Journal Collabits
Vol. 3 No. 1 January 2026 : 9-15
E-ISSN : 3046-6709, P-ISSN : 3062-8601
https://publikasi.mercubuana.ac.id/index.php/collabits

Implementation Of DBSCAN Clustering and Random Forest Algorithm for Mapping and Predicting Shooting Incidents in New York

Azka Niaji Rangkuti[1*], Samoedra Cakra Arifin[2], Muhammad Ramadansyah Kurnia Putra[3], Nila Natalia[4]

[1, 2, 3] Computer Science, Universitas Mercu Buana, Indonesia
[4] Computer Engineering, Politeknik Sukabumi, Indonesia

*Coressponden Author: ramadansyahx12@gmail.com

Abstract—Shooting incidents in crowded, heavily populated areas of cities cause serious threats to public safety and social security. New York State, which includes large metropolitan areas and suburban regions, experiences complex spatial and temporal crime patterns that are difficult to identify using traditional crime analysis methods that rely only on descriptive statistics and manual hot spot identification. This study proposes a data-driven quantitative approach to mapping and predicting shooting incidents by integrating spatial clustering and machine learning techniques. Density-based clustering methods are applied to the geographic coordinates of shooting incidents to identify areas with high incident concentrations while filtering out isolated events as noise. The resulting spatial clusters are then interpreted as hotspot locations and used as reference labels for a supervised classification model. A Random Forest algorithm is then used to predict hotspot and non-hotspot locations using spatial and temporal features, including geographic position and time of occurrence. The model is evaluated using standard classification performance measures, including accuracy, precision, recall, F1 score, and confusion matrix analysis.

Keywords :
Crime Mapping;
Machine Learning;
New York;
Spatial Clustering;

Article History:
Received: 26-11-2025
Revised: 23-12-2025
Accepted: 07-01-2025

Article DOI : 10.22441/collabits.v3i1.37587

INTRODUCTION

Shooting incidents in crowded and large urban areas are a serious threat to public safety and social security. New York state is one of the largest and most densely populated states and also the center of the metropolitan life in the United States. With major counties like Manhattan, Brooklyn, Bronx, Queens, and other cities across the state, it is not a surprise that the number of crime activities that occur in the state is really high. With high number of crime activities, it is also generating a high number of crime related data that is useful for analytical purposes. Traditional crime analysis methods often rely on descriptive statistics and manual identification of hot spots, which may overlook complex spatial and temporal patterns.

The advances of technology in machine learning and spatial data analysis are providing the ability to improve crime mapping and prediction. Unsupervised learning such as clustering algorithms can unveil hidden spatial patterns, while supervised learning can be applied to predict future events based on historical trends. This research focuses specifically on the implementation of DBSCAN for spatial clustering of shooting incidents and Random Forest for predicting possibility of shooting incident in a certain location.

LITERATURE REVIEW

Crime mapping has developed rapidly with the implementation of Geographic Information System (GIS) and machine learning technology. Spatial clustering algorithms such as k-means clustering and

hierarchical clustering are widely used to identify crime hotspots. However, these methods typically require predefined parameters, such as the number of clusters, which may not reflect the distribution of crime in the real world.

DBSCAN has become a popular method in crime mapping since it can detect clusters with various shapes and handle data noise effectively. Research shows that the DBSCAN clustering method works really well in detecting crime hot spots without forcing all data points into a cluster.

To predict criminal activity, machine learning algorithms such as Random Forest have proved to be very reliable. Random Forest combines multiple decision trees to improve prediction accuracy and reliability. According to previous studies, Random Forest performs particularly well when predicting crime compared to single classification, especially on datasets with multiple interactions and mixed data types.

METHODOLOGY

This study uses a statistical and computational method that apply spatial data analysis and machine learning algorithms. The methodology consists of data extraction, data preprocessing, spatial clustering using DBSCAN, model building using the Random Forest algorithm, and model evaluation.

3.1 Data Source

The dataset we used in this study is the NYPD Shooting Incident Data, which recorded reported shooting incidents throughout New York State. Each record contains information about a specific shooting incident, including geographic, temporal, administrative region, and personal information about the victim.

