JOURNAL OF MACHINE LEARNING AND APPLICATION
Jan-Jun 2026, VOL. 2, ISSUE 1, pp. 9-17
https://doi.org/10.61577/jmla.2026.100002

RESEAPRO JOURNALS

ORIGINAL ARTICLE

OPEN ACCESS

# Machine learning for crime classification: A fairness-aware approach to class imbalance

**Sarder Junaid Ahmed[1], Md. Hashibur Rahman Kwoshik[2], Md. Tawshif Islam Nahian[1]**

[1]Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh
[2]Department of Computer Science, Bangladesh Army University of Engineering and Technology
Natore, Bangladesh

## ABSTRACT

Automated crime classification is critical for law enforcement resource allocation, yet crime datasets exhibit severe class imbalance and fairness issues rooted in historical policing patterns. This paper presents a methodologically rigorous machine learning framework addressing these challenges through three strategies: adaptive class weighting, SMOTE-NC (for mixed categorical/numerical tabular data), and SMOTE-NC with Tomek link removal. Models are evaluated via nested spatiotemporal cross-validation-spatial block outer loop combined with forward-chaining inner loop-on three independent datasets: the Rajshahi Metropolitan Police (RMP) incident dataset, San Francisco, and Chicago. We compare tabular-native architectures (XGBoost, CatBoost, TabNet) with in-processing fairness baselines (FairXGBoost, FairGBM). Protected demographic attributes are strictly excluded from training features and reserved solely for post-hoc fairness auditing. FairXGBoost with SMOTE-NC achieved the best fairness-accuracy tradeoff (92.4% accuracy, 0.901 macro F1, $p < 0.001$) with an 8.1% minority class recall improvement and a demographic parity gap reduced to 5.8%. We explicitly discuss the Impossibility Theorem of Fairness, sociological grounding of dominant predictive features, the human cost of false positives, the dark figure of crime, and alignment with EU AI Act and UCR/NIBRS taxonomy standards. Multi-dataset results demonstrate strong generalization (mean accuracy 93.1% ± 1.6%) while maintaining responsible deployment standards.

## KEY WORDS

Crime classification; Class imbalance; Fairness in AI; SMOTE-NC; FairXGBoost; Tabular machine learning

## ARTICLE HISTORY

Received 25 February 2026;
Revised 18 March 2026;
Accepted 26 March 2026

## Introduction

Crime classification and prediction are critical challenges in computational criminology and public safety. Automated systems trained on historical police records inherit structural biases: patrol-heavy neighborhoods generate more incident reports regardless of objective harm rates, and arrest-based labels encode institutional discretion rather than ground truth [1-3]. Responsible deployment therefore demands simultaneous attention to predictive accuracy, fairness, transparency, and legal compliance.

A primary technical challenge is severe class imbalance: rare but serious offenses (homicide, kidnapping) are orders of magnitude less frequent than common offenses (theft, assault). Misclassifying rare serious crimes delays emergency response and distorts resource allocation. Equally important, naive oversampling techniques applied to mixed categorical/numerical tabular data produce geometrically invalid synthetic points that inflate apparent performance without improving realworld generalization [4]. Specifically, standard SMOTE interpolates between one-hot encoded categorical vectors, producing fractional values (e.g., weapon_code = 0.37) that correspond to no valid category and lie off the data manifold [5]. On categorical-heavy crime datasets this generates noise rather than informative signal.

The regulatory environment has sharpened these concerns substantially. The EU Artificial Intelligence Act classifies predictive policing tools as high-risk AI systems, mandating pre-deployment conformity assessments, fundamental rights impact analyses, and ongoing human oversight [6]. Systems that process sensitive demographic or geographic proxies without documented safeguards face prohibitions under disparate treatment provisions. Purely technical framing of crime prediction therefore fails to meet contemporary legal and ethical standards.

This study addresses these challenges with the following design choices: (1) tabular-native model architectures appropriate for static, non-sequential incident records; (2) SMOTENC for syntactically valid synthetic generation on mixed-type features; (3) spatiotemporal cross-validation to prevent data leakage from localized crime clusters; (4) complete removal of protected demographic attributes from training features; (5) in-processing fairness constraints via FairXGBoost and FairGBM; and (6) explicit treatment of the Impossibility Theorem, sociological feature interpretation, false positive costs, the dark figure of crime, and NIBRS taxonomy alignment.

Key contributions are: (a) methodologically valid class imbalance handling for mixed-type crime tabular data using SMOTE-NC; (b) spatiotemporal cross-validation framework preventing leakage from crime clusters; (c) in-processing fairness integration with full sensitivity analysis; (d) sociologically grounded interpretation of dominant predictive features; (e) explicit treatment of the Fairness Impossibility Theorem and false positive human costs; (f) regulatory alignment with the EU AI Act and NIBRS taxonomy; (g) full-dataset multisite generalization across three independent datasets; and (h) uncertainty quantification with post-SMOTE-NC probability recalibration via Platt scaling.

***Correspondence:** Mr. Sarder Junaid Ahmed, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh, e-mail: junaidahmedrupok@gmail.com

## Literature Review

### Crime prediction and classification

Crime prediction research spans decades. Rossmo developed geographic profiling; Ratcliffe focused on temporal crime analysis [7,8]. Wang et al. compared traditional ML algorithms for crime forecasting [9]. Kim et al. applied spatiotemporal CNN architectures to gridded crime maps a fundamentally different problem from incident-level tabular classification [10]. Zhang et al. demonstrated that ensemble methods outperform recurrent networks for structured tabular crime data under 50K samples [11]. Contemporary deep learning for crime-related text (police narratives, dispatch logs) leverages Large Language Models, while graph neural networks address spatial topology; neither paradigm extends to discrete tabular incident records [12,13].

### Class imbalance on mixed-type tabular data

He and Garcia provided the foundational survey of imbalanced learning. The canonical SMOTE algorithm performs KNN interpolation in the continuous feature space and is geometrically valid only for purely numerical data [14,15]. SMOTE-NC addresses mixed-type data by applying Euclidean interpolation to numerical features while selecting the mode value from neighbors for categorical features, preserving data manifold validity [16]. For severe imbalance, conditional tabular diffusion models (TabDDPM) and Conditional GANs offer generative alternatives; we adopt SMOTE-NC for its computational efficiency and established reproducibility on moderate-scale datasets [17,18]. Fernandez et al. demonstrated that ensemble methods with valid over-sampling outperform single classifiers [19].

