Harmonizing Mass-Shooting Databases for Cross-Source Risk Classification

Neha Sharma
Department of Computer Science
Virginia Commonwealth University
Richmond, Virginia, United States
sharman20@vcu.edu

Ritesh Sharma
Department of Electrical And Computer Engineering
Virginia Commonwealth University
Richmond, Virginia, United States
sharmar33@vcu.edu

***Abstract*—Predictive modeling of mass-shooting risk has largely been evaluated within individual curated datasets, leaving open the question of how well such models generalize across sources. We address this gap by harmonizing three widely used U.S. databases—Kaggle (1966–2017), Mother Jones (1982–2026), and Stanford MSA (1966–2016)—into a unified 12-variable schema covering 806 incidents. Using this schema, we evaluate a simple pipeline comprising three stratification strategies and three classifiers under stratified 5-fold cross-validation and a multi-seed leave-one-dataset-out (LODO) protocol with per-fold label recomputation to prevent leakage. Across 27 paired feature-set comparisons (Wilcoxon $W = 123$, $p = 0.117$), we do not detect a statistically significant difference between contextual and full feature sets; however, because the pairs are non-independent, this result should be interpreted as inconclusive. The best LODO configuration—a depth-3 decision tree with quartile labels, with depth selected under the same criterion so reported figures are upper bounds—achieves recall of 0.819 on Kaggle and 0.886 on Stanford MSA for the highest-severity risk class (VeryHigh). The higher recall on Mother Jones (0.989) reflects its 4+ fatality inclusion criterion rather than genuine cross-dataset generalization. Within-dataset VeryHigh precision remains low (0.147), limiting practical interpretability. Our primary contributions are: (i) a harmonized cross-source schema, (ii) a leakage-aware LODO evaluation protocol with documented seed-randomness scope, and (iii) an initial cross-dataset and temporal baseline showing that LODO and post-2010 evaluation select different optimal configurations. We frame these results as an auditing study rather than a deployable classifier.**

***Index Terms*—data integration, cross-dataset generalization, imbalanced classification, public safety, risk stratification**

I. Introduction

Mass-shooting risk is a serious public-safety concern, yet predictive modeling in this domain remains fragmented. Journalists, academic groups, and advocacy organizations maintain independent databases with differing inclusion criteria, variable definitions, and vetting processes. As a result, a model trained on one source may fail silently on another, but this cross-source gap has not been measured under a controlled protocol. This limitation affects both researchers seeking results that generalize beyond a single dataset and practitioners deciding which data sources to rely on when building risk-stratification tools.

Code and harmonized schema available at https://github.com/nehasharmacs/harmonizing-mass-shooting-db.

Prior machine-learning work on mass-shooting risk has focused on single datasets, including fatality prediction and descriptive profiling [1]. These studies model incident characteristics within one database at a time, leaving cross-source transfer unexamined. More broadly, ML approaches to violence risk assessment have similarly relied on single-source evaluations [2], with cross-source stability remaining unexamined across the field. Leave-source-out protocols are standard practice in adjacent domains where source heterogeneity dominates—most notably clinical imaging, where they have exposed within-source overestimates of generalization [11]. Our work adapts this protocol to the mass-shooting domain; the method itself is borrowed, while the contribution is its domain instantiation. Unlike adjacent domains, prior ML work in mass-shooting risk has not produced publicly available classifier systems that could serve as direct baselines; our LODO protocol therefore evaluates feature and stratification choices rather than benchmarking against existing models. At the same time, aggregate counts diverge substantially across databases due to definitional differences [3], [4]. No prior study, to our knowledge, has trained classifiers on one mass-shooting database and evaluated them on another under a controlled leave-one-dataset-out (LODO) protocol. The available datasets are small (hundreds of rows) and severely imbalanced, so within-dataset validation alone risks overstating real-world performance [5]. We therefore use random oversampling, per-class Youden's J thresholding [6], and stratified cross-validation [7], and we triangulate feature importance across permutation and Gini-based methods. Our ethics discussion draws on prior work showing how socially constructed categories, when encoded as features, can act as contested proxies [8].