3.2 Data Attributes

The main attributes used for this study include latitude, longitude, administrative region, date of the event, and time of the event. Additional temporal features such as month, day of the week, and hour were obtained from the original date and time attributes to provide support for predictive modeling.

Table 1. Description of variable used in the study

Variable	Description
Latitude	Geographic latitude of incident
Longitude	Geographic longitude of incident
BORO	Borough or administrative area
Month	Month of incident happen
Day of week	Day of week (0-6)
Hour	Hour of occurrence (0-23)

3.3 Data Preprocessing

Data preprocessing consists of processing date and time formats, filtering out records that do not have geographic coordinates, and generating temporal features. Only incidents with valid latitude and longitude values are kept to assure accurate spatial analysis.

3.4 DBSCAN Clustering Method

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a method for identifying the clusters of shooting incidents. DBSCAN was used because it does not require a predetermined number of clusters and can detect clusters of various shapes while effectively handling noise. Formulation for define the spatial dataset are :

$$D=\{x_1,x_2,\ldots,x_n\}$$

Where each data point $x_i = (lat_i,lon_i)$ represent the geographic coordinates of a shooting incident. The

ε-neighborhood of a point x_i is defined as :

$$N\varepsilon(x_i)=\{x_j \in D | dist(x_i,x_j) \leq \varepsilon\}$$

A point is considered a core point if :

$$|N\varepsilon(x_i)| \geq MinPts$$

Clusters are formed by connecting points that are reachable by a certain density. Points that aren't reachable by a certain density from any core point are classified as noise.

3.5 DBSCAN Parameter Configuration

In this study, DBSCAN is applied using Euclidean distance with ε = 0.003 and MinPts = 4. Shooting incidents included in a cluster are labeled as hotspots, while incidents labeled as noise are isolated events.

Table 2. DBSCAN Parameter Configuration

Parameter	Value
ε (epsilon)	0.003
MinPts	4
Distance Metric	Euclidean

3.6 Random Forest Prediction Model

Random Forest is applied to predict if a shooting location belongs to a hotspot cluster. This model is trained as a binary classification, where the target variable indicates the hotspot or non-hotspot classification from the DBSCAN results. To predict an event, features are needed to determine which events fit into the hotspot. The features included are spatial attributes (latitude and longitude), temporal features (month, day of the week, hour), and area information encoded using the one-hot encoding method.

Table 3. Features used in Random Forest

Feature Type	Variables
Spatial	Latitude, Longitude
Temporal	Month, DayofWeek, Hour
Administrative	Borough

The training dataset is also important in making prediction models. As for the training dataset, it can be defined as:

$$\{(x_i,y_i)\}_{i=1}^{n}$$

Where x_i is the feature vector and y_i is the hotspot label.
Each decision tree T_k produces a prediction $h_k(x)$. The final prediction is obtained through majority voting from each decision tree. The formulation is the following:

$$y=mode\{h_1(x),h_2(x),\ldots,h_K(x)\}$$

Feature selection at each split maximizes information. The formulation is the following:

$$IG(S,A)=H(S)-\sum_{v \epsilon values(A)} \frac{|Sv|}{|S|} H(S_v)$$

3.7 Model Training and Evaluation

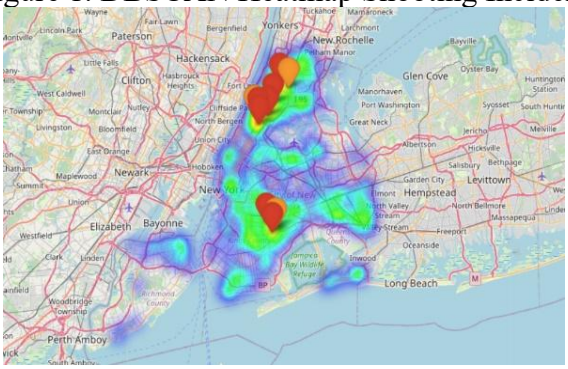
The dataset is divided into training and testing subsets using a 70:30 ratio. The Random Forest is trained by using 100 decision trees. For the model performance, it was evaluated using accuracy, precision, recall, and F1- score metrics.

