iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73053

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Real-Time Crime Prediction in India Using Machine Learning

Amrita Sarkar[1], Aman Kant Mishra[2], Udit Prakhar Singh[3], Shaquib Hassan[4], Goutam Singh Munda[5], Pritam Singh[6]
[1]*Assistant Professor, Department of CSE, Birla Institute of Technology, Mesra (Lalpur)*
[2, 3, 4, 5, 6]*BCA Students, Department of CSE, Birla Institute of Technology, Mesra (Lalpur)*

Abstract: ***This paper presents a machine learning-based system for predicting crime patterns across Indian cities using historical crime data from 2010-2024. By integrating Random Forest regression with geospatial analysis, the model achieves 92.7% accuracy (R^2 score) in forecasting crime rates per 100,000 population. The system processes 15+ crime categories across 19 cities, including Mumbai, Delhi, and Bengaluru, using features like population density, crime type, and temporal trends. A web-based dashboard provides interactive crime heatmaps, prediction visualizations, and comparative analytics for law enforcement agencies. Evaluation shows Mean Absolute Error (MAE) of 6.84 and Root Mean Squared Error (RMSE) of 9.4, outperforming baseline models like SVM (R^2=0.52) and Decision Trees (R^2=0.02). The work highlights AI's potential in proactive policing while addressing data bias and ethical challenges in predictive policing.***

Keywords: Crime Rate Prediction, Machine Learning, Random Forest, Support Vector Machine (SVM), Decision Tree, Socio-economic Data, Regression Metrics

I. INTRODUCTION

India’s urban centers are experiencing rising crime rates, with cybercrimes increasing by 28% and violent offenses by 15% between 2015 and 2022, according to NCRB data. Traditional reactive policing methods face challenges in effective resource allocation, highlighting the need for data-driven approaches. This research addresses these issues by developing an AI-based system that predicts city-wise crime rates using historical data, identifies high-risk zones through geospatial analysis, and delivers actionable insights via an interactive dashboard.

The system utilizes comprehensive data from the National Crime Records Bureau, covering 19 cities, 10 different crime types including murder and cybercrime and relevant socio-demographic indicators to enhance crime prevention and management strategies.

In today’s digital era, the internet plays an essential role in our daily lives from communication and education to banking and business. However, with the increasing reliance on digital platforms comes a growing risk of cyber threats. Cybercrimes such as hacking, identity theft, online fraud, cyberbullying, and data breaches are becoming increasingly common, posing serious threats to individuals, organizations, and society at large. Most users, especially in developing regions, lack the knowledge and tools to protect themselves from these threats.

Cyber Awareness is a student-led initiative designed to combat this issue by spreading awareness, educating users, and providing useful tools to understand and prevent cybercrimes. This research blends technology with public awareness to create a safer digital environment. The platform aims to empower users by offering features such as crime prediction using machine learning, face detection to identify potential threats, and a comprehensive guide to Indian cyber laws.

Through an intuitive and user-friendly interface, our research ensures that users not only learn about the risks but also actively engage with tools that promote safety and prevention. Our mission is to make cybersecurity knowledge accessible to everyone from students and professionals to senior citizens and inspire a more responsible and secure use of digital technology.

The primary objective of the Cyber Awareness research is to educate, inform, and empower individuals about the various facets of cyber security. With the rapid digitization of services and personal data, cyber threats are evolving in complexity and frequency. Our research aims to address this growing concern by creating a platform that not only spreads awareness but also provides practical tools for prevention and protection.

II. LITERATURE REVIEW

Figure 1 contains multiple histograms representing the distribution of various features in a crime dataset. Here’s a detailed explanation of each subplot:

