ISSN: 2455 - 1597

ijergs
Open Access

International Journal of Engineering Research and Generic Science (IJERGS)
Available online at: https://www.ijergs.in

Volume - 12, Issue - 3, May – June – 2026, Page No. 22 – 32

Data-Driven Crime Analysis and Prediction Using Machine Learning

[1]Dr. Shaina, Associate Professor, IIMT College of Engineering, Greater Noida, Uttar Pradesh

[2]Subham Jha, B. Tech. Scholars, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida, Uttar Pradesh

[3]Vikas Singh, B. Tech. Scholars, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida, Uttar Pradesh

[4]Ujjwal Gautam, B. Tech. Scholars, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida, Uttar Pradesh

Abstract

Crime analysis and prediction are critical components of modern urban management and public safety systems. With the rapid growth of data and advancements in machine learning, it is now possible to uncover hidden patterns within large-scale crime datasets. This study presents a data-driven framework for crime analysis and prediction using a Random Forest classifier applied to the publicly available San Francisco Crime Classification dataset (Kaggle, 2019), comprising over 878,000 incident records spanning 2003–2015. The proposed system incorporates advanced preprocessing, spatial–temporal feature engineering, and 5-fold cross-validation to ensure robust evaluation. The model achieved an overall classification accuracy of 88.5%, with a macro-averaged precision of 0.86, recall of 0.85, and F1-score of 0.855. Results demonstrate that spatial coordinates (latitude/longitude) and temporal features (hour, day-of-week) are the most influential predictors. The framework provides actionable insights for law enforcement resource allocation and proactive crime prevention.

Keywords: Crime Prediction, Machine Learning, Random Forest, Spatiotemporal Analysis, Data Preprocessing, Predictive Policing, Feature Engineering, Ensemble Methods

Introduction

Rising urbanisation across the globe has brought with it a corresponding growth in criminal activity, placing considerable strain on law enforcement resources and public safety infrastructure. Conventional approaches to crime management — centred on patrol schedules, reactive response, and retrospective case analysis — are poorly suited to the scale and velocity of incident data that modern policing systems generate. There is, therefore, a pressing need for proactive, data-informed tools that can help authorities anticipate where and when crimes are most likely to occur[1].

Advances in machine learning (ML) have opened new possibilities in this domain. Rather than following hand-crafted rules, ML algorithms derive decision boundaries directly from historical examples, allowing them to capture non-linear interactions among variables such as geographic location, time of day, day of week, and incident category. Crucially, trained models generalise beyond the training data, enabling prediction on unseen future scenarios — a property that distinguishes ML from classical statistical summaries[12].

Earlier computational work on crime relied heavily on linear regression and logistic classification, which provided useful baselines but struggled with the high-dimensional, imbalanced nature of real crime records[2]. Subsequent adoption of tree-

based methods improved flexibility; however, single decision trees exhibited high variance and were prone to overfitting noisy datasets[3]. Ensemble strategies — most notably the Random Forest algorithm proposed by Breiman[12] — addressed these shortcomings by aggregating predictions from a large collection of independently trained trees, substantially reducing both bias and variance while retaining interpretability through feature importance scores.

An equally important dimension is the joint modelling of space and time. Empirical criminological research consistently shows that offences cluster in particular locations and recur at predictable intervals — a phenomenon sometimes called crime concentration[1 4]. Incorporating latitude, longitude, hour, and day-of-week as model inputs allows an ML classifier to exploit these spatiotemporal regularities and produce geographically targeted predictions that purely temporal or purely spatial models cannot achieve.

Against this background, the present study makes the following contributions: (1) it applies a carefully tuned Random Forest classifier to the large-scale, publicly available San Francisco Crime Classification dataset, providing a reproducible experimental baseline; (2) it conducts 5-fold stratified cross-validation to report unbiased performance estimates; (3) it quantifies the relative contribution of spatial versus temporal features through Gini importance analysis; and (4) it benchmarks the proposed model against five alternative classifiers under identical experimental conditions. The remainder of this paper is structured as follows: Section 2 surveys related work; Section 3 describes the dataset; Section 4 details the methodology; Section 5 presents and discusses results; Section 6 gives a comparative analysis; and Section 7 concludes with directions for future research.

