Ngr. J. Multi. vol-2, issue-6 (Jun-2026) pp-204-211

ISSN: 3108-2750 (Online)

Nexus Global Research Journal of Multidisciplinary

Journal Homepage: https://nexusglobalresearch.com/journal/ngrjm/

Research Article

Bridging Data Scarcity in Criminology Synthetic Crime Datasets and Baseline Predictive Modelling in Nigeria

Anderline Amaogbo[1], Biralatei Fawei[2]

[1-2] *Department of Computer Science, Niger Delta University, Wilberforce Island, PMB 581, Bayelsa State, Nigeria*

ABSTRACT:

Crime prediction in developing nations faces a fundamental challenge: the absence of structured, incident-level data prevents the application of modern machine learning methods that have transformed predictive policing in data-rich environments. This paper addresses this challenge directly by proposing and evaluating a synthetic dataset methodology for criminological research in Nigeria. A dataset of 3,150 synthetic crime incidents spanning Yenagoa, Bayelsa State (2017 to 2023) was constructed from aggregate official statistics and validated against National Bureau of Statistics records. A hybrid ensemble model comprising XGBoost, Random Forest, and Gradient Boosting trained on 44 engineered demographic, spatial, and temporal features achieved an overall accuracy of 43.0 percent and balanced accuracy of 31.0 percent, exceeding random expectation but falling substantially below operational thresholds reported in data-rich settings. Complete classification failure on Fraud and Deception categories, combined with 226 mutual misclassifications between Property and Violent Crimes, empirically demonstrates that demographic and spatial features lack the discriminative power necessary for typological crime classification. Feature importance analysis revealed a diffuse distribution in which the top ten predictors explained only 36.4 percent of predictive variance. These findings quantify Nigeria's data infrastructure gap, establish a reproducible synthetic data methodology for data-scarce criminological research, and directly motivate contextual feature augmentation as a priority for future work.

Keywords: *Crime prediction, synthetic data, predictive policing, machine learning, data scarcity, feature engineering.*

1. INTRODUCTION

The application of machine learning to crime prediction has produced demonstrable operational benefits in data-rich jurisdictions, enabling proactive resource allocation, hotspot identification, and risk-stratified policing strategies (Perry et al., 2013; Wheeler & Steenbeek, 2021). However, this progress has been geographically uneven. In sub-Saharan Africa, and Nigeria in particular, the adoption of predictive policing technologies has been severely constrained by a persistent and well-documented data infrastructure deficit: police incident records remain predominantly paper-based, aggregated at divisional level, and lack the granular incident-level attributes victim demographics, suspect characteristics, precise geolocation, and situational context that form the input layer of effective predictive models (Maciver & Igbojinwaekwu, 2019; Adedamola & Ogundunmade, 2025). The consequences of this deficit are twofold. First, existing crime prediction models developed in high-resource settings cannot be directly transferred to the Nigerian context without substantial retraining on locally representative data. Second, the absence of ground-truth incident records prevents empirical validation of data-driven approaches, creating a circular barrier: models cannot be developed without data, and the case for data collection cannot be made without model evidence. Breaking this impasse requires methodological innovation.

This paper proposes synthetic dataset generation as a principled and ethically grounded methodological approach for criminological machine learning in data-scarce environments. Drawing on published aggregate statistics from the Nigerian National Bureau of Statistics (NBS), Nigerian Police Force annual reports, and international crime survey benchmarks, a dataset of 3,150 synthetic crime incident records was constructed for Yenagoa, Bayelsa State, spanning 2017 to 2023. The dataset

Corresponding author: Anderline Amaogbo, (Email: anderlineamaogbo@gmail.com)

Received: 10 Jun 2026; **Accepted:** 18 Jun 2026; **Published:** 24 Jun 2026

preserves the distributional properties of real crime data while containing no personally identifiable information, complying fully with the Nigeria Data Protection Regulation (NDPR, 2019). A hybrid ensemble model comprising XGBoost, Random Forest, and Gradient Boosting trained on 44 engineered temporal, spatial, and demographic features was evaluated on this dataset. The model serves a dual purpose: as a functional baseline that demonstrates machine learning feasibility under current data conditions, and as a diagnostic instrument that quantifies the data infrastructure gap by identifying precisely which categories of information are missing and what their absence costs in predictive accuracy.