We make three contributions. First, we harmonize three public U.S. mass-shooting databases—Mother Jones [9], Stanford MSA [10], and Kaggle—into a common 12-variable schema designed as a reusable preprocessing module, lowering the barrier for multi-source studies. Second, we introduce a multi-seed LODO evaluation protocol with labels recomputed per training fold to prevent leakage. We document precisely which quantities the seeds exercise (oversampling RNG, Youden validation split, tree tie-breaking) and which are deterministic given the LODO partition, providing the

TABLE I
HARMONIZED DATASET SOURCES (TOTAL_VICTIMS ≥ 3, LAS VEGAS 2017 EXCLUDED). SD = STD. DEV. OF TOTAL_VICTIMS.

Source	n	Years	Median	Mean	SD
Kaggle	315	1966–2017	5.0	8.56	8.8
Mother Jones	157	1982–2026	10.0	13.80	12.3
Stanford MSA	334	1966–2016	5.0	7.53	7.8
Total (pooled)	806	1965–2026	6.0	9.15	9.5

first empirical baseline for cross-source generalization in this domain. Third, we report a conservative empirical finding: across 27 paired feature-set comparisons, we cannot reject the null of no difference between contextual and full features (Wilcoxon $p = 0.117$; pairs are non-independent, so the result is inconclusive), and the highest-mean configuration partially exploits a distributional artifact in the Mother Jones dataset. We argue that this null result is itself informative at current sample sizes: cross-source ranking of feature sets and stratification strategies is not reliable on $n = 806$ rows, and within-source evaluation can mask this effect.

II. DATA AND HARMONIZATION

A. Sources and distributional shift

The Kaggle source is the *Mass Shootings in America* dataset, originally compiled by the Stanford Geospatial Center. The Mother Jones dataset exhibits a substantially higher mean victim count (13.80) than Kaggle (8.56) and Stanford MSA (7.53). This difference follows from Mother Jones's historical inclusion criterion of 4+ fatalities (vs. 3+ victims), which leads to systematic over-representation of higher-casualty incidents in earlier records. As a result, the distribution is highly asymmetric; in particular, VeryHigh incidents become nearly separable under quartile thresholds within the Mother Jones fold. This behavior is an artifact of dataset construction rather than evidence of true cross-source separability (Section V).

B. Common schema

We map each source into a shared set of twelve variables: `source`, `year`, `fatalities`, `injured`, `total_victims`, `incident_area` (eight normalized categories), `open_close`, `age`, `gender`, `race`, `mental_health`, and `multiple_shooters`. Several encoding decisions are required for alignment: for example, Stanford MSA's "Place Type" and Mother Jones's "location" are collapsed into eight venue categories. The `open_close` feature is only present in Kaggle and is marked as `Unknown` in the other sources; it is retained in the schema but subsequent importance analysis confirms it carries negligible predictive weight (Section IV-C), attributable to single-source coverage rather than genuine signal. Similarly, `mental_health` is mapped to {Yes, No, Unclear, Unknown}, with additional discussion in Section V.

C. Risk stratification

We construct three risk stratification schemes over `total_victims`, assigning Low, Medium, High, and Very-High labels. The first uses fixed rule-based thresholds (10, 20, 40). The second is based on standard deviations ($\mu \pm \sigma$, $\mu + 2\sigma$), and the third uses quartiles (Q_1, Q_2, Q_3). Under the LODO protocol, the standard deviation and quartile thresholds are computed only on the training fold, ensuring that no label information leaks into evaluation.

III. CLASSIFICATION PIPELINE

For each combination of stratification strategy, feature set, and classifier, we recompute labels using training-fold statistics (standard deviation and quartiles), apply random oversampling to balance classes, and generate class probabilities on the held-out fold. Under the default decision rule, predictions are assigned as $\hat{y}_i = \arg\max_k P_{ik}$. Under the Youden-based variant, a 20% stratified validation split is taken from the training data before oversampling; class-specific thresholds t_k are selected by maximizing $\text{TPR}_k(t) - \text{FPR}_k(t)$ on this split, and predictions follow $\hat{y}_i = \arg\max_k \{P_{ik} - t_k\}$.

A. Experimental setup

Feature sets. We evaluate two feature configurations. The *contextual* set (ctx) includes `incident_area`, `open_close`, and `gender`. The *full* set extends this with `age`, `mental_health`, `race`, and `multiple_shooters`.

Classifiers. We use a depth-3 decision tree selected via a pre-experiment LODO sweep over depths {3–7, unconstrained}. Because this sweep optimizes the same evaluation criterion as the main experiments, reported LODO results should be interpreted as upper bounds. We also evaluate L2-regularized multinomial logistic regression (MLR; 2000 iterations, `lbfgs`) and Gaussian naïve Bayes (GNB).

