

Interpretable by Construction, Identifiable by Construction: Privacy Leakage of Shapelet-Based Anti-Money-Laundering Models

Anonymous Author(s)

ABSTRACT

Supervisors expect anti-money-laundering (AML) alerts to be explainable, and example-based classifiers such as shapelets offer explanations by construction: a decision is justified by a short pattern in an account’s activity. Most shapelet methods select these patterns by cutting them out of training accounts; learned shapelets instead optimise them from random initialisation. We show that, for the selection-based methods, including the state of the art, the explanatory pattern is often a specific customer’s own transaction history, together with the customer’s label. Across two AML benchmarks (SAML-D, AMLworld) and real bank data (Berka), we hand a fitted model to an adversary and measure what it reveals. For the state-of-the-art Random Dilated Shapelet Transform (RDST), 86% of released shapelets match exactly one of ~30k SAML-D accounts (~827 customers per model), and each carries its source’s class, so the model discloses who was flagged and lets anyone holding the model reconstruct exact weekly amounts. Uniqueness persists against 249,946 candidate accounts (100% of unique shapelets stay unique), on real customers (75% on Berka), when amounts are perturbed by $\pm 5\%$ (96% precision), and when normalisation statistics are withheld (75%); it fails only when the adversary misses much of an account’s activity. The standard aggregate membership audit reports AUC ≈ 0.5 , whereas a tailored per-explanation audit flags 5.5% of training members at 0.1% false-positive rate. Learned shapelets, which never copy training data, single out no customer and resist memorisation; they match RDST accuracy on SAML-D but cost accuracy on AMLworld. Example-based explanations need their own privacy review.

1 INTRODUCTION

AML transaction monitoring is high-stakes, regulated, and increasingly machine-learned; deployed systems pair gradient-boosted trees with post-hoc explanations so investigators can act on alerts [7]. A different route to explainability is *example-based* models, whose explanation is a stored piece of training data, such as prototype networks [14] and shapelets. Shapelets were introduced by Ye and Keogh [1] as an interpretable primitive for time-series classification: a short subsequence whose distance to a series decides its class, so the model can show *which* pattern triggered a decision. They remain current: the shapelet transform [2] is a core component of HIVE-COTE 2.0 [11], a leading time-series ensemble, RDST [4] is among the strongest single classifiers, both ship in standard toolkits, and shapelets are applied to interpretable early classification [12] and financial forecasting [13]. We know of no published production AML deployment, but laundering typologies such as structuring and fan-in/fan-out are shapes in time, which makes shapelets a natural candidate wherever regulators demand explanations.

Shapelet methods come in two families: *selection-based* methods (the classic transform, RDST) keep subsequences cut from training series, while *learned* shapelets [3] are free parameters optimised by gradient descent and need not resemble any one series. That selected shapelets are copied subsequences is known; federated shapelet learning protects them with secure computation [8]. What has not been measured is how much a *released* model exposes in practice, against which adversaries, and whether standard audits notice. We find that the answer is worse than the construction suggests. The strongest method, RDST [4], also stores each shapelet’s normalisation statistics and source label; normalising every shapelet does not prevent re-identification; uniqueness holds against much larger candidate lists, imperfect adversary data, and real customers; and the aggregate membership audit is blind to all of it.

Contributions. (i) A threat model separating what a data-holding party and a model-only party learn from a released shapelet model. (ii) A measurement study on two AML benchmarks and real bank data, including bank-scale candidate lists and an imperfect adversary, mapping where the leak holds and where it stops. (iii) Evidence that aggregate membership inference misses the leak while a per-explanation audit catches it. (iv) A privacy–utility comparison of mitigations; learned shapelets remove the leak, at a dataset-dependent accuracy cost.

2 RELATED WORK

ML for AML. Production monitoring relies on boosted trees over aggregates with post-hoc explanations [7]; public benchmarks include SAML-D [6] and AMLworld [9]. **Privacy of models and explanations.** Membership inference [16] is the standard audit of training-data leakage, and strong audits report true-positive rates at low false-positive rates [17]. Feature attributions leak membership [18]. For time series, federated shapelets [8] and privacy-aware matrix profiles [10] defend shared patterns but do not quantify what a released model reveals about individuals.