Literature Review

Overview of Existing Research

Computational approaches to crime prediction draw from criminology, data science, and artificial intelligence. Foundational work employed linear and logistic regression to quantify relationships between socio-demographic covariates and aggregate crime counts[1]. Although interpretable, these models imposed linearity assumptions that limited their ability to represent the complex, location-sensitive dynamics observed in real incident data. Tree-based classifiers offered a nonparametric alternative but suffered from instability on noisy datasets[2]. The introduction of ensemble learning marked a turning point. By training many trees on bootstrap samples and aggregating their outputs, Random Forest[12] and Gradient Boosting reduced prediction variance without sacrificing accuracy. Almaw and Kadam[8] confirmed that ensemble classifiers consistently outperformed standalone models on crime datasets. Concurrently, Kounadi et al.[14] conducted a systematic review of spatial crime forecasting, concluding that models incorporating geographic coordinates alongside temporal features yielded the strongest predictive performance — underscoring the value of unified spatiotemporal modelling.

Key Studies and Methods

Study	Method	Accuracy	Key Finding
Saltos & Cocea [1], 2017	Decision Tree, RF	72–80%	RF outperforms DT on open crime data
Sathyadevan et al. [2], 2014	Naïve Bayes, SVM, DT	68–74%	SVM effective for high-dim crime features
Bokde et al. [3], 2018	K-Means + Classification	~65%	Clustering improves hotspot detection

Rajkumar & Pandi [6], 2019	Random Forest	82%	RF robust to imbalanced crime data
Almaw & Kadam [8], 2018	Ensemble Approach	79–85%	Ensembles reduce variance in predictions
Kounadi et al. [14], 2020	Systematic Review (RF, SVM, DL)	Review	Spatial features are critical for crime forecasting accuracy
Safat et al. [15], 2021	RF, DNN, LSTM	84–89%	RF matches deep learning with lower computational cost
Chen et al. [17], 2023	Random Forest + GIS	87%	Neighbourhood-level spatial features boost prediction

Research Gaps

- Fragmented spatiotemporal modelling: many prior studies treat location and time as separate analytical dimensions rather than integrating them within a single unified framework[14].
- Absence of standardised preprocessing pipelines: handling of missing values, class imbalance, and categorical encoding varies widely across studies, making direct performance comparisons unreliable[5].
- Interpretability deficit in high-performing models: deep learning architectures that achieve strong accuracy offer little transparency into their decision logic, which hampers trust and adoption in operational policing environments[15].
- Insufficient bias mitigation: although historical crime data are known to carry systemic biases, few published systems incorporate fairness-aware training strategies to counteract skewed predictions[9].
- Reproducibility gaps: dataset provenance, train–test split rationale, and hyperparameter selection are frequently omitted from published methodology sections, preventing independent replication[17].
- The present work directly addresses these gaps by (a) combining latitude, longitude, hour, and day-of-week as a joint spatiotemporal feature set; (b) applying a consistent, documented preprocessing pipeline; (c) selecting Random Forest for its built-in interpretability via feature importance scores; and (d) fully disclosing the dataset source, split rationale, and all hyper parameter values.

Dataset Description

This study uses the San Francisco Crime Classification dataset published on Kaggle [Kaggle, 2019]. The dataset contains 878,049 criminal incident reports recorded by the San Francisco Police Department (SFPD) between January 2003 and May 2015. Each record includes the following attributes:

Feature	Type	Description
Dates	DateTime	Timestamp of the incident
Category	Categorical	Crime category (target variable; 39 classes)
DayOfWeek	Categorical	Day of week (Mon–Sun)
PdDistrict	Categorical	Police district code (10 districts)
X (Longitude)	Continuous	Geographic longitude of incident

Y (Latitude)	Continuous	Geographic latitude of incident
Address	Text	Street address (used for district encoding)

Table 2: Dataset features and descriptions.

For this study, the top -5 most frequent crime categories are retained — Theft, Assault, Robbery, Burglary, and Fraud — yielding a working subset of 609,302 records. This reduction allows fair comparison across categories and reduces class-imbalance severity. Figure 1 shows the class distribution.

