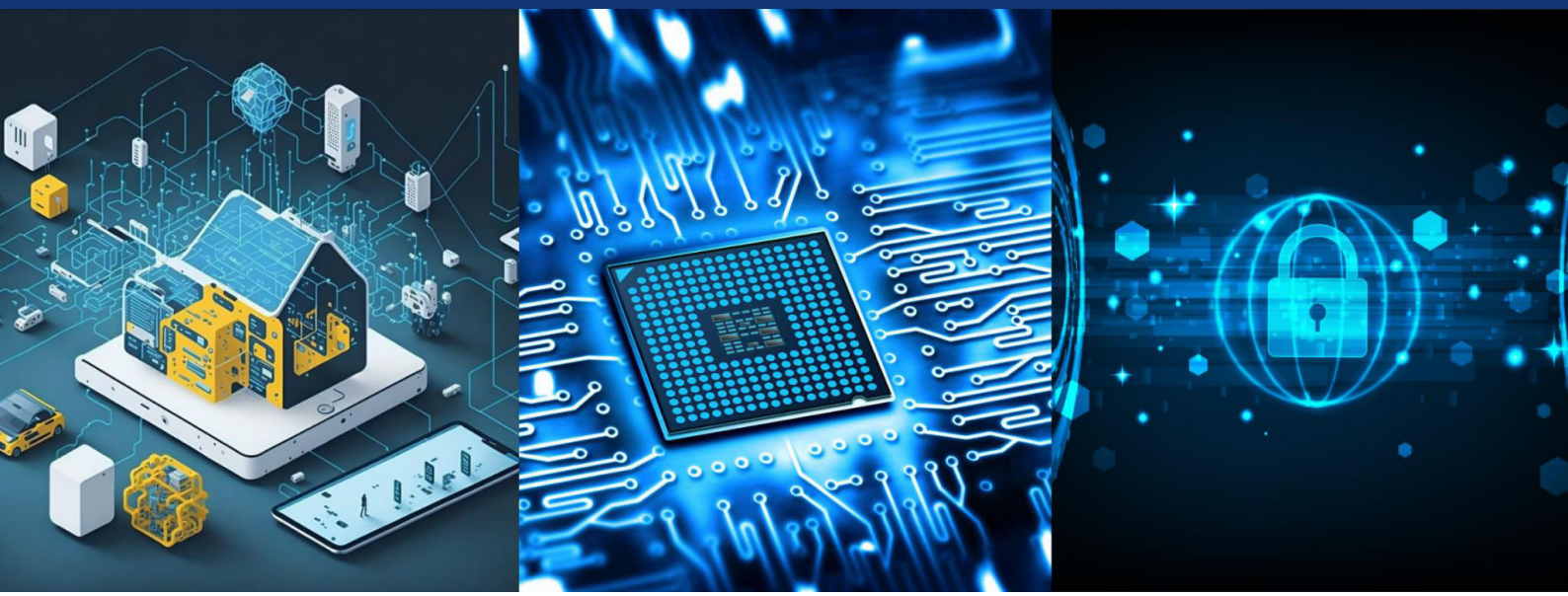
IJIRCCE

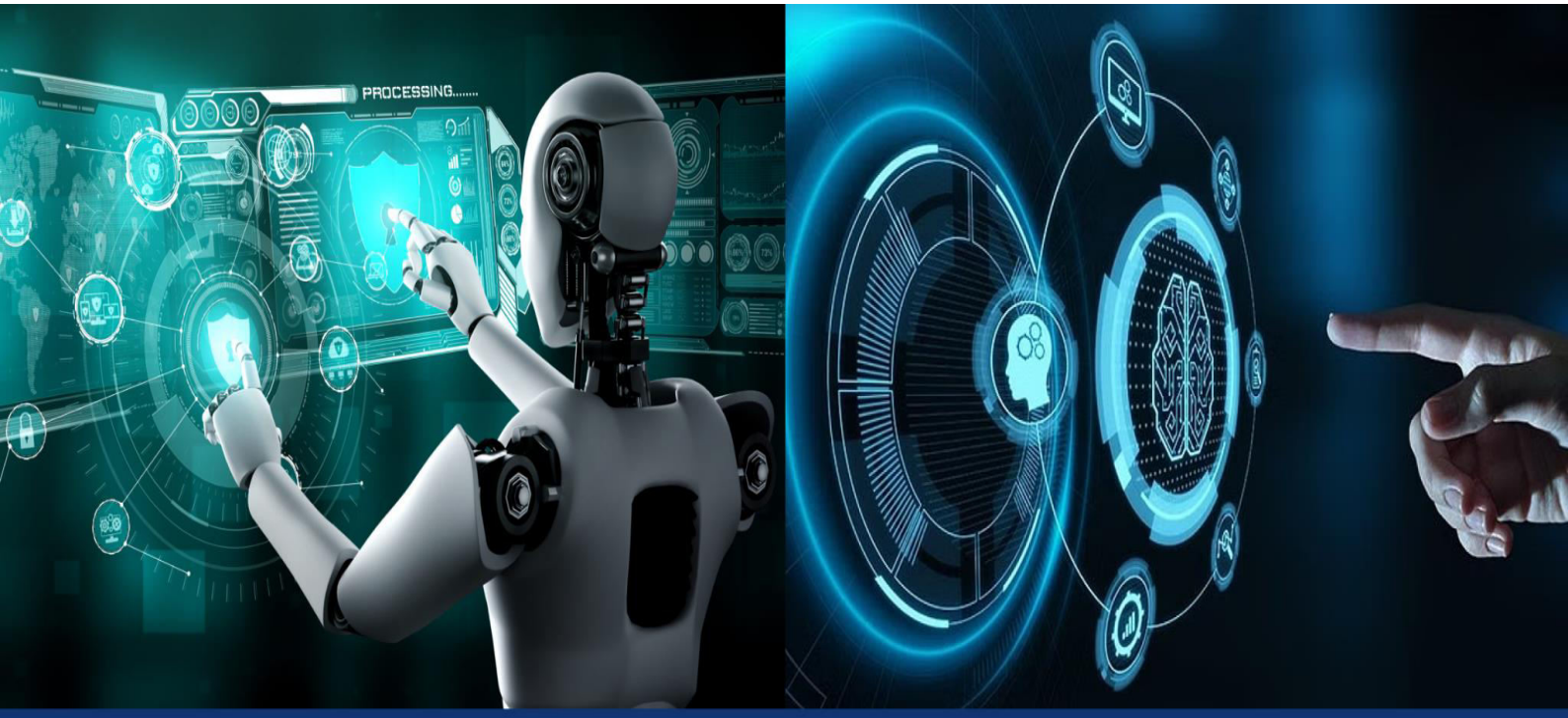
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 7, July 2026

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407055

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Secure Blockchain–NLP Integrated System for Privacy Preserving Crime Reporting and Evidence Management

Dr. R. Sridevi, Sindam Akshaya

Professor, Department of Computer Science Engineering, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India

Post-Graduate Student, Department of Computer Science Engineering, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India

ABSTRACT: The increasing prevalence of cyber-enabled and conventional crimes has created a growing demand for secure, intelligent, and privacy-preserving digital crime reporting systems. Existing centralized complaint management platforms often suffer from limitations such as data tampering, inadequate evidence protection, lack of transparency, delayed investigations, and privacy concerns that discourage victims from reporting incidents. To address these challenges, this paper proposes a Blockchain and Natural Language Processing (NLP) Integrated Crime Reporting and Evidence Management System that enables anonymous crime reporting while ensuring secure evidence handling and transparent investigation workflows. The proposed framework employs NLP techniques for complaint preprocessing and Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction. Four supervised machine learning algorithms—Multinomial Naive Bayes, Logistic Regression, Linear Support Vector Machine (Linear SVM), and Random Forest—are evaluated to automatically predict Crime Type, Indian Penal Code (IPC) Sections, Bharatiya Nyaya Sanhita (BNS) Sections, and Risk Level from complaint descriptions. Experimental analysis identifies Multinomial Naive Bayes as the most effective classifier in terms of overall predictive