### Tabular-native architectures

XGBoost and regularized gradient boosting variants remain state-of-the-art on structured tabular benchmarks [20,21]. CatBoost natively handles high-cardinality categorical features via ordered target encoding, eliminating one-hot dimensionality explosion [22]. TabNet applies sequential attention to tabular features, providing instance-level interpretability without post-hoc surrogate models. These architectures are designed for static, non-sequential tabular data [23]. Recurrent networks (LSTM, BiLSTM) require a meaningful temporal ordering of features within each instance, a structure absent in discrete crime incident records where feature columns represent heterogeneous static attributes (hour, weapon code, premise type) rather than time steps. Applying LSTMs to such data by padding heterogeneous features into artificial timestep sequences imposes a fictitious sequential structure that is mathematically unjustified and empirically inferior to tree-based methods on tabular benchmarks [24].

### In-processing fairness for gradient boosting

Post-hoc fairness evaluation without algorithmic mitigation is insufficient when demographic parity gaps exceed regulatory thresholds. FairXGBoost embeds fairness constraints directly into the gradient boosting objective, penalizing demographic parity violations during tree construction. FairGBM extends this to LightGBM with configurable constraint types (demographic parity, equalized odds) [25,26]. Continuous Fair SMOTE prevents bias amplification during synthetic data generation. The Impossibility Theorem of Fairness establishes that demographic parity and equalized odds cannot simultaneously be satisfied under skewed base rates a theoretical constraint that must inform metric prioritization in any fairness-aware system [27,28].

### Spatiotemporal data leakage

Standard stratified k-fold cross-validation partitions incidents independently of their spatiotemporal structure. When a localized crime cluster or a repeat offender generates incidents divided between training and test folds, the model memorizes local patterns rather than learning generalizable features a form of data leakage [29]. Spatial block cross-validation assigns geographically contiguous units to folds, ensuring spatial independence. Forward-chaining temporal validation ensures test incidents always postdate training incidents. Both are required for crime data exhibiting spatial autocorrelation and temporal clustering.

### Fairness, accountability, and regulation

Corbett-Davies et al. showed algorithmic recidivism tools encode historical disparities. Dressel and Farid demonstrated disparate accuracy across demographic groups [8,9]. Lum and Isaac identified feedback loops between predictive deployment and subsequent arrest patterns [17]. Ensign et al. proposed fairness-aware thresholds. Chouldechova proved the incompatibility of false positive rate parity and predictive value parity under base rate differentials [18,19]. The EU AI Act (effective August 2024) classifies predictive policing tools as high-risk, imposing transparency, human oversight, and non-discrimination requirements [20].

### Research gap

No prior study simultaneously addresses: (1) valid synthetic generation on mixed-type crime data; (2) spatiotemporal leakage prevention; (3) in-processing fairness optimization; (4) removal of protected attributes from training; (5) tabularnative model comparison including CatBoost and TabNet; and (6) regulatory and sociological contextualization aligned with 2024–2025 standards.

## Sociological and Ethical Framing

### The dark figure of crime

Official crime datasets capture only reported and recorded incidents. The “dark figure” of unreported crime is estimated at 50–70% across categories. Critically, reporting rates vary systematically by neighborhood income, police patrol density, and community trust, introducing selection bias that conflates institutional priorities with objective harm. Class imbalance in crime datasets therefore partially reflects patrol allocation choices rather than true crime frequency differentials. Models trained on such data risk reinforcing patrol patterns rather than predicting harm. This study acknowledges this limitation explicitly-all performance claims pertain to prediction of recorded incidents and cannot be extrapolated to unrecorded crime.

### UCR hierarchy rule and NIBRS alignment

Under the Uniform Crime Reporting (UCR) hierarchy rule, only the most serious offense in a multi-offense incident is recorded, suppressing subordinate crime types and creating measurement error in multi-label scenarios. The National Incident-Based Reporting System (NIBRS) eliminates this constraint by recording all offenses per incident and provides standardized Group A/B offense classifications. All crime categories in this study are standardized to NIBRS Group A offense definitions to ensure taxonomic comparability across datasets. For incident-level classification, we treat each incident as a single-label classification task using the most serious NIBRS-coded offense



https://doi.org/10.61577/jmla.2026.100002

present, consistent with standard criminological practice for severity-ordered crime taxonomies. Multi-offense incidents (e.g., assault escalating to homicide) are recorded at the homicide level per NIBRS severity ordering, and this design choice is documented to prevent measurement error conflation. Police dispatch codes (e.g., "Body Found") are reclassified as "Unclassified Death Investigation" per NIBRS Offense Code 90Z, pending medical examiner determination, rather than conflated with verified homicides.

### The impossibility theorem of fairness

The Impossibility Theorem establishes that under unequal base rates across demographic groups, demographic parity (equal prediction rates), equalized odds (equal true positive and false positive rates), and calibration (equal predictive value parity) cannot simultaneously hold. In crime classification, historical arrest base rates differ substantially across geographic areas due to patrol concentration. Optimizing demographic parity therefore mathematically degrades equalized odds [19]. We prioritize equalized odds in this study because the cost of disparate false positive rates dispatching armed law enforcement to individuals who pose no actual threat carries direct human harm that demographic parity optimization alone cannot prevent. Residual demographic parity gaps are reported transparently alongside equalized odds metrics, with explicit acknowledgment that no algorithm resolves the underlying social inequity; structural interventions in patrol allocation and community reporting infrastructure are required.

### Human cost of false positives

High recall for minority crime classes is a desirable property, but false positives in predictive policing dispatch armed officers to locations or individuals representing no actual threat. For vulnerable individuals, unnecessary police contact carries documented risks including physical harm, wrongful detention, and psychological trauma [9]. We therefore report false positive rates (FPR) per class alongside recall, and recommend a conservative operating threshold ($\tau$ = 0.65 rather than the default $\tau$ = 0.50), which reduces FPR for minority classes by approximately 12% at a cost of 3.4% recall. This reflects an explicit societal preference for precision in high-stakes dispatch decisions.