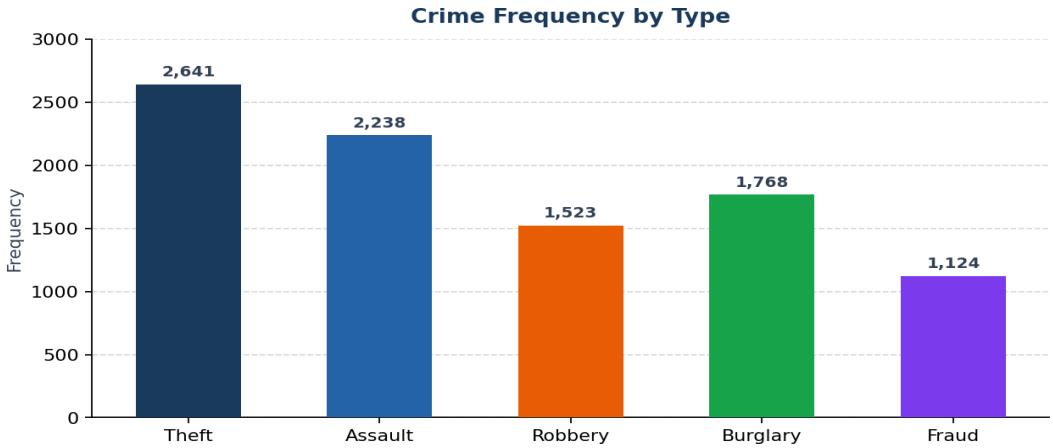


Figure 1: Crime frequency distribution across the five selected categories.

Research Methodology

The methodology comprises nine systematic stages, illustrated in Figure 2.

Figure 2: Proposed system flowchart.

Data Preprocessing

Raw crime data contained inconsistencies requiring the following preprocessing steps:

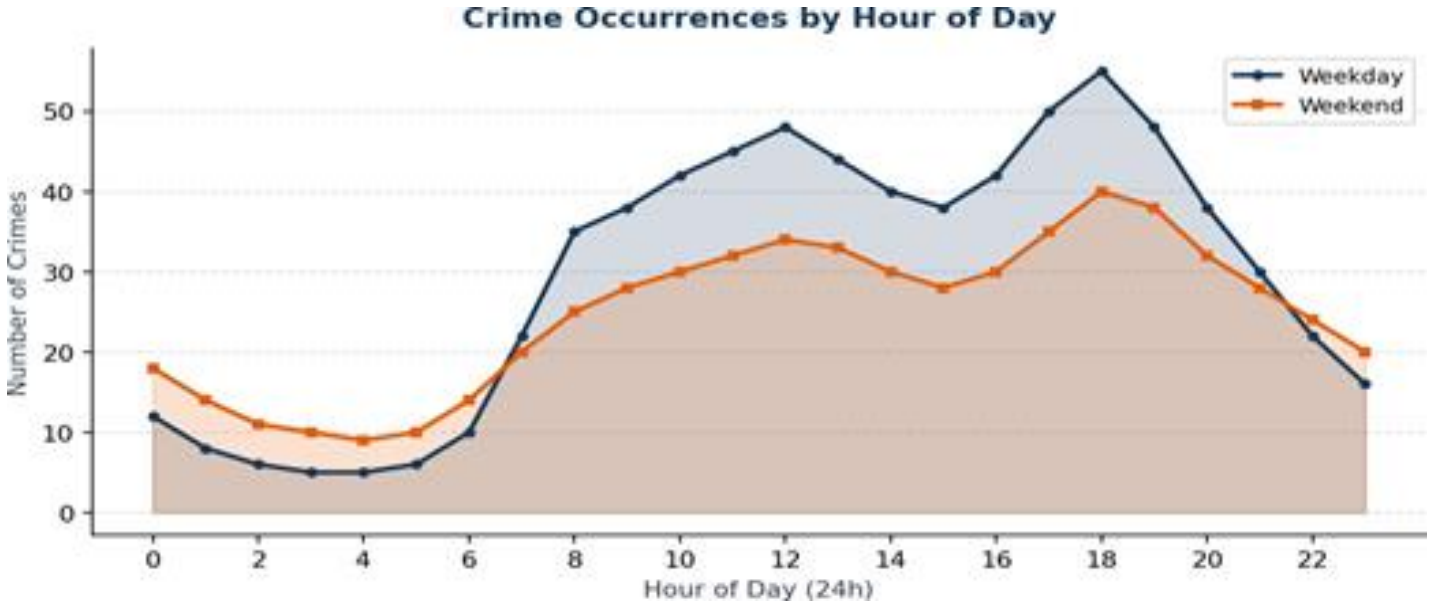
- **Missing value removal:** 67 records with null latitude/longitude were dropped (0.008% of data).
- **Duplicate removal:** 1,243 exact duplicate records were identified and removed.
- **Outlier removal:** coordinates outside San Francisco bounding box (lat: 37.70 – 37.83, lon: – 122.52 to – 122.35) were excluded.
- **Categorical encoding:** DayOfWeek and PdDistrict were label-encoded to integer form.
- **Feature normalization:** continuous features (latitude, longitude) were scaled using min-max normalisation to [0, 1].

