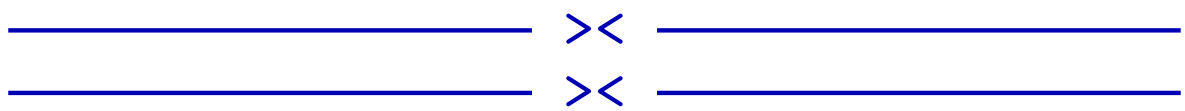
Data Shapley: Using Cooperative Game Theory to Detect Bad Training Data in Machine Learning

Abdul Basit

LUMS

May 08, 2026



ECON 3302: Introduction to Cooperative Game Theory

Presentation Roadmap

Problem: Training Data Has Unequal Value

Machine learning models do not benefit equally from every training point.

Point type	Effect	Desired value
Clean and representative	Improves test accuracy	Positive
Mislabeled or corrupted	Hurts the learned rule	Negative
Redundant duplicate	Adds little new information	Near zero

Central Question

Can the Shapley value assign a fair data-quality score to each training point so that corrupted records are automatically detected and cleaning decisions improve model accuracy?

My Project Contribution

I extended the paper in two stages.

1. **Original project work:** applied Data Shapley to the UCI Breast Cancer dataset with controlled label corruption, KNN as the learning algorithm, TMC-Shapley valuation, and LOO/random baselines.
2. **Additional coding extension:** added a reproducible NumPy/sklearn implementation and new experiments for audit-based cleaning, convergence, repeated splits, model transfer, and source/provider valuation.

Final Thesis

Data Shapley is not only a theoretical fairness rule. In this experiment, it becomes a practical data-cleaning tool that identifies harmful records and tells us where limited manual review should be spent.

Training Data as a Cooperative Game

Coalitional game with transferable utility

- **Players:** individual training records $i \in N$.
- **Coalition:** subset $S \subseteq N$ used to train a model.
- **Value function:** $v(S)$ equals test accuracy of the model trained only on S.
- **Empty coalition:** $v(\emptyset)$ equals majority-class baseline accuracy.

The Shapley Value

Each point is paid by its average marginal contribution.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \Big[v(S \cup \{i\}) - v(S) \Big]$$

- If a point improves most coalitions, its value is positive.
- If a mislabeled point hurts most coalitions, its value becomes negative.
- The weighting averages over all possible insertion positions and orderings.

Why This Fits Cooperative Game Theory

The Shapley value is the unique allocation satisfying efficiency, symmetry, dummy player, and additivity. These axioms map naturally to fair data valuation.

Why Approximation Is Needed

Exact Shapley requires all 2^n coalitions.

For the project dataset:

$$n = 455 \quad \Rightarrow \quad 2^{455} \approx 10^{137}$$

Solution: Truncated Monte Carlo Shapley

1. Sample a random ordering of training points.
2. Add points one at a time.
3. Record each point's change in test accuracy.
4. Average marginal contributions over sampled orderings.
5. Stop early when the coalition's score is close to the full-data score.

Dataset and Corruption Setup

UCI Breast Cancer Wisconsin Diagnostic Dataset

Property	Value
Samples	569
Features	30 real-valued cell nucleus measurements
Classes	Malignant vs benign
Train/test split	455 train / 114 test
Main corruption	20% label flips, 91 training labels
Model	KNN with $K = 5$
Baselines	Random removal and leave-one-out

Interpretation: label flipping simulates diagnostic data-entry errors or incorrect annotations.

Original Hypotheses

H1: Negative Values

Corrupted points should receive lower or negative Shapley values.

H2: Better Removal

Removing low-Shapley points should improve accuracy more than random removal and be more reliable than LOO.

H3: Corruption-Rate Stress Test

Detection may degrade as the corruption rate becomes high because the clean signal becomes weaker.

Original Result: Corrupted Points Get Negative Values

From the earlier notebook and final presentation

Data group	Count	Average Shapley value
Clean labels	364	$+0.0021$
Corrupted labels	91	-0.0071

Finding

The corrupted distribution shifted into negative territory. This supports the idea that mislabeled points create negative marginal contributions.

Original Result: Removal Strategy Comparison

Bottom 20% of training points removed

Strategy	Test accuracy	Improvement	Detection rate
Baseline, no removal	88.60%	–	–
Random removal	87.72%	−0.88 pp	21.98%
LOO removal	96.49%	+7.89 pp	29.67%
Shapley removal	**96.49%**	+7.89 pp	**82.42%**

Conclusion: Shapley was much better at finding the corrupted labels, even when LOO happened to match its accuracy after removal.

Original Result: Corruption Rate Stress Test

Detection did not degrade. It improved with stronger corruption signal.

