts.

Convolutional neural networks

To test the performance of a hybrid intelligence approach, we first identified an artificial agent that performs close to the current state-of-the-art (SOTA) in the same skin lesion classification tasks as that of the human experts. The AI technology that achieves SOTA performance in these tasks is convolutional neural networks (CNNs).

We made use of existing datasets to train our CNN for melanoma classification. Specifically, we used two that are publicly available on Kaggle: that of the 2019 International Skin Imaging Collaboration (ISIC) Challenge (Tschandl, Rosendahl, & Kittler, 2018; Codella et al., 2017; Combalia et al., 2019), and that of the 2020 Melanoma Classification Challenge organized by the Society for Imaging Informatics in Medicine (SIIM) and ISIC (Rotemberg et al., 2021).[1] Combined, they total 58,000 images that have been carefully selected and curated for computer-aided image classification. These data are highly suitable because training CNNs requires large, high-quality datasets to achieve good performance.

We identified and fine-tuned a SOTA CNN proposed by Chris Deotte. Deotte's model scored in the top 2% in the SIIM-ISIC Challenge and the training code and final weights of his model are publicly available. Unfortunately, the training set of Deotte's model overlaps with the images used in the MClass benchmark dataset, making it unsuitable for unbiased evaluation. Instead, we trained a model on the combined ISIC 2019 dataset and SIIM-ISIC 2020 dataset scrubbed of any images overlapping with the MClass benchmark dataset. In order to match Deotte's approach, the unbiased model is a finetuned EfficientNet B6 (Tan & Le, 2019) pretrained on the Imagenet dataset (Deng et al., 2009). The pretraining on

[1] For an overview of the competition see https://challenge.isic-archive.com/

Imagenet was done by Ross Wightman (2019). The model was trained using the Adam optimizer (Kingma & Ba, 2014) with default momentum parameters and the 1-cycle learning rate schedule (Smith & Topin, 2019).

In contrast to the binary decisions of human decision makers, CNNs produce a probability estimate (e.g. a 10% probability that a lesion is a melanoma). This allows us to modify the final categorization by adjusting the prediction threshold used to categorize images into malignant or benign tumors. For example, in some settings a low probability estimate of 10% may be deemed sufficient to categorize a lesion as malignant, whereas other settings might warrant a much higher threshold. Thus, instead of using the maximum likelihood threshold (at p >= 0.5) we consider the full family of classifiers obtained by varying the probability threshold of categorizing a tumour as malignant. Using such an analysis routine and calculating Area Under the Curve (AUC) metric (Brinker, Hekler, Enk, et al., 2019; Haenssle et al., 2018) we find that Deotte's model achieves an AUC score of 0.902.

Confirmation trees

In the *hybrid confirmation trees* approach, a decision is elicited from both a human expert and an algorithm. If they agree, that decision is accepted. If they disagree, a second human expert makes the final decision (Figure 1). Note that confirmation trees lead to the same decisions as 3-agent majoritarian decision-making and follow a long tradition that highlights the benefits of information aggregation (De Condorcet, 2014; Csaszar & Eggers, 2013; Hastie & Kameda, 2005; Herzog, Litvinova, Yahosseini, Tump, & Kurvers, 2019). Our approach has the key advantage that the branches of the decision tree are pruned once sufficient votes for a decision have been gathered, which reduces overall cost. Thus, similar to fast-and-frugal decision strategies and other simple decision-tree algorithms, part of the cost reduction comes from how the tree is structured (Christensen & Knudsen, 2010; Luan et al., 2011). A second important pathway for cost reduction stems from the fact that neural networks are extremely cheap once trained. Thus, part of the novelty comes from deciding where to place artificial agents in existing decision-tree algorithms. We also implemented the algorithm with only human decision makers as a baseline comparison.

Simulation procedures

We evaluate the performance of confirmation trees by sampling medical experts at random for the top and bottom node positions in the decision tree (Figure 1), and placing a decision from the CNN at the intermediate level. Because the binary CNN prediction depends on the threshold used, we repeated this process for each unique probability prediction value that the CNN generates over the set of images. This generates the prediction for each image in our test set, which we used to evaluate the true positive rate (TPR) and the false positive rate (FPR) for each threshold.