performance. Additionally, Cosine Similarity is utilized to retrieve semantically similar historical complaints, assisting investigators in case analysis and decision-making. To preserve digital evidence integrity, uploaded files are protected using the SHA-256 cryptographic hashing algorithm, while blockchain smart contracts maintain immutable records of complaint metadata, evidence hashes, investigation updates, and audit trails. The proposed system further incorporates role-based access control and anonymous communication between complainants and investigating officers, enhancing privacy, accountability, and operational efficiency. By integrating machine learning, NLP, cryptographic hashing, and blockchain technology into a unified framework, the proposed solution provides a secure, transparent, and scalable platform for modern digital crime reporting and forensic evidence management.

KEYWORDS: Crime Reporting, Blockchain, Natural Language Processing, Machine Learning, TF-IDF, Multinomial Naive Bayes, SHA-256, Smart Contracts, Digital Evidence Management, Cosine Similarity, Cyber Forensics, Privacy Preservation.

I. INTRODUCTION

The rapid growth of information technology and digital communication has significantly transformed the methods through which crimes are committed and reported. Along with the increase in cyber-enabled crimes and conventional criminal activities, there is a growing need for intelligent digital platforms capable of managing complaints, preserving evidence, and supporting law enforcement agencies throughout the investigation process. Existing crime reporting systems generally depend on manual documentation and centralized databases, making them susceptible to unauthorized modifications, limited transparency, delayed investigations, and inefficient evidence management. In addition, victims often hesitate to report crimes because of concerns regarding privacy and identity disclosure.