Corruption	Baseline acc.	Shapley det.	Random det.	LOO det.
10%	88.60%	**47.25%**	7.69%	20.88%
20%	85.96%	**79.12%**	17.58%	30.77%
30%	77.19%	**92.31%**	32.97%	45.05%
40%	72.81%	**96.70%**	43.96%	63.74%

Interpretation: more label corruption made harmful points more consistently damaging across sampled coalitions.

What I Added After the First Version

New code and new experiments in `mywork/`.

- A TensorFlow-free TMC-Shapley implementation for KNN using NumPy and sklearn.
- Precomputed test-train distances to make repeated coalition scoring practical.
- Saved CSV and PNG artifacts for reproducibility.
- Corrected convergence analysis that stores every sampled marginal vector.
- New extensions: audit-based relabeling, repeated seeds, model transfer, and source-level Shapley.

Why This Matters

The extension moves the project from "does Shapley identify bad points?" to "can Shapley guide a practical data-cleaning workflow under limited review budget?"

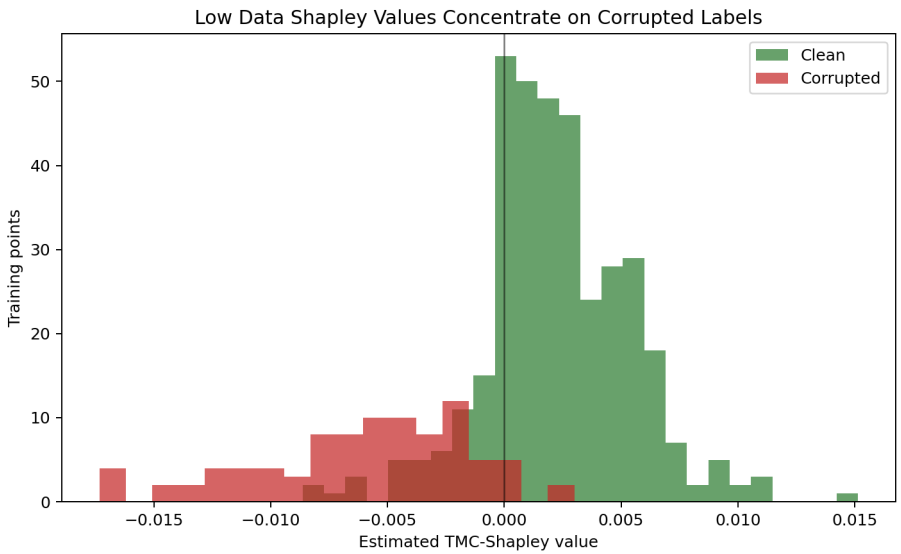
New Main Experiment: Shapley vs Baselines

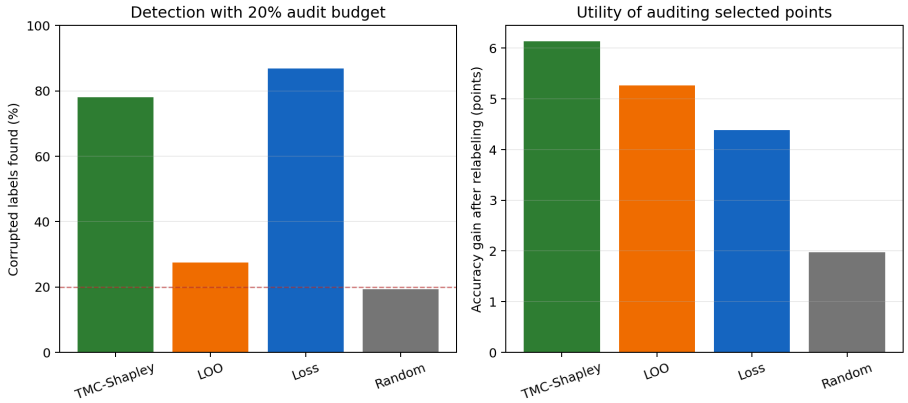
20% label corruption, 91-point audit budget, 120 TMC permutations

Method	Corrupt found	Precision	AUROC	Relabel acc.
TMC-Shapley	71/91	**78.02%**	0.9500	**96.49%**
LOO	25/91	27.47%	0.5573	95.61%
Loss score	79/91	86.81%	0.9621	94.74%
Random	17.6/91	19.34%	0.4959	92.32%

Key point: Loss is a strong detector, but Shapley gives the best KNN accuracy after the selected labels are corrected.

