A COMPREHENSIVE REVIEW OF ML AND DL-BASED CRIME PREDICTION FROM SOCIAL MEDIA DATA

Mr. Dandu Jaya Bharath Reddy
Sambhram University Jizzax bharath55.edu@gmail.com , +998 971278050

Annotation.

The rise in criminal activities has made crime prediction a crucial task for law enforcement, aimed at preventing future incidents by identifying potential threats. However, accurately predicting and classifying crimes remains challenging due to the evolving and adaptable nature of criminal behaviour. Social media, a platform where individuals frequently express their thoughts and emotions, offers valuable data for crime detection, but sentiment analysis of this data is

hindered by challenges such as linguistic diversity and imbalanced datasets. Recent advancements in Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL), have provided powerful tools for early crime prediction by identifying potential crime hotspots and forecasting future criminal activities. This paper provides a comprehensive review of crime prediction techniques based on social media data, focusing on various ML and DL frameworks. It reviews key models developed by researchers, evaluates their strengths and limitations, and presents a comparative analysis. Additionally, the paper proposes alternative approaches to improve the accuracy and efficiency of crime prediction using social media, aiming to enhance decision-making for law enforcement and contribute to public safety.

Keywords: Crime prediction, Machine Learning, Deep learning, Social media, Decision Making.

1. Introduction

Crime is a widespread issue, with incidents occurring daily, contributing to public anxiety. Predicting the timing and location of crimes is complex, as factors like poverty and unemployment significantly influence crime rates, making crime neither random nor evenly distributed. Traditional crime prediction approaches, such as demographic data, hot-spot mapping, and socio-economic indices, are limited, especially for crimes like taxi robberies that involve wide geographical areas. These methods often rely heavily on past crime statistics and overlook socio-behavioral neighborhood data.

With the rise of social media platforms like Instagram, Facebook, and Twitter, there is a new opportunity to use user-generated content for crime prediction. Sentiment analysis, a subfield of Natural Language Processing (NLP), has emerged as a tool for analyzing emotions and opinions expressed on social media to identify public attitudes toward crime and safety. This method has shown promise, allowing law enforcement to gain deeper insights into crime trends and allocate resources more effectively. However, sentiment analysis faces challenges like language diversity and imbalanced datasets, which make extracting reliable insights from large-scale data difficult.

Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL), plays a crucial role in crime prediction. ML models like Decision Trees, Random Forests, and Support Vector Machines have been used to identify crime hotspots and predict future incidents. However, they often struggle with large and complex datasets. DL models, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks, are more advanced, capable of analyzing real-time crime patterns and enhancing law enforcement efforts through surveillance technologies. Despite their potential, both ML and DL models face challenges related to accuracy, scalability, and incorporating socio-behavioral data.

This paper reviews the latest advancements in ML and DL-based crime prediction models, focusing on the integration of social media data to identify crime hotspots and predict future criminal activity. A comparative analysis of these models is presented, highlighting their strengths, limitations, and potential for future development.

2. Survey of Machine and Deep Learning Applications in Crime Detection and Classification

Recent advancements in machine learning (ML) and deep learning (DL) have revolutionized crime detection and classification, particularly through the analysis of social media data, such as tweets from Twitter. One method utilizes Kernel Density Estimation (KDE) to process tweets, identifying patterns related to criminal activities, enabling automated detection in urban areas and predicting instances of violence. Another innovative model integrates various data types, including environmental factors, through a multi-modal data fusion approach that uses a Deep Neural Network (DNN) for accurate crime predictions.

The Crime Telescope model is another advancement, predicting crime hotspots by merging urban data with social media insights. It employs Latent Dirichlet Allocation (LDA) to extract relevant crime topics, and a regression model assesses these hotspots, visualizing crime-prone areas on interactive maps. Additionally, a DNN-based framework analyzes historical crime

patterns to predict individuals' likelihood of committing crimes, utilizing a Data Pooling Method (DPM) for accurate crime category predictions.

The use of the BERT model for detecting crime-related Twitter posts marks significant progress, where pre-processing techniques transform tweets into feature vectors for analysis. BERT enhances crime detection accuracy by understanding the context of textual data. Another hybrid model combines Artificial Neural Networks (ANN) with BERT, integrating meteorological data and tweets for crime detection and prediction.