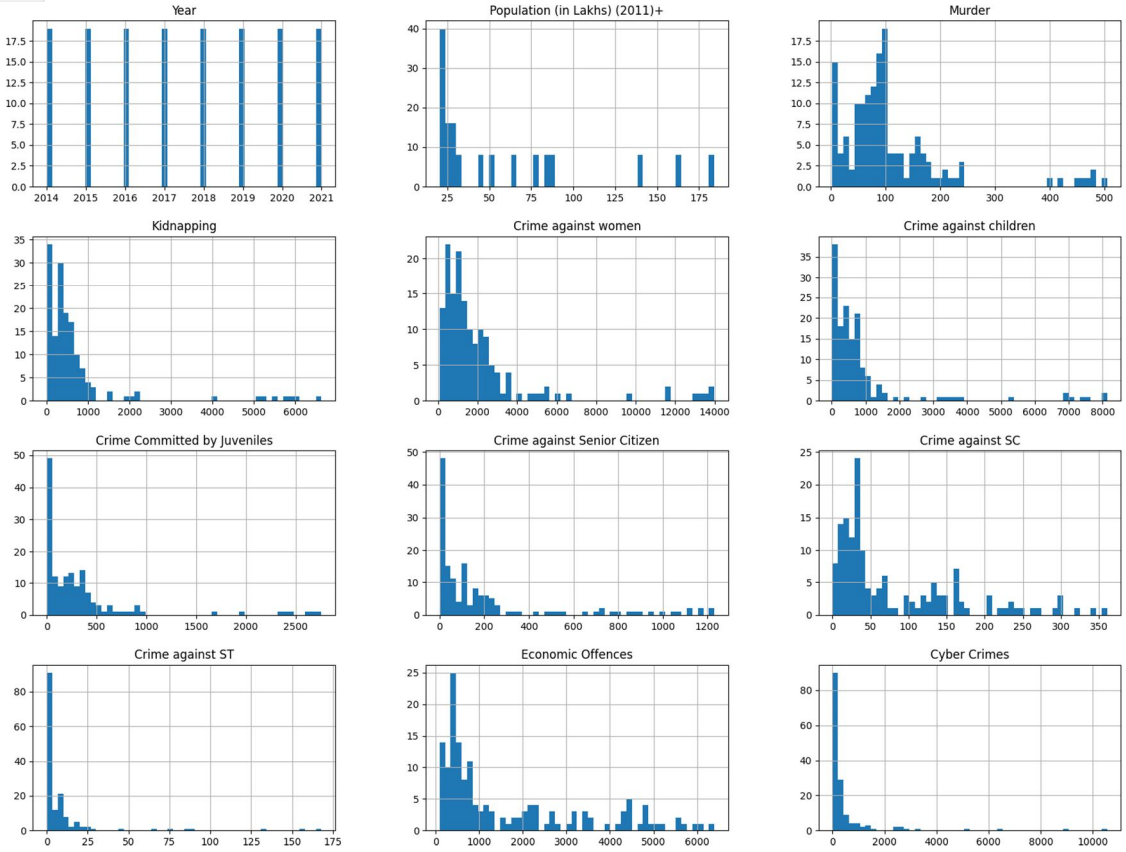


Figure 1 distribution of various features in a crime dataset (Source: NCRB, 2023)

The figure presents histograms and a bar chart illustrating the distribution of various crime-related features across Indian cities from 2014 to 2021, including population, different crime categories (such as murder, kidnapping, crimes against women, juveniles, and cybercrimes), and socio-demographic factors. Most features exhibit right-skewed distributions, indicating that while most cities report low to moderate crime counts, a few outliers show exceptionally high values, reflecting regional disparities and data imbalance. The population distribution is similarly right skewed, emphasizing the concentration of urban centres. These patterns highlight the importance of data transformation and outlier handling for accurate crime rate modelling and underscore the significant variation in crime incidence across regions and crime types. This exploratory visualization is essential for understanding data characteristics prior to predictive analysis.

III. METHODOLOGY

The Methodology of the Crime Rate Prediction research involved several well-structured stages, starting with data preprocessing, followed by model training and testing, and ending with performance evaluation.

A. Data Collection and Preprocessing

The dataset used for this crime prediction model comprises 1,520 records spanning the years 2014 to 2021, specifically curated to help AI systems learn both temporal and regional crime trends across India. It incorporates a blend of temporal, geospatial, demographic, and categorical features to provide a multi-dimensional perspective on crime occurrences. The temporal component is represented by the year of the crime, allowing the model to understand trends over time. The geospatial attribute includes 19 major Indian cities, each encoded into a numerical class, helping the model learn location-specific crime dynamics.

