

The graph displays the performance of 18 different models over a period of 140 time units. The models are grouped into three categories: 'raw' (noisy lines), 'avg' (smoother lines), and 'thrs' (threshold lines). The performance generally fluctuates around a baseline of 0.5, with a notable sharp decline for all models starting around time 120 and reaching a minimum near time 130.