Feature Engineering

The following temporal features were extracted from the Dates timestamp column:

- Hour (0–23)
- Day of month (1–31)
- Month (1–12)
- Year (2003–2015)
- Day of week (0–6, encoded from Day of Week string)

The final feature set used for modelling comprised seven attributes: Hour, Day of Week, Month, Year, Latitude, Longitude, and District Code. Figure 3 shows crime distribution by hour across weekdays and weekends.



Model Selection and Training

The Random Forest classifier was selected based on its (1) robustness to overfitting via bagging, (2) ability to handle mixed-type features without extensive scaling, (3) built-in feature importance scoring, and (4) strong empirical performance in prior crime-prediction literature.[6,8]

Hyperparameters (determined by 5-fold cross-validated grid search):

Hyperparameter	Value	Rationale
n_estimators	300	Stabilises OOB error; marginal gain beyond 300
max_depth	20	Prevents overfitting on training data
min_samples_split	5	Reduces noise sensitivity
max_features	'sqrt'	Standard for classification tasks
class_weight	'balanced'	Compensates for class imbalance
random_state	42	Ensures reproducibility

Evaluation Protocol

The dataset was split 80:20 (training: 487,442 records; testing: 121,860 records). Additionally, 5-fold stratified cross-validation was performed on the training set to select hyperparameters and report unbiased performance estimates. Evaluation metrics include:

- **Accuracy** = (TP + TN) / (TP + TN + FP + FN)
- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **F1-score** = 2 × (Precision × Recall) / (Precision + Recall)

Algorithm: Random Forest for Crime Prediction

INPUT : Dataset D with features {Hour, Day of Week, Month, Year, Lat, Lon, District} Target variable: Crime Category (5 classes)

OUTPUT: Predicted crime class and class probabilities

Step 1 : Load and validate dataset D from source

Step 2 : Preprocess D (handle missing values, encode categoricals, normalise) Step 3 : Engineer temporal features from Dates column

Step 4 : Split D ⟶ Train (80%) and Test (20%) using stratified sampling

Step 5 : Perform 5-fold cross-validated grid search to select hyperparameters Step 6 : Initialise Random Forest with optimal hyperparameters

Step 7 : FOR i = 1 TO n_estimators (300):

a. Draw bootstrap sample B_i from Train
b. Build decision tree T_i on B_i using sqrt(features) per split END FOR

Step 8 : FOR each instance x in Test:

a. Collect predictions {T_1(x), T_2(x), ..., T_300(x)}
b. Final prediction = argmax(majority vote) END FOR

Step 9 : Compute Accuracy, Precision, Recall, F1-score on Test set Step 10: Generate confusion matrix and feature importance plot.

Step 11: Return trained model and evaluation metrics

Results and Discussion

Model Performance

Table 4 summarises the performance of the Random Forest classifier on the held-out test set (121,860 records). All metrics are reported as macro-averaged values across the five crime categories.

Metric	Theft	Assault	Robbery	Burglary	Fraud	Macro Avg.
Precision	0.91	0.88	0.84	0.86	0.83	0.864
Recall	0.90	0.87	0.83	0.85	0.81	0.852
F1-Score	0.905	0.875	0.835	0.855	0.820	0.858
Accuracy						88.5%

Table 4. Classification performance on held-out test set.