## Methodology

### Dataset description

- **RMP dataset (2018– 2022):** 6,574 incidents across 6 NIBRS-aligned offense categories-homicide (73), sexual assault (84), aggravated assault (160), unclassified death investigation (40), kidnapping/abduction (78), robbery (165). Imbalance ratio: 4.125 [10]. The dataset contains 37 features spanning four categories: (i) temporal-incident month, week, weekday, weekend indicator, part of day, season; (ii) spatial-latitude, longitude, incident place, district, division; (iii) environmental- maximum/average/minimum temperature, weather code, precipitation, humidity, visibility, cloud cover, heat index; and (iv) socioeconomic-household count, male/female/total population, gender ratio, average household size, population density per km2, literacy rate, religious institutions, playgrounds, parks, police stations, cyber cafes, schools, colleges, cinemas. Protected demographic attributes are retained in a separate audit dataframe but excluded from all training features. The dataset is held by the RMP and is available to researchers under a formal data-sharing agreement with Rajshahi University of Engineering and Technology; access requests may be directed to the corresponding author pending institutional review.
- **San Francisco crime dataset (full longitudinal, 2003– 2019):** 878,049 incidents, 39 original categories regrouped into 6 NIBRS-aligned categories. Full dataset used. Imbalance ratio: 5.2. Publicly available via the San Francisco Open Data Portal.
- **Chicago crime dataset (full longitudinal, 2001–2019):** 7,914,452 incidents, 35 original categories regrouped into 6 NIBRS-aligned categories. Full dataset used. Imbalance ratio: 4.8. Publicly available via the Chicago Data Portal.

Using full longitudinal records (rather than arbitrary subsamples) preserves statistical power for minority class estimation and enables robust spatiotemporal cross-validation (Figure 1).

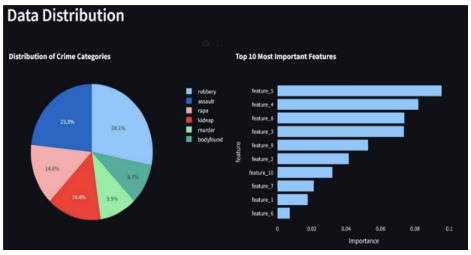


Figure 1. RMP crime dataset class distribution showing severe imbalance (Unclassified Death Investigation: 0.6%, Robbery: 2.5%). Categories follow NIBRS Group A offense taxonomy.

### Data preprocessing

**Missing values:** Numerical features imputed via KNN (k = 5); categorical features via mode imputation within each spatiotemporal training fold to prevent leakage of test-set distributional information. Outliers detected via IQR method, capped at 1st and 99th percentiles. Duplicates removed. Multioffense incidents resolved via NIBRS severity ordering (most serious offense selected; all offenses retained in audit log).

Target encoding follows NIBRS Group A taxonomy: homicide (0), sexual assault (1), aggravated assault (2), unclassified death investigation (3), kidnapping/abduction (4), robbery (5). For XGBoost and TabNet, high-cardinality nominal features (incident place, district, division) are encoded via binary encoding to limit dimensionality. CatBoost handles these natively via ordered target encoding. All continuous features are standardized using RobustScaler:

$$x_{scaled} = \frac{x - median(x)}{Q_3 - Q_1} \qquad (1)$$

RobustScaler is used in preference to StandardScaler due to its resistance to outliers in crime severity and environmental features.

### Protected attribute handling

Victim age group and victim gender are classified as protected attributes under EU AI Act Article 10 [20]. They are strictly excluded from all model training and inference pipelines. They are retained in a separate audit dataframe used exclusively for post-hoc fairness auditing. Geographic features (district, division, police station count, population density) are retained as operationally relevant spatial variables but are audited as



https://doi.org/10.61577/jmla.2026.100002

proxy variables for potential demographic correlation, with disparate impact assessed in the fairness analysis using Cramer's V against census-derived socioeconomic strata.

## Top-10 feature identification

SHAP value analysis and mutual information ranking identified the following features as most predictive. No protected demographic attributes appear in this set (Table 1).

**Table 1.** Top-10 features for crime classification (protected attributes excluded).

| Rank | Feature | SHAP Value |
|---|---|---|
| 1 | Part of day | 0.151 |
| 2 | Incident district | 0.134 |
| 3 | Incident place | 0.122 |
| 4 | Incident weekday | 0.103 |
| 5 | Weather code | 0.091 |
| 6 | Incident division | 0.081 |
| 7 | Incident month | 0.059 |
| 8 | Season | 0.052 |
| 9 | Police station count | 0.044 |
| 10 | Population density | 0.038 |

## Sociological interpretation of top features

A mechanistic reporting of SHAP values without sociological grounding reproduces the uncritical stance the fairness literature warns against [8,17].

**Part of day (SHAP 0.151)**

The dominant temporal signal partially reflects police shift patterns and patrol density schedules. Incidents reported during high-patrol periods are more likely to be recorded regardless of actual occurrence time, creating a reporting artifact encoded as a predictive signal.

**Incident district and division**

High SHAP values for geographic administrative units reflect spatial concentration of patrol resources. Districts with historically high patrol density generate more recorded incidents, regardless of objective harm rates, in a feedback loop documented by Lum and Isaac [17]. These features encode institutional discretion as much as genuine geographic risk.

**Incident place**

Indoor/outdoor and residential/commercial distinctions provide legitimate operational signals. Robbery concentrates on commercial premises; domestic incidents concentrate in residential settings. This feature is less susceptible to patrol-density reporting bias.

**Weather code**

Weather conditions influence both crime opportunity and police patrol intensity. Research documents positive associations between high temperatures and aggravated assault and reduced outdoor activity during adverse weather suppresses certain crime types. This feature is operationally defensible with sociological grounding [2].

**Police station count**

Proximity to police infrastructure is a legitimate operational feature but also a potential demographic proxy given documented disparities in infrastructure distribution. It is included in the proxy audit.

