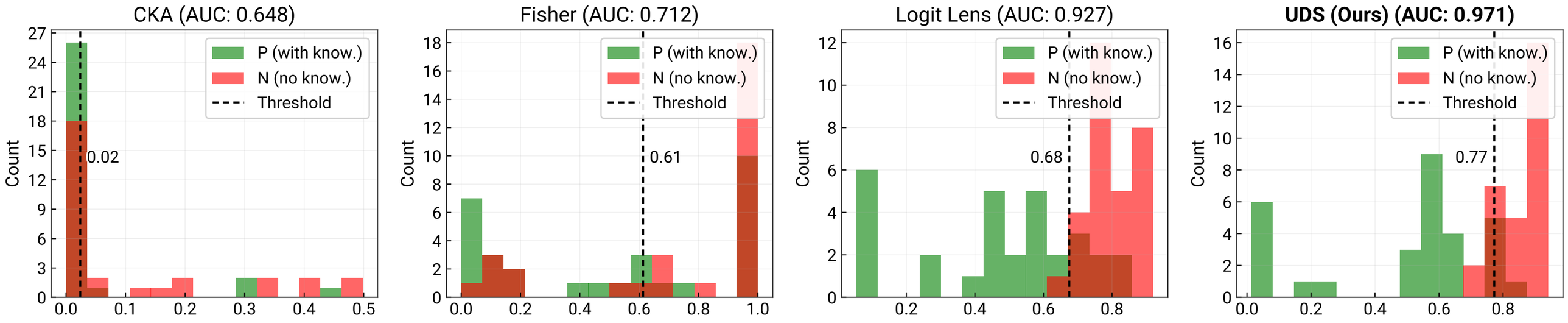
CKA (AUC: 0.648)
Fisher (AUC: 0.712)
Logit Lens (AUC: 0.927)
UDS (Ours) (AUC: 0.971)
P (with know.)
N (no know.)
Threshold
Count
0.02
0.61
0.68
0.77