3 THREAT MODEL

A bank trains a shapelet detector on labelled accounts (labels: past suspicious-activity findings) and the fitted model leaves the bank: shared with a vendor, auditor, regulator sandbox or consortium partner, or stolen. The model contains its shapelets; RDST also stores each shapelet’s normalisation mean and standard deviation and the class of the account it was taken from. We consider two adversaries. *Data-holding party* (vendor, auditor, partner, insider) holds transaction histories for many accounts but not the bank’s labels or training set. Matching shapelets to its accounts reveals *which customers were in the training set* and, via the stored class,

which were flagged; disclosing the latter conflicts with suspicious-activity-report confidentiality and anti-tipping-off rules. We measure this adversary both with an exact copy of the data and with a realistic one whose amounts are perturbed, who sees only part of each account's transactions, and whose list holds only some accounts. *Model-only party* (model theft, model marketplace) holds no data. From RDST it reconstructs raw windows (values $\times$ std + mean), i.e. exact weekly amounts and counts of specific customers, which it can link to them later.

4 EXPERIMENTAL SETUP

Data. *SAML-D* [6]: 9.5M synthetic transactions over 46 weeks, 0.10% suspicious, 28 typologies. Each account becomes a 6-channel weekly series (log outgoing/incoming amount, outgoing/incoming count, cross-border amount, cash amount). Accounts in any laundering transaction are positive (7,902); normal accounts are sampled 3:1 matched on number of active weeks (unmatched normals are mostly single-week accounts that make detection trivial), giving 29,946 accounts. *AMLworld* HI-Small [9] spans ~ 10 days, so we use 6-hour bins (40 steps; 23,409 matched accounts, 5,862 positive). *Berka* [15]: real, anonymised accounts of a Czech bank (1993–1998; 4,500 accounts, monthly bins, 72 steps). It has no laundering labels, so we use a proxy task (loan holder, 682 positive) with all loan repayments removed; the privacy measurement does not depend on the task. Accounts are split 50/50 into members (train) and non-members, stratified; results average 3 seeds.

Detectors. CatBoost on 40 aggregates plus the flattened series (industry baseline); MiniRocket [5] with logistic regression; a binary shapelet transform (5,000 sampled non-constant z-normalised candidates, quality = class separation of the min-distance profile, i.e. AUC in place of information gain, top-200 with self-similarity removal) with ExtraTrees; RDST (aeon, 1,000 shapelets) with ExtraTrees; learned shapelets [3] (144, random initialisation, soft-min distance).

Attacks. *Re-ID*: min (dilated) sliding distance from each released shapelet to every candidate account; a shapelet singles out a customer if exactly one account lies within $\epsilon=10^{-3}$. *Realistic adversary*: its series are rebuilt from raw transactions with log-normal amount noise ($\sigma=0.05$), each transaction kept with probability 0.8 or 0.5, and a subsampled candidate list; it claims an identity only if the nearest account is $\geq 2\times$ closer than the runner-up (300 sampled RDST shapelets per seed). *Bank scale*: we add 220k further SAML-D accounts (all with ≥ 5 transactions) to the candidate list and re-test 150 sampled shapelets that were unique among the 29,946 study accounts (seed 0). *Membership*: the standard aggregate audit (score = $-\min$ distance; confidence for baselines), and a per-explanation audit that flags accounts containing a released shapelet exactly, reported as TPR at low FPR [17].

5 RESULTS

Leakage is large but concentrated, and the standard audit misses it (Table 1). On SAML-D, 86% of RDST shapelets match exactly one account, ≈ 827 customers per model, and the classic shapelet transform singles out its source in 15% of cases from shape alone. Only the few hundred source accounts among $\sim 15k$ members are exposed, so the aggregate membership AUC stays at 0.52.

Table 1: SAML-D: detection on held-out accounts and leakage of the released model (mean $\pm$ std over 3 seeds). Re-ID: share of released shapelets that match *only* their source customer among 29,946 candidate accounts.

Model	PR-AUC	R@1%FPR	MIA AUC	Re-ID (%)
CatBoost (aggregates)	0.983 ± 0.001	0.842 ± 0.031	0.504 ± 0.002	n/a
MiniRocket	0.955 ± 0.002	0.637 ± 0.044	0.508 ± 0.003	n/a
Shapelet transform	0.878 ± 0.007	0.493 ± 0.033	0.500 ± 0.002	15.0 ± 2.2
RDST	0.948 ± 0.006	0.631 ± 0.021	0.520 ± 0.001	85.6 ± 1.6
RDST, norm.-only, no stats	0.925 ± 0.005	0.515 ± 0.015	0.512 ± 0.000	74.9 ± 1.9
Learned shapelets	0.957 ± 0.004	0.667 ± 0.017	0.500 ± 0.000	0.0

Table 2: Mitigations (shapelet transform; mean over seeds). Unique: shapelet matches only its source; Top-1: source is the nearest account.