Recent advancements in Artificial Intelligence (AI), Natural Language Processing (NLP), blockchain technology, and cryptographic techniques provide new opportunities to overcome these challenges. NLP enables automatic understanding of textual complaint descriptions, reducing manual analysis while assisting investigators in identifying crime categories

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

and applicable legal provisions. Blockchain technology provides decentralized, immutable, and transparent storage of complaint information, preventing unauthorized modifications and ensuring accountability throughout the investigation process.

The proposed system integrates these technologies into a unified web-based platform that enables anonymous crime reporting, intelligent complaint analysis, secure digital evidence management, and transparent case tracking. Complaint descriptions undergo text preprocessing followed by TF-IDF feature extraction. The generated feature vectors are used to train and evaluate four supervised machine learning algorithms: Multinomial Naive Bayes, Logistic Regression, Linear Support Vector Machine (Linear SVM), and Random Forest. The best-performing model is selected to predict Crime Type, IPC Sections, BNS Sections, and Risk Level. Additionally, Cosine Similarity is employed to retrieve similar historical complaints that may assist investigators during case analysis.

To ensure evidence authenticity, every uploaded digital evidence file is processed using the SHA-256 cryptographic hashing algorithm before storage. Rather than storing large evidence files directly on the blockchain, the system records their hash values together with complaint metadata and investigation updates using blockchain smart contracts. This approach ensures tamper-resistant storage while reducing blockchain overhead. The platform further provides role-based access for reporters, administrators, and investigating officers, enabling secure communication, transparent workflow management, and efficient digital evidence handling.

By integrating Natural Language Processing, machine learning, blockchain technology, and cryptographic verification into a single framework, the proposed system improves complaint classification accuracy, protects evidence integrity, preserves reporter anonymity, and enhances transparency throughout the digital investigation lifecycle

II. RELATED WORK

The increasing incidence of cyber-enabled and conventional crimes has highlighted the need for secure, privacy-preserving, and intelligent crime reporting systems. Traditional crime reporting platforms rely on centralized databases, making them susceptible to unauthorized modifications, privacy breaches, and inefficient evidence management. Recent research has explored the integration of blockchain technology, Natural Language Processing (NLP), machine learning, and cryptographic techniques to address these challenges.

Chaurasia et al. proposed an anonymous crime reporting system based on blockchain and smart contracts that enables users to report crimes without revealing their identities. Smart contracts automate complaint registration and maintain immutable records, thereby improving transparency and trust. However, the system mainly focuses on secure complaint registration and does not provide intelligent complaint analysis, legal section prediction, or similar case retrieval [1].

Ashar et al. introduced KhabriChain, a blockchain-based anonymous crime reporting platform integrating privacy-preserving NLP techniques. The framework employs DistilBERT and Sentence-BERT (SBERT) to classify complaints and identify semantically similar reports while ensuring user anonymity through blockchain technology. Although the system achieves high classification accuracy, transformer-based models require considerable computational resources, limiting deployment on resource-constrained systems [2]. Banaeian Far and Rajabzadeh Asaar proposed a blockchain-based anonymous reporting architecture without relying on a central authority. Their protocol utilizes virtual blockchains, ring signatures, and secret-sharing mechanisms to enhance anonymity, fault tolerance, and system reliability. While the approach provides strong privacy guarantees, it does not include automated complaint classification, evidence verification, or investigation workflow management [3].

Zhu et al. presented a secure anonymous reporting mechanism for public blockchain networks using ring signatures, covert communication, and smart contracts. Their framework strengthens reporter anonymity and protects against identity disclosure during blockchain transactions. However, the work primarily focuses on communication security and does not integrate Natural Language Processing or machine learning for automated complaint analysis [4]. Bird et al. described fundamental Natural Language Processing techniques, including tokenization, stop-word removal, and feature extraction, which form the basis of many text classification systems [5]. Devlin et al. introduced BERT, a transformer-based language model capable of learning contextual representations for various NLP tasks [6]. Subsequently, Sanh et al. proposed DistilBERT, which significantly reduces computational complexity while maintaining most of BERT's performance [7]. These transformer-based approaches provide excellent classification accuracy but require substantial

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407055

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

computational resources. Consequently, traditional TF-IDF-based machine learning models remain attractive for lightweight real-time crime reporting applications due to their lower computational cost and competitive performance [8]. it is evident that existing studies mainly focus on individual aspects such as anonymous reporting, blockchain security, privacy preservation, or NLP-based complaint classification. Very few systems integrate anonymous crime reporting, automated prediction of Crime Type, IPC Sections, BNS Sections, and Risk Level, similar case retrieval, SHA-256-based evidence verification, blockchain-backed immutable audit trails, and role-based investigation workflows within a unified framework. The proposed system addresses these limitations by integrating TF-IDF-based Natural Language Processing, supervised machine learning, Cosine Similarity, SHA-256 cryptographic hashing [9], blockchain technology, and smart contracts [10] into a comprehensive platform for secure, transparent, and intelligent crime reporting and evidence management.