The simulations select the decisions of two new human experts for every image, remaining agnostic about the differences in skill among the doctors. Thus, the predictions fed into the confirmation tree strategy can be constructed in a multitude of ways (i.e. depending on who is currently available at the hospital) and can vary to some degree every time the algorithm is run. We ran the simulation 1,000 times for each threshold value and each image and averaged the results. We compared the performance of the confirmation tree against three baselines: i) The average individual performance of human experts; ii) the performance of the CNN on its own; and iii) a strong collective intelligence baseline, by running the *confirmation tree* strategy using only medical experts (i.e. replacing the intermediate AI node with a human expert).

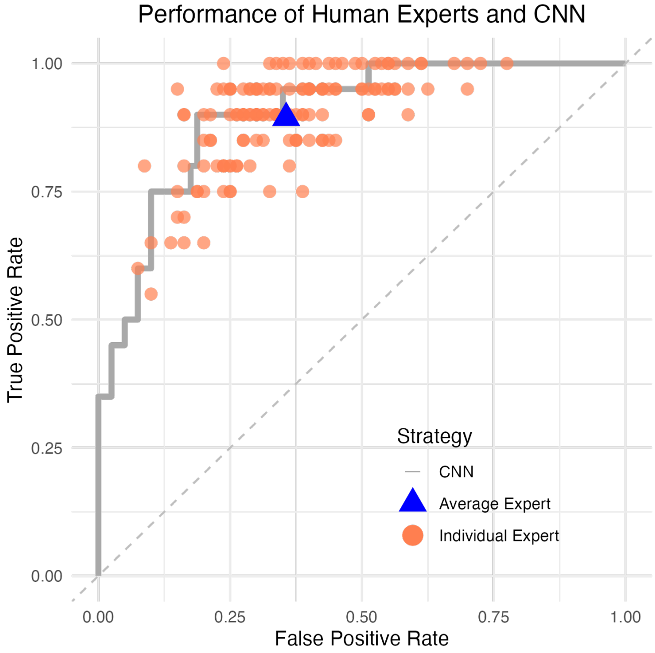

Figure 2: The true and false positive rate of the 145 medical experts (orange dots) and their average individual performance (blue triangle). The grey line shows the performance of the CNN model for different categorization thresholds.

Results

Medical expert performance

We first assess the performance of the 145 medical experts by comparing their diagnoses to the ground truth. There is substantial individual variability in the performance of medical experts in terms of both the achieved TPRs (range 0.55-1) and FPRs (range 0.075-0.775, Figure 2). The average individual performance across all medical experts amounts to a TPR of 0.894 and a FPR of 0.356. That is, if we were to sample a random medical expert, we would expect to correctly detect 89.4% of the melanomas (and miss 10.6%), and correctly reject 64.4% of the non-malignant cases (and in 35.6% of the cases predict a melanoma when there is none).

Medical experts vs. baseline CNN

One key difference between the human experts and the CNN model is that we can make the model more lenient or strict by manipulating the acceptance threshold, which is not possible for the binary responses of the medical experts. Thus, comparing medical experts and the CNNs on the same criteria is not straightforward. Nonetheless, we can compare the performance of medical experts to the ROC curve of the CNN (Figure 2). This visual inspection shows that a minority of the medical experts achieve a combined TPR and FPR that is better than the model's performance (i.e., above the grey CNN curve in the ROC-space), although many points reside on the curve. When investigating the average expert (blue triangle) we find the performance just below the curve with a TPR of 0.894 and a FPR of 0.356. The closest CNN results reside at a TPR of 0.900 and a FPR of 0.350, being marginally better.

Properties and performance of confirmation trees

Flexibility of hybrid confirmation trees The performance of this hybrid approach in terms of trading off TPR and FPR can be regulated by setting different thresholds—akin to the performance of the CNN itself. Figure 3 shows the performance of the confirmation trees for different thresholds (dots). The possible range in the trade-off space of the confirmation tree is smaller than that of the CNN. However, it is much larger than that of a human-only decision tree given the binary nature of the human decisions. Importantly, the possible combinations of TPRs and FPRs cover the region that is likely the most relevant for medical diagnostics.

Hybrid confirmation trees nest 2-person hierarchies and polyarchies For an acceptance threshold of 0 for the CNN (i.e. all skin lesions are categorized as malignant by the CNN), hybrid confirmation trees nest 2-person polyarchies (Sah & Stiglitz, 1988; Christensen & Knudsen, 2010). Polyarchies are a social fast-and-frugal decision-tree strategy where it suffices for one person to approve a binary categorization decision [see Martignon, Katsikopoulos, and Woike (2008), fast and frugal trees have a conclusive exit at each node, i.e. the first or second medical expert can make a conclusive decision and categorize a case as malignant]. At its polyarchy extreme, the hybrid confirmation tree achieves the highest TPR possible for the hybrid confirmation tree, but also the highest FPR (see the rightmost purple square in Figure 3 and the second to bottom row in Table 1).

