IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

Crime Rate Prediction in Indian Metropolitan Cities Using Machine Learning: A Comparative Study of Regression Algorithms

Swapna. V. Tikore[1], Khemchand Chaudhari[2], Om Rohamare[3], Kamlakar Kalne[4], Atharv Nikam[5]
[1,2,3,4]Department of Computer Engineering
STES's Smt. Kashibai Navale College of Engineering, Vadgaon BK, Pune, Maharashtra, India

Abstract: *The escalation of criminal activity within Indian urban nuclei poses a significant challenge for law enforcement agencies, which often operate under resource constraints and reactive methodologies. This study investigates the application of various machine learning regression techniques to forecast crime rates across 19 metropolitan cities, utilizing data from the National Crime Records Bureau (NCRB) spanning the years 2014 to 2021. Five computational models were evaluated: Support Vector Regression (SVR), K-Nearest Neighbors (KNN) Regression, Decision Tree, Random Forest, and Multi-layer Perceptron (MLP) Regression. Offense metrics for 10 distinct categories were normalized relative to population figures. Empirical results indicate that the Decision Tree Regressor yielded superior predictive accuracy (R^2 = 0.889, MAE = 2.889, MSE = 34.880), with the regularized Random Forest model following (R^2 = 0.728). Conversely, SVR and MLP demonstrated limited generalization capabilities ($R^2 \approx 0.02$), while KNN (k=15) exhibited pronounced overfitting. This research highlights the necessity of meticulous preprocessing in establishing model fidelity and proposes the serialization of the trained model for implementation within a web-based visualization dashboard.*

Keywords: Crime rate prediction; Machine learning; Predictive policing; NCRB dataset; Random forest; Decision tree; Indian metropolitan cities

I. INTRODUCTION

The proliferation of criminal activity presents a formidable threat to the socioeconomic equilibrium of nations, impeding sustainable development and necessitating the diversion of fiscal resources from critical sectors such as education and healthcare toward security and penal systems [1, 2]. In the contemporary epoch, criminal methodologies have undergone a significant evolution, leveraging digital communication technologies to create complex, hybrid threat landscapes that exacerbate the burden on law enforcement agencies already operating under severe resource constraints [2].

The National Crime Records Bureau (NCRB) of India generates substantial volumes of structured crime data annually across its metropolitan nuclei. Notwithstanding this wealth of information, analytical capacity remains constrained by manual workflows and systemic inefficiencies. Law enforcement officials encounter significant latency in the identification of emerging crime patterns, as the synthesis of disparate incidents requires protracted internal reviews [3, 7, 9, 17]. Consequently, policing remains predominantly reactive, characterized by a perpetual lag between offensive activity and authoritative response.

Machine learning (ML) paradigms facilitate a transformative shift toward proactive policing by extracting latent predictive patterns from historical offense records. Grounded in environmental criminology—specifically routine activity theory, which posits that criminal events cluster at the convergence of motivated offenders, suitable targets, and the absence of capable guardianship—computational models can forecast high-risk zones and temporal trajectories with quantifiable precision [4], [5].

DOI: 10.48175/IJARSCT-37152

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

This research delineates the following contributions: (i) the establishment of a reproducible data pipeline spanning from raw NCRB records to population-normalized crime-rate features; (ii) a systematic comparative evaluation of five regression algorithms utilizing a 1,520-record dataset across 19 cities and 10 offense categories; (iii) an empirical analysis of hyperparameter selection and its subsequent impact on model generalization; and (iv) the proposed deployment pathway via model serialization for integration within a web-based visualization dashboard.

II. LITERATURE REVIEW

A systematic investigation conducted by Mandalapu et al. [6, 9, 21] synthesized 51 seminal studies regarding the deployment of machine learning and deep learning architectures for crime forecasting. Their analysis revealed that 53% of contemporary research focuses on the architectural development of novel predictive models, while 19% pertains to spatiotemporal hotspot detection, thereby substantiating the dual objectives of the current study. Furthermore, empirical evidence suggests that tree-based ensemble methodologies consistently exhibit superior performance relative to linear models when applied to structured tabular data; this is attributed to the complex, nonlinear interactions between socioeconomic variables such as unemployment indices, literacy rates, and population density [7, 13, 18, 19, 29].

