Supplementary Information

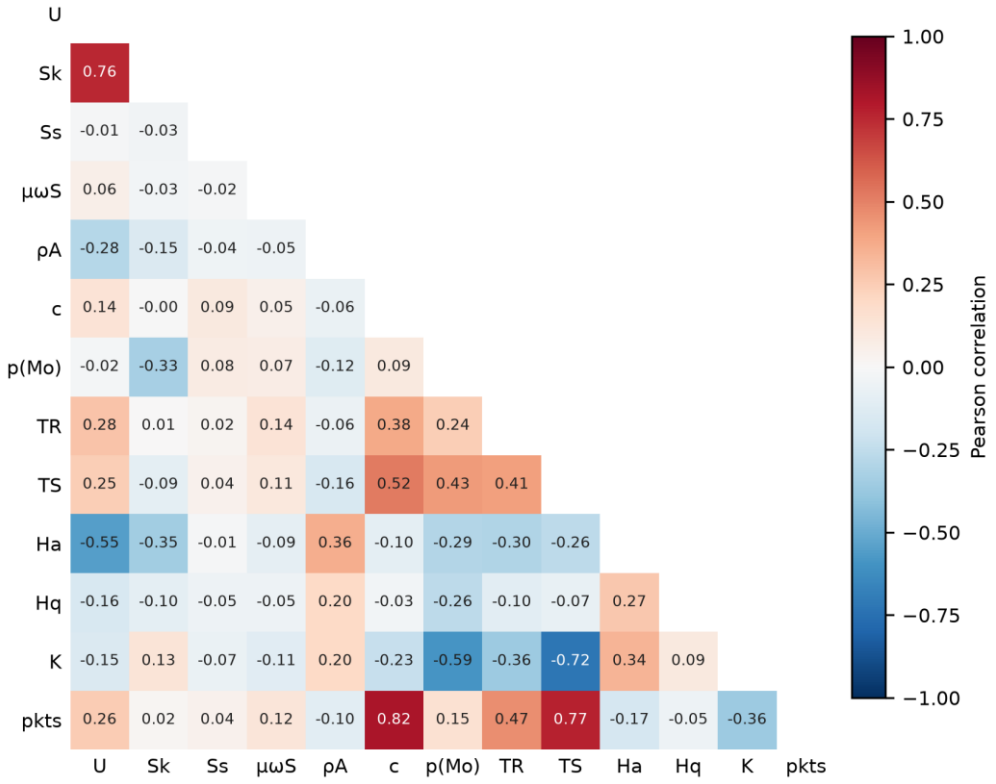
Cascaded Deep Stacking of State-Space and Tabular Deep Models for Detecting and Attributing Network Covert Timing Channels

Mohammed Youssef, Mohammed Al Khatib and Fahman Saeed

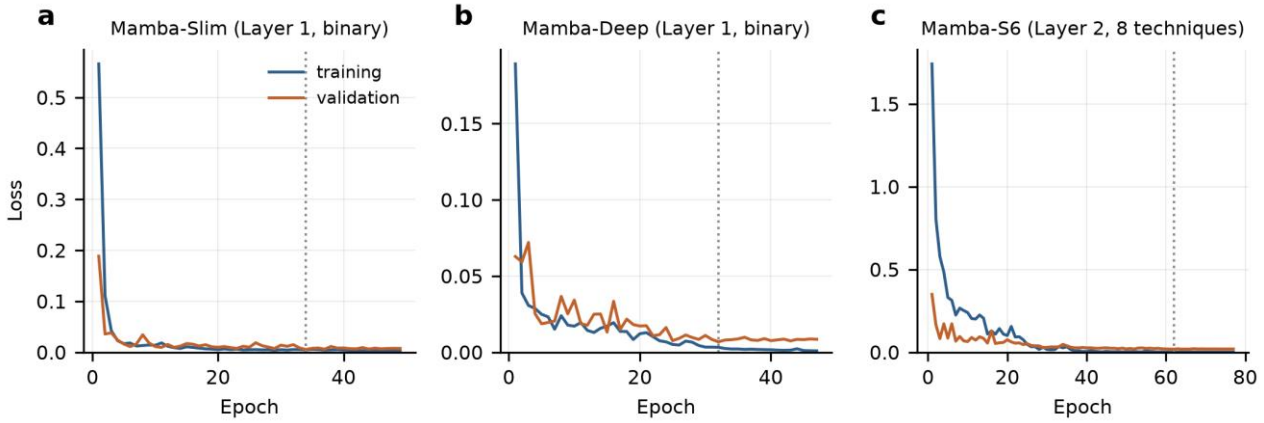
Contents: Supplementary Tables S1-S2, Figures S1-S7 and Note 1, placed in the order the main text first cites them.