III. PROPOSED ALGORITHM

The proposed Blockchain and NLP Integrated Crime Reporting and Evidence Management System integrates Natural Language Processing (NLP), Machine Learning (ML), SHA-256 cryptographic hashing, and blockchain technology to provide a secure, transparent, and intelligent crime reporting platform. The methodology consists of three major phases: complaint analysis and prediction, secure evidence management, and blockchain-enabled investigation workflow.

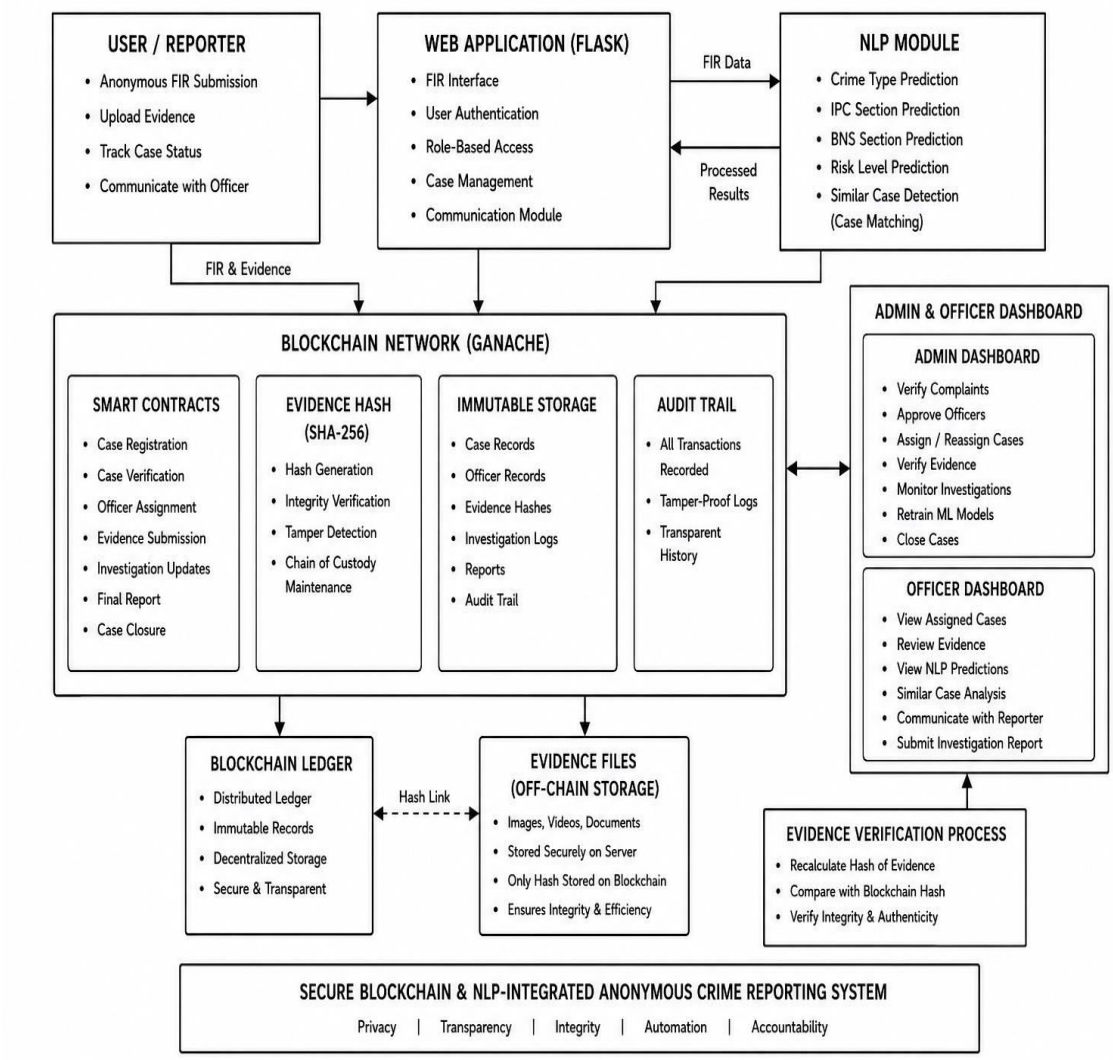


Fig. 1. System Architecture

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. *Complaint Analysis and Prediction*

Initially, a crime complaint is submitted through the reporting interface. The complaint text is preprocessed using NLP techniques including tokenization, stop-word removal, punctuation elimination, and lowercase conversion. The processed text is transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) method.

The Term Frequency (TF) of a term t in a document d is computed as

$$TF(t,d) = f(t,d) / \Sigma f(k,d)$$

where f(t,d) is the number of occurrences of term t in document d.