For an acceptance threshold of 1 for the CNN, by contrast, hybrid confirmation trees nest 2-person hierarchies (Sah & Stiglitz, 1988), a social fast-and-frugal decision-tree strategy where both individuals need to agree for a malignant decision to occur. In other words, if they disagree, the case is categorized as benign, and the first expert can already categorize a case as benign on their own. At its hierarchy extreme, the hybrid confirmation tree achieves the lowest possible FPR, but at the cost of a low TPR (see the green square in Figure 3 and the bottom row in Table 1).

Comparison to medical experts and artificial agents Comparing the performance of the hybrid confirmation tree to that of single medical experts and the SOTA CNN, we observe that hybrid confirmation trees perform substantially better than the average medical expert (the colored dots versus the blue triangle in Figure 3), and achieve slightly better TPR and FPR combinations than the CNN (the colored dots versus the grey line in Figure 3). Finally, we compare the performance of the hybrid confirmation trees to the performance of human confirmation trees (in which the intermediate AI nodes in Figure 1 are replaced by a human rater). Crucially, hybrid confirmation trees perform slightly better than human confirmation trees (the colored dots versus the red triangle in Figure 3). For example, the human confirmation tree achieves an average TPR of 0.930 and a FPR of 0.336. At roughly the same level of TPR, the hybrid confirmation tree achieves a FPR of 0.322 which is slightly lower (better) than that of the human confirmation tree. Importantly, this can be achieved at a much lower cost as we will discuss next.

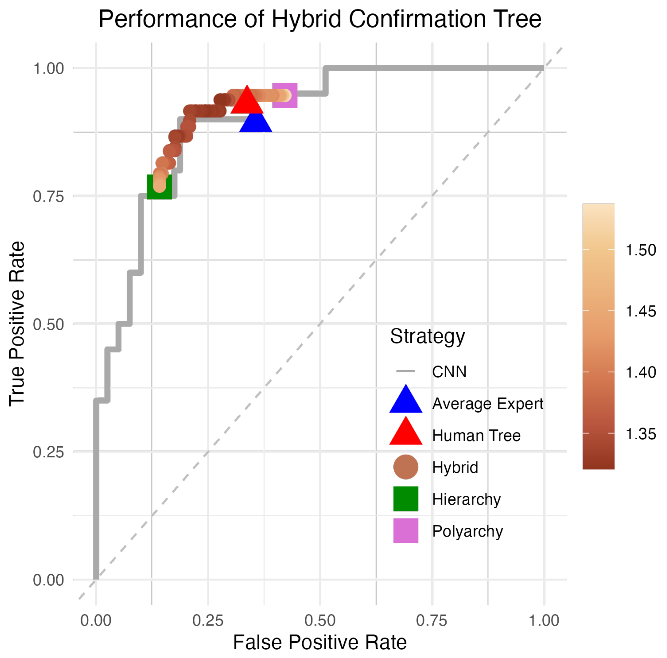

Figure 3: Performance of the confirmation tree (dots), the average expert performance (blue-triangle), the human confirmation tree (red triangle), the nested models of both the 2-person hierarchy (green square) and 2-person polyarchy (purple square), and the CNN (grey line). The light- to dark-brown dots represent the performance of our confirmation tree algorithm for each threshold, with the coloration indicating the average number of raters consulted.

Frugality of confirmation trees Using a CNN instead of a human decision maker in the decision tree allows for a substantial reduction in the number of human raters without any loss of performance. Figure 3 shows the average number of raters the hybrid confirmation tree used for each threshold. Figure 4 presents this in more detail, showing the likelihood of eliciting a second human rater per threshold. We compare the frugality of the hybrid confirmation tree to a human confirmation tree by calculating the difference in the number of human raters used at the threshold values of the hybrid confirmation tree that match the TPR or FPR of the human confirmation tree. The human confirmation tree achieves a TPR of 0.930 and a FPR of 0.337 using 2.27 raters (Table 1). When