Supplementary Table S1. The three public datasets released by the Iglesias group at TU Wien and the studies that report results on them, as reported by each study. Rows on different datasets are not comparable. The last two rows report results on the same ten-technique set of flows, which contains two CTC techniques (ED1, ED2) that neither study used for training. To our knowledge no study by other authors uses the feat_analysis set.

Dataset (TU Wien release)	Study	CTC techniques	Features	Method	Task and split	Reported result
ARES-2017 (outlier_det)	Iglesias & Zseby 2017 (ARES)	7 (no GIF)	8 DAT features; 5 for the Bayesian network	Outlier detectors (COF, INFLO, HBOS); Bayesian network	Outlier ranking; one flat 8-class classifier (7 techniques + overt), random 80/20 split	Covert flows not separable as outliers; per-technique test recall 60.71-100%
ARES-2017 (outlier_det)	Mao et al. 2026 (ICCCS)	7	Features of the released set	Transformer encoder, LSTM and additive attention; nine baselines	Binary, random 80:20 split	Accuracy 99.51%, precision 94.14%, recall 96.76%, F1 95.43%
CD-MAKE-2017 (classification)	Iglesias et al. 2017 (CD-MAKE)	8	8 DAT features	Decision tree, rules generalised by hand	Binary; separate training and testing sets	Testing: accuracy 99.18%, precision 90.95%, recall 95.95%
feat_analysis-2017	Iglesias et al. 2017 (JCSM)	8 in training and testing; 10 in a third, held-out set that adds ED1 and ED2	13 (Set A); 11 (Set B); 6 (Set C)	Decision tree, best of six shallow learners; outlier detectors	Binary; trained on the first set, tested on the second, then evaluated on the ten-technique set	Ten-technique set, Set A: accuracy 99.38%, recall 84.09%; per-technique table (Set B): ED1 29.29%, ED2 1.79%
feat_analysis-2017	This study	8 in training and validation; 10 in the held-out test set that adds ED1 and ED2	13 (Set A)	Cascaded deep stacking of Mamba and TabNet base learners	Binary detection, then 8-technique attribution; validation set for early stopping and model selection; test set used once	Ten-technique set: accuracy 99.71%, recall 95.45%, ED1 62.86%, ED2 96.43%; attribution macro-F1 0.9736



Supplementary Figure S1. Pearson correlation between the 13 DAT features on the training split (31,022 flows, overt and covert together; missing autocorrelation values filled with the training median). U, unique IAT values; Sk and Ss, multimodality scores; μωS, mean peak width; ρA, autocorrelation; c, covert-capacity estimate; p(Mo), share of the most common IAT; TR, runs test; TS, sign test; Ha, approximate entropy; Hq, Hurst exponent; K, Kolmogorov complexity; pkts, packet count.



Supplementary Figure S2. Training and validation loss per epoch of the three Mamba base learners in the cascade: (a) Mamba-Slim and (b) Mamba-Deep, the Layer 1 binary detectors, and (c) Mamba-S6, the Layer 2 technique classifier. The dotted line marks the epoch with the lowest validation loss, whose weights are kept; training stops 15 epochs later. TabNet's per-epoch losses were not recorded and are not shown.

Supplementary Note 1. Evaluation metrics

Binary metrics treat covert as the positive class. TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives. All metrics are computed with scikit-learn.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{4}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{6}$$

Accuracy (1) is dominated by the overt class at this imbalance, so the paper gives more weight to F1 (3), MCC (5) and the false-positive rate (4). MCC uses all four cells of the confusion matrix and stays defined when a class never appears among the predictions; for the eight-technique case it is computed in its multi-class form.