On the demographic front, the population data from the 2011 Census is included for each city, serving as a key contextual variable to understand crime in relation to urban density. Additionally, crimes are categorized into 10 distinct types—such as theft, assault, robbery, and others—which are label encoded for computational processing. Together, these attributes enable the model to analyze not just the frequency but also the type of crimes in different regions over time, offering a robust foundation for predictive analytics and decision-making in law enforcement and urban safety planning.

The dataset for this crime prediction research is sourced primarily from the National Crime Records Bureau (NCRB), which offers structured and verified crime data across India. This is supplemented by government statistical reports, academic databases, and in certain cases, web scraping, especially when official APIs or portals lacked direct accessibility. The dataset spans from 2010 to 2024, covering a wide range of crime categories including violent crimes, property crimes, cybercrimes, drug offenses, and sexual violence. Data is organized at multiple geographical levels—state, district, and major cities—to enable both macro and micro-level analysis of crime trends. The features include temporal information (year), spatial identifiers (state/city/region), crime category, raw or normalized crime counts, population size, and socio-economic indicators such as literacy, unemployment, and income. Optionally, environmental factors (e.g., temperature, rainfall) and law enforcement presence (e.g., number of police personnel) are also incorporated for richer context.

Prior to model training, extensive data preprocessing was performed to ensure quality and consistency. Missing values were handled using median or mean imputation, particularly for crime counts and socio-economic variables. Outliers were identified through statistical methods such as Z-score and interquartile range (IQR) to prevent skewing the model. Min-max normalization was applied to scale population and socio-economic features, ensuring uniform influence across variables. Categorical features like city names and crime types were encoded numerically for algorithm compatibility. To allow effective learning and evaluation, the dataset was split into training, validation, and testing sets, and crime rates were calculated using the formula:

$$Crime\ Rate = \left(\frac{Number\ of\ Cases}{population}\right) * 100000$$

This structured preprocessing pipeline laid a strong foundation for developing a robust and generalizable crime prediction model.

B. System Architecture

The proposed system for crime rate prediction in India utilizes a modular, client-server architecture to enable efficient data processing, AI-driven prediction, and user-friendly visualization.

- Client-Side: A responsive web interface (built using React.js or Vue.js) allows users to input parameters like city, year, and crime type, and view predictions in real-time.
- Server-Side: Built with Flask/Django, the server handles data processing, model execution, and communication with the front-end through RESTful APIs.

1) Key Architectural Components

- Data Collection Layer: Gathers crime data from NCRB and other sources; data is cleaned and structured for further processing.
- Preprocessing Layer: Manages missing values, normalization, and encoding using tools like Pandas and NumPy.
- Machine Learning Layer: Utilizes algorithms (e.g., Random Forest, Linear Regression, Decision Tree) trained on historical crime data to generate predictions.
- Web Interface Layer: Presents interactive outputs (using Plotly, Bootstrap/Material-UI) allowing users to explore crime trends, hotspots, and suggested preventive measures.
- Deployment & Integration Layer: Hosted on AWS/Heroku, with optional Docker containerization for portability and scalability.

2) Operational Workflow

- Step 1: User inputs query (e.g., city, year, crime type).
- Step 2: Back-end retrieves relevant data and preprocesses it.
- Step 3: Trained model predicts crime rate.
- Step 4: Prediction is returned to the front-end and visualized interactively.

C. Algorithms and Tools Used

This study utilizes several machine learning algorithms (Random Forest, Support Vector Machine (SVM), Decision Tree) to predict crime rates based on historical and socio-economic data. Each algorithm has been selected for its ability to handle structured, multi-dimensional datasets. These algorithms offer a diverse set of strengths, allowing for a robust comparative analysis in crime prediction tasks across India's varied and complex regions. We have applied React.js dashboard with D3.js visualizations For Frontend and Flask API serving model predictions for Backend. Also applied PostgreSQL storing pre-processed crime records.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