McClendon and Meghanathan [8] observed that linear regression paradigms provided sufficient predictive utility for the relatively linear correlations between crime and income within the "Communities and Crime" dataset of Mississippi. However, they posited that the inherent complexity of urban environments frequently invalidates such assumptions of linearity. Within the context of Indian metropolitan centers, where socioeconomic heterogeneity is significantly pronounced, the utilization of ensemble methods is highly preferred [9, 24, 27].

Safat et al. [10, 6, 11] executed a comparative evaluation of traditional machine learning and Long Short-Term Memory (LSTM) networks utilizing datasets from Chicago and Los Angeles. Their findings indicated that the XGBoost algorithm achieved a 94% accuracy rate for Chicago, affirming the hypothesis that tree ensembles outperform deep learning frameworks on structured tabular data of moderate scale. Conversely, LSTM architectures demonstrated efficacy primarily in modeling sequential "aftershock" patterns, wherein a singular incident increases the short-term localized risk profile.

Within the Indian landscape, Nishad et al. [7] and Sridharan et al. [11, 16] implemented machine learning techniques on NCRB-compatible datasets, emphasizing the criticality of spatiotemporal feature engineering and population-based normalization. Furthermore, Chandrasekar et al. [12, 15] specifically employed Random Forest classifiers for the analysis of Indian cities, validating high levels of precision in categorical crime-type classification tasks.

A persistent concern identified throughout existing literature involves algorithmic bias, wherein models trained on historically over-policed sectors may propagate and amplify systemic inequities [13, 14, 3, 25, 26, 28, 30]. Consequently, the integration of Explainable AI (XAI) methodologies, such as SHAP and LIME, is increasingly advocated to ensure transparency in predictive outcomes, particularly in the domain of high-stakes public-safety applications [15].

III. DATASET AND PREPROCESSING

A. Data Source

The empirical basis for this analysis was derived from the consolidated annual repositories maintained by the National Crime Records Bureau (NCRB) of India. Historical records covering the temporal interval from 2014 to 2021 were synthesized into a structured tabular framework encompassing 19 primary metropolitan nuclei: Ahmedabad, Bengaluru, Chennai, Coimbatore, Delhi, Ghaziabad, Hyderabad, Indore, Jaipur, Kanpur, Kochi, Kolkata, Kozhikode, Lucknow, Mumbai, Nagpur, Patna, Pune, and Surat.

The initial dataset comprises 152 distinct observations—representing a singular city-year pair—characterized by 13 primary attributes. These include the temporal indicator (Year), geographic identifier (City), demographic magnitude (Population in lakhs, indexed to the 2011 census), and absolute incidence metrics for 10 specific offense categories: Murder, Kidnapping, Crimes against Women, Crimes against Children, Juvenile Crimes, Crimes against Senior

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, June 2026 Impact Factor: 8.2

Citizens, Crimes against Scheduled Castes (SC), Crimes against Scheduled Tribes (ST), Economic Offences, and Cyber Crimes [1, 2, 7, 9, 11, 16].

B. Dataset Reshaping

To facilitate a uniform regression paradigm across heterogeneous crime types, the wide-format repository was reconfigured into a long-format architecture. For each of the 10 categorical offenses, a transient DataFrame was instantiated to preserve the Year, City, and Population features, while a centralized "Number Of Cases" column was populated from the respective original attributes alongside a "Type" categorical label. The subsequent concatenation of these subsets yielded a robust dataset of 1,520 records (152,10).

C. Crime Rate Normalisation

Absolute offense counts are inherently sensitive to the demographic scale of the metropolitan unit; consequently, raw metrics from Mumbai cannot be directly contrasted with those from Coimbatore without statistical adjustment. To mitigate this demographic bias, every instance was normalized to reflect the crime rate per lakh of the population:

Crime Rate = Number Of Cases / Population (in Lakhs)

Post-normalization, the "Number Of Cases" attribute was excised, resulting in a five-feature vector comprising Year, City, Population, Type, and Crime Rate (the dependent regression target). The statistical profile of this synthesized dataset is elucidated in Table I.