For technique attribution, precision, recall and F1 are computed for each technique against the rest, and macro-F1 averages them with equal weight over the K = 8 known techniques, so the small JIN class counts as much as the others:

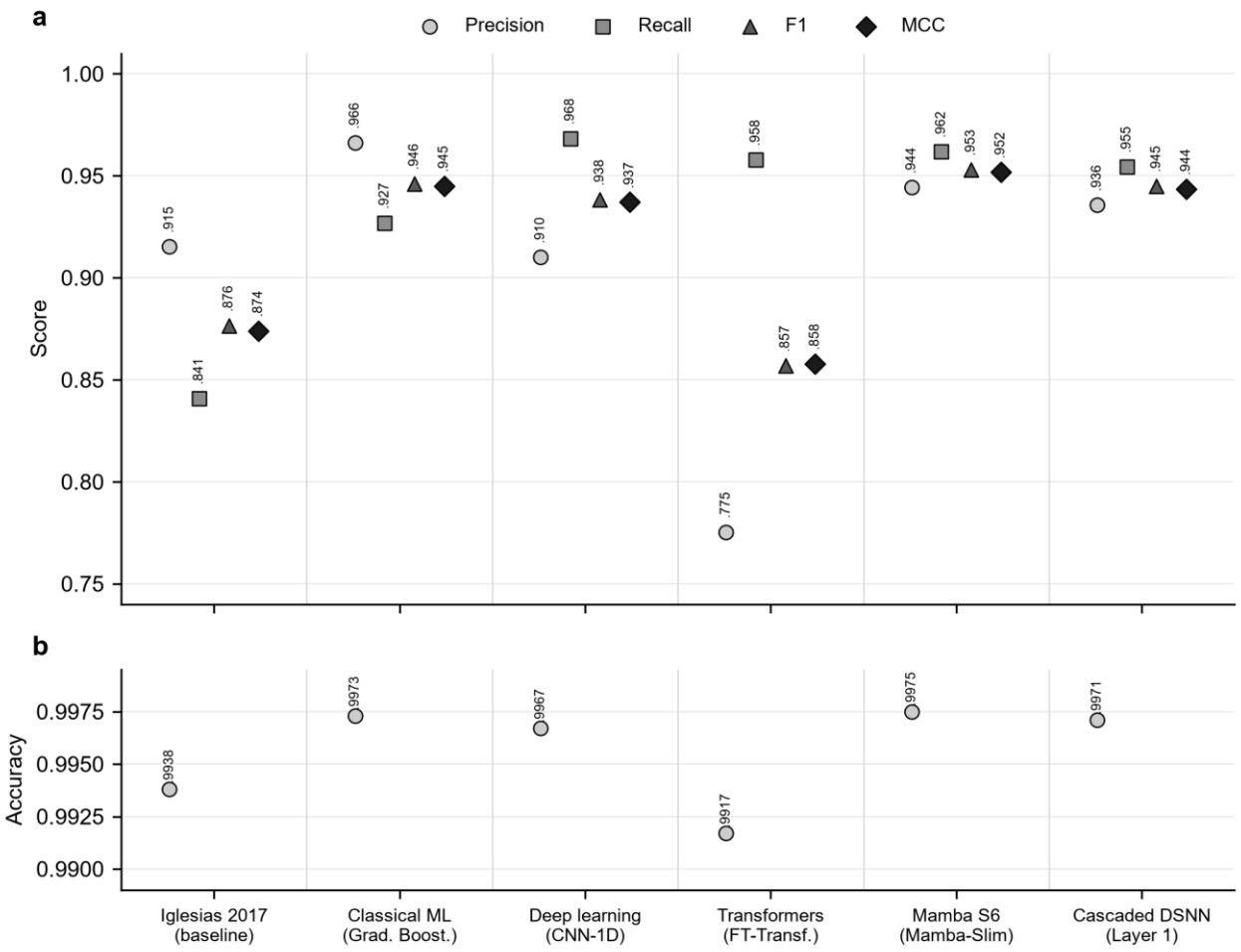
$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \text{F1}_k \tag{7}$$

