9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Confirmation trees: A simple strategy for producing hybrid intelligence

Keywords: hybrid intelligence, neural networks, decision trees, collective intelligence, medical decision-making

Introdution Deep neural networks have the potential 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 [3]. At the same time, however, professionals and lay people have both been reluctant to adopt algorithms in their daily routines [2, 1]. 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. Here, we present hybrid confirmation trees, a simple sequential decision-making strategy that does exactly that. 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 is consulted to break the tie and make the final decision (see Figure 1). None of the decision agents (human or AI) have access to the predictions of the others, to ensure maximum independence between predictions. We show that hybrid confirmation trees have the potential to improve the 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. Furthermore, this approach always has at least a human decision-maker on-board approving a decision.

Dataset and human experts To assess the performance of individual medical experts, groups of medical experts, and human-algorithm hybrids, we used dermoscopic images and diagnostic decisions by medical experts from a study by Zalaudek and colleagues [4]. In that study, diagnosticians were presented with 165 dermoscopic images of skin lesions, and asked to decide whether the images were benign or malignant. The images were drawn from a large database of medical images to create a diverse data set and consisted of 116 benign and 49 malignant images (incidence rate of 29.7 %), where the first fifteen images were used as training cases to familiarize diagnosticians with the task (we thus limited our analyses to the remaining 150 test images). The diagnosticians in the study varied in their function and level of dermoscopic experience: most were dermatologists, but a few were general physicians or other medical experts (see also Figure 3). The ground truth of each skin lesion was verified through histopathological examination.

AI decision-makers The AI technology that achieves state-of-the-art (SOTA) performance in detecting skin lesions is that of convolutional neural networks (CNNs). In contrast to the binary decisions of human decision makers, CNNs produce a probability estimate (e.g. 10 % probable that a lesion is a melanoma). This allows for some flexibility in the final categorizations, as it is possible to modify the decisions by adjusting the prediction threshold used to categorize images into malignant or benign tumors, but also calls for a different method for evaluatng performance. We used receiving operating characteristic (ROC) curve analysis and calculated the area under the curve for several contender CNNs. We identified a CNN originally trained by Chris Deotte, which also performed exceptionally well at the International Skin Imaging Collaboration (ISIC) Challenges of 2019 and 2020, as the best performing CNN in our dataset. We thus used this CNN as our artificial agent in the confirmation trees.

Simulating confirmation trees We simulated the performance of confirmation trees by sampling medical experts at random for the top and bottom node positions in the decision tree (remaining agnostic about the differences in skill among the doctors). Because the binary CNN prediction depends on the threshold that is used, we repeated this process for each of the 52 thresholds that were identified as inflection points in the receiving operating characteristic (ROC) curve analysis of the CNN. 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).

Results 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 orange dots versus the blue dot in Figure 4), and achieve better true positive rate (TPR) and and false positive rate (FPR) combinations than the CNN algorithm (the orange dots versus the red line in Figure 4). Further, hybrid confirmation trees offer more flexibility than human confirmation trees, as it is possible to manipulate the threshold value use to categorize tumors to achieve a range of FPR and TPR combinations. Hybrid confirmation trees encompass the performance of human confirmation trees within that space of possibililities (orange dots versus the green dot in Figure 4). For example, the human confirmation tree achieves an average TPR of 0.787 and a FPR of 0.15. At roughly the same level of TPR, a hybrid confirmation tree can achieve a FPR of 0.139 which is slightly lower (better) than that of the human confirmation tree. Importantly, improved performance can be achieved at a lower cost. To demonstrate that, 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 for which the hybrid confirmation tree can match the true positive or false positive rates of the human confirmation tree. The human confirmation tree achieves a TPR of 0.787 and a FPR of 0.150 using 2.24 raters (see Table 1 and Figure 4). When we match that TPR with the hybrid confirmation tree, we only have to engage 1.30 human experts. Likewise, when matching the FPR, we again have to engage only 1.30 human raters (see Table 1 and Figure 4).

Conclusion Overall, our results demonstrate that hybrid confirmation trees have the potential to improve diagnostic performance, while at the same time improving flexibility, reducing the cost of decision-making and maintaining responsibility with human decision-makers.

References

[1] Robyn Dawes, David Faust, and Paul Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

[2] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[3] Eric Topol. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

[4] I Zalaudek, G Argenziano, HP Soyer, Roberto Corona, F Sera, A Blum, RP Braun, H Cabo, G Ferrara, AW Kopf, et al. Three-point checklist of dermoscopy: an open internet study. *British journal of dermatology*, 154(3):431–437, 2006.