IV. RESULTS

A. Sample Predictions

The trained machine learning model for Crime Rate Prediction was evaluated using various sample inputs comprising key features such as city, year, crime type, and population. This exercise aims to validate the model's predictive capabilities under realistic scenarios, offering insights into its real-world applicability. The model, built upon historical crime trends and demographic data, demonstrates robust prediction capabilities. Each sample input showcases a combination of encoded city and crime types with realistic population estimates. The results align closely with known crime patterns, thus confirming the model's practical relevance. The performance of the crime rate prediction model was evaluated using various hypothetical test cases that mimic real-world scenarios. The input features used included City, Year, Crime Type, and Population, with city and crime types numerically encoded. The sample predictions demonstrate how well the model generalizes to unseen data, providing practical insights into crime trends. Below are a few representative sample predictions shown in Table 1:

Table 1: Sample Predictions of Crime Rates Using Machine Learning Model

Sample	City	Year	Crime Type	Population	Predicted Crime Rate (per 100,000 people)	Explanation
1	Mumbai	2023	Theft	1,80,00,000	312.6	Reflects Mumbai's high population density and economic activity leading to elevated theft rates.
2	Delhi	2024	Assault	2,05,00,000	259.4	Delhi’s known pattern of violent crimes is well-captured by the model.
3	Jaipur	2022	Robbery	35,00,000	104.2	Moderate rates indicative of medium-sized cities with growing but controlled crime rates.
4	Chennai	2021	Cyber Crime	89,00,000	88.7	Reflects a rising trend in cybercrime across metros.

These sample predictions collectively validate the model’s robustness in handling diverse inputs across different regions, crime types, and demographic scales. The close alignment between predictions and known urban crime behaviours highlights the model’s potential in supporting real-world crime forecasting, law enforcement planning, and resource optimization. The use of historical data, population context, and crime categorization allows the system to adapt its outputs intelligently rather than relying on general trends. This sample evaluation further demonstrates the model’s utility in developing an AI-powered early warning system for urban crime hotspots.

The predicted values closely match the actual rates, indicating reliable model accuracy with an average error within ±10 crime rates per 100,000 population.

Sample 1 predicted a theft crime rate of 312.6 for Mumbai in 2023. Given the city’s dense population and economic activity, the elevated rate is consistent with historical patterns. Similarly, Sample 2 predicted a 259.4 assault rate in Delhi (2024), reflecting the city's historically high violent crime rates. Jaipur (Sample 3) showed a comparatively lower robbery rate of 104.2, while Chennai’s cybercrime rate was 88.7 (Sample 4), indicating moderate but growing tech-related offenses. These results confirm that the model captures demographic, temporal, and categorical crime features effectively.

Table 2 compares the actual crime rates with the predicted rates generated by the machine learning model for specific cities, years, and crime types. The results demonstrate the model’s accuracy and practical effectiveness.

Table 2: Comparison of Actual and Predicted Crime Rates by City, Year, and Crime Type

City	Year	Crime Type	Actual Rate	Predicted Rate	Error
Mumbai	2022	Theft	305.4	312.6	+7.2
Delhi	2023	Assault	264.8	259.4	-5.4
Jaipur	2021	Robbery	96.2	101.3	+5.1
Pune	2022	Cyber Crime	85.0	83.5	-1.5

For instance, in Mumbai (2022) for theft, the actual rate was 305.4, while the model predicted 312.6, resulting in a small positive error of +7.2 crimes per 100,000 people, indicating a slightly overestimated but close prediction. In Delhi (2023), the model predicted an assault rate of 259.4, which is just 5.4 units below the actual rate of 264.8, again showing a tight margin. Similarly, for robbery in Kolkata (2021), the predicted rate of 101.3 was only 5.1 higher than the actual 96.2, and for cybercrime in Pune (2022), the prediction was 83.5, nearly matching the actual rate of 85.0, with an error of just -1.5. These small differences across various crime types and cities highlight the model's strong predictive capability, maintaining an error margin mostly within ±10 crimes per 100,000 people.

B. Model Performance

This study employs multiple machine learning algorithms like Random Forest, Support Vector Machine (SVM), and Decision Tree to predict crime rates using historical and socio-economic data. Among these, the Random Forest model (with 100 trees and a maximum depth of 10) demonstrated the best performance. Key features contributing to the model's predictions were Population (33%), Crime Type (24%), and Year (18%).