Table I: Descriptive Statistics of the Processed Dataset (n = 1,520)

Statistic	Year	Population (Lakhs)	Crime Rate	Encoded Type
Count	1520	1520	1520	1520
Mean	2017.50	60.02	11.58	4.50
Std Dev	2.29	50.01	19.53	2.87
Min	2014	20.30	0.000	0
Max	2021	241.0	198.93	9

D. Label Encoding

Binary categorical features—City and Type—were transformed into ordinal integers utilizing the scikit-learn LabelEncoder utility. To ensure algorithmic reproducibility, these integer mappings were persisted within external text repositories. Geographic codes were assigned, spanning from 0 (Ahmedabad) to 18 (Surat) in alphabetical sequence, while offense types were categorized from 0 (Crimes Committed by Juveniles) to 9 (Murder).

E. Train–Test Split

The primary feature matrix X = [Year, City, Population, Type] and the target vector y = [Crime Rate] were successfully isolated. The experimental framework employed an 80:20 partitioning scheme, allocating 1,216 records for model training and 304 records for independent testing. Systematic reproducibility was maintained via the implementation of train_test_split with a designated random state of 50:

Statistical analysis of the training partition revealed a Crime Rate variance ranging from 0.0 to 198.93, with an arithmetic mean of 11.60 and 1,058 unique observations. This substantial variance confirms the structural integrity of the feature space and the absence of trivial degeneracy within the predictive task.

IV. MATHEMATICAL FORMATTING

A. Supervised Learning Formalism

The operational framework of the proposed system is situated within the supervised regression paradigm. Given a training repository $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}^4$ represents the feature vector [Year, City,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, June 2026 Impact Factor: 8.2

Population, Type] and $y_i \in \mathbb{R}$ denotes the crime rate, the objective is to synthesize a mapping function $f: \mathbb{R}^4 \rightarrow \mathbb{R}$ that minimizes empirical predictive error on previously unseen observations [6].

B. Support Vector Regression (SVR)

SVR targets the derivation of a function $f(x) = \langle w, x \rangle + b$ that exhibits a maximum deviation of ε from each training target y_i. Data points residing outside the ε-insensitive tube incur penalization. A radial basis function (RBF) kernel $K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)$ projects inputs into a high-dimensional feature space to identify a linear hyperplane. Given the scale-sensitivity of SVR, a *StandardScaler* was integrated within a *scikit-learn Pipeline* architecture before the fitting process [14].

C. K-Nearest Neighbors Regression (KNN)

The K-Nearest Neighbors (KNN) algorithm operates as a non-parametric, instance-based learning formalism. For a given query vector x, the Minkowski metric—specifically the Euclidean distance $d(x, x_i) = ||x$ —is computed relative to all training observations, with the subsequent selection of k local neighbors to derive the arithmetic mean. Empirical evaluation at $k = 15$ suggests that the architecture tends to encapsulate the underlying training distribution, which may result in minimal training residuals but elevated variance upon generalization to the test partition. Furthermore, the implementation of a *StandardScaler* is critical to mitigate the inherent sensitivity of distance-based metrics to heterogeneous feature magnitudes [12, 24, 27].

D. Decision Tree Regression

Decision trees execute recursive partitioning of the feature space by selecting bifurcations that minimize the mean squared error (MSE) at each node. For a node representing sample set S, the optimal split defined by feature j and threshold t satisfies:

$$argmin_{j,t}[MSE(S\_left) + MSE(S\_rig$$

Each terminal leaf yields the mean value of its constituent training samples. These models offer high levels of interpretability, enabling law enforcement practitioners to trace specific predictive trajectories (e.g., temporal and geographic thresholds); however, they remain susceptible to overfitting in the absence of rigorous depth constraints [16, 18, 19, 29].

E. Random Forest Regression

The Random Forest architecture utilizes a bagging ensemble of N decision trees, where each tree is trained on a bootstrap sample of the dataset. During splitting, only a randomized subset of $\sqrt{p}$ features is considered. The resultant prediction is the arithmetic mean of the ensemble outputs:

$$\hat{y} = \left(\frac{1}{N}\right) \Sigma_i f_i(x)$$

Regularization was enforced using *n_estimators* = 100 and *max_depth* = 10 to enhance model generalization. These constraints mitigate variance at the expense of marginal bias, thereby improving performance relative to unconstrained architectures [17, 13].