Release	PR-AUC	Unique (%)	Top-1 (%)
As trained	0.878 ± 0.007	15.0 ± 2.2	19.3 ± 1.6
Noise $\sigma=0.1$	0.879 ± 0.007	0.0 ± 0.0	10.8 ± 2.5
Noise $\sigma=0.25$	0.877 ± 0.007	0.0 ± 0.0	3.2 ± 0.5
Noise $\sigma=0.5$	0.876 ± 0.005	0.0 ± 0.0	0.2 ± 0.2
Noise $\sigma=1.0$	0.870 ± 0.004	0.0 ± 0.0	0.0 ± 0.0
k -anon. mean, $k=5$	0.877 ± 0.007	0.3 ± 0.2	13.5 ± 0.4
k -anon. mean, $k=20$	0.876 ± 0.005	0.2 ± 0.2	12.5 ± 2.1
k -anon. mean, $k=50$	0.876 ± 0.006	0.0 ± 0.0	10.7 ± 1.3

Table 3: Re-identification by a realistic adversary (SAML-D, 300 sampled RDST shapelets, mean over seeds). The adversary's copy of the data has perturbed amounts and/or misses transactions, and holds all or half of the accounts; it claims an identity only if the nearest account is at least $2\times$ closer than the runner-up. Cust.: customers correctly re-identified; Prec.: share of claims that are correct.

Adversary's view	All accounts		Half of accounts	
	Cust.	Prec.	Cust.	Prec.
Exact copy	251	94%	125	86%
Amounts $\pm 5\%$	227	96%	115	86%
80% of txns, $\pm 5\%$	71	78%	36	62%
50% of txns, $\pm 5\%$	10	27%	5	17%

A per-explanation audit that flags accounts matched by a shapelet and by no other account identifies 5.5% of members at 0.1% FPR (827 members, 0 non-members), against $\approx 0.1\%$ for the aggregate score: auditors of example-based models must test explanations individually.

What each adversary learns. Every identified RDST shapelet carries its source's class, so a data-holding party learns which customers were flagged: 217 per model on SAML-D. A model-only party reconstructs exact windows: a median of 13 non-zero weekly amount/count cells per identifying shapelet (median peak $\approx$ £25k). Forcing every shapelet to be normalised and withholding the statistics removes the amounts but not the identity: 75% of shapelets still single out one customer (PR-AUC 0.925), because a multi-channel dilated shape is itself a fingerprint.

Table 4: AMLworld HI-Small (23,409 accounts, 6-hour bins) and Berka (4,500 real accounts, monthly bins, proxy task); mean over seeds. MIA: aggregate membership AUC. Re-ID: share of shapelets matching only their source; Cust.: customers singled out per model.

Model	PR-AUC	MIA AUC	Re-ID (%)	Cust.
<i>AMLworld</i>				
CatBoost	0.619 ± 0.004	0.625 ± 0.004	n/a	n/a
MiniRocket	0.475 ± 0.003	0.586 ± 0.003	n/a	n/a
Shapelet transform	0.408 ± 0.016	0.499 ± 0.002	6.0 ± 1.1	–
RDST	0.571 ± 0.001	0.507 ± 0.002	64.5 ± 1.2	621 ± 10
RDST, norm.-only	0.546 ± 0.005	0.498 ± 0.002	40.4 ± 1.7	391 ± 14
Learned shapelets	0.389	0.500	0.0	0
<i>Berka (real)</i>				
CatBoost	0.583 ± 0.013	0.632 ± 0.009	n/a	n/a
MiniRocket	0.625 ± 0.012	0.561 ± 0.003	n/a	n/a
Shapelet transform	0.470 ± 0.014	0.502 ± 0.002	43.5 ± 2.1	–
RDST	0.421 ± 0.021	0.530 ± 0.003	75.4 ± 0.8	637 ± 6
RDST, norm.-only	0.391 ± 0.016	0.501 ± 0.004	50.2 ± 1.0	445 ± 17
Learned shapelets	0.357 ± 0.017	0.499 ± 0.001	0.0	0

Table 5: Pushing learned shapelets to memorise (SAML-D, seed 0). Nearest-is-member: share of shapelets whose nearest account is a training member, vs. the members' share of the pool (chance). Claims: confident re-identifications ($d_2 \geq 2d_1$).

Training	PR-AUC	Nearest-is-member (chance)	Claims
30 epochs, full train	0.957	50% (50%)	0
200 epochs, full train	0.952	50% (50%)	0
200 epochs, 1k train	0.928	4% (3%)	0
400 epochs, 300 train	0.895	0% (1%)	0

Real customers are more identifiable (Table 4). On Berka's real accounts, 75% of RDST shapelets single out one customer (≈ 637 of 4,500 customers per model) and even shape-only shapelets do so 44% of the time, against 15% on SAML-D. Synthetic generators therefore understate, not inflate, the problem. AMLworld, with only 40 time steps, gives 64% for RDST and 6% for the classic transform.