RESULTS AND DISCUSSION

4.1 DBSCAN Result

The DBSCAN algorithm was able to identify several spatial clusters of shooting incidents across New York State. Most incidents were grouped into dense clusters, which showed spatially concentrated shooting hot spots. Incidents classified as noise were isolated events that didn't form consistent spatial patterns. This result shows that DBSCAN was able to capture the spatial distribution of crime in the real world without requiring a certain predetermined number of clusters.

Figure 1. DBSCAN Heatmap Shooting Incident



4.2 Random Forest Classification Result

The Random Forest model shows solid performance in predicting shooting locations into hotspot and non-hotspot categories. The overall classification accuracy is 88.28%, which means the model accurately classifies most of the test examples.

For hotspot predictions, the model reached a precision of 0.88 and a high recall of 0.99, resulting in an F1 score of 0.94. The high recall indicates that the model is highly effective at identifying actual hotspot locations, which is very important in the context of crime analysis, where failure to detect high-risk areas can lead to serious public safety problems.

For non-hotspot predictions, the model achieved a precision of 0.83 but a relatively low recall of 0.21, resulting in an F1 score of 0.33. This result indicates that a significant number of non-hotspot locations were misclassified as hotspots. This outcome was primarily influenced by class imbalance in the dataset, where hotspot cases far outnumbered non-hotspot cases. As a result, the model showed conservatively biased classification behavior that prioritized reducing false negatives in hotspot detection at the expense of increasing false positives.

The weighted average F1-score of 0.85 indicates that the model maintains a strong overall performance while accounting for class distribution. These results show that Random Forest effectively utilizes spatial and temporal features to predict the risk of shooting incidents when trained using hotspot labels generated by DBSCAN, making it suitable for risk-oriented crime prevention and decision support applications.

Figure 2. Random Forest result

```
Training Random Forest...

Random Forest Results:
   Accuracy: 0.8828 (88.28%)

Classification Report:
              precision    recall  f1-score   support

 Not Hotspot       0.83      0.21      0.33        72
     Hotspot       0.88      0.99      0.94       440

    accuracy                           0.88       512
   macro avg       0.86      0.60      0.63       512
weighted avg       0.88      0.88      0.85       512
```

4.3 Random Forest Confusion Matrix

Figure 3. Random Forest Confusion Matrix

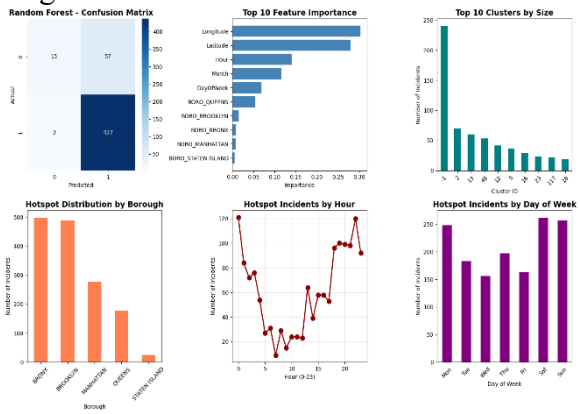


Figure 3 shows the confusion matrix of Random Forest classification used to categorize shooting incident locations into hotspots and non-hotspots. This matrix provides a detailed overview of the model's classification behavior apart from overall accuracy.

The model successfully identified 437 hotspot cases (true positives) while misclassifying only 3 hotspot cases as non-hotspots (false negatives). This result corresponds to a hotspot recall of 0.99, indicating that the model is highly effective in detecting high-risk locations. This kind of performance is very important in crime analysis applications, where failure to identify actual hotspots can lead to inadequate preventive measures and increased public safety risks.

For non-hotspot locations, the model correctly classified 15 cases as non-hotspots (true negatives) but misclassified 57 cases as hotspots (false positives). This resulted in a low recall for the non-hotspot class. The unbalanced distribution of hotspot and non-hotspot samples in the dataset also contributed to this result, as the classification was trained on a dataset dominated by hotspot instances.

The resulting classification performance reflects a more conservative prediction strategy, where the model prioritizes reducing false negatives in hotspot detection at the expense of increasing false positives. From a public safety perspective, this exchange is reasonable, as overestimating risk is generally better than underestimating the possibility of dangerous locations.