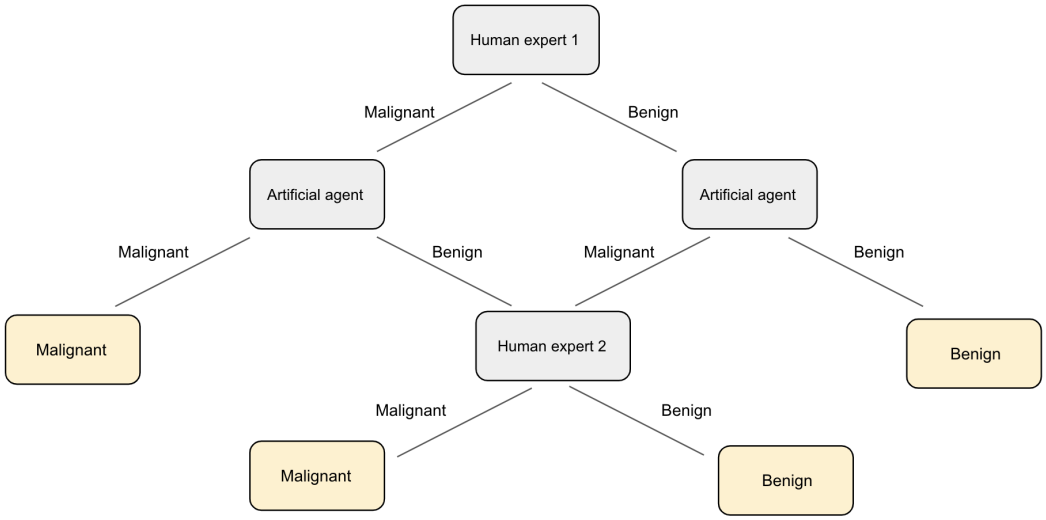

Figure 1: Visual representation of the confirmation tree decision process. The decision of a human expert is compared to the decision of an algorithm. In cases of agreement, that decision is adopted. In cases of disagreement, a second human expert breaks the tie. Yellow boxes indicate the final decision (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 tree structure in a graph-theoretical sense

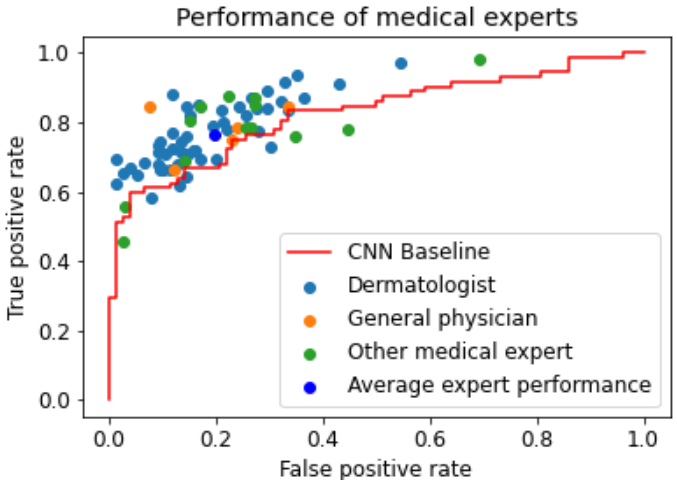

Figure 2: The true and false positive rate of the 69 medical experts (shown per profession) and their average individual performance (dark blue dot). The red line shows the performance achieved by Chris Deotte's CNN for different thresholds.

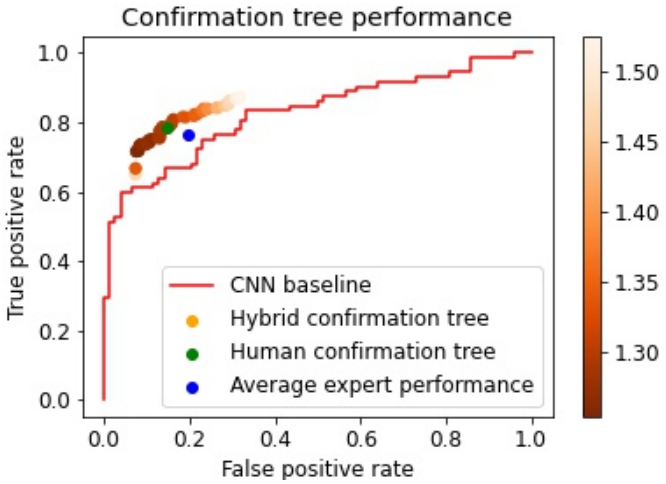

Figure 3: Performance of the confirmation tree (white-to-orange- colored dots), the average individual performance (blue dot), the CNN (red line), and, finally, a human confirmation tree (green dot). The 52 orange dots represent the performance of our confirmation tree algorithm for each threshold, with the coloration indicating the average number of raters consulted.

Strategy type	Strategy	TPR	FPR	Cost	θ
2*Human baselines	Average medical expert	0.765	0.196	1	-
	Human confirmation tree	0.787	0.150	2.24	-
3*Hybrid trees	Minimum cost hybrid tree	0.729	0.083	1.25	0.0693
	Matching FPR of human tree	0.789	0.153	1.30	0.0358
	Matching TPR of human tree	0.785	0.139	1.30	0.0365

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.

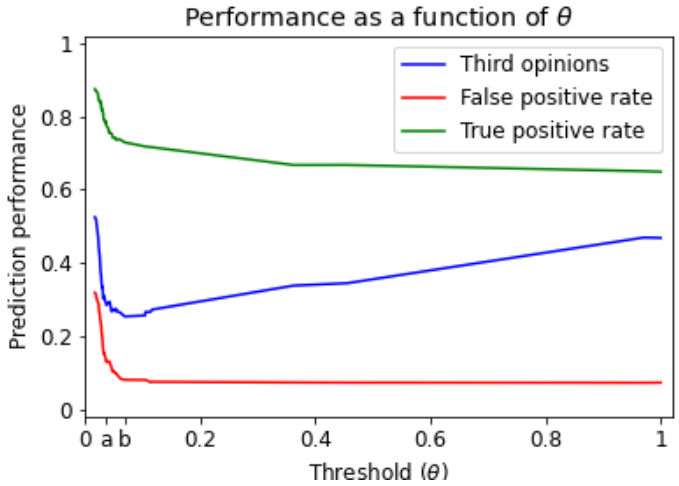

Figure 4: he TPR (green line) and FPR (red line) achieved by the hybrid confirmation tree for different threshold values (x-axis). The blue line shows the likelihood that a second human rater is consulted. The letters on the x-axis represent the thresholds corresponding to matching the TPR and FPR of the human tree (a), or the minimum cost tree (b) (see Table 1)