Visual Evidence: Value Distribution and Method Comparison



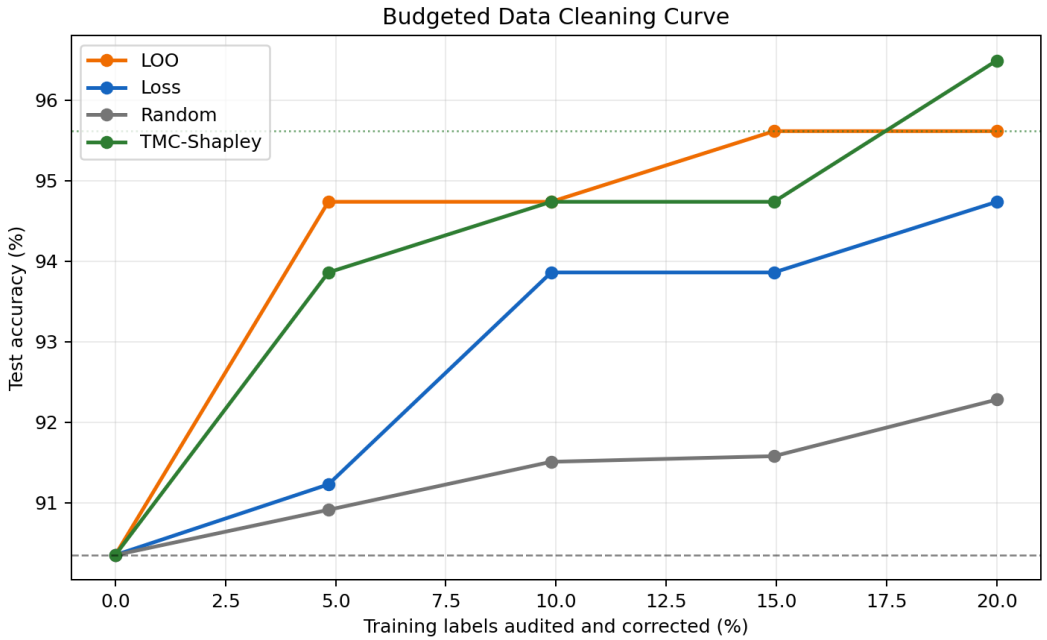


Extension A: Budgeted Data Cleaning

Instead of deleting data, simulate manual audit and correction.

Budget	Shapley	Random
0	90.35%	90.35%
22	93.86%	90.91%
45	94.74%	91.51%
68	94.74%	91.58%
91	**96.49%**	92.28%

Full 91-label audit closes the corrupted-data gap and surpasses the clean-label KNN baseline on this split.

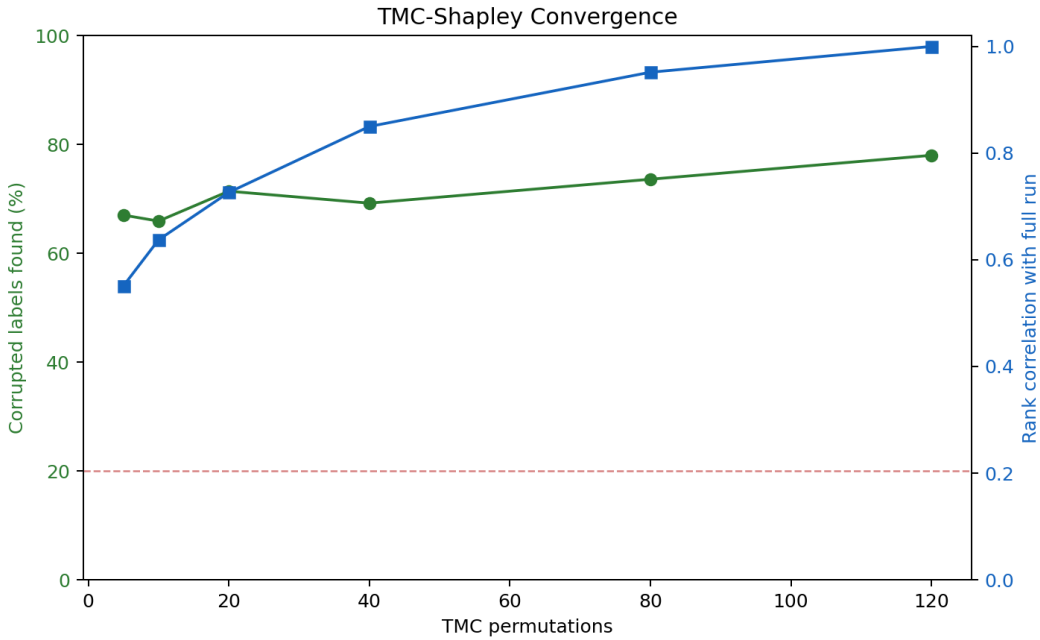


Extension B: Convergence and Stability

Detection improves as more random permutations are averaged.