The Inverse Document Frequency (IDF) is calculated as

$$IDF(t) = \log(N / df(t))$$

where N is the total number of complaint documents and df(t) is the number of documents containing the term t.

The TF-IDF weight is obtained as

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

The extracted feature vectors are supplied to supervised machine learning classifiers. Among the evaluated algorithms, Multinomial Naive Bayes achieved the best overall performance for predicting Crime Type, Indian Penal Code (IPC) Sections, Bharatiya Nyaya Sanhita (BNS) Sections, and Risk Level.

The Multinomial Naive Bayes classifier estimates the posterior probability as

$$P(c|d) = P(c) \times \prod P(w_i|c) / P(d)$$

where c denotes the predicted class, w_i represents the words in the complaint, and P(c) is the prior probability of class c.

To retrieve similar historical complaints, Cosine Similarity is computed as

$$Similarity(A,B) = (A \cdot B) / (|A| \times |B|)$$

where A and B represent the TF-IDF vectors of two complaint documents.

B. *Secure Evidence Management*

Digital evidence uploaded by users is protected using the SHA-256 cryptographic hash function. For every uploaded evidence file E, the hash value is computed as

$$H = SHA\text{-}256(E)$$

where H is the unique 256-bit hash corresponding to the evidence file.

Evidence integrity is verified by comparing the stored and newly generated hash values.

If

$$Hstored = Hcomputed$$

the evidence is considered Valid; otherwise,

$$Hstored \neq Hcomputed$$

indicates that the evidence has been Tampered.

Only the SHA-256 hash and evidence metadata are stored on the blockchain, while the original evidence files remain in secure application storage.

C. *Blockchain-Based Investigation Workflow*

Following complaint registration, the administrator verifies the complaint and assigns it to an investigating officer. The officer reviews complaint details, verifies digital evidence, analyzes machine learning predictions, retrieves similar historical cases, and updates investigation progress.

Each blockchain block is represented as

$$B_i = (Index, Timestamp, Previous\ Hash, Data, Hash)$$

where Data contains complaint metadata, evidence hash values, investigation updates, and audit records.

The hash of the current block is computed as

$$Hash_i = SHA\text{-}256(Index \,||\, Timestamp \,||\, Previous\ Hash \,||\, Data)$$

where || denotes concatenation of all block fields before applying the SHA-256 algorithm.

Since each block stores the hash of its previous block, any modification to a block changes its hash value and breaks the blockchain structure, thereby ensuring immutability, transparency, and tamper resistance throughout the investigation lifecycle.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407055

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. RESULTS

The proposed Blockchain and NLP Integrated Crime Reporting and Evidence Management System was evaluated using four supervised machine learning algorithms: Multinomial Naive Bayes (MNB), Logistic Regression (LR), Linear Support Vector Machine (Linear SVM), and Random Forest (RF). The models were trained using TF-IDF feature vectors extracted from the complaint dataset and evaluated on four prediction tasks: Crime Type, IPC Sections, BNS Sections, and Risk Level.

The evaluation metrics include Accuracy, Macro Precision, Macro Recall, Macro F1-score, and Training Time. Among the evaluated models, Multinomial Naive Bayes demonstrated the best balance between predictive performance and computational efficiency and was therefore selected for deployment in the proposed system.

A. Crime Type Prediction

Figure 2 illustrates the performance comparison of different machine learning classifiers for crime type prediction. Multinomial Naive Bayes, Logistic Regression, and Linear SVM achieved an identical classification accuracy of 93.35%, whereas Random Forest obtained 91.90%. Similarly, the macro precision, recall, and F1-score values remained consistent among the first three models. Although all three algorithms achieved comparable predictive performance, Multinomial Naive Bayes required only 0.392 seconds for training, making it the most computationally efficient classifier.