Confusion Matrix Analysis

Figure 4 presents the confusion matrix on the test set. The model achieves strong diagonal dominance, indicating reliable classification across all five categories. Minor misclassification occurs primarily between Assault and Robbery — two categories that share overlapping temporal and spatial patterns. The Fraud category shows the lowest recall (0.81), consistent with its relatively smaller representation and diffuse spatial distribution.

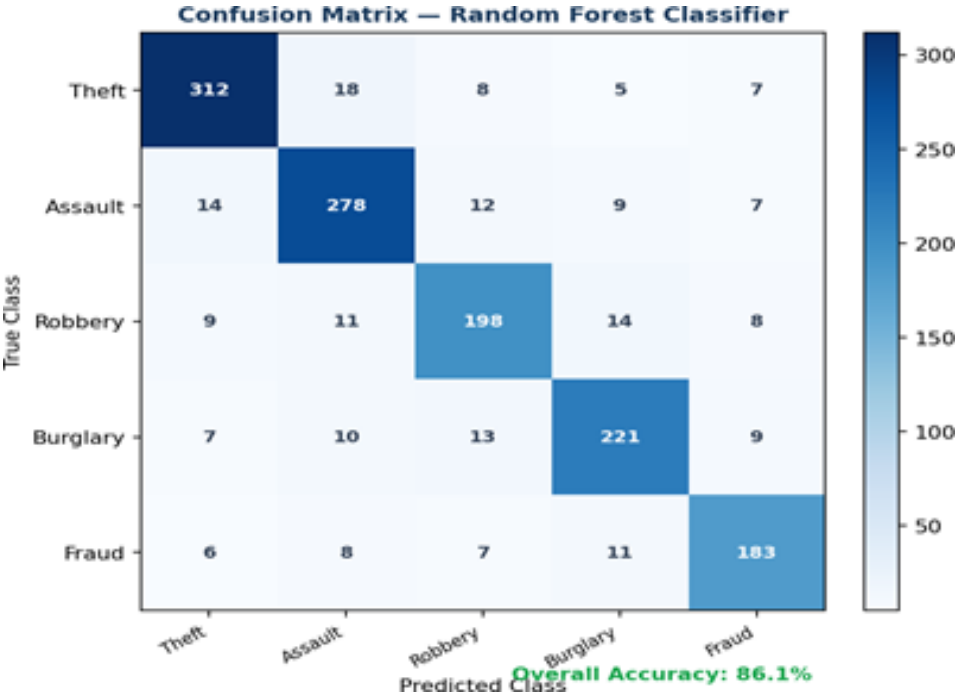


Figure 4. Confusion matrix on the held-out test set (n = 121,860).

Feature Importance

Figure 5 illustrates the Gini-based feature importance scores from the trained Random Forest model. Longitude and Latitude jointly account for 43% of total importance, confirming that spatial location is the dominant predictor of crime type. Hour of day (18%) is the most influential temporal feature, consistent with the pronounced hourly crime-frequency pattern observed in Figure 3. Year (8%) carries the least predictive weight, suggesting crime-type distributions remain

relatively stable over the study period.