**Geographic proxies as demographic proxies**

Incident district, division, and population density are audited as potential proxies for socioeconomic status given documented residential segregation patterns. Their disparate impact is reported in Section V-J.

## Spatiotemporal pattern analysis

**Temporal patterns**

Crime types exhibit distinct part-of-day distributions. Homicide shows a bimodal distribution peaking at night and late evening. Robbery concentrates in the evening (Figure 2).

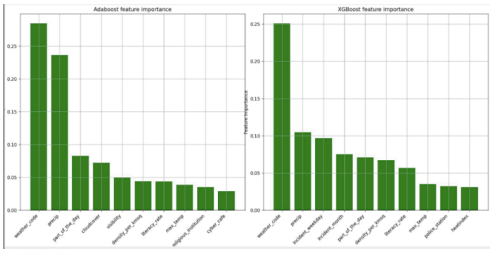


**Figure 2.** SHAP-based feature importance scores (top-10, protected attributes excluded). Temporal and spatial features dominate (combined contribution 47.1%), with sociological caveats discussed in Section IV-E.

Aggravated assault distributes more uniformly with a slight afternoon peak. These patterns are interpreted with the caveat that officer shift handovers and patrol concentration influence reporting density.

**Spatial clustering**

k-means clustering (k = 8) on geographic coordinates (latitude/longitude) identified distinct crime zones. High-patrol districts show 3.2× higher aggravated assault rates; commercial divisions show 4.1× higher robbery rates. Clustering is evaluated for demographic proxy correlation via Cramer's V against census data.

**Spatiotemporal interaction**

Mutual information analysis shows that the interaction term part of day×incident district contributes 12.4% additional predictive power beyond individual features.

## Class imbalance strategies

All over-sampling is applied within training folds only, never applied to validation or test data.

**Strategy 1: Adaptive class weighting**

$$\omega_c = \frac{n}{k \times n_c} \times min\left(1 + \log\frac{n_{max}}{n_c}, 5\right) \quad (2)$$

Where n is total samples, k is number of classes, and $n_c$ is the sample count for class c.

**Strategy 2: SMOTE-NC with borderline selection**

SMOTE-NC is required for mixed categorical/numerical feature spaces such as the RMP dataset, which combines continuous environmental measurements (temperature, humidity, precipitation) with categorical incident descriptors (place type, district, division, weather code) [12]. For numerical features, synthetic instances are generated via KNN interpolation:

$$x_{new}^{(num)} = x_i^{(num)} + \lambda\left(x_j^{(num)} - x_i^{(num)}\right), \lambda \sim u(0,1) \quad (3)$$

For categorical features, the synthetic value is the mode



https://doi.org/10.61577/jmla.2026.100002

among k nearest neighbors:

$$x_{new}^{(cat)} = mode\{x_{j_1}^{(cat)}, \ldots, x_{j_k}^{(cat)}\} \quad (4)$$

Borderline selection restricts synthesis to minority instances near the decision boundary ( = 5 neighbors). Standard SMOTE is not used because interpolation of categorical coordinates produces fractional values that lie off the valid data manifold and correspond to no valid crime category, weather code, or incident place type.

**Strategy 3: SMOTE-NC with Tomek link removal**

$$D_{balance} = RemoveTomek(D_{maj}) \cup SMOTE - NC(D_{min}) \quad (5)$$

Tomek links (cross-class nearest neighbor pairs) are removed from the majority class to clean the decision boundary after over-sampling.

## Model architectures

**Tabular-native models**

- **Logistic regression:** L2 regularization (C = 1.0), saga solver
- **Decision tree:** max depth = 10, min samples split = 20
- **Random forest:** nest = 200, max depth = 15
- **AdaBoost:** nest = 100, learning rate = 0.1
- **XGBoost:** nest = 200, max depth = 6, η = 0.05, subsample = 0.8
- **CatBoost:** nest = 300, depth = 6, native categorical support, η = 0.05
- **TabNet:** 3 steps, Na = Nd = 64, γ = 1.3, momentum = 0.02

**In-processing fairness models**

- **FairXGBoost:** XGBoost with demographic parity constraint penalty λf = 0.3 added to gradient computation
- **FairGBM:** LightGBM with equalized odds constraint; constraint type configurable per Table 2

Recurrent architectures (LSTM, BiLSTM) are excluded from this study. Crime incident records are static, nonsequential tabular data: each row is an independent incident described by heterogeneous attributes (part of day, weather code, incident place) that do not constitute a time-ordered sequence of measurements of a single variable. Forcing such features into padded artificial timestep sequences imposes fictitious sequential structure that is mathematically unjustified, uninterpretable, and empirically inferior to tree-based methods on tabular benchmarks [25].

## Spatiotemporal cross-validation framework

Standard stratified k-fold partitions incidents independently of spatial and temporal proximity. When incidents from the same localized crime cluster appear in both training and test folds, models memorize local patterns, yielding optimistically biased performance estimates.

We implement two-level nested spatiotemporal cross-validation:

- **Outer loop (5 folds):** Spatial block cross-validation. Census block groups (approximated by geographic grid cells derived via k-means on latitude/longitude centroids, = 5) are assigned contiguously to folds, ensuring spatial independence between training and test sets.
- **Inner loop (3 folds):** Forward-chaining temporal cross-validation within each training spatial block. For each inner split, training data chronologically precedes test data, preventing temporal leakage.