The principal contributions of this paper are threefold. First, a reproducible synthetic dataset methodology for criminological research in low-income countries is presented and validated. Second, a rigorous baseline model evaluation establishes the performance ceiling achievable under Nigeria's current data conditions. Third, the empirical analysis of feature importance and classification failure patterns provides a data-driven prescription for the specific incident-level attributes that law enforcement agencies must begin systematically recording to enable operational predictive analytics. The rest of the paper is organised as follows: Section 2 presents the background and related work. Section 3 describes synthetic dataset construction. Section 4 details the methodology. Section 5 reports results. Section 6 discusses findings and their implications. Section 7 concludes the paper.

2. RELATED WORK

Machine learning has been applied to crime prediction across a wide range of jurisdictions and algorithmic approaches (Mandalapu et al., 2023; Jenga et al., 2023). Studies in high-resource settings consistently report accuracy figures of 78–99 percent on classification tasks (Safat et al., 2021; Kshatri et al., 2021). However, direct comparisons with developing nation contexts are methodologically problematic: models trained on rich operational data from the United States or European countries embed assumptions about data availability, feature diversity, and reporting culture that do not transfer readily to other institutional contexts (Bappee et al., 2021; Zhou et al., 2023). The limited body of ML-based crime prediction research in Nigeria reflects this constraint directly. Akinwumi (2023) noted that data unavailability and quality remain the main challenges in developing predictive models for Nigerian institutions. Fawei et al. (2024) were compelled to use the Denver crime dataset precisely because of the absence of a comprehensive dataset in the Nigerian police department, which emphasizes a methodological workaround that highlights the structural deficit. Adedamola and Ogundunmade (2025) and Eke (2025) further document the pattern of small sample sizes, poor data quality, and incomplete records that characterise Nigerian crime data, with Eke additionally noting infrastructure constraints, including epileptic power supply as impediments to AI-based policing. The present study differs from prior Nigerian work by addressing the data problem directly through synthetic generation rather than working around it. Synthetic data generation has a well-established history in machine learning, most notably through the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) and its variants. More broadly, synthetic dataset construction has been employed in medical research, financial modelling, and social simulation when individual-level data is unavailable for privacy or access reasons (Hernandez et al., 2022; Liu et al., 2024). Research demonstrates that synthetic data can maintain utility comparable to real data while preserving privacy, making it a robust privacy enhancing technology (Razi et al., 2025). Specifically, in criminological research, Fan et al. (2025) developed a Multi-modal Generative Adversarial Network (MM-GAN) framework to generate synthetic crime data that closely matches real crime patterns across geographical distribution, temporal patterns, and category information directly by addressing data scarcity and privacy limitations in criminological ML. The present study extends this tradition by constructing a fully synthetic primary dataset from aggregate statistics, and using the resulting baseline performance as a measure of the data infrastructure gap. Routine Activity Theory (Cohen & Felson, 1979) posits that crime occurs at the convergence of a motivated offender, a suitable target, and the absence of a capable guardian in space and time. This framework has been operationalised extensively in contemporary ML feature engineering process. Zhang et al. (2022) demonstrated that ambient population, temporal patterns, and spatial covariates grounded in criminological theory

significantly improve predictive accuracy, while Zhang et al. (2020) showed that models incorporating built environment covariates derived from criminological theories consistently outperform those relying on historical crime data alone. Social Disorganisation Theory (Shaw & McKay, 1942) further justifies socioeconomic features including unemployment rates, population density, and poverty indices as structural crime predictors. Critically, neither theory provides direct guidance on crime type classification. They are good at explaining where and when crime occurs but not their typological manifestation. Berg and Schreck (2021) demonstrate that the victim offender relationship is central to understanding crime type. Offenders act principally in response to targets, and any theory of crime that cannot account for the victim offender overlap may be subject to falsification. This limitation is central to the baseline model's findings. Features derived from opportunity theory predict crime occurrence but not crime category.

3. SYNTHETIC DATASET CONSTRUCTION

The synthetic dataset was constructed from four primary sources: the National Bureau Statistics (NBS) 2017 Crime Statistics Report, which documented 1,543 reported offences in Bayelsa State with a breakdown by crime type; Nigerian Police Force divisional reports providing seasonal and geographic distribution patterns; OpenStreetMap and Google Earth for spatial coordinates of Yenagoa's 18 neighbourhoods; and United Nation Office on Drugs and Crime (UNODC) comparative crime statistics for Nigeria used to validate distributional assumptions against international benchmarks. The synthesis process preserved three properties of real crime data which are distributional realism (crime type frequencies matching aggregate NBS statistics), spatial realism (incident coordinates reflecting population density gradients derived from OpenStreetMap), and temporal realism (daily and seasonal patterns consistent with published Nigerian crime reports). The dataset was validated by comparing generated distributions against published aggregate statistics and against crime patterns in comparable Niger Delta cities (Port Harcourt, Warri) for which partial operational data are available.