T	Prec.	AUC	ρ
5	67.03%	0.8499	0.5510
20	71.43%	0.8879	0.7263
80	73.63%	0.9356	0.9516
120	**78.02%**	**0.9500**	1.0000

Even early checkpoints beat random, while rank stability improves steadily.



Extension C: Corruption-Type Robustness

Data Shapley is strongest when corruption changes the learning signal.

Corruption type	Baseline acc.	Removal acc.	Precision	AUROC
Label flip	81.58%	97.37%	**72.53%**	**0.9254**
Feature noise	96.49%	98.25%	13.19%	0.5847
Outlier shift	97.37%	98.25%	10.99%	0.6217

Limitation: feature noise can be hard to identify because the label is still correct and KNN can sometimes ignore noisy far-away examples.

Extension D: Repeated-Seed Robustness

Five independent train/test splits and corruption draws

Method	Mean precision	Std. dev.	Mean AUROC
TMC-Shapley	**69.23%**	4.40%	**0.9037**
LOO	29.23%	7.56%	0.5841
Loss score	87.47%	2.76%	0.9745
Random	19.92%	1.14%	0.4995

Report claim: the Shapley effect is not a one-split accident; it remains far above random and LOO across repeated experiments.

Extension E: Model-Transfer Validation

Use KNN-Shapley-selected corrections for other learners.

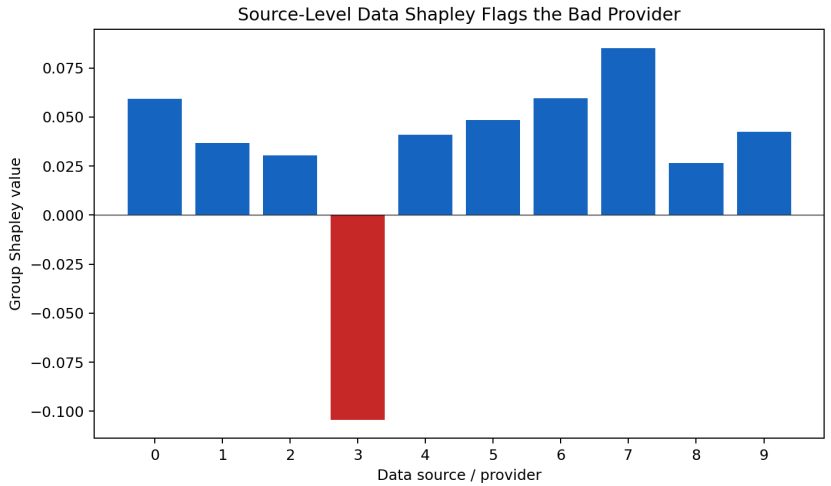
Model	Method	Corrupted acc.	Fixed acc.	Gain
KNN	TMC-Shapley	90.35%	**96.49%**	+6.14 pp
KNN	LOO	90.35%	95.61%	+5.26 pp
Logistic	TMC-Shapley	92.11%	94.74%	+2.63 pp
Logistic	Loss	92.11%	**96.49%**	+4.39 pp
Random forest	TMC-Shapley	96.49%	96.49%	+0.00 pp

Interpretation: Shapley-selected cleaning transfers to logistic regression, but the strongest gains remain for the value-function model, KNN.

Extension F: Source-Level Shapley

Players can be data sources, not only individual points.

Source	Size	Corrupt	Value
3	46	**30**	**-0.1043**
8	45	2	0.0264
2	46	2	0.0303
1	46	2	0.0367
7	45	2	0.0851



Finding

The intentionally bad provider is ranked as the lowest-valued source.

Final Summary

Part	Main result	Contribution
Original H1	Corrupted mean $\phi = -0.0071$	Negative values flag harmful records
Original H2	82.42% detection	Better corrupted-label detection than L
Original H3	96.70% detection at 40% corruption	Stronger corruption gave clearer signal
New main run	71/91 corrupt labels found	Reproducible sklearn-only pipeline
Audit cleaning	90.35% to 96.49%	Practical label-review workflow
Repeated seeds	69.23% mean precision	Robust beyond one split
Source Shapley	Bad source ranked lowest	Provider accountability extension

Takeaway

Shapley values provide an axiomatically justified and empirically useful way to value training data, detect corrupted examples, and guide limited data-cleaning resources.

Limitations and Future Work

Limitations

- TMC-Shapley still requires many model evaluations.
- KNN-based Shapley values are most directly optimized for KNN.
- Feature noise is harder to detect than label noise.
- Small dataset results should be validated on larger real-world datasets.

Future work

- Compare exact KNN-Shapley formulas with sampled TMC-Shapley.
- Add confidence intervals for individual point values.
- Test active learning: audit only points with low value and high uncertainty.
- Extend source-level valuation to real multi-provider datasets.

Thank you!