ROC-AUC is the area under the curve of true-positive rate against false-positive rate as the decision threshold τ moves; it equals the probability that a randomly chosen covert flow scores higher than a randomly chosen overt flow. PR-AUC is computed as average precision, where P_n and R_n are the precision and recall at the n-th threshold:

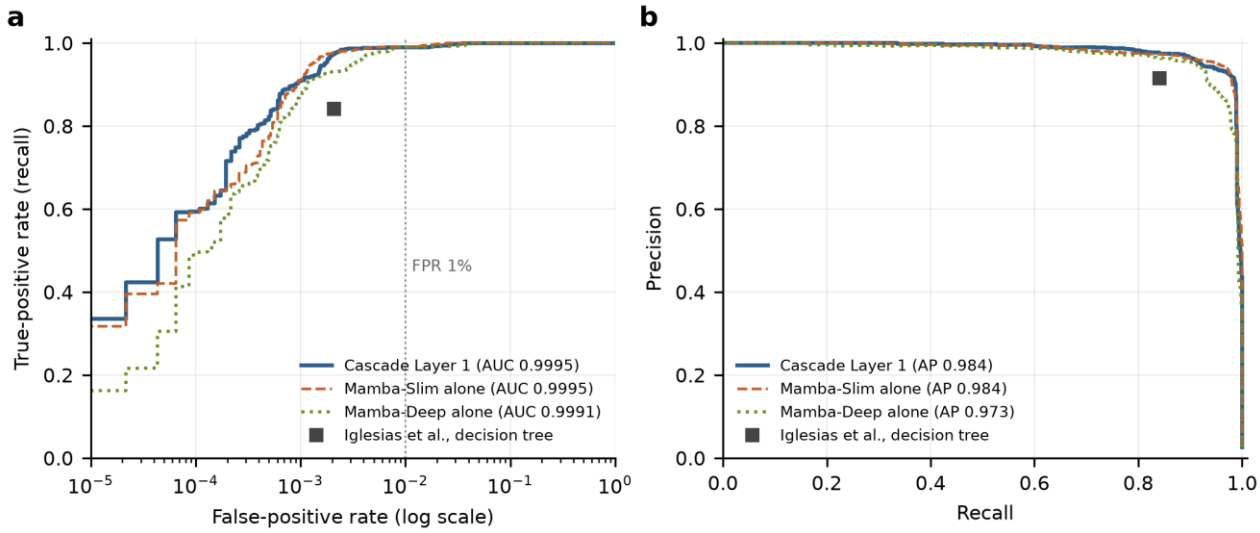
$$\text{AP} = \sum_{n} (R_n - R_{n-1}) P_n \tag{8}$$

The operating point reported for monitoring is the highest recall reachable while the false-positive rate stays at or below 1%:

$$\text{Recall}_{\text{FPR} \le 1\%} = \max_{\tau} \{\, \text{TPR}(\tau) \;:\; \text{FPR}(\tau) \le 0.01 \,\} \tag{9}$$