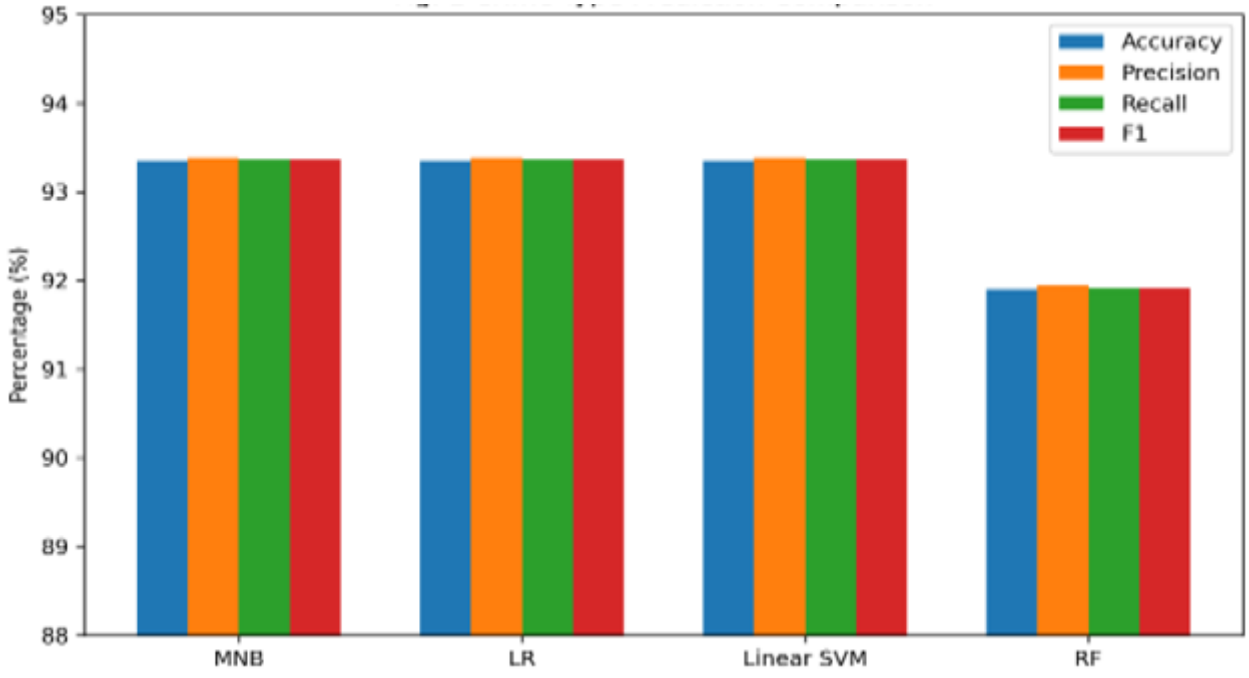


Fig. 2. Performance comparison of machine learning models for Crime Type prediction.

B. IPC Section Prediction

Figure 3 presents the comparative performance of the evaluated classifiers for IPC Section prediction. Multinomial Naive Bayes, Logistic Regression, and Linear SVM each achieved an accuracy of 91.90%, significantly outperforming Random Forest, which achieved 90.50%. Among all evaluated classifiers, Multinomial Naive Bayes exhibited the shortest training time (0.295 seconds) while maintaining competitive precision, recall, and F1-score, making it suitable for real-time deployment.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407055

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

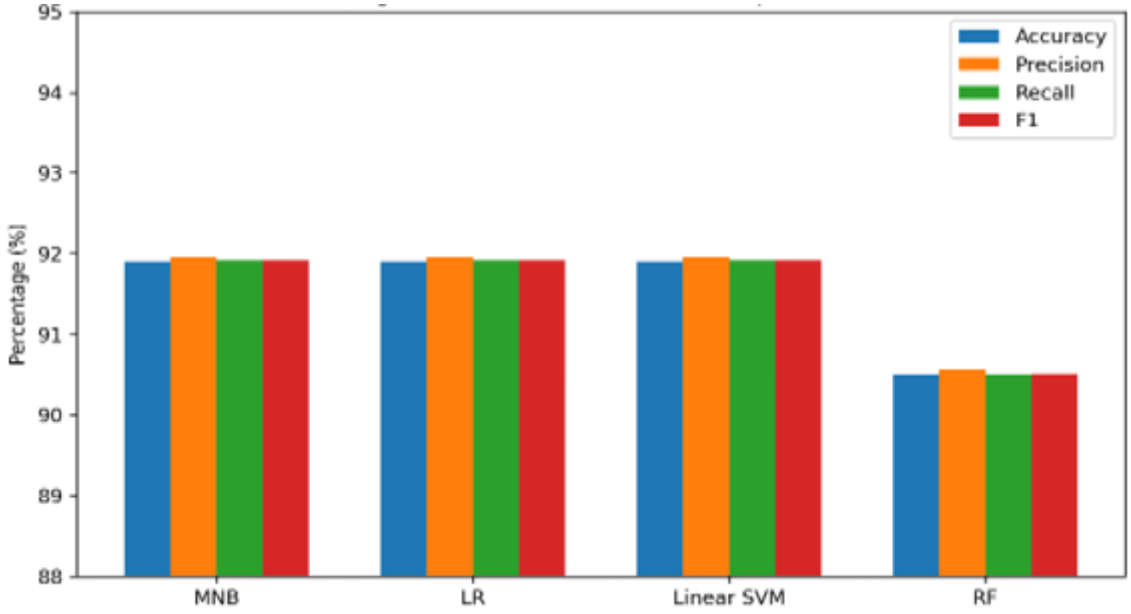


Fig. 3. Performance comparison of machine learning models for IPC Section prediction.

C. BNS Section Prediction

Figure 4 illustrates the experimental results obtained for BNS Section prediction. Multinomial Naive Bayes, Logistic Regression, and Linear SVM achieved an identical accuracy of 91.65%, while Random Forest obtained 90.50%. The proposed system selected Multinomial Naive Bayes because it required only 0.262 seconds for model training, the lowest computational time among all evaluated classifiers without compromising prediction performance.