Strategy type	Strategy	TPR	FPR	Cost	Agreement rate	Threshold (θ)
Human baselines	Average medical expert	0.894	0.356	1	-	-
	Human confirmation tree	0.930	0.337	2.27	0.733	-
Hybrid trees	Minimum cost hybrid tree	0.938	0.277	1.32	0.679	0.0163
	Matching FPR of human tree	0.946	0.334	1.40	0.603	0.00386
	Matching TPR of human tree	0.941	0.322	1.39	0.608	0.00635
Nested models	2-person Polyarchy (nested for $\theta = 0$)	0.946	0.422	1.54	0.463	0
	2-person Hierarchy (nested for $\theta = 1$)	0.768	0.142	1.46	0.537	1

Table 1: Overview of the FPR, TPR, cost (i.e., average number of human raters consulted), and agreement between the first two decision-making agents (either human or CNN) for different decision strategies. For the hybrid confirmation trees we selected CNN thresholds which either minimized the cost, or matched the FPR (or TPR) performance of the human confirmation tree.

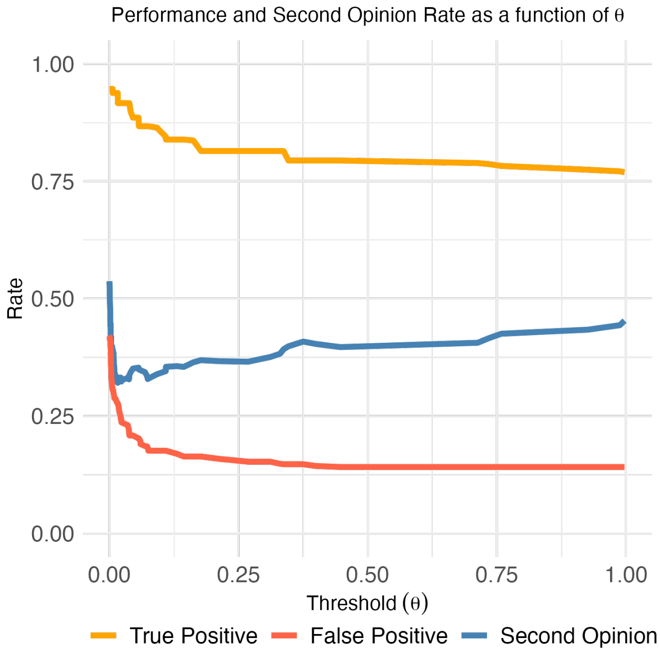

Figure 4: The TPR (yellow line) and FPR (red line) achieved by the hybrid confirmation tree for different threshold values (x-axis). The blue line shows the likelihood that a second human rater is consulted.

we match that TPR with the hybrid confirmation tree, we only need 1.39 human raters. Likewise, when matching the FPR, we only need 1.40 human raters (Table 1). Thus, to match the performance of the human confirmation tree, the hybrid confirmation tree reduces the number of human raters by 0.88. Hybrid confirmation trees are also substantially more frugal than 2-person social fast and frugal tees (polyarchies and hierarchies, see bottom rows of Table 1).

Agreement rate How do the human and the hybrid confirmation trees fare in terms of the rates of agreement between the first and the second medical experts, and medical experts and the CNN? Medical experts have an overall agreement rate of 0.733 (i.e. when randomly sampling two expert decisions for a particular image, they agree in 73.3% of cases). When matching for the TPR and FPR of the human confirmation tree, for example, the hybrid tree exhibits agreement rates between the first medical expert and the CNN of 0.608 and 0.603 respectively. This lower agreement rate could explain why the hybrid approach works so well. The medical experts and the CNN may rely on partly different informational cues, and therefore their judgments may be more independent and complementary as compared to between-human judgments. Note that even a hybrid confirmation tree that minimizes the decision-making cost has a lower agreement rate than the human confirmation tree (see Table 1 and the blue line in Figure 4), providing evidence that medical experts and the CNN are relatively independent.

General Discussion

We investigated *hybrid confirmation trees*, a simple approach for producing hybrid intelligence and improving the prediction rates in a challenging diagnostic task. Our approach comes at a low decision-making cost, while maintaining human agency and control. It can be grounded in the language of decision trees, and it is expressive enough to nest some well-established decision processes as boundary conditions.