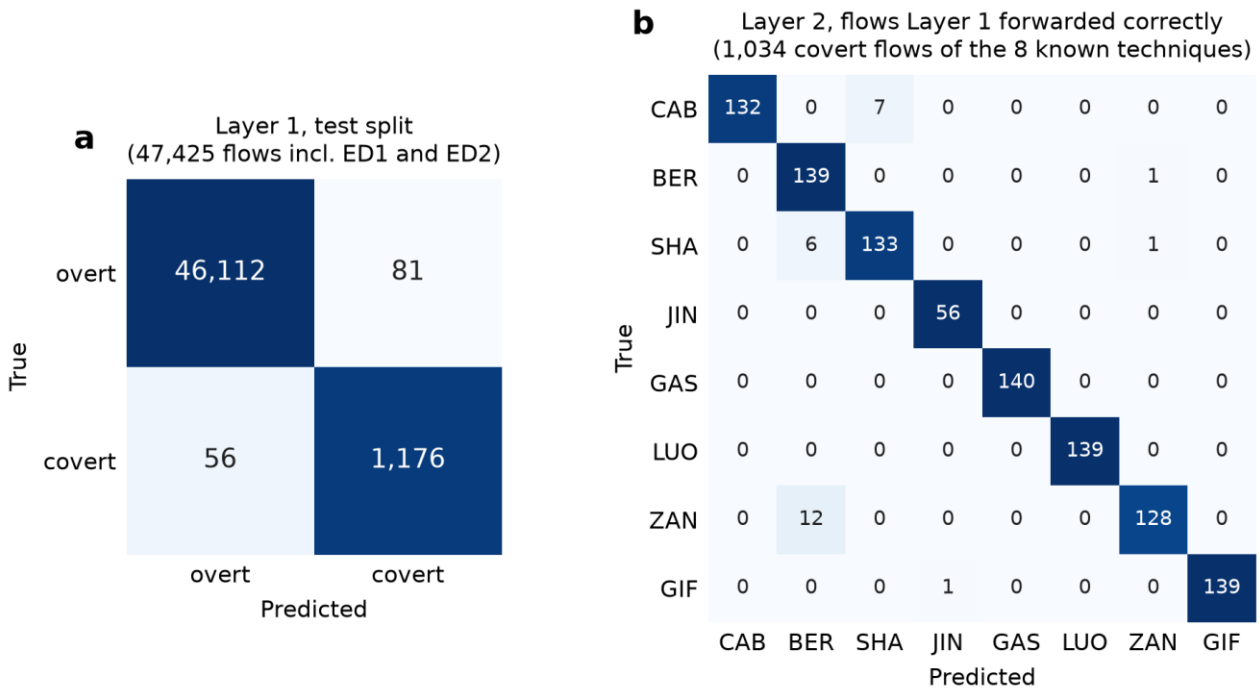
Supplementary Figure S3. Binary covert-versus-overt detection across model families on the test partition, plotting the values of Table 2 of the main text. The published baseline of Iglesias et al. is the leftmost group and the cascade's Layer 1 the rightmost. (a) Precision, recall, F1 and MCC, shown together because no single metric is adequate at this class imbalance. (b) Accuracy on a magnified scale: every approach exceeds 99%, so accuracy does not separate them.



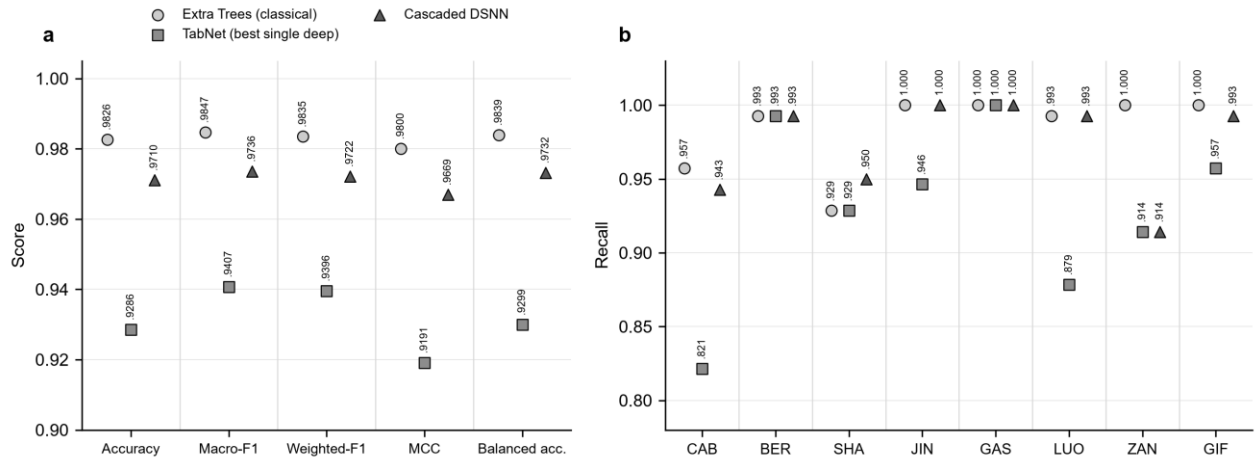
Supplementary Figure S4. Receiver-operating-characteristic (a) and precision-recall (b) curves on the held-out test partition (47,425 flows) for the cascade's Layer 1 and its two base learners used alone. The square marks the single operating point of the Iglesias et al. decision tree on the same flows (recall 84.09%, precision 91.52%, false-positive rate 0.21%). The false-positive axis in (a) is logarithmic so that the low-alarm region that matters for monitoring is visible; the dotted line marks a 1% false-positive rate, where Layer 1 recalls 99.03% of covert flows. AUC, area under the ROC curve; AP, average precision.