Evaluation. Within-dataset evaluation uses stratified 5-fold cross-validation on the full 806-row dataset. For cross-source evaluation, we use leave-one-dataset-out (LODO): models are trained on two sources and evaluated on the third, repeated over five random seeds (42–46). One-hot encoding is fit on the union of training and test categorical levels to avoid unseen-category failures while preserving label separation.

Seed-randomness scope. Under LODO the test fold is fully determined by the held-out source. Seeds 42–46 therefore vary only (a) the random-oversampling bootstrap on the training fold, (b) the 20% stratified validation split used for Youden threshold selection, and (c) tie-breaking in decision-tree split choice. For configurations using default decisions on training-fold quartile labels, the resulting depth-3 tree is invariant to the oversampling bootstrap at this sample size, producing SD$= 0.000$ across seeds; non-zero variance appears in Youden-validated configurations (e.g., std/MLR/Youden-val on MJ: 0.61 ± 0.29). The protocol therefore characterizes oversampling and validation-split sensitivity but does not exercise test-fold randomness, which would require repeated stratified subsampling within each LODO fold.

TABLE II

WITHIN-DATASET 5-FOLD CV, TOP 5 BY VERYHIGH RECALL (DEPTH-3 DT). RECALL = MEAN±STD; PREC = MEAN PRECISION. Q=QUARTILE, S=STD, R=RULE. $^{\times}$=BELOW 0.25 RANDOM BASELINE.

Strat.	Model	Thr.	Feat.	Acc.	Recall	Prec
S	DT	def.	full	0.655	0.889 ± 0.141	0.147
Q	DT	def.	ctx	0.360	0.865 ± 0.109	0.325
Below-chance accuracy ($^{\times}$); not recommended						
R	*GNB*	*def.*	*ctx*	*0.119*$^{\times}$	0.858 ± 0.149	*0.062*
R	*GNB*	*def.*	*full*	*0.168*$^{\times}$	0.851 ± 0.112	*0.063*
Q	MLR	Youden	ctx	0.347	0.708 ± 0.138	0.331

$^{\times}$ Accuracy $<$ 0.25 random baseline.

TABLE III

LODO VERYHIGH RECALL (MEAN±STD, 5 SEEDS, DEPTH-3 DT), TOP-4 BY MINIMUM RECALL PER FEATURE SET. CI_{95}: $[0.724, 0.862]$, MEAN 0.814. †MJ = DISTRIBUTIONAL ARTIFACT (SECTION V).

Strategy	Model	Thr.	Kag.	MJ	Stan.	min	min src
Contextual features only							
quartile	DT	default	0.82±0.09	0.99±0.00†	0.89±0.06	**0.82**	Kag.
std	MLR	Youden-val	0.60±0.17	0.61±0.29	0.91±0.13	0.60	Kag.
quartile	MLR	default	0.51±0.00	0.61±0.00	0.90±0.00	0.51	Kag.
std	MLR	Youden	0.45±0.00	0.50±0.00	0.58±0.00	0.45	Kag.
Full features (+ age, race, mental_health, mult. shooters)							
std	DT	default	0.84±0.04	0.73±0.02	0.87±0.04	**0.73**	MJ
std	GNB	Youden	0.73±0.00	0.73±0.04	0.56±0.22	0.56	Stan.
quartile	DT	default	0.51±0.00	0.61±0.00	0.12±0.00	0.12	Stan.
std	MLR	default	0.35±0.00	0.84±0.00	0.43±0.00	0.35	Kag.

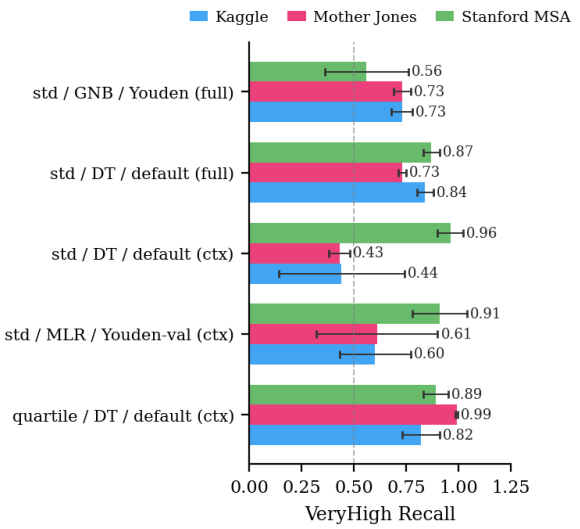

Fig. 1. Per-source VeryHigh recall with 95% bootstrap CIs for the five top configurations. Dashed line = 0.5. Mother Jones's near-ceiling recall (0.99) for the top configuration is a distributional artifact, not genuine generalization (Section V).

IV. RESULTS

A. Within-dataset (pooled 5-fold CV)

Table II reports the top configurations by mean VeryHigh recall. With depth 3, std+DT+default+full leads (acc = 0.655, VH rec = 0.889 ± 0.141), followed by quartile+DT+default+ctx (acc = 0.360, VH rec = 0.865 ± 0.109). Rows 3–4 (rule+GNB) reach high recall with accuracy well below the 0.25 random baseline, an artifact of minority-class over-prediction.

B. Cross-dataset generalization

Figure 1 shows per-source VeryHigh recall with 95% bootstrap CIs for the five top configurations. The best contextual configuration (quartile + DT + default, depth 3) achieves mean recalls of 0.819 (Kaggle, CI $[0.727, 0.865]$), 0.989 (Mother Jones, CI $[0.989, 0.989]$; distributional artifact — see Section V), and 0.886 (Stanford MSA, CI $[0.857, 0.943]$). The per-seed minimum recall is 0.814 (SD $= 0.090$, CI $[0.724, 0.862]$).

Across all 27 paired feature-set comparisons, 12 favor contextual and 15 favor full features. We initially computed Wilcoxon $W = 123$, $p = 0.117$ but report the sign-split descriptively rather than as inference: the 27 pairs share classifiers, strategies, and datasets, so the independence assumption underlying the Wilcoxon test is violated and the p-value is not interpretable. A directional claim would require an inference method that respects the dependence structure—a clustered bootstrap over configurations, a linear mixed-effects model with random effects for classifier/strategy/dataset, or a Bayesian estimate with a credible interval. We defer this to future work with larger samples and treat the present result as descriptive evidence that neither feature set can be ranked above the other at $n = 806$ rows. The contextual configuration's high-mean point estimate is dominated by Mother Jones recall, which Section V identifies as a distributional artifact.

Paired ablation. Figure 2 shows the full Δ surface. The strongest contextual win is quartile/DT/default ($+0.695$); the strongest full-feature win is std/DT/youden (-0.689). Rule-strategy rows mostly favor full features, confirming the feature-set effect is configuration-specific rather than systematic.

C. Feature importance (triangulated)

We triangulate importance for the best full-feature configuration (std, DT, default) across three methods: permutation on the primary depth-3 DT, permutation on a 200-tree RF, and mean Gini decrease from the same RF. `age` ranks first in all three (permutation-primary: 0.005; permutation-RF: 0.012; Gini-RF: 0.299) — the single most robust signal. Secondary rankings diverge: RF permutation surfaces `incident_area_School` (0.010) and `incident_area_Home` (0.008), while the primary DT assigns zero importance to all features except `age`, reflecting the constraint imposed by depth 3. `mental_health_Yes` ranks second under Gini-RF (0.138) but substantially lower under permutation methods, reflecting Gini's sensitivity to impurity over predictive transfer. All `open_close` encodings receive zero permutation importance on the primary DT; the small RF permutation weight (0.006) is attributable to Kaggle being the sole source encoding this field.

D. Temporal holdout

Training on pre-2010 incidents ($n = 292$) and testing on post-2010 incidents ($n = 514$) reveals that LODO

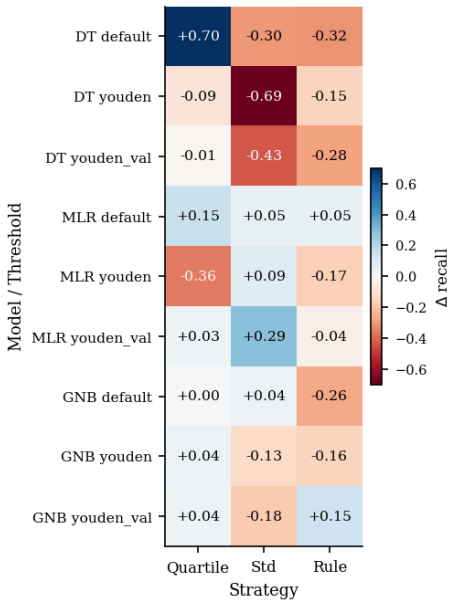

Fig. 2. All 27 paired feature-set comparisons (Δ = ctx_min − full_min). Blue = contextual features win; red = full features win. Pairs are non-independent; result is descriptive, not confirmatory.

and temporal rankings diverge. The LODO best configuration (quartile/DT/default/ctx) collapses to VeryHigh recall of 0.047 post-cutoff, while the best temporal configuration (quartile/GNB/Youden/ctx) achieves 0.820 ($\pm$0.026). Per-source, Kaggle shows the sharpest temporal drop (recall 0.000 for the LODO-best config), consistent with its role as the hardest transfer target under LODO. These results suggest that cross-source and cross-time evaluations are complementary: a model robust to source shift is not necessarily robust to temporal shift.

V. Discussion, Limitations, and Ethics

What integration buys. The most defensible interpretation of our LODO results is a negative one: even with three harmonized sources, we cannot reliably separate feature-set choices, stratification strategies, or threshold rules at this sample size. The highest-mean configuration (contextual features, quartile labels, depth-3 DT) appears to benefit from near-trivial separation of Mother Jones's high-casualty tail under quartile thresholds, rather than reflecting a stable cross-source signal. Within-dataset performance is also limited: the best configuration achieves a VeryHigh precision of only 0.147 (Table II), meaning that approximately 85% of VeryHigh predictions are false positives under severe class imbalance. Taken together, these findings suggest that the primary value of integration is methodological: per-fold label recomputation, multi-seed LODO, and explicit accounting for distributional effects, rather than selection of any specific model or feature set.

Limitations. The pooled dataset contains 806 observations, and bootstrap 95% confidence intervals span [0.724, 0.862]. The near-ceiling Mother Jones recall (0.989, SD= 0.000) reflects its 4+ fatality inclusion criterion; effective cross-source evaluation is therefore driven primarily by the Kaggle and Stanford MSA subsets. The Stanford MSA collapse under full features (recall 0.12 vs. 0.89 contextual) reflects mismatched demographic encodings across sources, leading to negative transfer. Youden-based thresholding improves Kaggle performance ($+0.218$) but slightly degrades Stanford MSA (-0.044). All datasets inherit reporting bias, and mental-health variables should not be interpreted causally [8].

Ethics and deployment. This pipeline is a research artifact, not a deployable system. Operational use would risk reinforcing coverage bias and encoding demographic proxies as risk signals; we will release the work under a non-commercial research license prohibiting operational deployment.

VI. Conclusion

Integrating three heterogeneous public mass-shooting databases into a common schema enables the first leakage-free assessment of cross-source generalization in this domain. Under this protocol, neither feature set can be ranked above the other across 27 paired comparisons; we report the sign-split descriptively because dependence among configurations precludes a valid pooled test on $n = 806$ rows. The highest-mean configuration is partly driven by a known distributional artifact in Mother Jones, and within-dataset VeryHigh precision of 0.147 renders the pipeline unsuitable for operational use. At current sample sizes, model performance is dominated more by dataset structure than by feature or model choice. The temporal holdout further reveals that LODO and temporal rankings select different models, suggesting cross-source and cross-time generalization are complementary evaluation axes.

References

[1] J. Peterson and J. Densley, "The Violence Project Database of Mass Shootings in the United States," The Violence Prevention Project Research Center, 2024.

[2] G. Parmigiani, B. Barchielli, S. Casale, T. Mancini, and S. Ferracuti, "The impact of machine learning in predicting risk of violence: A systematic review," *Front. Psychiatry*, vol. 13, p. 1015914, 2022. doi: 10.3389/fpsyt.2022.1015914.

[3] J. A. Fox and M. J. DeLateur, "Mass shootings in America: Moving beyond Newtown," *Homicide Studies*, vol. 18, no. 1, pp. 125–145, 2014.

[4] M. Rocque and G. Duwe, "Rampage shootings: An historical, empirical, and theoretical overview," *Current Opinion in Psychology*, vol. 19, pp. 28–33, 2018.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[6] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[7] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[8] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, pp. 671–732, 2016.

[9] M. Follman, G. Aronsen, and D. Pan, "U.S. Mass Shootings, 1982–2026," *Mother Jones*, 2026, motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/.

[10] Stanford Geospatial Center, "Stanford Mass Shootings in America (MSA) Database, 1966–2016," 2016, github.com/StanfordGeospatialCenter/MSA.

[11] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Medicine*, vol. 15, no. 11, e1002683, 2018.