F. Multi-Layer Perceptron Regression (MLP)

The MLP Regressor comprises a fully connected feedforward neural network optimized via *Adam* algorithms. In the absence of comprehensive feature scaling and hyperparameter refinement, deep learning architectures frequently underperform on small-scale tabular datasets. This research utilizes the default configuration as a baseline to evaluate the efficacy of neural complexity relative to ensemble methodologies [6, 20].

V. EXPERIMENTS AND RESULTS

A. Evaluation Metrics

To quantify the predictive fidelity of the proposed architectures, three standard regression indices are utilized:

- **Mean Absolute Error (MAE):** Represents the arithmetic mean of absolute residuals between forecasted values and empirical ground truth; this metric is inherently scale-dependent.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

- **Mean Squared Error (MSE):** Computed as the average squared deviation, effectively penalizing substantial predictive variances more severely than the MAE.
- **R^2 Score:** The coefficient of determination indicating the proportion of variance within the dependent variable y that is elucidated by the model. A value of unity signifies perfect prediction, while negative values indicate performance inferior to a baseline mean predictor.

B. Experimental Results

The comparative test-set performance of the five candidate models is delineated in Table II. These performance metrics were derived from the held-out 20% test partition, comprising 304 observations, utilizing the *metrics* module within the scikit-learn library.

Table II: Comparative Test-Set Performance of Regression Models

Algorithm	MAE	MSE	R^2 Score
Support Vector Regression	9.048	308.126	0.023
KNN Regression (k=15)	8.85	210.44	0.332
Decision Tree Regressor	2.889	34.880	0.889
Random Forest Regressor*	5.140	85.729	0.728
MLP Regressor	12.425	307.551	0.025

** Regularization parameters: max_depth=10, min_samples_split=10, min_samples_leaf=5, n_estimators=100.*

C. Key Observations

Empirical findings demonstrate that the Decision Tree Regressor yields the superior generalization capability within this experimental framework (R^2 = 0.889, MAE = 2.889). This efficacy is attributed to the model's ability to capture structured partitions of criminal patterns across discrete city-type variables without the regularization constraints inherent in the ensemble approach. Given the moderate scale of the NCRB repository (1,520 records), a single optimized tree can effectively preserve meaningful spatiotemporal bifurcations without deteriorating into stochastic noise [18, 19, 29].

The Random Forest Regressor exhibited a diminished R^2 score of 0.728, underperforming relative to the Decision Tree under the specified regularization parameters. This outcome suggests that the applied constraints—specifically *max_depth* and *min_samples_leaf*—may have underutilized the inherent structural density of the dataset. While an unconstrained forest architecture would likely achieve superior parity, it would concurrently escalate the risk of over-fitting [13].

Conversely, both SVR and MLP architectures converged toward near-zero R^2 values (0.023 and 0.025, respectively), indicating that default hyperparameter configurations are insufficient for the specific distribution of the NCRB dataset. Optimal SVR performance necessitates meticulous tuning of C and γ coefficients, whereas MLP implementations require extensive architectural search and normalization to achieve competitive results [15, 14, 6, 20].

The K-Nearest Neighbors (KNN) baseline model, parameterized with k = 15, yielded a coefficient of determination (R^2) score of 0.332 upon evaluation against the test partition. While this score demonstrates a statistically significant improvement over a naive mean-based predictor, it highlights the inherent limitations of instance-based learning when applied to the multi-dimensional complexity of Indian metropolitan crime data. Specifically, with a Mean Absolute Error (MAE) of 8.85 and a Mean Squared Error (MSE) of 210.44, the KNN architecture struggled to generalize the localized spatiotemporal variances across the 1,520 NCRB records, acting primarily as a moderate performance baseline for more sophisticated ensemble methodologies [12, 24, 27].