The dataset encompasses 19 distinct crime types grouped into four analytically stable categories. Violent Crimes include Armed Robbery, Assault, Murder, Kidnapping, Attempted Murder, Grievous Harm and Wounding, and Rape and Indecent Assault. Property Crimes include Theft and Stealing, House Burglary, Arson, and Unlawful Possession. Fraud and Deception includes False Pretence, Cheating, Fraud, and Forgery. Other Offences includes Breach of Peace, Child Stealing, Unnatural Offences, and Attempted Suicide. This aggregation balances predictive granularity with statistical stability, ensuring adequate class representation for model training. The class imbalance ratio between the largest category (Property Crimes, n = 1,248) and the smallest (Fraud and Deception, n = 271) is approximately 4.62:1, necessitating specialised class-balancing strategies during model training, addressed through Borderline Synthetic Minority Oversampling Technique (SMOTE) oversampling and algorithmic class weighting. Moreover, the final dataset comprises 3,150 synthetic crime incident records spanning from January 2017 to December 2023, covering 18 locations within Yenagoa Local Government Area. Each record includes 19 raw attributes: crime type, date, time, location, geographic coordinates, suspect age and gender, victim age and gender, population density, socioeconomic index, and unemployment rate (see dataset summary in Table 3.1).

Table 3.1: Dataset Summary Statistics

Attribute	Value
Total Records	3,150
Temporal Coverage	January 2017 – December 2023 (7 years)
Geographic Scope	18 locations, Yenagoa LGA, Bayelsa State
Crime Types (raw)	19 distinct types
Crime Categories	4: Violent, Property, Fraud & Deception, Other
Property Crimes	1,248 (39.6%)
Violent Crimes	1,246 (39.6%)
Other Offences	386 (12.3%)

Fraud & Deception	271 (8.6%)
Imbalance Ratio	4.62:1 (Property : Fraud)
Raw Features per Record	19

4. METHODOLOGY:

The 19 raw features were expanded into 44 engineered predictors through six systematic feature construction procedures, each grounded in criminological theory. Temporal features (13) operationalised Routine Activity Theory's emphasis on time-of-day and day-of-week crime rhythms (Zhang et al., 2020; Zhang et al., 2022). Raw hour and day values are supplemented with sine-cosine cyclical encodings to preserve temporal continuity: $Hour\_\sin = \sin(2\pi \times {Hour}/{24})$ and $Hour\_\cos = \cos(2\pi \times {Hour}/{24})$, ensuring that 23:00 and 00:00 hours are represented as adjacent rather than maximally distant values. Binary flags for night-time incidents (22:00–06:00) and weekends are also included. Location historical features (4) provide the model with spatial memory: total crime frequency per location, dominant crime type per location, Shannon entropy of crime-type distribution as a diversity index, and violent crime rate per location. Recency features (2) operationalise Repeat Victimisation Theory: days elapsed since the most recent crime at each location, and a 30-day rolling crime count per location. Demographic features (8) encode victim-offender relational patterns, including absolute age gap, gender interaction pattern, and vulnerability flags for young suspects and elderly or child victims. Geographic features (4) operationalise distance-decay principles from Environmental Criminology. Socioeconomic interaction features (5) capture structural crime risk through composite interaction terms incorporating unemployment rate, population density, and a weighted Composite Risk Score.

All numerical features were imputed using K-Nearest Neighbours (KNN) Imputation with k = 5, preserving local spatial and demographic coherence. Standardisation was applied using StandardScaler (zero-mean, unit-variance). Categorical variables were encoded using Ordinal Encoding with unknown-value handling. Class imbalance was addressed through Borderline-SMOTE (Han et al., 2005), which was applied after preprocessing within the training pipeline to prevent leakage. The dataset was partitioned 80:20 into training (n = 2,520) and held-out test (n = 630) subsets using stratified sampling technique.

