Supplementary Material

1. Evaluation Based on Observer Consensus Regions

To analyse the algorithm performance in isolation from interobserver differences we repeat the ROC analysis from the main article, this time using only regions where there is consensus between the observers for evaluation. For clarity, it is noted that the training data remained unchanged for this experiment, only the evaluation procedure was altered. The results of this consensus analysis are shown in Figure 1. The median line from the previous result, as well as the previous performance of subject 8 are overlaid on this graph for reference. It is clear that the algorithm performance on subject 8 is markedly improved, as well as the overall median performance, which is to be expected since the excluded regions (non-consensus) are likely to be the most difficult to label correctly. Having eliminated the possibility of algorithm 'error' which coincides with observer disagreement, we now look at the subjects where the algorithm has poorest performance on the consensus data, in terms of sensitivity and specificity on the binary result. (Subject 14 continues to illustrate the consistently best curve in the highest specificity ranges in this experiment).

The worst result in terms of sensitivity at is obtained for subject 5. (sensitivity=0.36, specificity=0.998). The curve for this subject continues to demonstrate worst performance, or second from worst at lower specificity levels. Interobserver agreement was moderate for this subject with sensitivity=0.87 and specificity=0.96. In terms of specificity the worst performance was on subject 2 (sensitivity=0.94, specificity=0.90), although the curve for that subject is close to the median performance overall. For this subject interobserver sensitivity is 0.82 and specificity is 0.99. Results for a single slice from each of subjects 2 and 5 are shown in Figure 2.

Subject 5 (lowest sensitivity) showed a pattern of mainly small and diffuse ischemic regions, according to observer annotations, which were difficult for the algorithm to identify, possibly because such small and/or weakly contrasting lesions were not well represented in other data the algorithm had been trained on. As an example, the green region marked by both observers, and denoted by an arrow in Figure 2 has a median ADC value of 0.99x$10^{-3}mm^2$/sec, which is relatively low, but also not atypical for healthy tissue in the cortical region. For subject 2 (worst specificity) the algorithm segmented substantial quantities of tissue in the white matter which were not identified by either observer. (It should be noted that the algorithm results are more similar to the observers at higher probability thresholds, t_{prob}.) Although it is not easily visible on the

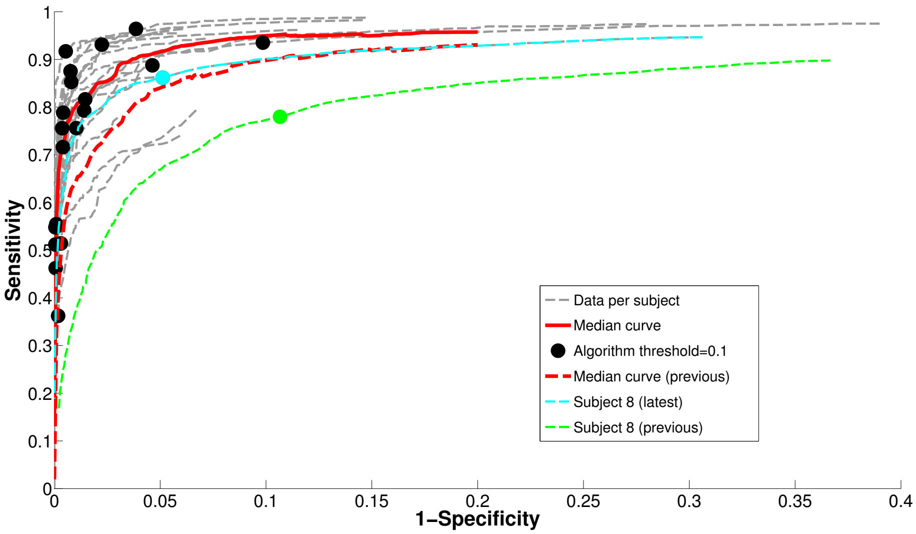

Figure 1: ROC curves for the algorithm performance considering only regions where the two observers have consensus. For comparison, we also plot the median line and subject 8 result from the previous analysis against observer 1 only (see ROC analysis in main article)

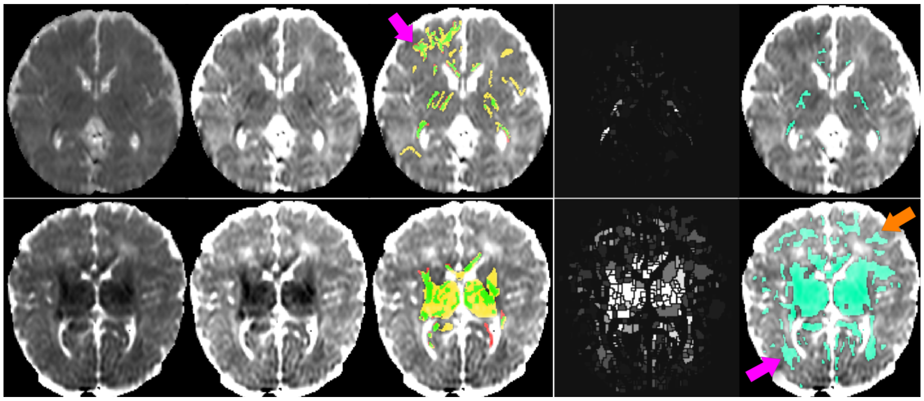

Figure 2: Examples from the scans where the algorithm has worst performance according to evaluation on consensus data. Top Row: Subject 5. Bottom Row: Subject 2. In each case a single slice from subject showing from left to right: 1) and 2) The ADC map seen with two different brightness and contrast settings. 3) The observer annotations (red=observer 1 only, yellow=observer 2 only, green=agreement). 4) The probabalistic outcome from the algorithm. 5) The final binary result from the algorithm at threshold $t_{prob} = 0.1$. Regions denoted by arrows are discussed in the text.

ADC map, since the basal ganglia injury is very dominant, the white matter in this subject showed unusually low ADC values throughout the image. For example, the regions detected by the algorithm, indicated by arrows in Figure 2, have median ADC values of $0.87\text{x}10^{-3}mm^2$/sec (pink arrow, left occipital) and

0.95x10$^{-3}mm^2$/sec (orange arrow, right frontal), which are exceptionally low values for this area, and notably lower than the region marked by both observers in subject 5 (Figure 2, upper row). Restrospective analysis by expert clinicians revealed that there was, in fact, white-matter injury in the regions denoted by the algorithm. The affected regions were overlooked by both observers due to the dominance of the basal-ganglia/thalamus injury which made it difficult to visualise the abnormally low ADC values elsewhere. This example illustrates the importance of the type of exhaustive, objective and quantitative analysis which can be provided by an automated system.