Alternative examples of hybrid intelligence design One of the main approaches advanced in decision support is for artificial agents to act as advisers or recommenders by providing their predictions to human decision makers, who then make the final decision (Lai & Tan, 2019). This approach has shown some promise in improving diagnostic performance (Groh, Epstein, Firestone, & Picard, 2022; Han et al., 2020). However, it comes with limitations. First, people might lack the ability to determine when the artificial agent is more knowledgeable and should be followed (Fügener, Grahl, Gupta, & Ketter, 2022). Second, human decision makers might be influenced by the artificial agents, and therefore contribute less new information to the decision-making process than when producing their own judgment independently of the AI adviser (Fügener, Grahl, Gupta, & Ketter, 2021). And, in a worst-case scenario, they may even become less attentive and gradually lose their skills on specific tasks, a process known as deskilling (Parasuraman & Riley, 1997). Researchers have been exploring ways to address some of these limitations, for example by providing explanations of the AI adviser scheme to help humans build meta-knowledge about the artificial agents (Bansal et al., 2019; Mozannar, Satyanarayan, & Sontag, 2022). In the near future we aim to directly compare hybrid confirmation trees and the "AI as adviser scheme" in terms of performance.

Circumventing algorithmic aversion Going beyond performance, however, people have been particularly reluctant to adopt the predictions of algorithms for expert tasks, even when there has been substantial evidence that they could lead to superior performance. This holds true both for seasoned professionals who could use actuarial decision algorithms to improve their own diagnostic performance (Dawes et al., 1989) and end-consumers of algorithms (i.e. patients, investors) who appear unwilling to adopt algorithmic predictions (Dietvorst et al., 2015) and are less forgiving when they observe algorithms to err (Dietvorst & Bharti, 2020) than they are to humans. We believe that hybrid confirmation trees can largely circumvent these issues by (i) having professionals and artificial agents provide independent judgments and (ii) always having a human decision maker onboard approving a decision. Thus, future comparisons should not only be in terms of performance, but also in terms of people's willingness to adopt these hybrid decision-making procedures.

CNN potential and hybrid confirmation tree performance In the coming years, CNNs (and other AI models) are bound to get better as the relevant neural network architectures and training methods improve, there are more data available to train the models (i.e. more images and metadata), and the cost of computation further decreases. Thus, we expect that in many tasks these models will eventually outperform most human experts in visual tasks. In many scenarios, using hybrid confirmation trees would likely achieve better diagnostic outcomes than the average medical expert or the human confirmation tree baseline, but worse outcomes than CNNs. Nevertheless, using hybrid confirmation trees might be advisable for avoiding deskilling, addressing algorithm aversion or for legal reasons.

For now, there are still cases for which CNNs (and other AI models) perform worse than human experts, especially when these models have not been trained appropriately. For example, CNNs' diagnostic performance is worse than that of most medical experts when standard models are evaluated on people with diverse skin colours (Daneshjou et al., 2022). In scenarios where CNNs are lagging behind the average human expert, could it be reasonable to deploy hybrid confirmation trees (as opposed to human confirmation trees)? Even in such cases, hybrid confirmation trees could improve diagnostic outcomes if AI models bring about new and relatively independent information, compared to that of another human expert.

An important task for future research is to outline the conditions under which hybrid confirmation trees outperform their constituent parts (human experts and AI alike), and identify settings where we would expect them to outperform only the human experts.

Applications to other domains We demonstrated the benefits of our approach in a visual diagnostic task in dermatology. Hybrid confirmation trees, however, are generic and can be applied in other cutting-edge domains where CNNs and other algorithms have been recently deployed in medicine and beyond, for example in screening for breast cancer (McKinney et al., 2020), predicting recidivism (Dressel & Farid, 2018) and deciding when to release people after trial (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018), in detecting misinformation (Marcellino et al., 2020), or in identifying deepfake videos (Groh et al., 2022). Hybrid confirmation trees can be directly deployed in these scenarios by choosing the SOTA algorithm for these specific domains as the artificial agent. Such applications constitute promising avenues for future research.

Acknowledgments

We thank the members of the Strategic Organization Design group for their valuable comments, Iris Zalaudek and Giuseppe Argenziano for sharing a dataset of dermoscopic images and diagnostic decisions that we used earlier to develop our approach, Pietro L. Nickl for designing Figure 1, and K. Rhett Nichols for editing the manuscript.

References

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 7, pp. 2–11).

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., ... others (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, *113*, 47–54.