A hybrid voting ensemble comprising XGBoost, Random Forest, and Gradient Boosting served as the baseline model. This architecture was chosen in accordance with the No Free Lunch Theorem (Wolpert, 1996), which motivates combining diverse inductive biases rather than committing to a single algorithm. Soft probability-based voting was used, with an initial weight assignment of 3:2:1 (XGBoost:RF:GB). Model performance was assessed using Overall Accuracy, Balanced Accuracy, macro and weighted F1-Score, Receiver Operating Characteristics Area Under Curve (ROC-AUC) on a one-vs-rest multi-class extension, Cohen's Kappa, and five-fold stratified cross-validation.

5. RESULTS

The baseline model achieved an overall accuracy of 43.0 percent and balanced accuracy of 31.0 percent on the held-out test set of 630 incidents. While this exceeds random expectation (25 percent for four equally probable categories) by a factor of 1.72, it falls substantially below the 78–99 percent accuracy range reported in comparable studies from data-rich settings (Safat et al., 2021; Kshatri et al., 2021; Mandalapu et al., 2023). Table 5.1 summarises the complete baseline performance metrics.

Table 5.1: Baseline Model Performance Summary

Metric	Score	Interpretation
Overall Accuracy	43.0%	Below operational threshold
Balanced Accuracy	31.0%	Weak cross-class generalisation
Macro F1-Score	0.310	Poor precision-recall balance
Weighted F1-Score	0.407	Marginally better under class weighting
ROC-AUC (Macro)	0.605	Modest discrimination above chance
Cohen's Kappa	0.309	Fair agreement beyond chance

The per-class breakdown reveals the mechanism of baseline failure with precision. Fraud and Deception exhibited complete classification failure: precision = 0.00, recall = 0.00, F1 = 0.00. Of 54 fraud test cases, zero were correctly identified; 34 were misclassified as Violent Crimes and 18 as Property Crimes. This reflects the model's inability to distinguish deception-based offences from confrontational crimes using only demographic and temporal signals. Property Crimes achieved the highest per-class performance (F1 = 0.51), attributable to their majority-class representation. Violent Crimes (F1 = 0.42) showed substantial confusion with Property Crimes, consistent with their shared opportunity structures in mixed residential-commercial zones. Here, the confusion matrix is dominated by off-diagonal elements: only 271 of 630 predictions were correct (43.0 percent). The 226 mutual misclassifications between Property and Violent Crimes represent the most prominent error cluster, accounting for 62.9 percent of total errors. Only 8 cases were predicted as Fraud across the entire test set, confirming the complete collapse of minority-class identification. Table 5.2 presents the full confusion matrix.

Table 5.2: Baseline Model Confusion Matrix (Actual × Predicted)

Actual ↓ / Predicted →	Fraud	Other	Property	Violent
Fraud & Deception (n=54)	**0**	2	18	34
Other Offences (n=77)	3	**17**	19	38
Property Crimes (n=250)	3	3	**141**	103
Violent Crimes (n=249)	2	11	123	**113**

Feature importance analysis of the XGBoost component reveals a strongly diffused distribution. The leading predictor, Victim Gender, contributes only 5.6 percent of total importance. The top ten features collectively account for only 36.4 percent of predictive variance which is an exceptionally low concentration compared to well-performing models in which dominant features typically account for 60–80 percent of importance. Six of the ten most important features are demographic, suggesting the model relies almost entirely on gender and age patterns. Table 5.3 presents the top ten features.

Table 5.3: Top 10 Features by Importance (Baseline Model)

Rank	Feature	Importance	Category
1	Victim Gender	0.0560	Demographic
2	Latitude Zone	0.0453	Geographic
3	Suspect Youth Risk	0.0397	Demographic
4	Suspect Gender	0.0389	Demographic
5	Same Gender	0.0363	Demographic
6	Day Cosine (cyclical)	0.0357	Temporal
7	Is Night Time	0.0315	Temporal
8	Is High Crime Hour	0.0283	Temporal
9	Gender Pattern	0.0261	Demographic
10	Weekend High Crime Hour	0.0260	Interaction
Total (Top 10)		**0.3638**	**36.4% of variance**