Supplementary Table S2. Binary detection on the held-out test partition: 47,425 flows, of which 1,232 are covert flows of ten CTC techniques, including the unseen ED1 and ED2, and 46,193 are overt. These are the same flows on which Iglesias et al. report their Feature Set A results, shown here in the confusion-matrix format of that study. Baseline counts, accuracy, precision, recall and ROC-AUC are from Table 3 (Feature Set A) of Iglesias et al.; F1 and MCC for the baseline are computed from those counts. Covert is the positive class.

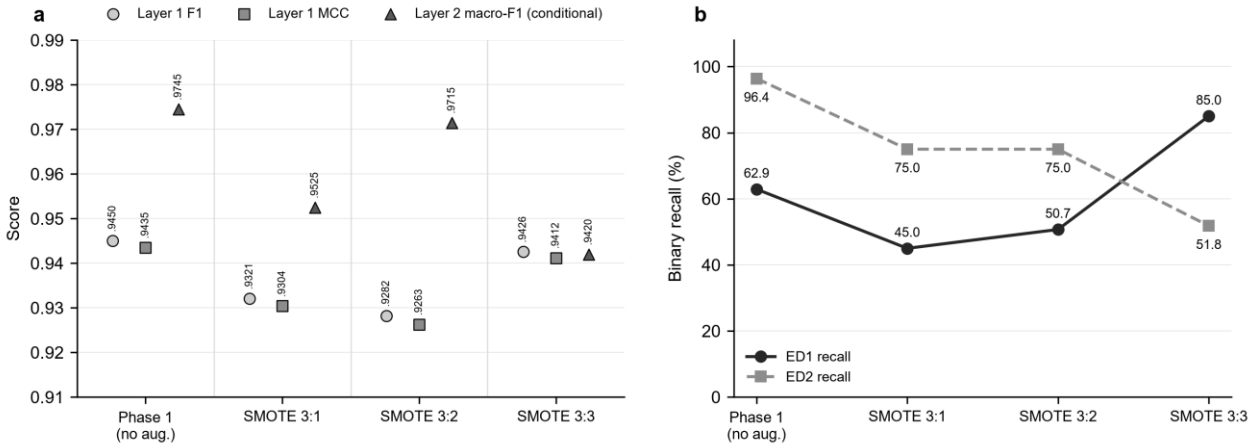
Detector	TP	FN	FP	TN	Accuracy (%)	Precision (%)	Recall (%)	F1	MCC	ROC-AUC
Iglesias et al., decision tree (Feature Set A)	1,036	196	96	46,097	99.38	91.52	84.09	0.876	0.874	0.903
This study, cascade Layer 1	1,176	56	81	46,112	99.71	93.56	95.45	0.945	0.944	0.9995



Supplementary Figure S5. Confusion matrices of the cascade on the held-out test partition. (a) Layer 1 on all 47,425 test flows, including the unseen ED1 and ED2 flows: 1,176 of the 1,232 covert flows are detected and 81 of the 46,193 overt flows are flagged. (b) Layer 2 on the 1,034 covert flows of the eight known techniques that Layer 1 forwarded correctly. The largest remaining errors are ZAN flows labelled BER (12), CAB flows labelled SHA (7) and SHA flows labelled BER (6). Shading is the share of each true class; numbers are flow counts.



Supplementary Figure S6. Technique attribution over the eight known techniques, plotting the values of Table 3 of the main text. All three models are scored on one basis: the held-out test partition, all 1,036 covert flows of those techniques, unconditional. (a) Aggregate metrics: the cascade leads the strongest single deep model on every metric, and the classical Extra Trees ensemble leads both. (b) Recall for each of the eight techniques.



Supplementary Figure S7. Effect of SMOTE rebalancing, plotting the values of Table 4 of the main text. (a) Cascade metrics at each augmentation ratio: the unaugmented configuration is best on every aggregate. (b) Binary recall on the two unseen techniques, showing that augmentation never moves both in the same direction.