Bank-scale candidate lists. Adding 220,000 further SAML-D accounts to the adversary's list, 100% of shapelets that were unique among 29,946 accounts remain unique among 249,946: a larger customer base does not hide the source.

Imperfect adversaries (Table 3, Figure 1). Perturbing every amount by $\pm 5\%$ barely matters: 227 correct identifications (62 flagged) at 96% precision per 300 RDST shapelets, 193 at 94% with statistics withheld. With half the account list the adversary finds about half as many, mostly correctly. What stops it is missing activity: with 80% of transactions visible it still makes 71 correct identifications at 78% precision, with 50% precision falls to 27%. On Berka's real customers the pattern is sharper: 207 re-identified at 98% precision with $\pm 5\%$ amounts, but few once a fifth of transactions is missing.

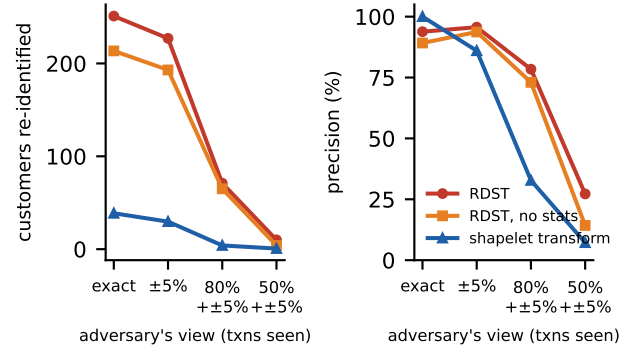


Figure 1: Realistic adversary on SAML-D (full account list; mean over seeds). Amount noise barely matters; missing transactions are what stops the attack.

The threat is real for parties that see an account's full activity (vendors and auditors given bank data, insiders, breached data) and weak for outsiders with a partial view.

Conventional models leak differently. CatBoost and MiniRocket store no training subsequences and show no membership signal on SAML-D, but on AMLworld and Berka they overfit and leak membership (CatBoost AUC 0.63 on both, MiniRocket 0.59 and 0.56). They reveal *whether* a customer was used; shapelet models reveal *who* the customer is, what they did, and how they were labelled.

Mitigations (Table 2). Gaussian noise on shapelet values removes exact fingerprints at no accuracy cost, but nearest-neighbour linkage persists until $\sigma \geq 0.5$. Replacing each shapelet by the mean of its k nearest training windows from distinct accounts stops unique matching but keeps the source's neighbourhood linkable. Learned shapelets copy no one: no learned shapelet lies within 10^{-2} of any account on SAML-D, and pushing them to memorise (Table 5; up to 400 epochs on 300 accounts) leaves the share of shapelets whose nearest account is a member at chance, with 0 confident claims. Their accuracy cost depends on the data: on SAML-D they are the best shapelet model (PR-AUC 0.957 vs. RDST 0.948), on AMLworld they trail RDST (0.39, or 0.42 with shorter patterns and longer training; the near-copies this creates match nearly every account, never one customer), and on Berka they reach 0.36 vs. RDST 0.42.

6 DISCUSSION AND LIMITATIONS

Scope. Banks may not deploy shapelets for AML today; we use them as the clearest case of example-based explanation, and the mechanism (an explanation that is a stored training record plus its label) applies equally to prototype and case-based models. **Data.** The AML benchmarks are synthetic and their positives are true laundering, not the investigator-filed reports a bank would train on; Berka is real but uses a proxy task. Activity-matched 3:1 sampling makes detection easier than at real prevalence, but our bank-scale test shows that a much larger pool of normal accounts does not dilute re-identification. **Methods.** Our shapelet transform uses AUC rather than information gain for candidate quality because a standard implementation degenerated on our sparse series; RDST is the off-the-shelf aeon degeneration. The realistic-adversary