Further developments include frameworks that classify and geolocate criminal tweets using Natural Language Processing (NLP) models, like Naïve Bayes, and techniques for spatio-temporal crime detection using Long Short-Term Memory (LSTM) models. These models analyze tweets' geographical properties to forecast crime patterns over time.

Additionally, deep learning architectures have been created to classify crimes and identify urban crime hotspots. By ranking crime frequency in specific neighborhoods, law enforcement can allocate resources efficiently. Hybrid sentiment analysis models utilizing BERT and Spark Natural Language Toolkit (SNLTK) have also been introduced to improve crime prediction accuracy.

Other contributions include Federated Long Short-Term Memory (F-LSTM) models for time-series crime prediction, Unsupervised Domain Adaptation Models (UDAC) for crime risk prediction across cities, and Modified Capsule Networks (MCN) for cybercrime sentiment analysis. These approaches, along with CNN-based models for crime intention detection and agent-based simulations, highlight the growing sophistication of AI in crime prediction and public safety analysis.

The multimodal deep learning crime prediction system integrates tweets and historical crime data using ConvBiLSTM models, improving prediction accuracy by combining semantic and contextual knowledge from various sources. This comprehensive framework represents a cutting-edge approach to crime prediction using social media data and advanced machine learning techniques.

10. **Comparative Analysis** This section presents a comparison of the crime prediction models previously discussed, which utilize social media data and are founded on machine learning (ML) and deep learning (DL) techniques. Table 1 outlines the various crime prediction models that employ ML and DL algorithms.

Table. I Comparison of Different ML and DL Based Crime Prediction Models

	Techniques	Advantages	Disadvantages	Performance Metrics
1	KDE	Enhances resource allocation, reduces wasted effort, and decreases crime response times.	Does not properly account for temporal effects like trends, lags, and periodicity.	AUC=0.65
2	DNN	Effectively analyzes the relationship between environmental context information.	Limited to regions with insufficient data, potentially causing performance degradation.	Accuracy = 34.25%, Precision = 74.35%, Recall=80.55%
3	LDA (CrimeTelescope)	Reduces ambiguity and reconsiders factors.	Insufficient to handle large crime databases.	Accuracy = 0.82, AUC=0.79
4	SVM-TD-IDF	Robust performance on large datasets	Slower inference time	Accuracy = 82.6%, Precision = 88.89%
5	ANN-NLP	Better convergence rate	Complexity and computational issues	Accuracy = 82%, Precision = 91.5%

6	BERT	Improved performance on crime prediction	Dependency on multiple datasets	Accuracy = 81.5%, Precision = 87.2%
7	DNN-BERT	High scalability and handles high-dimensional data	Overfitting issues	Accuracy = 82%, Precision = 92.8%
8	LSTM	Robust performance on simple or large datasets	Slower convergence	Accuracy = 82.7%, Precision = 85.4%
9	SVM-TD-IDF	Effective for crime prediction	Takes time to train on large datasets	Accuracy = 82.6%, Precision = 88.89%
10	AI	High scalability and better performance	Inefficient handling of overlapping and imbalance problems	Accuracy = 84%, Recall = 85.4%
11	CNN	Effective for crime prediction	Doesn't capture the semantic features and long-term dependencies	Accuracy = 91.7%, Recall = 92.4%
12	BERT-SLTK	Doesn't effectively capture the semantic features and long-term dependencies	Large dataset size and computational requirements	Accuracy = 91.0%, Precision = 92.4%
13	Federated LSTM	Optimizes parameters and reduces loss	Trained with limited high-dimensional data	Accuracy = 91.0%, Precision = 92.4%
14	Federated convolutional neural network with domain adaptation	Eliminates noisy, irrelevant data and devises quick response time	This model collected data from various sources, which might lead to bias and target leakage	Accuracy = 92.8%, Precision = 97.5%
15	NLP, SUM, RF	Eliminates noisy, irrelevant data and devises quick response time	Takes long time to train the collected data	Accuracy = 86%
16	LSTM	Collectively determines the most likely locations for crime at any given time throughout the country	Doesn't consider tweet contents to determine accurate interaction among users, which might impact predictive decision	RMSE=778, MAE=6.29
17	MCN-NLP, CCO	Better Generalizability and can be applied for real time applications	Slower training time and prone to overfitting issues	Accuracy = 92%, Precision = 94.2%
18	VGG-16, ResNet, GoogleNet, YOLO	Rapid response time, early warning, lower memory space	Optimization algorithms were highly necessary for fine-tuning pre-trained CNN models, leading to uncertainty issues	Accuracy = 0.91, Recall = 0.91, F1-Score = 0.91, Inference speed = 68, MSE = 0.103