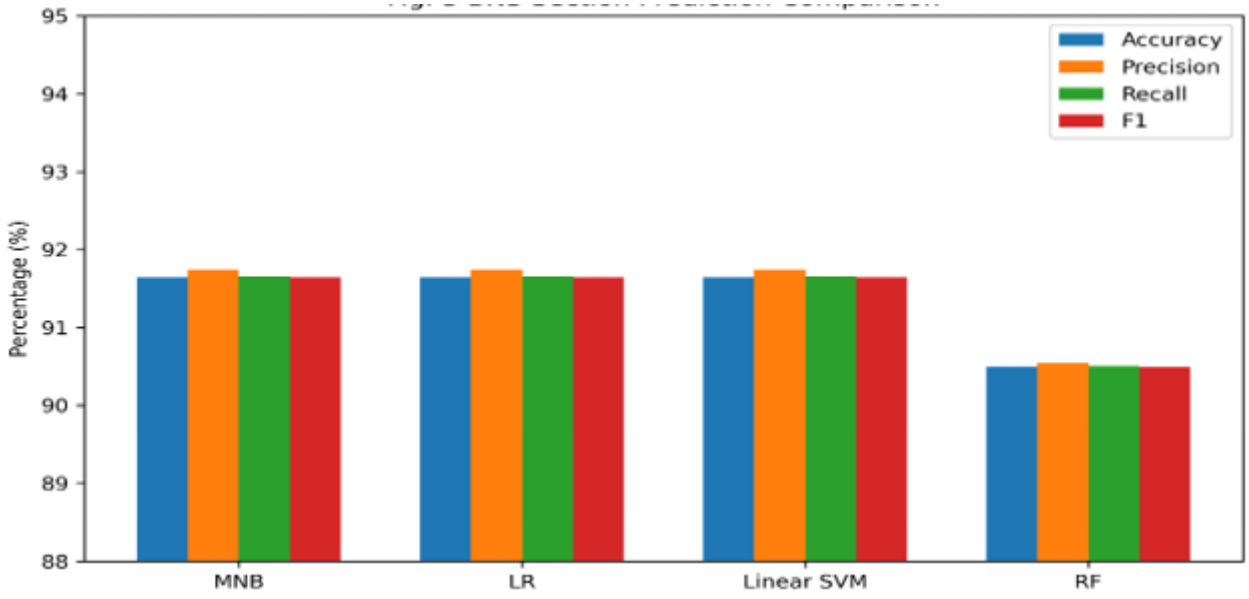


Fig. 4. Performance comparison of machine learning models for BNS Section prediction.

D. Risk Level Prediction

Figure 5 presents the classification performance for Risk Level prediction. Multinomial Naive Bayes, Logistic Regression, and Linear SVM each achieved an accuracy of 92.30%, whereas Random Forest achieved 90.70%. Although

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407055

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Logistic Regression exhibited a comparable training time, Multinomial Naive Bayes consistently provided the best trade-off between computational efficiency and predictive performance across all evaluation metrics.