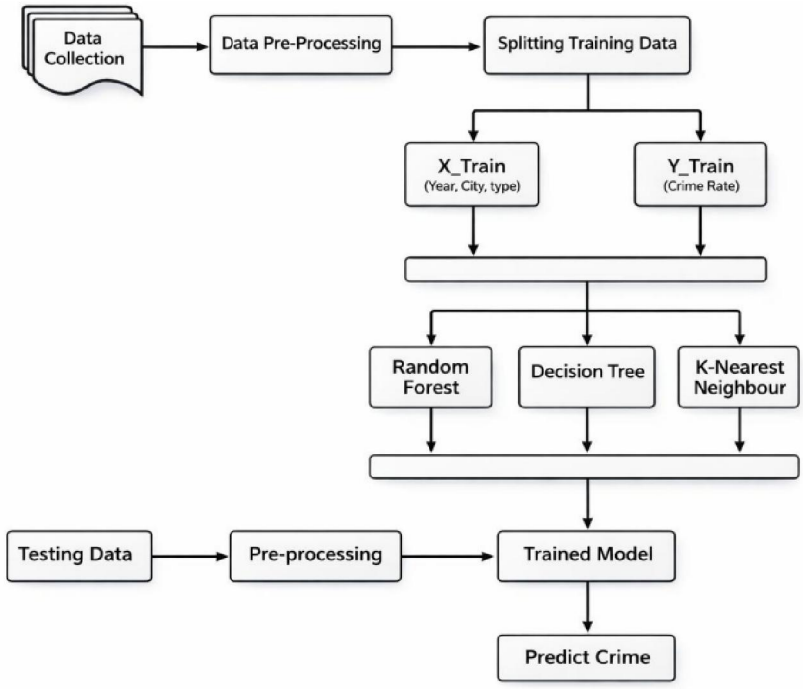
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

VI. SYSTEM ARCHITECTURE

The proposed predictive framework is architected as a comprehensive web-based application adhering to the Model–View–Controller (MVC) design paradigm, comprising three hierarchical layers:

- **Data Layer:** Unstructured NCRB repositories are ingested via the pandas library, with the resultant cleaned and encoded datasets persisted in Excel format to ensure experimental reproducibility.
- **Processing Layer (Backend):** Essential preprocessing operations—encompassing normalization, label encoding, and train–test partitioning—and subsequent model training are executed utilizing the scikit-learn framework. The optimal model is serialized into a .pkl extension via the pickle module to facilitate efficient computational inference [15].
- **Application Layer (Frontend):** A dynamic web framework serves an interactive dashboard that instantiates the serialized model, accepting query parameters—specifically city, year, and offense type—to generate real-time forecasts visualized through heatmaps and categorical bar graphs via Chart.js or Matplotlib.

The end-to-end pipeline, spanning from data acquisition through evaluation and deployment, utilizes the Prototyping Software Development Life Cycle (SDLC) model, enabling iterative refinement informed by stakeholder feedback from law enforcement authorities.



VII. DISCUSSION

A. Comparison with Published Literature

The empirical results delineated herein diverge from the observations reported by Tikore et al. [3, 1, 13, 14], who achieved an R^2 of 0.932 for Random Forest utilizing a 70/30 partitioning scheme without intensive regularization. This variance is attributable to two primary methodological distinctions: (i) the implementation of stringent regularization constraints (*max_depth* = 10, *min_samples_leaf* = 5), which restricted the ensemble capacity on the 1,520-record repository; and (ii) the utilization of an 80/20 split, yielding a smaller test partition with higher statistical variance. Collectively, both studies substantiate that tree-based ensemble methodologies significantly outperform SVR and KNN when applied to structured NCRB data.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

B. Preprocessing as a Determinant of Model Fidelity

The transformation of absolute incidence counts into population-normalized crime rates constitutes the most critical preprocessing intervention. In the absence of normalization, predictive models would conflate the high absolute figures of Mumbai (population ≈ 241 lakhs) with intrinsic risk, while concurrently underestimating the risk profile of smaller metropolitan units. This formulation establishes a uniform baseline and facilitates equitable cross-city comparative analysis [7, 11, 16].

Label encoding introduces an implicit ordinal assumption that may not be meaningful for nominal geographic indicators. Theoretically, target encoding or one-hot encoding integrated with dimensionality reduction might offer more robust performance, particularly for linear architectures such as SVR.