Overall, the confusion matrix confirms that the Random Forest model is suitable for crime predictions, providing strong detection of hotspots while maintaining solid overall prediction performance.

CONCLUSION

This study presents an integrated machine learning framework for analyzing shooting incidents across New York State by combining spatial clustering and predictive classification techniques. DBSCAN was

employed to identify spatial concentrations of shooting incidents without requiring predefined cluster structures, enabling the detection of naturally occurring crime hotspots while distinguishing isolated incidents as noise. The clustering results reveal that shooting incidents are highly concentrated in specific geographic areas, underscoring the importance of spatial context in urban crime analysis.

Building upon the spatial patterns identified by DBSCAN, a random forest classifier was implemented to predict hotspot and non-hotspot locations using spatial and temporal features. The model demonstrates strong predictive performance, particularly in identifying hotspot locations, achieving very high recall for high-risk areas. This conservative prediction behavior prioritizes the detection of potentially dangerous locations, which is appropriate for public safety–oriented applications despite the presence of increased false positives. Feature importance analysis indicates that geographic coordinates are the most influential factors in hotspot prediction, while temporal variables such as hour, day of week, and month also contribute meaningfully to model performance.

REFERENCES

[1] A. Y. Permana, H. N. Fazri, M. F. Nur Athoilah, M. Robi, and R. Firmansyah, "Penerapan data mining dalam analisis prediksi kanker paru menggunakan algoritma Random Forest," *Jurnal Ilmiah Teknik Informatika dan Komunikasi*, vol. 3, no. 2, pp. 27–41, 2023.

[2] M. Samantri and Afiyati, "Perbandingan algoritma Support Vector Machine dan Random Forest untuk analisis sentimen terhadap kebijakan pemerintah Indonesia terkait kenaikan harga BBM tahun 2022," *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, pp. 1–9, 2024.

[3] A. Salsabila and L. Iswari, "Identifikasi pengelompokan titik penjemputan dan titik pengantaran perjalanan taksi menggunakan algoritma DBSCAN," *EDUSAINTEK: Jurnal Pendidikan, Sains dan Teknologi*, vol. 11, no. 2, pp. 739–755, 2024.

[4] Y. Ruan, W. Liu, T. Wang, J. Chen, X. Zhou, and Y. Sun, "Dominant partitioning of discontinuities of rock masses based on DBSCAN algorithm," *Applied Sciences*, vol. 13, no. 15, p. 8917, 2023.

[5] L. Mochurad, A. Sydor, and O. Ratinskiy, "A fast parallelized DBSCAN algorithm based on OpenMP for detection of criminals on streaming services," *Frontiers in Big Data*, vol. 6, Article 1292923, 2023.

[6] A. Fauzan, A. Novianti, R. R. M. A. Ramadhani, and M. A. S. Adhiwibawa, "Analysis of hotels spatial clustering in Bali: Density-Based Spatial Clustering of Application Noise (DBSCAN) algorithm approach," *EKSakta: Journal of Sciences and Data Analysis*, vol. 3, no. 1, pp. 25–38, 2022.

[7] J. Xu, Y. Zhang, X. Li, and D. Liu, "Bank customer segmentation and marketing strategies based on improved DBSCAN algorithm," *Applied Sciences*, vol. 15, no. 6, p. 3138, 2025.

[8] M. F. Fadhillah, A. L. A. Suyoso, and I. Puspitasari, "Perbandingan algoritma DBSCAN dan K-MEANS dalam segmentasi pelanggan pengguna transportasi publik Transjakarta menggunakan metode RFM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 48–56, 2025.

[9] Y. Zhang, X. Liu, and H. Chen, "Enhanced stratified sampling density-based spatial clustering of applications with noise (SS-DBSCAN) for high-dimensional data," *Journal of Algorithms & Computational Technology*, 2025.

[10] J. Weng, J. Zhang, and L. Lin, "h-DBSCAN: A simple fast DBSCAN algorithm for big data," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 157, pp. 10680–10689, PMLR, 2021.