19	ABM	High scalability and applied as real-time mode to minimize crime frequency	High computational cost and necessities some advanced models to improve detection efficiency	Accuracy = 97.5%
20	ConvBiLSTM, Deep crime-semantic model	Effectively minimizes feature number to eliminate overfitting and parameter complexity	The study was constrained to English tweets specifically related to theft, battery, burglary, robbery, and motor vehicle theft	Accuracy = 97.5%

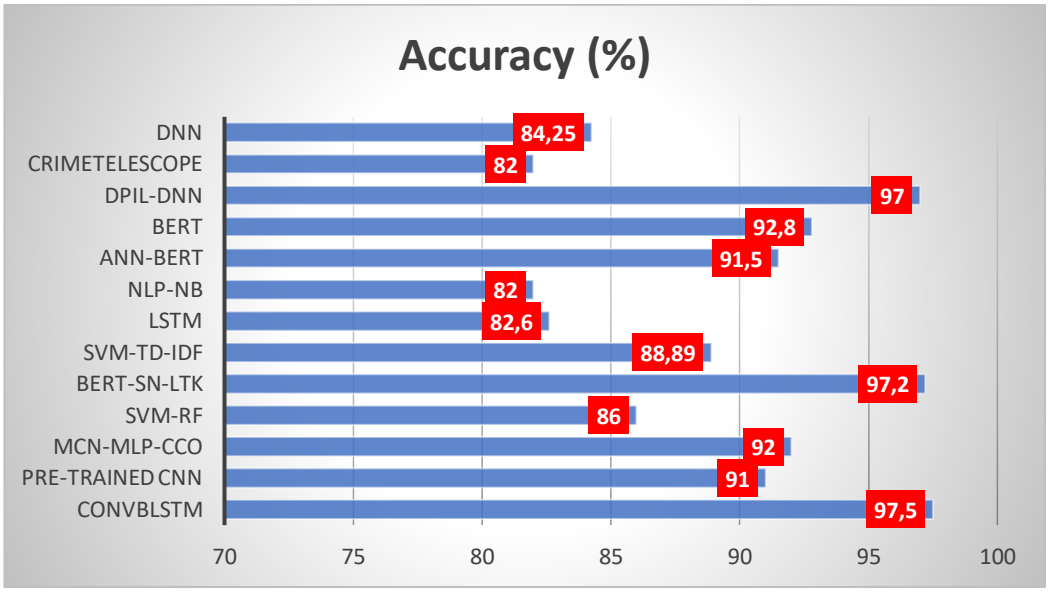

Figure 1 ML and DL Based Crime Prediction Using Various Social Media Data

4 Performance Analysis and Metrics

This section evaluates the performance of existing machine learning (ML) and deep learning (DL) methods in predicting and classifying crimes, as shown in Table 1. The comparison emphasizes the accuracy of different crime prediction models utilizing ML and DL techniques.

Figure 1 visually demonstrates the effectiveness of these models by analyzing data from social media posts and blogs. Notably, the study referenced in article [29] achieves superior crime prediction results using Twitter data. The researchers developed a multimodal deep learning model that integrates real-time tweets with historical crime data, using advanced data fusion techniques to overcome limitations in prior crime prediction research.

This approach analyzes tweets and crime data at the vector level, extracting key word features and merging them into a unified representation. The model uses a ConvBiLSTM architecture, which processes data in both directions (left-to-right and right-to-left), enabling it to capture long-term dependencies and contextually relevant information. Word embedding vectors further enhance the model's language comprehension, allowing for more accurate crime predictions.

The convolutional and pooling layers within the model distill meaning from both crime and Twitter datasets, integrating insights into a single representation. This comprehensive model effectively combines vector features from crime and social media data, significantly improving its crime prediction accuracy.

5. Conclusion

This research provides a thorough comparison of crime prediction models that utilize various machine learning (ML) and deep learning (DL) approaches. With the rapid increase in

crime rates, there is an urgent need for effective intervention strategies. Traditional methods of solving crimes are often slow and inefficient, highlighting the necessity for law enforcement agencies to adopt more effective techniques.