6. DISCUSSION:

The resulting output from the baseline models of 43.0 percent accuracy represents more than a performance shortfall. It is a measurement of institutional capacity. By constructing a model that uses every available feature derivable from Nigeria's current aggregate crime reporting system and evaluating its performance rigorously, this study produces a quantitative estimate of what current data infrastructure supports. The gap between 43.0 percent and the 78–99 percent achievable result in data-rich settings (Safat et al., 2021; Mandalapu et al., 2023) is not an algorithmic

gap but rather a governance gap. This is consistent with Buil-Gil et al. (2021) findings, which demonstrated empirically that police recording bias and data quality limitations directly constrain the accuracy of crime predictions at fine geographic scales with Yeung et al. (2025) observation. Which shows that poor data quality renders intervention strategies largely ineffective regardless of algorithmic sophistication. This idea has direct policy implications, which emphasizes that investment in algorithmic complexity without corresponding investment in data collection infrastructure will yield diminishing returns. The ablation analyses in comparable data-scarce contexts suggest that incorporating contextual fields such as victim-suspect relationship, weapon involvement, and location type could substantially elevate predictive accuracy. These features are contextual in nature but they encode the situational and relational dynamics of the crime event itself, which Berg and Schreck (2021) identified as essential to understanding crime type. Baek et al. (2021) further demonstrated that crime type prediction using contextual case data including crime severity can be achieved effectively with ML methods. The cost of improvement is institutional reform in data recording practice, not algorithmic innovation. Moreover, the construction of a synthetic dataset as the primary research instrument represents a methodological contribution that extends beyond the Nigerian context. The synthetic dataset was constructed from published aggregate statistics and validated against distributional benchmarks from comparable cities. The approach is replicable in any jurisdiction where individual level crime data is restricted, and is ethically grounded. It complies fully with the Nigeria Data Protection Regulation NDPR 2019 and avoids privacy risks associated with operational police records. Razi et al. (2025) confirm that synthetic data maintains higher predictive utility than differentially private or anonymised data reinforcing the validity of this methodological choice. Also, the baseline results provide an empirical support for a theoretical distinction that criminological literature has long drawn but rarely tested computationally; the distinction between crime occurrence prediction and crime type classification. Routine Activity Theory and Environmental Criminology effectively explain when and where crimes occur and the temporal and spatial features derived from these theories perform as expected in the feature importance rankings (Zhang et al., 2020; Zhang et al., 2022). However, these theories do not explain why a specific crime type manifests in a given context. The implication is that crime type classification is an inherently contextual problem requiring features that encode situational and relational dynamics, precisely those absent from aggregate statistical records.

Nonetheless, a careful investigation of the dataset used for the experiment shows that the constructed synthetic dataset cannot fully replicate the statistical properties of real incident-level data. Moreover, the modesty of the dataset (3,150 records) compared with crime prediction studies in data-rich settings, and performance estimates may deviate from what larger samples would yield. Finally, the baseline model combines three algorithms in a fixed-weight voting ensemble; the optimal weighting of individual classifiers remains an open question that future work should address through systematic ablation and hyperparameter optimization.

7. CONCLUSION

This paper has presented a synthetic dataset methodology for criminological machine learning in data-scarce environments, demonstrated through a baseline model evaluation for Yenagoa, Bayelsa State, Nigeria. The baseline model achieved 43.0 percent overall accuracy and 31.0 percent balanced accuracy on four crime categories, exceeding random chance but falling substantially below operational thresholds. Complete classification failure on Fraud and Deception categories, combined with 226 mutual misclassifications between Property and Violent Crimes, demonstrated that demographic and spatial features lack the discriminative power necessary for typological crime classification. Feature importance analysis confirmed the mechanism that the model relied on weak and diffused demographic signals because the richer contextual information such as relationship dynamics, weapon involvement, situational setting is systematically absent from Nigeria's current aggregate reporting system. This finding redirects the challenge of predictive policing in Nigeria from a technological problem to a data governance problem. The contributions of this paper are a reproducible synthetic dataset methodology, a rigorously quantified baseline, and a data-driven prescription for the specific incident-level attributes that law

enforcement agencies must begin recording to enable operational predictive analytics.

Three limitations require acknowledgment. First, the synthetic dataset, however carefully constructed, cannot fully replicate the statistical properties of real incident-level data. Second, the 3,150-record dataset is modest compared with crime prediction studies in data-rich settings, and performance estimates may deviate from what larger samples would yield. Third, the baseline model combines three algorithms in a fixed-weight voting ensemble; the optimal weighting of individual classifiers remains an open question that future work should address through systematic ablation and hyperparameter optimization.