[11] Y. Zhao, X. Liu, and H. Wang, "An AIS data-driven approach to analyze the pattern of ship trajectories in ports using the DBSCAN algorithm," *Applied Sciences*, vol. 11, no. 2, p. 799, 2021.

[12] S. Nugroho et al., "Internet traffic classification model based on A-DBSCAN algorithm," *International Journal of Advances in Soft Computing and Its Applications*, 2024.

[13] A. Pratama et al., "Clustering of seismicity in the Indonesian region for the 2018–2020 period using the DBSCAN algorithm," *Jurnal Geofisika*, 2021.

[14] Z. Wang et al., "An improved adaptive radar signal sorting algorithm based on DBSCAN by a novel CVI," *IEEE Access*, 2024.
[15] R. Siregar, "Clustering negara berdasarkan skor pengendalian konsumsi tembakau menggunakan algoritma DBSCAN," *Jurnal Teknologi Informasi dan Komunikasi*, 2022.
[16] S. Lestari et al., "Penggunaan algoritma DBSCAN dalam pengelompokan kabupaten/kota di Sulawesi Tenggara berdasarkan indikator pendidikan," *SIMTEK: Jurnal Sistem Informasi dan Teknologi*, 2022.
[17] A. Kurniawan, "Penerapan algoritma DBSCAN untuk clustering penjualan supermarket," *ZETA: Jurnal Ilmu Komputer dan Sistem Informasi*, 2023.
[18] M. Rahman et al., "Analisis dan evaluasi algoritma DBSCAN pada tuberkulosis," *Jurnal Ilmiah Teknik Elektro dan Informatika Terapan*, 2022.
[19] A. Firmansyah et al., "Analysis of Sulawesi earthquake data from 2019 to 2023 using DBSCAN clustering," *RESTI: Rekayasa Sistem dan Teknologi Informasi*, 2024.
[20] H. Li et al., "WOA-DBSCAN: Application of whale optimization algorithm in DBSCAN parameter adaption," *IEEE Access*, 2023.
[21] R. Wijaya et al., "Komparasi algoritma hierarchical, K-means, dan DBSCAN pada analisis data penjualan melalui Facebook," *Jurnal Ilmiah Informatika*, 2023.
[22] A. Nugraha, "Klasterisasi data penjualan toko perak J-Maskus menggunakan algoritma HDBSCAN," *Jurnal Locus*, 2024.
[23] D. Prabowo et al., "Implementasi DBSCAN dalam clustering data minat mahasiswa setelah pandemi Covid-19," *Jurnal Konstelasi*, 2023.
[24] I. Maulana et al., "Analisis pemilihan parameter pada algoritma DBSCAN untuk pengelompokan titik api di Indonesia," *JPTI: Jurnal Penelitian Teknologi Informasi*, 2022.
[25] R. Putri et al., "Pengelompokan data pendistribusian listrik menggunakan algoritma Density-Based Spatial Clustering of Applications with Noise (DBSCAN)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 2023.
[26] A. Hidayat et al., "Implementasi algoritma clustering DBSCAN terhadap pola navigasi pengguna di perpustakaan digital untuk mengungkap zona buta akses informasi dan optimalisasi antarmuka sistem," *SKANIKA*, 2024.
[27] R. Maulana et al., "Identifikasi hotspot kebakaran hutan Kalimantan Timur tahun 2023 menggunakan teknik spasial-temporal clustering ST-DBSCAN," *Jurnal Tematik*, 2024.
[28] A. H. Nugroho and R. A. Pratama, "Implementasi algoritma Random Forest untuk klasifikasi data," *Jurnal Tekno*, 2022.
[29] M. A. Putra, "Implementasi algoritma klasifikasi Random Forest pada data kesehatan," *Jurnal Ilmiah Sistem Informasi*, 2020.
[30] T. Kurniawan et al., "Analisis performa algoritma Random Forest dalam klasifikasi data," *Jurnal Teknologi Informasi*, 2019.
[31] A. Setiawan and D. Prabowo, "Penerapan algoritma Random Forest untuk prediksi dan klasifikasi data," *SISTEMASI: Jurnal Sistem Informasi*, 2023.