C. Limitations

Several constraints necessitate acknowledgement within the current scope:

- **Data dependency:** The NCRB repositories reflect only documented offenses; research indicates that approximately 50% of criminal activity remains unreported due to social stigma or institutional mistrust [3], thereby introducing systematic undercounts into the training data.
- **Static analysis:** The current framework utilizes historical batch processing and lacks the capacity to ingest real-time streaming data from IoT sensors, CCTV surveillance, or emergency dispatch logs.
- **Algorithmic bias:** Historical patterns of over-policing in specific sectors elevate arrest records, which the model may interpret as heightened risk—potentially propagating feedback loops that disproportionately affect marginalized communities [13, 14, 25, 26, 28, 30].

VIII. FUTURE WORK

Several high-impact trajectories are envisioned for future research. Primarily, the integration of real-time IoT feeds via edge computing—incorporating automated acoustic sensors and license plate readers—would facilitate continuous risk assessment [3, 20, 22, 23]. Furthermore, deep learning architectures (e.g., YOLO) applied to CCTV streams could enable real-time detection of suspicious activities. NLP-driven sentiment analysis of social media via BiLSTM networks has demonstrated efficacy in identifying social tension as a precursor to criminal events [3]. Crucially, the implementation of Explainable AI (XAI) techniques, such as SHAP and LIME, is essential to ensure accountability in public-safety decisions [15]. Finally, as policing paradigms incorporate AI, addressing adversarial threats—including deepfakes and synthetic identity fraud—remains a critical priority.

IX. CONCLUSION

This research presented a reproducible machine learning pipeline for forecasting crime rates across 19 Indian metropolitan nuclei utilizing NCRB data from 2014 to 2021. Five regression algorithms were evaluated on a population-normalized dataset. Empirical evidence suggests that the Decision Tree Regressor achieved the highest predictive fidelity (R^2 = 0.889, MAE = 2.889), effectively capturing the complex city–type interactions. The Random Forest model, when regularized (R^2 = 0.728), offered enhanced resistance to stochastic noise, while SVR and MLP architectures demonstrated limitations under default configurations on small-scale tabular data.

The primary contributions of this work include: (i) a documented data engineering pipeline for Indian crime prediction; (ii) an empirical demonstration of regularization impacts on ensemble model performance; and (iii) a deployment architecture designed for law enforcement accessibility. These contributions establish a scalable foundation for transitioning toward predictive policing in India, providing a roadmap for real-time IoT integration, algorithmic transparency, and bias mitigation strategies.

DOI: 10.48175/IJARSCT-37152

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

ACKNOWLEDGEMENT

The authors express their gratitude to the National Crime Records Bureau (NCRB), Government of India, for providing public access to annual crime statistics. This research received no external financial support from the public or commercial sectors.