REFERENCES:

1. Adedamola, A. O., & Ogundunmade, T. P. (2025). Predictive modelling of crime data using machine learning models: A case study of Oyo State, Nigeria. International Journal of Innovative Science and Research Technology. https://doi.org/10.38124/ijisrt/25apr851
2. Akinwumi, D. A. (2023). The use of K-Nearest Neighbour approach for crime rate prediction in public universities: Case study Adekunle Ajasin University, Akungba–Akoko (AAUA), Ondo State, Nigeria. Advances in Multidisciplinary and Scientific Research Journal Publication, 11(2), pp. 11–24. https://doi.org/10.22624/aims/maths/v11n2p1
3. Baek, M.-S., Park, W., Park, J., Jang, K., & Lee, Y.-T. (2021). Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation. IEEE Access, 9. https://doi.org/10.1109/access.2021.3112682
4. Bappee, F. K., Soares, A., Petry, L. M., & Matwin, S. (2021). Examining the impact of cross-domain learning on crime prediction. Journal of Big Data, 8. https://doi.org/10.1186/s40537-021-00489-9
5. Berg, M. T., & Schreck, C. J. (2021). The meaning of the victim-offender overlap for criminological theory and crime prevention policy. Annual Review of Criminology, 5. https://doi.org/10.21428/cb6ab371.322ebfb5
6. Buil-Gil, D., Moretti, A., & Langton, S. H. (2021). The accuracy of crime statistics: Assessing the impact of police data bias on geographic crime analysis. Journal of Experimental Criminology. https://doi.org/10.1007/s11292-021-09457-y
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, pp. 321–357.
8. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American Sociological Review, 44(4), pp. 588–608.
9. Eke, R. N. (2025). Adopting artificial intelligence-based predictive policing in Nigeria. African Journal of Advances in Science and Technology Research. https://doi.org/10.62154/ajastr.2025.018.010680
10. Fan, T., Hu, X., & Shi, Y. (2025). Leveraging multi-modal generative adversarial networks (GANs) for synthetic crime data simulation. International Journal of Artificial Intelligence Tools. https://doi.org/10.1142/s0218213025500101
11. Fawei, B., Amaogbo, A., & Okolai, B. D. (2024). Crime rate prediction system – an experiment with Denver crime dataset using machine learning technique. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/cseit24104100
12. Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of ICIC 2005 (pp. 878–887). Springer.
13. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. Neurocomputing. https://doi.org/10.1016/j.neucom.2022.04.053
14. Jenga, K., Catal, C., & Kar, G. (2023). Machine learning in crime prediction. Journal of Ambient Intelligence and Humanized

Computing. https://doi.org/10.1007/s12652-023-04530-y

15. Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M., & Sinha, G. R. (2021). An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach. IEEE Access. https://doi.org/10.1109/access.2021.3075140
16. Liu, Q., Khalil, M., Jovanović, J., & Shakya, R. (2024). Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics. Proceedings of the 14th Learning Analytics and Knowledge Conference. https://doi.org/10.1145/3636555.3636921
17. Maciver, M., & Igbojinwaekwu, C. (2019). Predictive policing challenges in Nigeria: Infrastructure and institutional constraints. African Security Review, 28(3), pp. 112–129.
18. Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. IEEE Access. https://doi.org/10.1109/access.2023.3286344
19. Nigerian Bureau of Statistics. (2017). Crime Statistics Report 2017. Abuja: NBS.
20. Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Santa Monica, CA: RAND Corporation.
21. Razi, Q., Datta, S., Hassija, V., Chalapathi, G., & Sikdar, B. (2025). Privacy utility tradeoff between PETs: Differential privacy and synthetic data. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/tcss.2024.3479317
22. Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE Access. https://doi.org/10.1109/access.2021.3078117
23. Shaw, C. R., & McKay, H. D. (1942). Juvenile Delinquency and Urban Areas. Chicago: University of Chicago Press.
24. UNODC. (2023). Global Study on Homicide: Homicide Data. Vienna: United Nations Office on Drugs and Crime.
25. Wheeler, A. P., & Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. Journal of Quantitative Criminology, 37, pp. 445–480.
26. Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7), pp. 1341–1390.
27. Yeung, W. N., Di Clemente, R., & Lambiotte, R. (2025). Garbage in garbage out? Impacts of data quality on criminal network intervention. EPJ Data Science. https://doi.org/10.1140/epjds/s13688-025-00553-x
28. Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. IEEE Access. https://doi.org/10.1109/access.2020.3028420
29. Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., & Chen, J. (2022). Interpretable machine learning models for crime prediction. Computers, Environment and Urban Systems. https://doi.org/10.1016/j.compenvurbsys.2022.101789
30. Zhou, B., Chen, L., Zhao, S., Li, S., Zheng, Z., & Pan, G. (2023). Unsupervised domain adaptation for crime risk prediction across cities. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/tcss.2022.3207987