Brinker, T. J., Hekler, A., Hauschild, A., Berking, C., Schilling, B., Enk, A. H., ... others (2019). Comparing artificial intelligence algorithms to 157 german dermatologists: the melanoma classification benchmark. *European Journal of Cancer*, *111*, 30–37.

Bruner, J. S., & Austin, G. A. (1986). *A study of thinking*. Transaction publishers.

Christensen, M., & Knudsen, T. (2010). Design of decision-making organizations. *Management Science*, *56*(1), 71–89.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... others (2017). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1710.05006*.

Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., ... others (2019). Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

Csaszar, F. A., & Eggers, J. (2013). Organizational decision making: An information aggregation view. *Management Science*, *59*(10), 2257–2277.

Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., ... others (2022). Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, *8*(31), eabq6147.

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674.

De Condorcet, N. (2014). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? merits and pitfalls of working with ai. *Management Information Systems Quarterly (MISQ)-Vol*, *45*.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. *Information Systems Research*, *33*(2), 678–696.

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1), e2110013119.

Grosz, B. J. (1996). Collaborative systems (aaai-94 presidential address). *AI magazine*, *17*(2), 67–67.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, *46*(3), 205–211.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., … others (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, *29*(8), 1836–1842.

Haggenmüller, S., Maron, R. C., Hekler, A., Utikal, J. S., Barata, C., Barnhill, R. L., … others (2021). Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer*, *156*, 202–216.

Han, S. S., Park, I., Chang, S. E., Lim, W., Kim, M. S., Park, G. H., … Na, J.-I. (2020). Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology*, *140*(9), 1753–1761.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, *25*(1), 65–69.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological review*, *112*(2), 494.

He, L., Analytis, P. P., & Bhatia, S. (2022). The wisdom of model crowds. *Management Science*, *68*(5), 3635–3659.

Hekler, A., Utikal, J. S., Enk, A. H., Berking, C., Klode, J., Schadendorf, D., … others (2019). Pathologist-level classification of histopathological melanoma images with deep neural networks. *European Journal of Cancer*, *115*, 79–83.

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., … others (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, *120*, 114–121.

Herzog, S. M., Litvinova, A., Yahosseini, K. S., Tump, A. N., & Kurvers, R. H. (2019). 13 the ecological rationality of the wisdom of crowds. *Taming uncertainty*, *245*.

Kamar, E. (2016). Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *Ijcai* (pp. 4070–4073).

Karelaia, N. (2006). Thirst for confirmation in multi-attribute choice: Does search for consistency impair decision performance? *Organizational Behavior and Human Decision Processes*, *100*(1), 128–143.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Kurvers, R., Nuzzolese, A. G., Alessandro, R., Barabucci, G., Herzog, S., et al. (2023). Automating hybrid collective intelligence in open-ended medical diagnostics.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., … Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 29–38).

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*(2), 316.

Marcellino, W., Cox, K., Galai, K., Slapakova, L., Jaycocks, A., & Harris, R. (2020). *Human-machine detection of online-based malign information.* RAND.

Marchetti, M. A., Liopyris, K., Dusza, S. W., Codella, N. C., Gutman, D. A., Helba, B., … others (2020). Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: Results of the international skin imaging collaboration 2017. *Journal of the American Academy of Dermatology*, *82*(3), 622–627.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*(6), 352–361.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., … others (2020). International evaluation of an ai system for breast cancer screening. *Nature*, *577*(7788), 89–94.

Mozannar, H., Satyanarayan, A., & Sontag, D. (2022). Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 5323–5331).

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., … others (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine*, *2*(1), 1–10.

Rigel, D. S., & Carucci, J. A. (2000). Malignant melanoma: prevention, early detection, and treatment in the 21st century. *CA: a cancer journal for clinicians*, *50*(4), 215–236.

Rogers, H. W., Weinstock, M. A., Feldman, S. R., & Coldiron, B. M. (2015, 10). Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the US Population, 2012. *JAMA Dermatology*, *151*(10), 1081-1086.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., … others (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, *8*(1), 34.

Sah, R. K., & Stiglitz, J. E. (1988). Committees, hierarchies and polyarchies. *The Economic Journal*, *98*(391), 451–470.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 1–28.

Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications* (Vol. 11006, pp. 369–386).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., … others (2019). Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA dermatology*, *155*(1), 58–65.

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, *5*(1), 1–9.

Wightman, R. (2019). *Pytorch image models.* https://github.com/rwightman/pytorch-image-models. GitHub.