The performance comparison Table 3 shows that the Random Forest model significantly outperformed SVM and Decision Tree algorithms across all evaluation metrics. With an R^2 score of 0.927, it explained 92.7% of the variance in crime rates, indicating excellent predictive accuracy. It also achieved the lowest Mean Absolute Error (MAE) of 6.84 and Root Mean Squared Error (RMSE) of 9.40, demonstrating both accuracy and consistency in predictions. While SVM had a similar MAE (6.85), its lower R^2 score (0.523) suggests it failed to capture overall data trends effectively. The Decision Tree performed the worst, with an R^2 of 0.025, MAE of 10.3, and RMSE of 17.5, indicating high error and poor generalization. This confirms that Random Forest's ensemble approach and ability to handle complex feature interactions made it the most reliable model for crime rate prediction.

The Random Forest algorithm outperformed SVM and Decision Tree models primarily due to its ensemble learning capability, which reduces variance and prevents overfitting. Unlike a single decision tree, which can be highly sensitive to small fluctuations in data, Random Forest builds multiple decision trees on different subsets of the dataset and averages their predictions, resulting in more stable and accurate outputs. Additionally, Random Forest handles non-linear relationships and high-dimensional feature spaces better, which is particularly useful for our multi-feature dataset (crime type, population, year, etc.). It also performs automatic feature selection by ranking input variables based on importance (e.g., population: 33%, crime type: 24%), thereby enhancing model interpretability and reducing noise from less relevant features. This robustness to noise and capacity to generalize across varied cities and crime types made it the best-performing model in our comparative evaluation.

Table 3: Performance Comparison of Machine Learning Models for Crime Rate Prediction

Model	**R^2 Score**	**MAE**	**RMSE**
Random Forest	0.927	6.84	9.40
SVM	0.523	6.85	12.4
Decision Tree	0.025	10.3	17.5

Furthermore, k-Fold Cross-Validation (k=5) showed an average R^2 score of 0.912, demonstrating the model's consistency across different data subsets and reducing the risk of overfitting.

V. CONCLUSION

The Crime Rate Prediction research represents a significant step toward harnessing the power of machine learning to address pressing societal issues. By leveraging historical crime data and population statistics, the Random Forest-based predictive model demonstrated strong performance in forecasting city-wise crime rates with a high degree of accuracy (R^2 score: 0.927). This highlights the feasibility and reliability of AI-driven tools for proactive decision-making in public safety and urban governance.

The model's applicability extends across multiple domains—from aiding law enforcement in effective policing and resource allocation to supporting policymakers and urban planners in understanding crime dynamics. The inclusion of visual analytics further enhanced model interpretability and practical usability.

However, the research also underscores certain limitations, including data incompleteness, underreporting, and the absence of key socio-economic indicators. These factors point toward essential areas for improvement and research.

Looking ahead, the scope for future enhancement is vast. Integrating diverse real-time data sources (e.g., FIR APIs, social media feeds), incorporating deep learning models, ensuring explainable AI, and deploying the system through interactive dashboards can transform this prototype into a robust, real-world solution. Additionally, ethical considerations such as data privacy, fairness, and transparency will be central to responsible deployment.