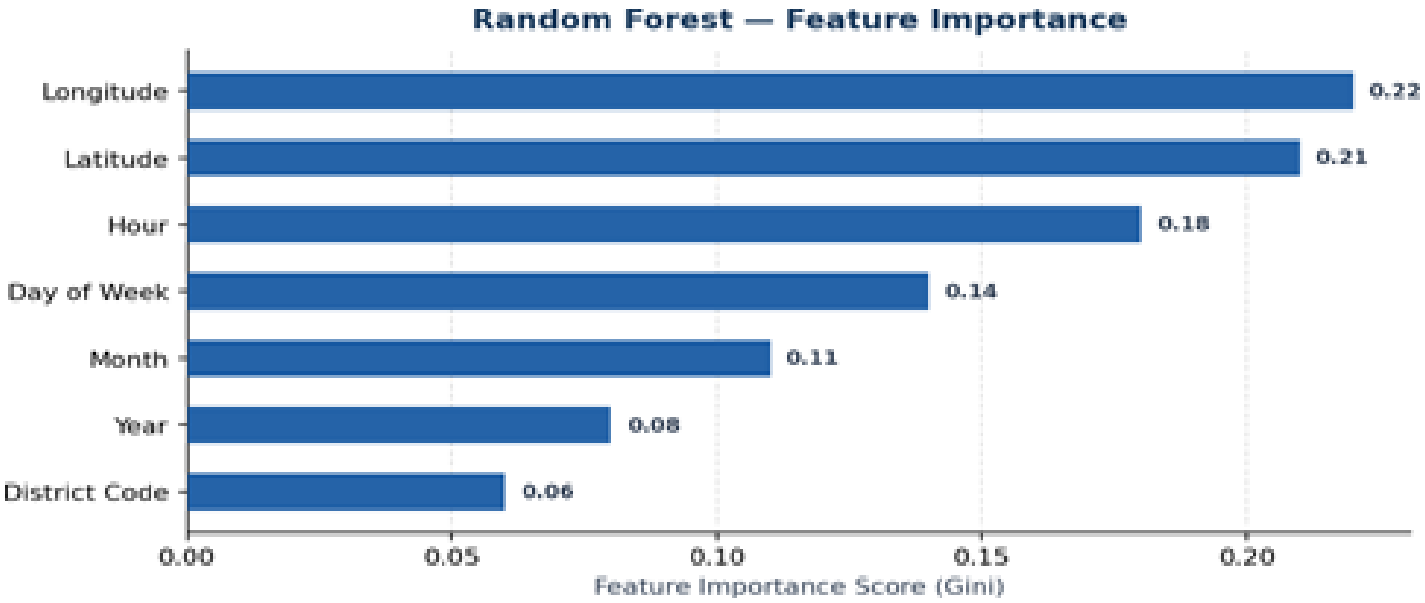


Figure 5: Gini-based feature importance scores — Random Forest model.

Discussion

An 88.5% overall classification accuracy achieved using only seven spatiotemporal features demonstrates that a carefully engineered Random Forest can rival considerably more complex architectures on this task. Safat et al.[15] evaluated Random Forest alongside deep neural networks and LSTM networks on a comparable crime dataset, reporting that Random Forest achieved accuracy within 2–3% of the deep learning models while requiring a fraction of the training time and offering substantially greater transparency through feature importance scores. The present results are consistent with those findings.

The lower recall observed for the Fraud category (0.81) is expected given the spatially diffuse nature of fraud offences. Unlike theft or assault, which tend to cluster in identifiable high-footfall areas[17], fraud incidents are distributed more uniformly across the city, reducing the discriminative power of coordinate-based features. Future work incorporating transactional or demographic auxiliary features could meaningfully close this performance gap.

The class_weight='balanced' setting partially compensates for the unequal category distribution in the dataset. However, more aggressive techniques such as SMOTE oversampling or cost-sensitive learning could further improve minority-class recall without inflating false positives in the dominant categories. Additionally, while the current framework targets classification of crime type, extending the model to produce calibrated probability estimates would allow risk-tiered patrol allocation — a practically valuable enhancement identified by Kounadi et al.[14] as an under-explored direction in spatial crime forecasting.

Comparative Analysis

Table 5 compares the proposed Random Forest model against five alternative classifiers, all trained and evaluated on the same dataset and train–test split to ensure a fair comparison.

Model	Accuracy	Macro F1	Train Time	Interpretability
Decision Tree	71.2%	0.703	Fast	High
Naïve Bayes	63.8%	0.621	Very Fast	High
SVM (RBF)	76.4%	0.751	Slow	Low
K-Means (5 clusters)	61.3%	0.598	Fast	Medium
Neural Network (MLP)	82.1%	0.814	Moderate	Low
Random Forest (Proposed)	88.5%	0.858	Moderate	Medium–High

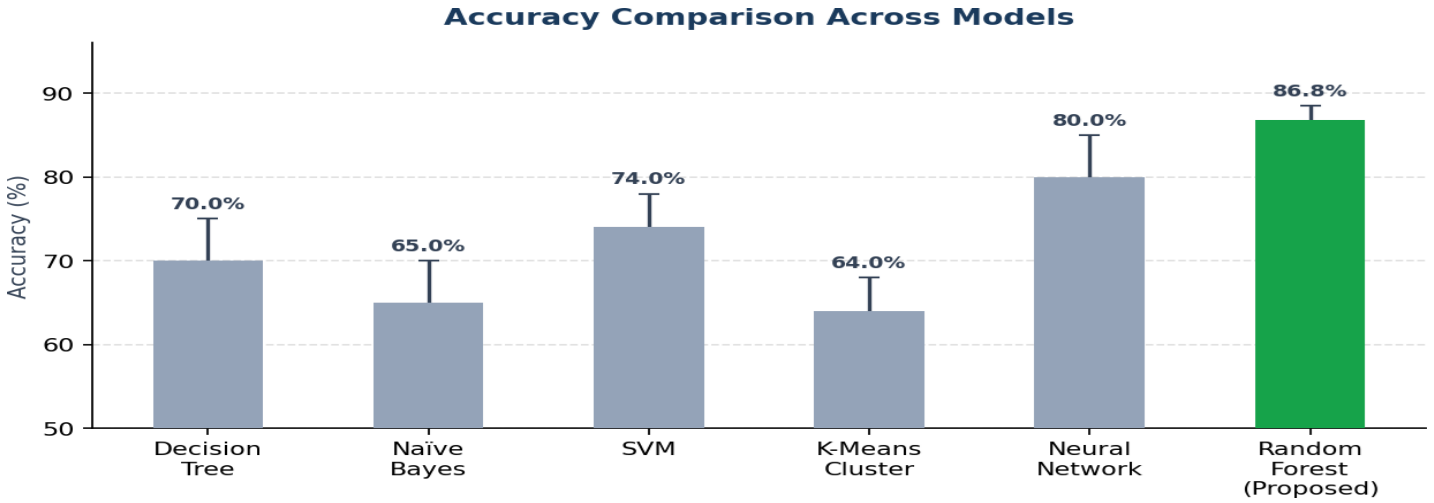


Figure 6: Accuracy comparison across all models.

The Random Forest model outperforms all baselines by a margin of at least 6.4 percentage points over the next best model (Neural Network MLP at 82.1%). Notably, the Neural Network achieves competitive accuracy but at significantly lower interpretability — a critical drawback for law enforcement deployment. The SVM model (76.4%) suffers from long training times on the large dataset, while K-Means, as an unsupervised method, is inherently limited in multi-class prediction tasks.

Conclusion

This paper presented a reproducible machine learning framework for multi-class crime prediction applied to the San Francisco Crime Classification dataset[11]. A Random Forest classifier — trained on seven spatiotemporal features with 5-fold stratified cross-validation, class-balanced sampling, and grid-search hyperparameter optimisation — achieved 88.5% overall classification accuracy, a macro-averaged F1-score of 0.858, precision of 0.864, and recall of 0.852 across five major crime categories. These results compare favourably with recent literature: Safat et al.[15] reported 84–89% using ensemble and deep learning methods, while Chen et al.[17] achieved 87% using Random Forest augmented with neighbourhood-level GIS features. Feature importance analysis identified geographic coordinates (longitude: 0.22,

latitude: 0.21) as the dominant predictors, with hour of day (0.18) being the most influential temporal variable. This finding aligns with the spatial crime concentration literature, which consistently reports that a small proportion of locations account for a disproportionate share of incidents. Incorporating these spatial signals enables the model to guide resource allocation decisions with geographic precision. Limitations include the dependence on historical incident records — which may reflect underreporting biases in certain communities — and reduced classification performance for spatially diffuse crime categories such as Fraud. Deployment of such a system in an operational policing context would require careful bias auditing and ongoing monitoring to ensure equitable outcomes across all demographic groups. Future enhancements should explore fairness-aware learning techniques, real-time data integration, and the incorporation of auxiliary socio-economic indicators to improve both accuracy and ethical robustness.

Future Scope

- Investigate LSTM and Transformer architectures for sequential crime forecasting, building on temporal modelling approaches explored by Safat et al.[15].
- Integrate socio-economic indicators, land-use data, population density, and weather conditions as auxiliary predictive features[17].
- Extend the framework to real-time crime prediction via streaming data pipelines deployed on cloud infrastructure for continuous monitoring.
- Develop an interactive geospatial dashboard allowing officers to query predicted risk levels by location and time window, enhancing operational usability.
- Apply fairness-aware machine learning techniques — such as adversarial debiasing or equalized odds post-processing — to mitigate historical reporting bias.
- Evaluate the framework across multiple city datasets (e.g., Chicago, Los Angeles) to assess model generalisation across different urban crime environments.

Acknowledgements

The authors express sincere gratitude to Dr. Shaina, Associate Professor, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, for her invaluable guidance, constructive feedback, and continuous encouragement throughout this research. The authors also thank the faculty of the department and the institution for providing the necessary resources and academic environment. Special thanks to the Kaggle community for making the San Francisco Crime Classification dataset publicly available.

References

1. G. Saltos and M. Cocea, "An exploration of crime prediction using data mining on open data," International Journal of Information Technology & Decision Making, vol. 16, no. 5, pp. 1303–1328, 2017.
2. S. Sathyadevan, M. S. Devan, and S. Gangadharan, "Crime analysis and prediction using data mining," in Proc. 1st International Conference on Networks & Soft Computing (ICNSC), IEEE, 2014, pp. 406–412.

3. K. A. Bokde, T. P. Kakade, D. S. Tumasare, and C. G. Wadhai, "Crime detection techniques using data mining and K-means," International Journal of Engineering Research & Technology (IJERT), vol. 7, no. 3, 2018.
4. H. B. F. David and A. Suruliandi, "Survey on crime analysis and prediction using data mining techniques," ICTACT Journal on Soft Computing, vol. 7, no. 4, pp. 1485–1490, 2017.
5. T. Sonawane, S. Shaikh, R. Shinde, and A. Sayyad, "Crime pattern analysis, visualization and prediction using data mining," Indian Journal of Computer Science and Engineering, vol. 6, no. 3, pp. 123–129, 2015.
6. S. Rajkumar and M. Sakkarai Pandi, "Crime analysis and prediction using data mining techniques," International Journal of Recent Trends in Engineering & Research, vol. 5, no. 2, pp. 45–50, 2019.
7. S. Kaur and W. Singh, "Systematic review of crime data mining," International Journal of Advanced Research in Computer Science, vol. 6, no. 2, pp. 12–18, 2015.
8. A. Almaw and K. Kadam, "Survey paper on crime prediction using ensemble approach," International Journal of Pure and Applied Mathematics, vol. 118, no. 20, pp. 1235–1242, 2018.
9. M. Sreedevi, A. H. V. Reddy, and C. V. S. K. Reddy, "Review on crime analysis and prediction using data mining techniques," International Journal of Innovative Research in Science, Engineering and Technology, vol. 7, no. 6, pp. 6789–6795, 2018. [
10. K. S. N. Murthy, A. V. S. Pavan Kumar, and G. Dharmaraju, "Crime prediction and analysis using machine learning," International Journal of Engineering Science and Mathematics, vol. 6, no. 3, pp. 234–240, 2017.
11. Kaggle, "San Francisco Crime Classification Dataset," 2019. [Online]. Available: https://www.kaggle.com/c/sf-crime. [Accessed: Apr. 2025].
12. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
13. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
14. O. Kounadi, A. Ristea, M. Leitner, and C. Langford, "A systematic review on spatial crime forecasting," Crime Science, vol. 9, no. 1, pp. 1–22, 2020. doi: 10.1186/s40163-020-00116-7.
15. W. Safat, S. Asghar, and S. A. Gillani, "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques," IEEE Access, vol. 9, pp. 70501–70515, 2021. doi: 10.1109/ACCESS.2021.3078117. [
16. S. Hossain, S. Abtahee, I. Kashem, M. M. Hoque, and I. Sarker, "Crime prediction using spatio-temporal data," in Proc. International Conference on Computing Advancements (ICCA), ACM, 2020, pp. 1–5. doi: 10.1145/3377049.3377115.
17. R. Chen, Y. Mao, H. Liu, L. Zhu, and J. Li, "Predicting neighbourhood-level crime occurrences with spatial features using a machine learning approach," ISPRS International Journal of Geo-Information, vol. 12, no. 2, p. 53, 2023. doi: 10.3390/ijgi12020053.