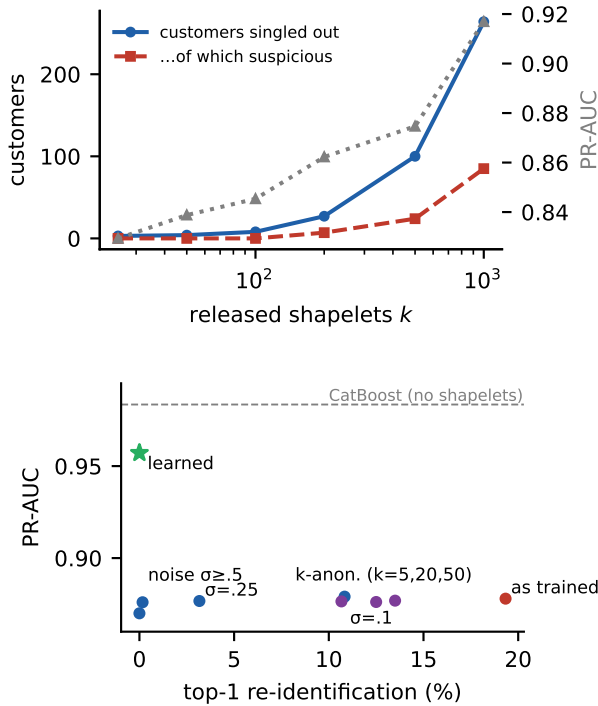


Figure 2: Top: customers uniquely exposed (and how many are flagged) grow with the number of released shapelets k , while accuracy saturates (shapelet transform, SAML-D, seed 0). Bottom: privacy-utility of releases (top-1 re-identification vs. PR-AUC; seed mean).

study samples 300 of 1,000 RDST shapelets per seed. We did not run shadow-model attacks such as LiRA; our per-explanation audit already detects the leak at zero false positives. **Defences.** Noise and learned shapelets carry no formal guarantee; differentially private shapelet selection and DP-SGD training are the natural next step.

Ethics. SAML-D and AMLworld are synthetic; Berka is a public, anonymised research release, and we report only aggregate statistics. We describe the attack to motivate audits and cheap defences.

Recommendations. Treat released shapelets, and the labels stored with them, as personal and confidential data; do not rely on normalisation or on withholding statistics; prefer learned shapelets and check their accuracy cost per dataset; and audit example-based models per explanation, at low false-positive rates, rather than with aggregate membership AUC.

REFERENCES

- [1] L. Ye and E. Keogh. 2009. Time series shapelets: a new primitive for data mining. In *KDD*.
- [2] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall. 2014. Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery* 28.
- [3] J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. 2014. Learning time-series shapelets. In *KDD*.
- [4] A. Guillaume, C. Vrain, and W. Elloumi. 2022. Random dilated shapelet transform: a new approach for time series shapelets. In *ICPRAI*.

- [5] A. Dempster, D. F. Schmidt, and G. I. Webb. 2021. MiniRocket: a very fast (almost) deterministic transform for time series classification. In *KDD*.
- [6] B. Oztas, D. Cetinkaya, F. Adedoyin, M. Budka, H. Dogan, and G. Aksu. 2023. Enhancing anti-money laundering: development of a synthetic transaction monitoring dataset. In *IEEE ICEBE*.
- [7] P. Tertychnyi, M. Godgildieva, M. Dumas, and M. Ollikainen. 2022. Time-aware and interpretable predictive monitoring system for anti-money laundering. *Machine Learning with Applications* 8.
- [8] Z. Liang and H. Wang. 2024. FedST: secure federated shapelet transformation for time series classification. *The VLDB Journal* 33.
- [9] E. Altman, J. Blanuša, L. von Niederhäusern, B. Egressy, A. Anghel, and K. Atasu. 2023. Realistic synthetic financial transactions for anti-money laundering models. In *NeurIPS Datasets and Benchmarks*.
- [10] L. Zhang, J. Ding, Y. Gao, and J. Lin. 2023. PMP: privacy-aware matrix profile against sensitive pattern inference for time series. In *SIAM SDM*.
- [11] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall. 2021. HIVE-COTE 2.0: a new meta ensemble for time series classification. *Machine Learning* 110.
- [12] Z. Xing, J. Pei, P. S. Yu, and K. Wang. 2011. Extracting interpretable features for early classification on time series. In *SIAM SDM*.
- [13] J. Kim, H. Lee, H. Jeon, S. Jin, and S. Ko. 2025. From patterns to predictions: a shapelet-based framework for directional forecasting in noisy financial markets. *arXiv:2509.15040*.
- [14] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. 2019. This looks like that: deep learning for interpretable image recognition. In *NeurIPS*.
- [15] P. Berka. 1999. Guide to the financial data set. In *PKDD'99 Discovery Challenge*.
- [16] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. 2017. Membership inference attacks against machine learning models. In *IEEE S&P*.
- [17] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. 2022. Membership inference attacks from first principles. In *IEEE S&P*.
- [18] R. Shokri, M. Strobel, and Y. Zick. 2021. On the privacy risks of model explanations. In *AIES*.