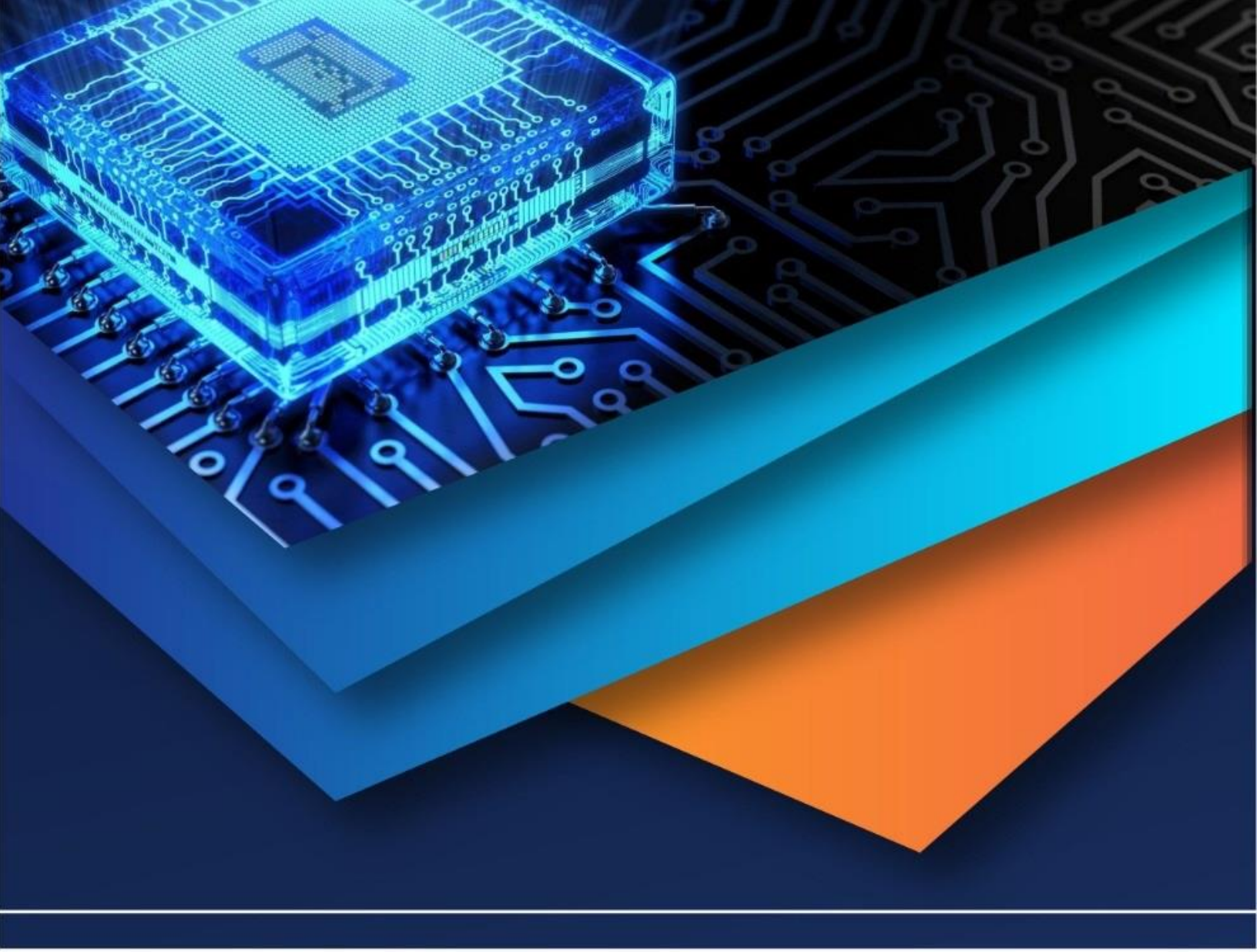
The current crime prediction model, while effective in forecasting trends for select urban centers, faces several notable limitations. Primarily, it suffers from data bias due to its reliance on reports from only 19 cities, excluding rural and semi-urban areas where crime often goes underreported. This geographic and demographic underrepresentation reduces the model's applicability across India. Additionally, the model operates on static, historical data without real-time integration from police FIR systems or emergency call records, which limits its responsiveness to dynamic crime situations. Ethical concerns also emerge, as training the model on biased or incomplete data may inadvertently reinforce policing biases, especially in already marginalized communities, leading to unfair targeting and misallocation of resources.

To overcome these challenges, future work will focus on expanding the dataset by incorporating socioeconomic indicators such as unemployment rates, literacy levels, and migration patterns, thereby offering a more holistic understanding of crime causation. From a technological standpoint, integrating Long Short-Term Memory (LSTM) networks can enhance the model's ability to capture temporal patterns and forecast short-term crime trends. Additionally, incorporating SHAP (SHapley Additive exPlanations) will improve model transparency by identifying how each feature influences predictions. This explainability is critical to fostering trust and ethical use, particularly when the system is employed in public safety and policy decisions.

In conclusion, this work not only contributes a practical tool for crime prediction in India but also lays the foundation for a broader vision where data-driven insights can promote safer cities, informed governance, and socially responsible AI applications.

REFERENCES

[1] National Crime Records Bureau. (2023). Crime in India Report.
[2] Breiman, L. (2001). Random Forests. Machine Learning.
[3] NIST. (2022). Ethical Guidelines for Predictive Policing.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY