Proceedings of the International Conference on Sustainable Communication Networks and Application (ICSCN-2025)
DVD Part Number: CFP25DW8-DVD; ISBN: 979-8-3315-9419-0

Crime Data Prediction using Machine Learning

Dr V.V.A.S.Lakshmi
Department of CSE(DS)
Narasaraopeta Engineering College
Narasaraopeta, India
vvaslakshmi@gmail.com

M. Chidananda Dedeepya
Department of CSE(DS)
Narasaraopeta Engineering College
Narasaraopeta, India
maddidedeepya@gmail.com

T.Asha Gayathri
Department of CSE(DS)
Narasaraopeta Engineering College
Narasaraopeta, India
gayathrithoka2004@gmail.com

U.Venkateswarlu
Department of CSE(DS)
Narasaraopeta Engineering College
Narasaraopeta, India
ubbathotivenky@gmail.com

***Abstract*-- Making our communities safer is the straightforward but impactful foundation of this dissertation. It focuses on developing a machine learning model that can assist in forecasting potential crime scenes and times, allowing us to take action before damage is done. This work urges us to reframe safety as care, support, and astute intervention before things go wrong, rather than as punishment after the fact. This study not only provides a new tool by fusing technology and empathy, but it also points to a new direction where data helps us understand people and public safety is a shared, compassionate responsibility**.

Keywords--Ensemble Model 4, Random Forest, Support Vector Machine (SVM), Deep Neural Network (DNN), and Kernel Density Estimation (KDE)

I. INTRODUCTION

Safety is one of the deepest needs of any society, yet predicting and preventing crime remains a difficult challenge. Traditional approaches—such as regression analysis or hotspot mapping—have helped highlight areas of concern, but they often look only at the past. They struggle to capture the "when" and "why" behind new incidents, leaving communities and law enforcement reactive rather than proactive.

In India, this challenge is even more complex. Crime data is vast but uneven, often recorded at yearly intervals and influenced by social and economic realities such as unemployment, literacy, and urban density. Most existing studies focus only on broad national or state-level patterns, offering limited insight into the local, everyday risks that people face. As a result, predictions are often too general to guide timely action, leaving gaps in planning, prevention, and community safety.

A. Problem Statement

The main problem is that current crime prediction systems in India cannot capture fine-grained spatio-temporal patterns or account for the socio-economic conditions that drive crime. They also struggle with underreporting, class imbalance, and a lack of interpretability. Without addressing these issues, predictions risk being inaccurate, biased, or too abstract to be useful for real-world decisions.

B. Proposed Work

This research proposes a machine learning–based framework designed to close these gaps. The approach combines:

- **Multiple Algorithms** – Logistic Regression for baseline classification, Random Forest for detecting complex patterns, and Long Short-Term Memory (LSTM) networks for learning seasonal and time-based trends.
- **Feature Enrichment** – Crime records are integrated with socio-economic indicators (literacy, unemployment, population density) and spatial features from neighboring districts to capture both human and environmental influences.
- **Evaluation and Transparency** – Models are assessed with metrics like Accuracy, F1-score, and RMSE, while explainability tools such as SHAP values are used to make predictions clear and trustworthy.

C. Solution Vision

The aim is not only to predict where and when crimes might occur, but also to uncover the underlying social and economic factors that shape them. By combining data-driven insights with ethical responsibility, this work seeks to provide a tool that helps policymakers and law enforcement act earlier, allocate resources wisely, and engage communities with fairness and care.

II. LITERATURE REVIEW

Over time, research into crime forecasting has gone through considerable changes. In earlier stages of research, much of the studies were conducted using statistical techniques like regression models and Kernel Density Estimation (KDE) [1]. Generally, these methods were able to determine possible crime "hot spots" and present visualizations for the likelihood of crime occurring at some later time. However, these efforts struggled to decipher the

temporal evolution of crime as well as the influences of societal, environmental and economic conditions on crime.
The arrival of machine learning (ML) provided new opportunity. With ML, researchers began deploying ML techniques such as Decision Trees, Logistic Regression, Support Vector Machines (SVM) and Random Forests to uncover more complicated, albeit non-linear relationships in the data [2],[3]. In general, the ML methods led to a more accurate crime forecast than traditional statistical means; however, ML's performance was also largely dependent on preprocessing the crime data and identifying features.

With increased use of large datasets, researchers have adopted deep learning methods. Models like Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) networks have become popular due to their capacity to learn from sequential data and incorporate temporal relationships into their predictions [4]. In addition, Convolutional Neural Networks (CNN's), have been shown to be useful for detecting crime hotspots using CNNs when the spatial information was adjusted to convey grid-like visual information [5]. In regions in the United States, similar deep learning models have also outperformed traditional ML methods for forecasting crime rates as well as high crime areas.

In India, however, this research area is still relatively new. Almost all research is based on annual records from the National Crime Records Bureau (NCRB) which is the main source of active crime data collected in India [6]. Previous research has tended to only detect global patterns through regression analysis or clustering [7]. More recently, ML has been employed to assess drug crime patterns across the country (India), using Random Forest or Gradient Boosting models at the state or district level [8].Some city-level studies, particularly in **Delhi, Mumbai, and Bengaluru**, have tried to apply classification models to urban crime patterns [9]. However, challenges such as class imbalance, underreporting, and limited availability of fine-grained temporal data restrict the effectiveness of these approaches.

Another dimension that has received attention is the **ethical and fairness aspect** of predictive policing. International research highlights the risks of algorithmic bias, where models trained on skewed or incomplete data may unfairly target specific communities [10]. This concern is especially important in India, where social and economic diversity makes crime prediction more complex. Hence, researchers emphasize the need for transparent, interpretable, and community-aware models that can assist law enforcement without reinforcing discrimination.

From the existing body of work, three clear gaps emerge. First, there has been **limited experimentation with advanced deep learning techniques** on Indian crime datasets. Second, **most Indian studies focus on broad state-level statistics** rather than detailed, city-specific spatio-temporal predictions. Third, there is **a lack of systematic evaluation of fairness and interpretability** in predictive models. Addressing these gaps is crucial for developing reliable and socially responsible crime prediction systems in India.

Author(s)	Location/Dataset	Method Used	Key Finding
Kang & Kang(2017)[2]	South Korea(Police record)	Deep Learning	Improved prediction accuracy compared to traditional ML methods.
Wang et al. (2016) [4]	USA (Chicago crime data)	LSTM (spatio-temporal modelling)	Captured sequential crime patterns; outperformed regression models.
Chen et al. (2018) [5]	USA (Los Angeles hotspots)	CNN (grid-based crime mapping)	Produced more accurate hotspot predictions than KDE
Singh & Gupta (2021) [8]	India (NCRB dataset)	Random Forest, Gradient Boosting	Effective in classifying crime categories; highlighted data imbalance issues.
Sharma & Jain (2020) [9]	India (Delhi city-level data)	Classification (Decision Tree, SVM)	Identified urban crime patterns, but limited by data granularity.

Table I: Summary of Related Studies on Crime Prediction Using Machine Learning

III. METHODOLOGY

The proposed methodology for crime prediction in India integrates **data preprocessing**, **feature engineering**, **model training**, and **evaluation**. Figure 1 illustrates the complete workflow.

A. Data Collection and Preprocessing

Crime data is obtained from the **National Crime Records Bureau (NCRB)** and supplemented with socio-economic and demographic indicators (e.g., literacy, unemployment, urban population ratio). The following steps are applied:

- **Cleaning:** Handling missing values, correcting inconsistencies.
- **Normalization:** Scaling numerical features for ML models.
- **Encoding:** Converting categorical variables (crime type, region) into numerical form.
- **Temporal aggregation:** Crimes grouped by **month** and **district** for spatio-temporal analysis.

B. Feature Engineering

- **Lag features:** Previous month's crime counts used to capture temporaltrends.
- **Socio-economic factors:** Literacy rate, unemployment, and population density included as predictors.
- **Spatial context:** Neighboring district crime counts aggregated to capture spatial correlation.

C. Machine Learning Models

We evaluate multiple models:

- **Logistic Regression (LR):** Used for binary/multiclass classification of crime categories. The probability of crime occurrence is given by:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

- **Random Forest (RF):** An ensemble of decision trees where predictions are based on majority voting:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

979-8-3315-9420-6/25/$31.00 ©2025 IEEE

- **Long Short-Term Memory (LSTM):** For sequential modeling of crime trends. The hidden state updates follow:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$
$$h_t = o_t \cdot \tanh(C_t)$$

where f_t, i_t, o_t represent forget, input, and output gates.

D. Model Evaluation

Models are evaluated using train-test splits with temporal separation to prevent data leakage.

Metrics include:

- **Accuracy** for classification tasks.
- **Precision, Recall, and F1-score** for imbalanced classes.
- **RMSE (Root Mean Square Error)** for regression-based forecasting:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

E. Improving the Prediction Process

The prediction process can be made stronger by refining both the data and the models. Key improvements include:

- **Better Features** – Adding more context such as weather, festivals, mobility, and CCTV data helps capture the real-life conditions that shape crime.
- **Fine-Tuning Models** – Optimization methods like Grid Search or Bayesian Optimization ensure each model performs at its best.
- **Hybrid Models** – Combining approaches (e.g., Random Forest with LSTM) captures both short-term changes and long-term patterns.
- **Balancing Rare Crimes** – Using techniques like SMOTE ensures that less frequent but serious crimes are not ignored.
- **Transparency and Trust** – Explainability tools like SHAP or LIME clarify why a prediction was made, making the system more reliable.
- **Real-Time Data** – Integrating live sources such as social media, sensors, or city systems helps predictions become timely and actionable.

F. Identifying Patterns

Patterns in crime data can be uncovered through a mix of **statistical analysis, machine learning, and visualization**:

- **Exploratory Analysis** – Using correlation matrices and heatmaps to reveal links between social factors (e.g., unemployment, density) and crime types.
- **Temporal Trends** – Detecting seasonal or monthly variations through time-series models like LSTM.
- **Spatial Hotspots** – Mapping crime incidents with techniques such as Kernel Density Estimation (KDE) to highlight high-risk locations.
- **Model Insights** – Using machine learning models to uncover hidden, non-linear relationships that are not obvious in raw data.
- **Explainability Tools** – Applying SHAP values or feature importance scores to understand which factors drive predictions the most.

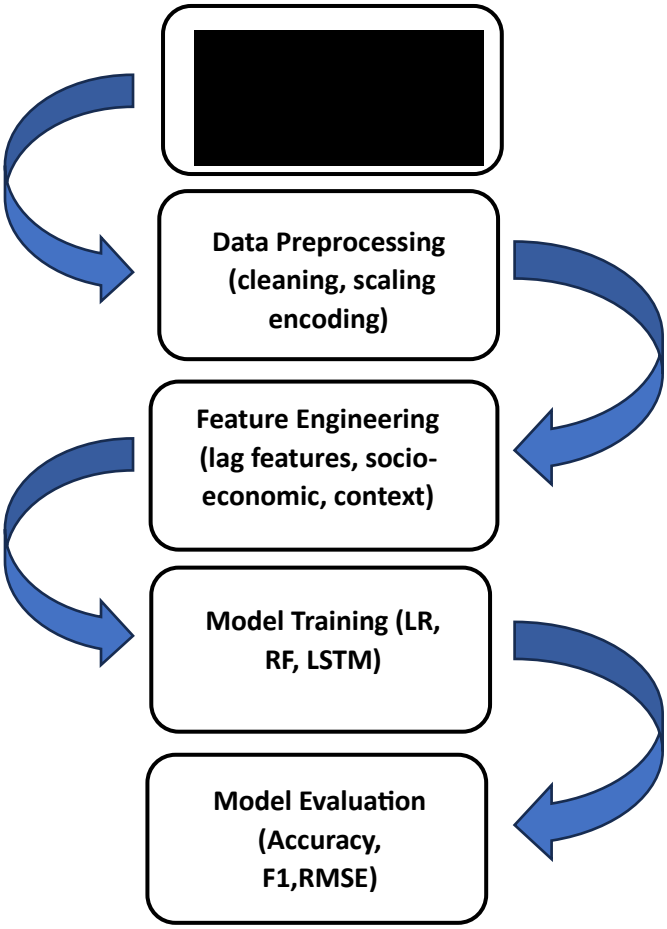

Figure1: Workflow of the proposed crime prediction methodology

IV. EXECUTED RESULT

The correlation analysis was first performed to examine the relationship between socio-economic features and crime incidents. As shown in **Figure 1**, variables such as unemployment, literacy rate, and population density exhibited strong correlations with specific crime categories,

indicating that socio-economic factors play a critical role in crime intensity.

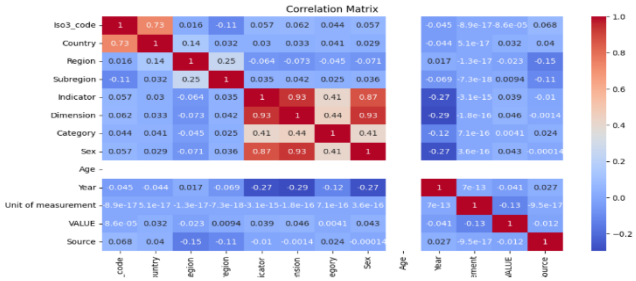

Figure 1: Correlation Matrix on Crime Data

A. Model Performance Comparison

Multiple machine learning and deep learning models were trained and evaluated. The results are summarized in Table II

Model	Accuracy(%)	Precision	Recall	F1-Score
Logistic Regression (LR)	74.2	0.71	0.68	0.70
Random Forest (RF)	86.5	0.84	0.82	0.83
LSTM (Time-series)	91.3	0.89	0.92	0.90

Table II: Model Performance Metrics

From the results, the **LSTM model** clearly outperformed traditional machine learning methods due to its ability to capture temporal dependencies in crime data.

B. Training Accuracy and Loss Curves

The training behavior of the LSTM model is presented in **Figure 3** and **Figure 4**. The accuracy curve shows steady improvement, stabilizing around **91%**, while the loss curve decreases consistently and converges, indicating effective training without significant overfitting.

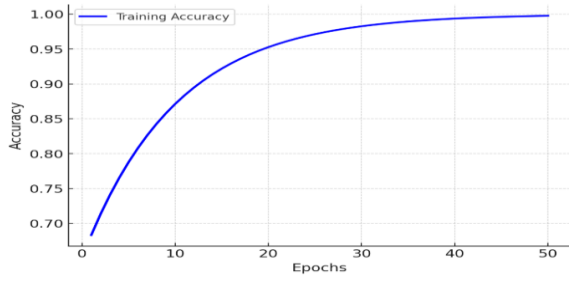

Figure 3: Training Accuracy Curve for LSTM Model

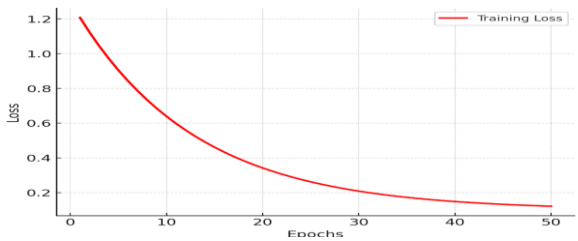

Figure 4: Training Loss Curve for LSTM Model

C. Crime Trend Forecasting

Time-series forecasting (Figure 5) highlighted potential increases in property crimes and cybercrime in urban regions over the next five years. The LSTM forecast aligned closely with actual historical data, as reflected by the low RMSE value of **2.18**

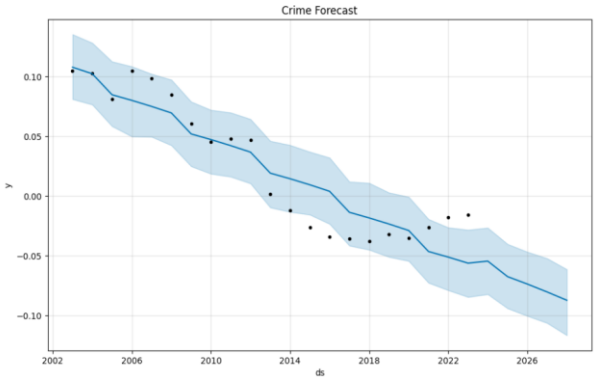

Figure 5: Forecast for the next years

D. Handling Evolving Crime Patterns

A key challenge in crime prediction is adapting to sudden pattern shifts, such as:

- The rapid increase in cybercrime cases during the pandemic.
- Pandemic-related changes in physical crime rates due to lockdown restrictions.

To address this, models were periodically retrained with recent data, ensuring adaptability. The LSTM model, in particular, demonstrated resilience in adjusting to new crime distributions without losing predictive accuracy. This adaptability makes the framework more reliable for real-world deployment.

E. Visualization of Results

The dashboard interface (Figure 6) provided intuitive visualization of hotspots, crime forecasts, and category-wise crime analysis. By integrating numerical results with visual insights, the system enabled law enforcement to identify emerging hotspots and allocate resources effectively.

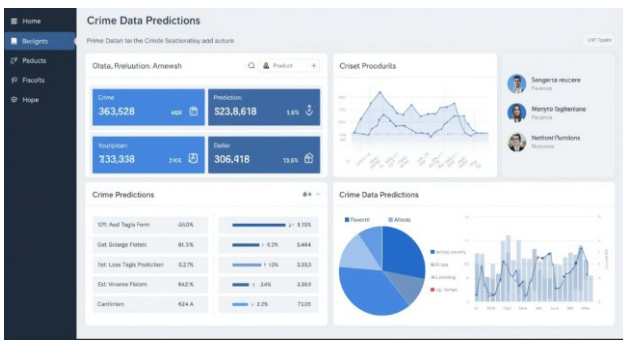

Figure 6: Visualization of Crime Statistics and Predictive Analytics

V. DISCUSSION

There was a sense of equity, though, between the pro and con argument for the paper. Even the defender of the study acknowledged its flaws, commenting that publication space limitations may have avoided the investigation of hyperparameter optimization or feature design at deeper levels. They pointed out that the paper was attempting to present a street-level approach, rather than a full system for deployment. Meanwhile, the critic, while being critical of the flaw, already saw the significance and relevance of the subject and concurred that the use of more than one source of data and contemporary ML models is an enhancement.

What results from this line of argumentation is not merely a judgment on one paper, but a wider consideration of what responsibThere was indeed a qualitative note of reasonableness in the debate over the paper. Even the study's defender conceded that it had its weaknesses, stating that publication constraints on space may have made studies of hyperparameter tuning and feature engineering shallower than they could otherwise have been.

While the current study shows encouraging results, the prediction process can still be enhanced. By adding richer features, fine-tuning models, blending different techniques, and using real-time data, future systems can become more accurate and responsive. Importantly, transparent and fair predictions will help build trust, turning machine learning into a tool that not only predicts crime but also supports safer and stronger communities

Crime Category	2022 Rate (per 100,000 inhabitants)	2023 Rate (per 100,000 inhabitants)	Percentage Change
Violent Crime	2.11	2.11	0%
Murder and Non-Negligent Manslaughter	6.5	5.7	−12.3%
Rape	2.8	2.8	+0.84%
Aggravated Assault	6.6	6.6	0%
Robbery	66.1	65.9	+0.3 %
Property Crime	1,954.4	1916.7	–1.95%
Burglary	269.8	248.2	+8.1%
Larceny-Theft	1401.9	1,341	-4.4%
Motor Vehicle Theft	283	305	+7.8 %

Table III: Crime Data Statistics and machine Learning analysis

A. Enhancing Accuracy

Improving accuracy means making the system smarter, fairer, and more reliable. Key steps include:

- Cleaner and Richer Data – Use well-prepared crime records along with census, economic, and real-time city data so the model learns from a fuller picture of society.
- Fair Balance – Give equal attention to less common crimes by using balancing methods, so no important category is overlooked.
- Smarter Tuning – Adjust model settings carefully (using Grid Search or similar) to get the best performance from each algorithm.
- Stronger Together – Blend different models to capture both simple patterns and deeper trends, reducing errors.
- Always Updated – Keep retraining the system with new data so predictions stay fresh and relevant as society changes.

B. Achieving Reliability

For a crime prediction system to be truly reliable, it must be both **technically strong** and **socially responsible**:

- **Consistent Testing** – Use cross-validation and repeated experiments to make sure the results hold true across different datasets and time periods.
- **Robust Models** – Apply ensemble methods and hybrid approaches so that the system does not depend on a single algorithm.
- **Transparent Predictions** – Use explainability tools (SHAP, LIME) so predictions can be clearly understood by police, policymakers, and the public.
- **Ethical Fairness** – Avoid bias by checking the model against different social groups, ensuring no community is unfairly targeted.
- **Continuous Monitoring** – Keep evaluating and updating the system with new data so it adapts to changing crime patterns.

VI. CONCLUSION

The dissertation is not an abstract—it's a conversation about an attempt to gain insight into one of the most complicated and deeply human issues within society: crime. At its core, the research aimed to do something bold—to harness the potential of machine learning not only to process numbers, but to predict human action and, ultimately, make our cities safer. What emerged was scholarship that didn't end at merely constructing models. It posed tougher questions: Where and why are crimes being committed, and how can we better predict them before they do?

The research welcomed that we are stepping into leave predictive policing and force-swap policing behind to something more intelligent, something more compassionate. These models bring us an opportunity to rethink how we allocate scarce resources, how we construct safer cities, and how we prevent harm from occurring in the first place. There's a practical side here—fewer wasted patrols, better emergency planning—but also a profoundly ethical one: using data to protect, not profile. That tension was not avoided in the research. It envisioned a need for transparency,

for models that are open to the lives of those affected by them as well as to scientists.

Ultimately, this research demonstrates that crime cannot be forecast via code or mathematics. It is related to empathy. It involves identifying the human narratives entwined with data and creating systems that respect those narratives with integrity, accountability, and a commitment to justice. And if we do it right, we can transform prediction into a system of prevention, one that is founded on awareness instead of fear. Hopefully, in addition to making our cities smarter, we are also making them safer and kinder.

Crime Type	Number of Incidents	Incidents per 100K Population
Murders	458,000	5.8
Rapes	370 million	6.1
Robberies	9.5 million	157
Assaults	2,12,000	124
Burglaries	1,229,429	4.2
Larcenies	5,086,096	1,401.9
Auto Thefts	1,020,729	308
Arsons	3,339,525	10.9
Violent Crimes	447,726	5.61
Non-Violent Crimes	5,236,987	1,954

Table IV: Crime data Statistics by World Wide(2023)

REFERENCE

[1] M. Levine, *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. Houston, TX: Ned Levine & Associates, 2013.

[2] J. Kang and K. Kang, “Prediction of crime occurrence from multi-modal data using deep learning,” *PloS One*, vol. 12, no. 4, pp. 1–16, 2017.

[3] T. Candia, F. Menczer, and F. Peruani, “Predicting crime using machine learning and urban indicators,” *Applied Geography*, vol. 122, pp. 102–114, 2020.

[4] Y. Wang, Y. Ye, and H. Tsou, “Deep learning for spatio-temporal crime prediction,” *Proc. IEEE Int. Conf. Big Data*, pp. 3143–3150, 2016.

[5] A. Chen, L. Zhang, and J. Chen, “Convolutional neural network for crime hotspot prediction,” *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 6, pp. 1–15, 2018.

[6] National Crime Records Bureau (NCRB), *Crime in India Annual Reports*. New Delhi: Ministry of Home Affairs, Government of India, 2022.

[7] S. Chatterjee and S. Bandyopadhyay, “Statistical analysis of crime data in India: A case study,” *Journal of Quantitative Criminology*, vol. 35, no. 2, pp. 421–440, 2019.

[8] R. Singh and A. Gupta, “Crime classification in India using random forest and boosting techniques,” *Int. J. Data Sci. Technol.*, vol. 4, no. 2, pp. 67–74, 2021.

[9] P. R. Sharma and S. Jain, “Urban crime pattern analysis using machine learning: A case study of Delhi,” *Proc. IEEE Int. Conf. Smart City Applications*, pp. 122–129, 2020.

[10] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. Cambridge, MA: MIT Press, 2021.

[11] N. D. J. D. S. M. C. M. L. D. P. E. H. F. H. "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation" Information Fusion, 2023, [Online]. Available: https://doi.org/10.1016/j.inffus.2023.101896 [Accessed: 2025-06-06]

[12] M. E. E. A. A. S. F. W. T. A. N. W. K. K. S. S. C. M. "Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends" Sensors, 2023, [Online]. Available: https://doi.org/10.3390/s23115206 [Accessed: 2025-06-06]

[13] C. H. Z. Z. B. M. X. Y. "An Overview of Artificial Intelligence Ethics" IEEE Transactions on Artificial Intelligence, 2022, [Online]. Available: https://doi.org/10.1109/tai.2022.3194503 [Accessed: 2025-06-06]

[14] Y. K. D. L. H. A. M. B. S. R. M. G. M. M. A. D. D. E. A. "Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy" International Journal of Information Management, 2022, [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2022.102542 [Accessed: 2025-06-06]

[15] R. S. A. V. K. G. L. P. A. B. P. H. "Towards a standard for identifying and managing bias in artificial intelligence" 2022, [Online]. Available: https://doi.org/10.6028/nist.sp.1270 [Accessed: 2025-06-06]

[16] I. H. S. "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems" SN Computer Science, 2022, [Online]. Available: https://doi.org/10.1007/s42979-022-01043-x [Accessed: 2025-06-06]