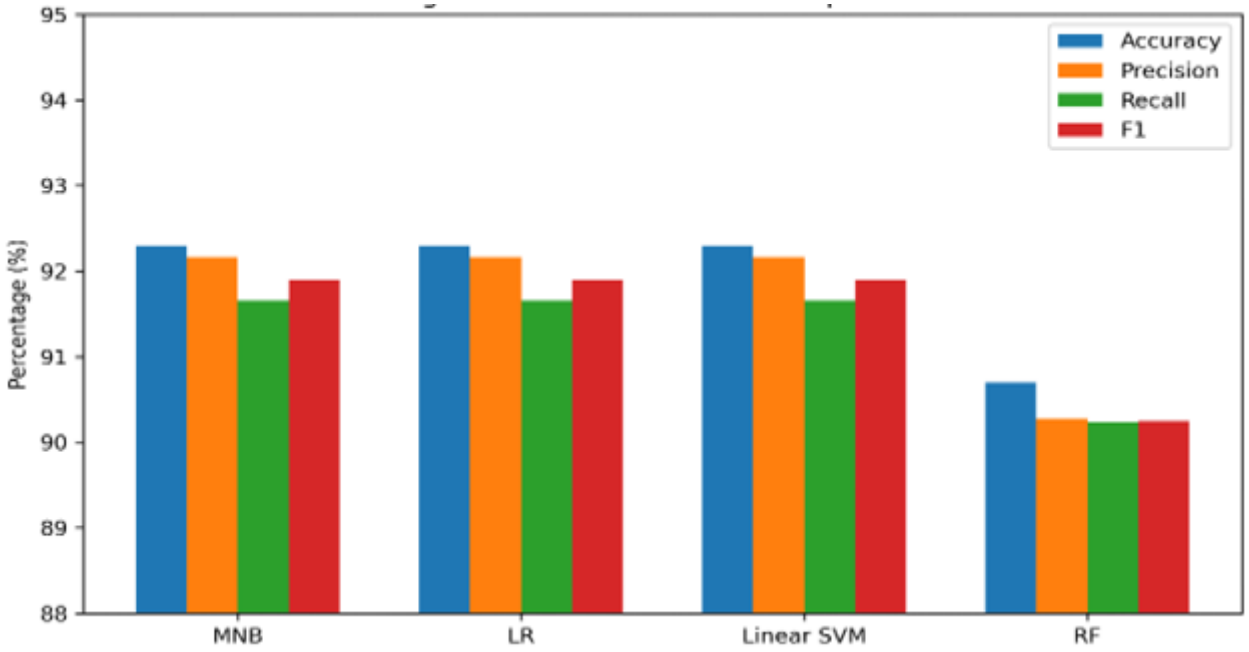


Fig. 5. Performance comparison of machine learning models for Risk Level prediction.

The graphical comparisons demonstrate that Multinomial Naive Bayes consistently achieved high classification performance across all four prediction tasks while requiring substantially lower computational resources than Logistic Regression, Linear SVM, and Random Forest. Although Logistic Regression and Linear SVM produced comparable accuracy values, their training times were relatively higher. Random Forest recorded the lowest performance in terms of accuracy, macro precision, macro recall, and macro F1-score, while also requiring the longest training time. Therefore, Multinomial Naive Bayes was selected as the final prediction model for the proposed blockchain-enabled crime reporting and evidence management system because it offers an optimal balance between prediction accuracy, computational efficiency, and real-time applicability.

V. CONCLUSION AND FUTURE WORK

The proposed Blockchain and NLP Integrated Crime Reporting and Evidence Management System provides a secure, transparent, and intelligent solution for modern crime reporting and digital evidence management. By integrating Natural Language Processing (NLP), Machine Learning (ML), SHA-256 cryptographic hashing, and blockchain technology, the system addresses key challenges associated with traditional crime reporting platforms, including data tampering, privacy concerns, inefficient complaint processing, and lack of evidence integrity. The use of TF-IDF for feature extraction and Multinomial Naive Bayes for complaint classification enables accurate prediction of Crime Type, IPC Sections, BNS Sections, and Risk Level, while Cosine Similarity assists investigators by retrieving semantically similar historical cases. Furthermore, the implementation of SHA-256 ensures the integrity of uploaded digital evidence, whereas blockchain-based smart contracts provide immutable storage of complaint metadata, evidence hashes, and investigation records. Experimental evaluation demonstrates that the proposed framework achieves high prediction accuracy with low computational overhead, making it suitable for real-time deployment. Overall, the integration of artificial intelligence and blockchain technology significantly enhances transparency, accountability, privacy preservation, and operational efficiency in crime reporting and forensic evidence management.

Future work will focus on extending the proposed framework by incorporating deep learning-based language models such as BERT and LegalBERT to improve complaint understanding and legal section prediction. The system can also be

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

enhanced by integrating multilingual complaint processing to support users from diverse linguistic backgrounds. Future versions may include speech-to-text complaint registration, image and video forensic analysis, and real-time threat detection using computer vision techniques. Additionally, deploying the framework on a permissioned blockchain network with advanced consensus mechanisms can improve scalability, transaction throughput, and energy efficiency. Integration with national law enforcement databases and judicial information systems can further strengthen inter-agency collaboration and facilitate faster, more reliable criminal investigations. These enhancements will enable the proposed system to evolve into a comprehensive, scalable, and intelligent digital crime reporting ecosystem for smart law enforcement applications.

REFERENCES

1. P. Chaurasia, D. Rana, R. V., and S. Srividhya, "Anonymous Crime Reporting using Blockchain and Smart Contract," 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN), IEEE, 2024.
2. D. Ashar, M. Dwivedi, P. Patil, R. Sawant, et al., "A Blockchain-Based Anonymous Crime Reporting Platform Integrating Privacy-Preserving NLP for Enhanced Public Safety," 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 2025.
3. S. Banaeian Far and M. Rajabzadeh Asaar, "A Blockchain-Based Anonymous Reporting System with No Central Authority: Architecture and Protocol," Cyber Security and Applications, vol. 2, 2024.
4. L. Zhu, J. Zhang, C. Zhang, F. Gao, Z. Chen, and Z. Li, "Achieving Anonymous and Covert Reporting on Public Blockchain Networks," Mathematics, vol. 11, no. 7, Art. 1621, 2023.
5. S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.
6. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019.
7. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," NeurIPS Workshop, 2019.
8. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
9. National Institute of Standards and Technology (NIST), Secure Hash Standard (SHA-256), FIPS PUB 180-4, 2015.
10. G. Wood, "Ethereum: A Secure Decentralised Generalised Transaction Ledger," Ethereum Yellow Paper, 2014.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details