In recent years, ML and DL models have been successfully employed to analyze social media data, aiding in the detection and prevention of crime while enhancing community safety and security. This paper offers a detailed review of different ML and DL methodologies for crime prediction, evaluating their strengths, weaknesses, and detection efficiencies.

By leveraging this review, researchers can better allocate resources, anticipate criminal activity in real time, and enhance service to the community. The findings of this study can guide the development of the most efficient and reliable techniques for identifying criminal incidents and recognizing patterns within them.

Looking ahead, future research will concentrate on advancing deep learning models that law enforcement agencies can implement for proactive interventions and on-the-spot actions. This approach aims to alleviate complications within the enforcement systems and improve their overall efficacy in maintaining public safety.

References

1. Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). Learning to detect patterns of crime. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, Proceedings* (Vol. 13, pp. 515-530). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40994-3_33
2. Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017). Crime pattern detection, analysis & prediction. In *2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA)* (Vol. 1, pp. 225-230). IEEE. https://doi.org/10.1109/ICECA.2017.8203676
3. Wang, Z., & Zhang, H. (2020). Construction, detection, and interpretation of crime patterns over space and time. *ISPRS International Journal of Geo-Information*, 9(6), 339. https://doi.org/10.3390/ijgi9060339
4. Yu, H., Liu, L., Yang, B., & Lan, M. (2020). Crime prediction with historical crime and movement data of potential offenders using a spatio-temporal cokriging method. *ISPRS International Journal of Geo-Information*, 9(12), 732. https://doi.org/10.3390/ijgi9120732
5. Chohlas-Wood, A., & Levine, E. S. (2019). A recommendation engine to aid in identifying crime patterns. *INFORMS Journal on Applied Analytics*, 49(2), 154-166. https://doi.org/10.1287/inte.2019.0985
6. Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505, 435-443. https://doi.org/10.1016/j.physa.2018.03.084
7. Aghababaei, S., & Makrehchi, M. (2016). Mining social media content for crime prediction. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 526-531). IEEE. https://doi.org/10.1109/WI.2016.0089
8. Mahajan, R., & Mansotra, V. (2021). Correlating crime and social media: Using semantic sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 12(3). https://doi.org/10.14569/IJACSA.2021.0120338
9. Lopez Torres, I. (2023). SafeCity: Predictive crime analysis integrating social media big data and official reports. Available at SSRN 4566143. https://doi.org/10.2139/ssrn.4566143
10. Sharma, S., & Jain, A. (2020). Role of sentiment analysis in social media security and analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5), e1366. https://doi.org/10.1002/widm.1366
11. Prichard, J., Watters, P., Krone, T., Spiranovic, C., & Cockburn, H. (2015). Social media sentiment analysis: A new empirical tool for assessing public opinion on crime? *Current Issues in Criminal Justice*, 27(2), 217-236. https://doi.org/10.1080/10345329.2015.12036042
12. Ristea, A., Al Boni, M., Resch, B., Gerber, M. S., & Leitner, M. (2020). Spatial crime distribution and prediction for sporting events using social media. *International Journal of*

Geographical Information Science, 34(9), 1708-1739. https://doi.org/10.1080/13658816.2020.1719495

13. Dakalbab, F., Talib, M. A., Waraga, O. A., Nassif, A. B., Abbas, S., & Nasir, Q. (2022). Artificial intelligence & crime prediction: A systematic literature review. *Social Sciences & Humanities Open*, 6(1), 100342. https://doi.org/10.1016/j.ssaho.2022.100342

14. Llaha, O. (2020). Crime analysis and prediction using machine learning. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 496-501). IEEE. https://doi.org/10.23919/MIPRO48935.2020.9245120

15. McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, 2(1), 1-12. https://doi.org/10.5121/mlaij.2015.21

16. Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3286344

17. Azeez, J., & Aravindhar, D. J. (2015). Hybrid approach to crime prediction using deep learning. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1701-1710). IEEE. https://doi.org/10.1109/ICACCI.2015.7275858

18. Wang, B., Yin, P., Bertozzi, A. L., Brantingham, P. J., Osher, S. J., & Xin, J. (2019). Deep learning for real-time crime forecasting and its ternarization. *Chinese Annals of Mathematics, Series B*, 40(6), 949-966. https://doi.org/10.1007/s11401-019-0168-y