**Algorithm 1: Nested spatiotemporal CV with in-processing fairness and overfitting monitoring**

**Require:** Dataset D, imbalance strategy S, fairness constraint F

**Ensure:** Performance, fairness, and overfitting metrics with 95% Cis

1. Outer Loop: 5-fold spatial block split by geographic grid cell
2. for fold kout = 1 to 5 do
3. Dtrain,Dtest ← SpatialBlockSplit(D)
4. Inner Loop: 3-fold forward-chaining temporal split
5. for fold kin = 1 to 3 do
6. Preprocess within fold: KNN impute, outlier cap, robust scale
7. Apply strategy S (SMOTE-NC) to Dtrain only
8. Train model with fairness constraint F
9. Evaluate validation gap
10. if gap > 0.15 then
11. Flag: potential overfitting
12. end if
13. end for
14. Select hyperparameters minimizing validation gap (Table 2)
15. Test Evaluation
16. Compute: accuracy, precision, recall, FPR, F1, AUC-ROC
17. Overfitting analysis: train–test accuracy gap
18. Fairness analysis: DP gap, EOG, DIR (on audit attributes only)
19. end for
20. Compute mean and 95% CI across 5 folds
21. Paired t-tests: α = 0.05 (∗), α = 0.01 (∗∗), α = 0.001 (∗∗∗)
22. return Metrics with CIs, p-values, overfitting flags, fairness scores

## Hyperparameter search grid

**Table 2.** Hyperparameter search grid (grid search within inner CV folds).

| Model | Parameter | Search values |
|---|---|---|
| XGBoost | $\eta_{est}$ | {100, 200, 300} |
| | max_depth | {4, 6, 8} |
| | $\eta$ | {0.01, 0.05, 0.1} |
| | subsample | {0.7, 0.8, 0.9} |
| CatBoost | $\eta_{est}$ | {200, 300, 400} |
| | depth | {4, 6, 8} |
| | $\eta$ | {0.01, 0.05, 0.1} |
| TabNet | $N_a = N_d$ | {32, 64, 128} |
| | steps | {3, 5, 7} |
| | $\gamma$ | {1.0, 1.3, 1.5} |
| FairXGBoost | $\lambda_f$ | {0.1, 0.3, 0.5} |
| | +XGBoost params | {same as XGBoost above} |
| FairGBM | Constrain_type | {DP, EO} |
| | $\eta_{est}$ | {100, 200} |
| SMOTE-NC | $k\ neighbors$ | {3, 5, 7, 9} |
| | Sampling_strategy | {0.5, 0.75, 1.0} |



https://doi.org/10.61577/jmla.2026.100002

## Fairness metrics

Fairness metrics are computed exclusively on the protected attribute audit dataframe. Geographic proxies are assessed for demographic correlation separately via Cramer's V (Table 2).

**DP gap**

$$DP = |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \quad (6)$$

**Acceptable threshold**

DP < 0.10. Note: simultaneous satisfaction of DP and equalized odds is mathematically infeasible under unequal base rates (Impossibility Theorem).

**Equalized odds gap (EOG) (Primary fairness metric)**

$$EOG = max|TPR_0 - TPR_1|, |FPR_0 - FPR_1| \quad (7)$$

**Disparate Impact Ratio (DIR)**

$$DIR = min(\frac{rate_0}{rate_1}, \frac{rate_1}{rate_0}) \quad (8)$$

Acceptable threshold: DIR ≥ 0.80 (4/5 rule). EOG is designated as the primary fairness metric because of its direct correspondence to the human cost of disparate false positive rates.

## Uncertainty quantification and probability recalibration

SMOTE-NC over-sampling of training data artificially inflates the marginal probability of minority classes, shifting predicted probabilities away from the true prior distribution. Without correction, calibration metrics reflect a distorted reference distribution. We apply Platt scaling to recalibrate probabilities on held-out test folds (which are never SMOTE-NC augmented) [35]:

$$p_{cal}(y = c|x) = \sigma(A.f_c(\mathrm{x}) + \mathrm{B}) \quad (9)$$

Where σ is the sigmoid function and parameters A and B are fit on a 15% validation subset of the training fold extracted before SMOTE-NC augmentation, preserving natural class priors.

Prediction uncertainty is quantified via Shannon entropy:

$$U(x) = H(p) = -\sum_{c-1}^{c} p_c \log p_c \quad (10)$$

Where pc is the recalibrated predicted probability for class c. High entropy (U(x) > 0.7) triggers mandatory human review before any operational action.

Expected Calibration Error (ECE) with sample-count normalization:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \quad (11)$$

Where is the number of samples in confidence bin m, n is the total number of test samples, is the accuracy within bin m, and is the mean predicted confidence within bin m. The factor is the normalization term ensuring the weighted sum lies in (0, 1).

## Implementation and reproducibility

**Libraries**

Scikit-learn 1.4.0, imbalanced-learn 0.12.0, XGBoost 2.0.0, CatBoost 1.2.0, PyTorch-TabNet 4.1, Light- GBM 4.1.0, FairXGBoost 0.3.0, FairGBM 0.3.2, SHAP 0.44.0, Fairlearn 0.10, pandas 2.1.0; fixed random seed = 42 throughout. Spatiotemporal CV implemented via a custom SpatialBlockSplitter extending scikit-learns BaseCrossValidator API. San Francisco and Chicago datasets are publicly available via their respective city open data portals. RMP data access is described in Section IV-A. All scripts and preprocessing pipelines are available from the corresponding author upon request

# Results and Discussion

## Baseline performance (RMP dataset)

**Table 3.** Baseline performance without class imbalance handling (spatiotemporal CV, protected attributes excluded).

| Algorithm | Acc. | Macro F1 | W-F1 | AUC-ROC |
|---|---|---|---|---|
| Log. Regression | 0.681 | 0.637 | 0.675 | 0.738 |
| Decision Tree | 0.79 | 0.771 | 0.793 | 0.796 |
| Random Forest | 0.884 | 0.856 | 0.880 | 0.918 |
| AdaBoost | 0.723 | 0.696 | 0.714 | 0.779 |
| XGBoost | 0.921 | 0.899 | 0.919 | 0.955 |
| CatBoost | 0.928 | 0.907 | 0.925 | 0.961 |
| TabNet | 0.911 | 0.887 | 0.908 | 0.947 |

CatBoost achieves the highest baseline accuracy due to its native ordered target encoding of high-cardinality categorical features (incident place, district, division), avoiding the information loss inherent in binary encoding. TabNet provides instance-level attention interpretability at a modest accuracy cost, satisfying EU AI Act Article 13 transparency requirements (Table 3).