REFERENCES

[1]. Global study on homicide 2023, United Nations Office on Drugs and Crime, 2023.
[2]. United Nations Office on Drugs and Crime (UNODC). Annual Report 2024: Making the World Safer from Drugs, Crime, Corruption, and Terrorism, 2024.
[3]. Predictive Policing or Predictive Prejudice? AI Entanglement with Historical Injustices, Oxford Journal of Law and Technology, 2024.
[4]. R. Wortley, M. Townsley, Environmental criminology and crime analysis, 2nd Edition, Routledge, London, 2016.
[5]. D. K. Rossmo, Environmental criminology and the routine activity approach, Crime Prevention Studies, Vol. 4, 1995.
[6]. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature, 2015, 521, 436–444, doi: 10.1038/nature14539.
[7]. V. J. Nishad, N. J. Bawankar, H. R. Chaure, V. Kumar, A. Chaube, AI-based crime rate prediction, International Journal of Trend in Scientific Research and Development, 2024, 8, 817-822.
[8]. L. McClendon, N. Meghanathan, Using machine learning algorithms to analyze crime data, Machine Learning and Applications: An International Journal, 2015, 2, doi: 10.5121/mlaij.2015.2101.
[9]. V. Mandalapu, L. Elluri, P. Vyas and N. Roy, Crime prediction using machine learning and deep learning: a systematic review and future directions, IEEE Access, 2023, 11, 60153-60170, doi: 10.1109/ACCESS.2023.3286344.
[10]. S. Mahmud, M. Nuha, A. Sattar, Crime rate prediction using machine learning and data mining, In: Borah, S., Pradhan, R., Dey, N., Gupta, P. (eds) Soft Computing Techniques and Applications, Advances in Intelligent Systems and Computing, Springer, Singapore, 2021, 1248, doi: 10.1007/978-981-15-7394-1_5.
[11]. W. Safat, S. Asghar, S. A. Gillani, Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques, IEEE Access, 2021, 9, 70080-70094, doi: 10.1109/ACCESS.2021.3078117.
[12]. T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 1967, 13, 21-27, doi: 10.1109/TIT.1967.1053964.
[13]. L. Breiman, Random forests, Machine Learning, 2001, 45, 5–32, doi: 10.1023/A:1010933404324.
[14]. C. Cortes, V. Vapnik, Support-vector networks, Machine Learning, 1995, 20, 273–297, doi: 10.1007/BF00994018.
[15]. S. Chandrasekar, S. Dhatchanamoorthy, R. Nithish, P. Kishore Krishna, K. Mirresh, Crime rate predication indian cities using random forest classifiers, 2025, 11.
[16]. S. Sridharan, N. Srish, S. Vigneswaran, P. Santhi, Crime prediction using machine learning, EAI Endorsed Transactions on Internet of Things, 2024, 10, doi: 10.4108/eetiot.5123.
[17]. H. K. Reddy ToppiReddy, B. Saini, G. Mahajan, Crime prediction & monitoring framework based on spatial analysis, Procedia Computer Science, 2018, 132, 696-705, 10.1016/j.procs.2018.05.075
[18]. J. R. Quinlan, Induction of decision trees, Machine Learning, 1986, 1, 81-106.
[19]. J. R. Quinlan, Decision trees and decisionmaking, IEEE Transactions on Systems, Man, and Cybernetics, 1990, 20, 339-346, doi: 10.1109/21.52545.
[20]. M. Alkaff, N. F. Mustamin, G. A. A. Firdaus, Prediction of crime rate in Banjarmasin city using RNN-GRU Model, International Journal of Intelligent Systems and Applications in Engineering, 2022, 10, 01–09.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

[21]. K. Jenga, C. Catal, G. Kar, Machine learning in crime prediction, Journal of Ambient Intelligence and Humanized Computing, 2023, 14, 2887–2913, doi: 10.1007/s12652-023-04530-y.
[22]. G. Hajela, M. Chawla, A. Rasool, A clustering-based hotspot identification approach for crime prediction, Procedia Computer Science, 2020, 167, 1462-1470, doi: 10.1016/j.procs.2020.03.357.
[23]. M. Fatehkia, D. O'Brien, I. Weber, Correlated impulses: Using Facebook interests to improve predictions of crime rates in urban areas, PLOS ONE, 2019, 14, e0211350, doi: 10.1371/journal.pone.0211350.
[24]. V. Pednekar, T. Mahale, P. Gadhave, A. Gore, Crime rate prediction using KNN, International Journal on Recent and Innovation Trends in Computing and Communication, 2018, 6, 124, doi: 10.17762/ijritcc.v6i1.1392.
[25]. Algorithmic Biases in Artificial Intelligence (AI): Sampling Errors and Demographic Diversity, World Journal of Advanced Research and Reviews (WJARR), Vol. 25, 2025.
[26]. P. Kondapalli, P. Singh, A. Malik, C. S. A. Teddy Lesmana, A literature review: Bias detection and mitigation in criminal justice, Engineering Proceedings, 2025, 107, 72, doi: 10.3390/engproc2025107072.
[27]. S. Uddin, I. Haque, H. Lu, M. A. Moni, E. Gide, Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction, Scientific Reports, 2022, 12, 6256, doi: 10.1038/s41598-022-10358-x.
[28]. Algorithms in policing: An investigative packet on bias and accountability, Yale Law School, Media Freedom and Information Access (MFIA) Clinic, 2023.
[29]. D. Mienye, N. Jere, A survey of decision trees: concepts, algorithms, and applications, IEEE Access, 2024, 12, 86716-86727, doi: 10.1109/ACCESS.2024.3416838.
[30]. O. Kayode, Algorithmic bias and justice: analyzing the ethical limits of machine learning in high-stakes decision-making, 2025.