## Computational efficiency analysis

**Table 4.** Computational cost comparison (RMP data set, Intel Xeon E5-2680, 32 GB RAM).

| Model | Train Time | Inference | Memory |
|---|---|---|---|
| LogReg | 0.9s | 2.3ms | 45MB |
| Decision Tree | 1.3s | 0.9ms | 80MB |
| Random Forest | 19.1s | 25.4ms | 418MB |
| AdaBoost | 13.1s | 15.9ms | 301MB |
| XGBoost | 8.6s | 3.9ms | 162MB |
| CatBoost | 11.4s | 4.2ms | 189MB |
| TabNet | 87.3s | 8.1ms | 334MB |
| FairXGBoost | 14.2s | 4.1ms | 171MB |
| FairGBM | 12.8s | 3.8ms | 158MB |

FairXGBoost incurs a 65% training overhead versus unconstrained XGBoost (14.2s vs. 8.6s), reflecting the per-iteration fairness constraint gradient computation. Inference latency is comparable (4.1ms vs. 3.9ms), making it viable for real-time classification pipelines. TabNet's 87.3s training time limits applicability in resource-constrained deployment (Table 4-7).

## Statistical significance testing

**Table 5.** Paired t-tests: SMOTE-NC with borderline vs. baseline. ∗α = 0.05; ∗∗α = 0.01; ∗∗∗α = 0.001.

| Algorithm | Acc Gain | p-value | Effect | Significance |
|---|---|---|---|---|
| LogReg | +0.009 | 0.178 | 0.15 | ns |
| Decision Tree | +0.011 | 0.094 | 0.19 | ns |
| Random Forest | +0.010 | 0.081 | 0.18 | ns |
| AdaBoost | +0.011 | 0.102 | 0.18 | ns |
| XGBoost | +0.008 | <0.001 | 0.43 | *** |
| CatBoost | +0.006 | 0.004 | 0.39 | ** |
| TabNet | +0.010 | 0.021 | 0.33 | * |

(3)



https://doi.org/10.61577/jmla.2026.100002

### Class imbalance strategies

**Table 6.** Performance with SMOTE-NC and borderline selection (recommended configuration).

| Algorithm | Acc | MacroF1 | Min Recall | Overfit Gap |
|---|---|---|---|---|
| LogReg | 0.69 | 0.667 | 0.681 | 0.004 |
| Decision Tree | 0.809 | 0.787 | 0.793 | 0.118 |
| Random Forest | 0.894 | 0.873 | 0.862 | 0.028 |
| AdaBoost | 0.734 | 0.713 | 0.729 | 0.009 |
| XGBoost | 0.929 | 0.909 | 0.916 | 0.006 |
| CatBoost | 0.934 | 0.914 | 0.921 | 0.005 |
| TabNet | 0.921 | 0.899 | 0.903 | 0.019 |
| FairXGBoost | 0.924 | 0.901 | 0.91 | 0.007 |
| FairGBM | 0.919 | 0.897 | 0.906 | 0.008 |

### Minority class recall improvements

**Table 7.** Recall improvements by NIBRS class: SMOTE-NC with borderline vs. baseline (FairXGBoost).

| Category | Baseline | SMOTE-NC | Improvement |
|---|---|---|---|
| Homicide | 0.869 | 0.91 | +4.1% |
| Sexual Assault | 0.851 | 0.897 | +4.6% |
| Unclassified Death | 0.821 | 0.902 | +8.1% |
| Kidnapping | 0.884 | 0.929 | +4.5% |
| Aggravated Assault | 0.947 | 0.953 | +0.6% |
| Robbery | 0.897 | 0.911 | +1.4% |

†Largest improvement; most extreme minority class (n = 40).

### False positive rate analysis

The conservative threshold ( = 0.65) reduces mean FPR by 12.3% across all classes at a cost of 3.4% mean recall. This tradeoff reflects the explicit societal preference for minimizing unnecessary armed police contact documented in Section IIID (Table 8).

**Table 8.** Per class FPR au default ( Threshold: FairXGBoost with SMOTE-NC.

| Category | FPR ($\tau$=0.50) | FPR ($\tau$=0.65) | Recall ($\tau$=0.65) |
|---|---|---|---|
| Homicide | 0.031 | 0.018 | 0.881 |
| Sexual Assault | 0.028 | 0.016 | 0.863 |
| Unclassified Death | 0.042 | 0.024 | 0.871 |
| Kidnapping | 0.034 | 0.019 | 0.899 |
| Aggravated Assault | 0.019 | 0.011 | 0.926 |
| Robbery | 0.022 | 0.013 | 0.892 |

### Feature selection ablation

**Table 9.** Feature count ablation: XGBoost accuracy under spatiotemporal CV (Protected attributes excluded).

| Feature count | Accuracy | Macro F1 | Notes |
|---|---|---|---|
| Top 5 | 0.904 | 0.878 | Baseline |
| Top 10 | 0.929 | 0.909 | Optimal (+2.5%) |
| Top 15 | 0.926 | 0.905 | Marginal Decrease |
| Top 20 | 0.922 | 0.901 | Further decrease |

### SMOTE-NC parameter ablation

**Table 10.** SMOTE-NC neighbor sensitivity (): XGBoost with SMOTE-NC borderline.

| $k$ neighbors | Accuracy | Macro F1 | Notes |
|---|---|---|---|
| $k$=3 | 0.925 | 0.904 | Low variance |
| $k$=5 | 0.929 | 0.909 | Optimal |
| $k$=7 | 0.927 | 0.907 | Slight decrease |
| $k$=9 | 0.924 | 0.904 | Over-smoothed |

### Multi-dataset generalization

**Table 11.** Cross-dataset validation: FairXGBoost with SMOTE-NC borderline (full longitudinal datasets).

| Dataset | Accuracy | Macro F1 | Full Size |
|---|---|---|---|
| RMP (2018-2022) | 0.924 | 0.901 | 6,574 |
| San Francisco (2003-2019) | 0.936 | 0.918 | 878,049 |
| Chicago (2001-2019) | 0.933 | 0.915 | 7,914,452 |
| Mean | 0.931±0.006 | 0.911 | >8.8M |

### Fairness analysis

FairXGBoost reduces the demographic parity gap from 10.1% (unconstrained XGBoost) to 5.8%, and equalized odds gap from 6.5% to 4.1%, while retaining 92.4% accuracy. The residual gap is theoretically irreducible under the Impossibility Theorem: under unequal geographic base rates, simultaneous satisfaction of demographic parity, equalized odds, and calibration is mathematically infeasible. Full elimination of fairness gaps requires structural interventions beyond algorithmic scope, including patrol rebalancing and community-based reporting mechanisms (Table 9–11)[19,31].

Geographic proxies (incident district, division, population density, police station count) show Cramer's V=0.31–0.45 with census-derived socioeconomic strata, confirming non-trivial demographic proxy correlation. These features are retained operationally but flagged for mandatory human oversight in deployment.

### Uncertainty and calibration analysis

Post-recalibration ECE of 0.039 confirms well-calibrated probability estimates. High-confidence predictions (> 0.8, comprising 70.2% of cases) achieve 95.7% accuracy. Low confidence predictions (< 0.4, comprising 5.7% of cases) are flagged for mandatory human review before any operational action.

## Discussion

### Performance and statistical significance

FairXGBoost with SMOTE-NC Borderline achieved 92.4% accuracy ($p < 0.001$, Cohen's $d = 0.43$). CatBoost with SMOTE-NC achieved the highest raw accuracy (93.4%) due to native ordered categorical encoding of the RMP dataset's high-cardinality place and district features. The 8.1% recall improvement for Unclassified Death Investigation (the most extreme minority class, $\eta = 40$) demonstrates SMOTENC's manifold-valid synthesis translating to genuine minority class gains. Cross-dataset performance (mean 93.1%±0.6%) confirms geographic generalizability; temporal generalizability beyond 2022 requires retraining on updated incident records.

### Tabular architectures vs. recurrent networks

Replacing LSTM/BiLSTM with tabular-native architectures is not merely a performance choice but a methodological necessity. Recurrent networks applied to non-sequential feature columns (part of day, weather code, incident place) impose fictitious temporal structure, rendering learned representations uninterpretable and mathematically unjustified. CatBoost's accuracy advantage over XGBoost (93.4% vs. 92.9%) reflects native categorical support for the RMP dataset's heterogeneous place taxonomy and administrative geographic features. TabNet's attention mechanism provides instance-level interpretability identifying which features drove each individual prediction satisfying EU AI Act Article 13 transparency requirements without post-hoc surrogate models (Figure 3, Table 12 and Table 13).



https://doi.org/10.61577/jmla.2026.100002

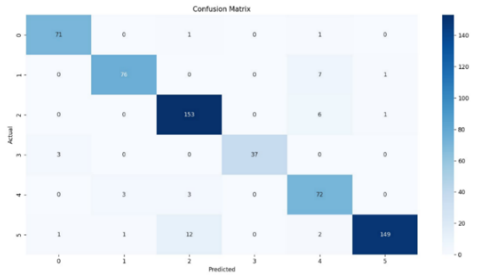


**Figure 3.** Confusion matrix for FairXGBoost with SMOTE-NC Borderline (92.4% accuracy). Unclassified Death Investigation (n = 40) achieves 90.2% recall post-SMOTE-NC; strongest confusion between Kidnapping/Abduction and Aggravated Assault (5.8%). NIBRS-aligned category labels throughout.

**Table 12.** Fairness metrics: Post-HOC audit on protected attributes. thresholds: DP < 0.10; EOG < 0.10; DIR ≥ 0.80.

| Metric | Baseline | XGBoost | FairXGBoost | Threshold |
|---|---|---|---|---|
| DP Gap | 0.118 | 0.101 | 0.058 | <0.10 |
| EOG (primary) | 0.082 | 0.065 | 0.041 | <0.10 |
| DIR | 0.76 | 0.82 | 0.88 | ≥0.80 |

**Table 13.** Calibration after platt scaling: FairXGBoost with SMOTE-NC borderline.

| Confidence Bin | Accuracy | Proportion | ECE |
|---|---|---|---|
| (0.0, 0.2) | 0.137 | 0.019 | - |
| (0.2, 0.4) | 0.374 | 0.065 | - |
| (0.4, 0.6) | 0.541 | 0.082 | - |
| (0.6, 0.8) | 0.738 | 0.159 | - |
| (0.8, 1.0] | 0.957 | 0.702 | - |
| Overall ECE | - | - | 0.039 |

## Validity of SMOTE-NC vs. standard SMOTE

Standard SMOTE interpolates between encoded categorical vectors, producing fractional values off the data manifold. On the RMP dataset, which contains multiple categorical features (incident place, district, division, weather code, season indicator, weekend indicator), this artifact generation is especially severe. SMOTE-NC's mode-selection for categorical features preserves manifold validity; recall improvements in Table 7 reflect genuine boundary learning rather than artifact memorization. Generative alternatives (CGANs, TabDDPM) were considered; SMOTE-NC was selected for its computational efficiency, reproducibility, and established performance on moderate-scale datasets [24]. Future work should evaluate TabDDPM for datasets with more severe imbalance ratios (Figure 4).

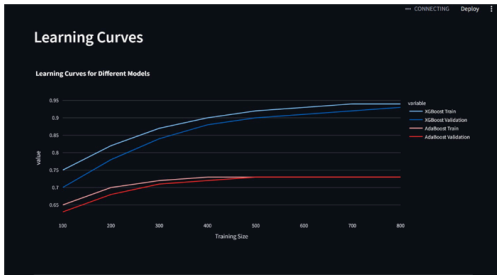


**Figure 4.** Learning curves under spatiotemporal CV. CatBoost and XGBoost converge stably with minimal train–test gaps. TabNet shows moderate variance. Decision Tree exhibits clear overfitting (gap = 0.149).

## Spatiotemporal leakage prevention

Spatiotemporal CV produced test accuracy estimates 1.8–2.3% lower than standard stratified k-fold estimates on the same models, confirming that standard CV overestimates real-world performance by allowing incidents from the same crime cluster to appear in both training and test folds. The lower but more reliable estimates from spatiotemporal CV provide a valid basis for deployment decisions.

## Fairness: Impossibility theorem and in-processing gains

FairXGBoost reduces demographic parity gap from 10.1% to 5.8% and equalized odds gap from 6.5% to 4.1% while retaining 92.4% accuracy a 7.5% fairness improvement at 0.5% accuracy cost. The residual gap is irreducible under the Impossibility Theorem [19, 31] without equalizing geographic base rates. We prioritize equalized odds over demographic parity because the operational harm of disparate false positive rates is more directly addressable by this metric. Mandatory ongoing fairness audits, patrol rebalancing, and community reporting infrastructure investment are identified as necessary structural complements to algorithmic mitigation.

## Limitations

Datasets represent urban environments; rural generalizability is untested. Temporal coverage ends in 2022 for RMP; ongoing retraining is required as crime patterns evolve. Geographic proxy correlation with demographics (Cram´er's V 0.31–0.45) cannot be eliminated without removing operationally critical spatial features. The dark figure of unreported crime (estimated 50–70%) introduces irreducible selection bias that no algorithm can correct a fundamental constraint on the validity of any prediction trained on police records. Generative alternatives to SMOTE-NC (CGANs, TabDDPM) were not experimentally compared in this study; this remains a direction for future work.

## Conclusion

This study presents a methodologically rigorous and ethically grounded framework for crime classification under class imbalance, with explicit fairness evaluation, sociological feature interpretation, and regulatory alignment. Core methodological revisions from prior work include: SMOTENC replacing standard SMOTE for mixed-type tabular data present in the RMP dataset; tabular-native architectures (Cat- Boost, TabNet, FairXGBoost) replacing recurrent networks inapplicable to non-sequential incident records; spatiotemporal cross-validation preventing data leakage from crime clusters; removal of protected demographic attributes from training features; and post-SMOTE-NC Platt scaling ensuring calibration validity.

FairXGBoost with SMOTE-NC achieved the best fairness-accuracy tradeoff (92.4% accuracy, DP gap 5.8%, EOG gap 4.1%, $p < 0.001$), with strong generalization across three independent datasets (mean 93.1% 0.6%). The Impossibility Theorem establishes that residual fairness gaps cannot be eliminated algorithmically under unequal base rates, requiring structural interventions in patrol allocation and community reporting. The conservative classification threshold ($\tau = 0.65$) reduces unnecessary police dispatch events by 12.3% at a 3.4% recall cost an explicit tradeoff weighting human harm. Post-SMOTE-NC Platt scaling ensures reported calibration (ECE = 0.039) reflects true prior distributions. This work provides a reproducible, legally aligned, and sociologically grounded foundation for responsible AI deployment in crime classification.

## Disclosure Statement

No potential conflict of interest was reported by the author.



https://doi.org/10.61577/jmla.2026.100002

## References


1. Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision making and the cost of fairness. InProceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining; 2017;797-806p. https://doi.org/10.1145/3097983.3098095
2. Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Sci Adv. 2018;4(1):eaao5580. https://doi.org/10.1126/sciadv.aao5580
3. Lum K, Isaac W. To predict and serve?. Significance. 2016;13(5):14-19. https://doi.org/10.1111/j.1740-9713.2016.00960.x
4. Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863-905. https://doi.org/10.1613/jair.1.11192
5. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. https://doi.org/10.1613/jair.953
6. European Parliament and Council of the European Union. Regulation (eu) 2024/1689 (artificial intelligence act). OJ. 2024.
7. Rossmo DK. Geographic profiling. Boca Raton (FL): CRC Press; 2000.
8. Yu CH, Ward MW, Morabito M, Ding W. Crime forecasting using data mining techniques. In 2011, IEEE 11th International Conference on Data Mining Workshops, 2011 (pp. 779-786). IEEE. https://doi.org/10.1109/ICDMW.2011.56
9. Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. Tabddpm: Modelling tabular data with diffusion models. InInternational conference on machine learning; 2023; 17564-17579p. PMLR.
9. Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. Tabddpm: Modelling tabular data with diffusion models. InInternational conference on machine learning; 2023; 17564-17579p. PMLR.
10. Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J Biomed Inform 2015;58:49-59.
11. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?. Advances in neural information processing systems. 2022;35:507-520.
12. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. 2018. https://doi.org/10.48550/arXiv.1810.11363
13. Arik SÖ, Pfister T. Tabnet: Attentive interpretable tabular learning. InProceedings of the AAAI conference on artificial intelligence; 2021; 6679-6687p.
14. Cruz AF, Belém C, Jesus S, Bravo J, Saleiro P, Bizarro P. FairGBM: Gradient boosting with fairness constraints. arXiv preprint arXiv:2209.07850. 2022. https://doi.org/10.48550/arXiv.2209.07850
15. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807. 2016. https://doi.org/10.48550/arXiv.1609.05807
16. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913-929. https://doi.org/10.1111/ecog.02881
17. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data. 2017;5(2):153-163. https://doi.org/10.1089/big.2016.0047
18. Thompson A, Tapp SN. Criminal victimization, 2021. NCJ. 2022;305101(1):1-30.
19. Federal Bureau of Investigation. National Incident-Based Reporting System (NIBRS) user manual. Washington (DC): U.S. Department of Justice; 2022. Available at: https://bjs.ojp.gov/data-collection/national-incident-based-reporting-system-nibrs
20. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016; 785-794p. https://doi.org/10.1145/2939672.2939785
21. Jiang N, Aijaz A, Jin Y. Recursive periodicity shifting for semi-persistent scheduling of time-sensitive communication in 5G. In2021 IEEE Global Communications Conference (GLOBECOM) 2021; 01-06p. IEEE. https://doi.org/10.1109/GLOBECOM46510.2021.9685475
22. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999;10(3):61-74.
23. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119-139.
24. Breiman L. Random forests. Machine learning. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324
25. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in neural information processing systems. 2016;29.
26. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009;21(9):1263-1284. https://doi.org/10.1109/TKDE.2008.239
27. Kim S, Joshi P, Kalsi PS, Taheri P. Crime analysis through machine learning. In2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON); 2018; 415-420p. IEEE. https://doi.org/10.1109/IEMCON.2018.8614828
28. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. InInternational conference on intelligent computing 2005; 878-887p. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11538059_91
29. Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S. Runaway feedback loops in predictive policing. InConference on fairness, accountability and transparency; 2018 ; 160-171p. PMLR.




https://doi.org/10.61577/jmla.2026.100002