Grenze International Journal of Engineering and Technology, June Issue

Spatio-Temporal Clustering and Forecasting of Urban Incident Patterns using Location-based Insights

Jane Rubel Angelina Jeyaraj[1], Mandepudi Sri Ram Sai[2], Kamatam Lokesh[3], Nalagatla Rithwin Reddy[4] and Talluru Rajaiah[5]
[1-5]Kalasalingam Academy of Researach and Education, Computer Science and Engineering, Krishnankoil, Tamilnadu, India
Email: j.janerubelangelina@klu.ac.in, sriramsaimandepudi@gmail.com, lokeshlokee52@gmail.com, nrithwinreddy@gmail.com, rajaiahtalluru@gmail.com

***Abstract*— Public safety is seriously threatened by the high rate of criminal activity. Because current prediction techniques typically rely on K-means clustering and basic classification models, evaluating large, complex crime datasets can be difficult for law enforcement agencies. The ability of these methods to produce precise, date-specific forecasts is limited because they largely ignore the dynamic, time-based behavioral patterns of various locations. To produce thorough and intelligible crime forecasts, this study suggests a novel hybrid methodology. Our method presents a two-phase system. Using a shape-based algorithm called k-Shape clustering, we first divide police patrol zones into discrete groups according to how similar their daily crime time-series "shape," as opposed to just volume. This enables us to distinguish and classify regions with distinct behavioral patterns (e.g., "stable low-level activity" versus "weekend-driven spikes"). Second, a group of specialized LightGBM (Light Gradient Boosting Machine) forecasting models can be developed thanks to this clustering, one for every recognized crime pattern. To accurately forecast future crime counts for both the overall total and ten distinct crime categories, these models are trained on a wealth of features, such as rolling averages, temporal lags, weather data, and holiday indicators. Additionally, SHAP (SHapley Additive exPlanations) is integrated into the system to offer location-based, explicable insights into the variables influencing each prediction. The final product is intended to provide law enforcement with precise, actionable intelligence for the purpose of allocating resources in a proactive manner.**

***Keywords*— Crime Rate Clustering, Time-Series Forecasting, k-Shape, LightGBM, Explainable AI (XAI), SHAP, Predictive Policing, Spatio-Temporal Analysis.**

I. INTRODUCTION

From small towns to large cities, criminal activity remains a serious issue that affects everyone. It not only threatens public safety but also creates significant social and economic challenges by instilling fear and placing a heavy burden on community resources. Law enforcement agencies are continuously seeking ways to move from reactive responses to proactive crime prevention. However, a major challenge lies in the vast and complex nature of modern crime data, where each incident is associated with time, location, and type, often resulting in datasets with millions of records. Identifying meaningful patterns within such data is extremely difficult. Traditional methods like K-Means clustering have been used to group areas based on average crime levels, such as “high crime” and “low crime” zones.

Grenze ID: 01.GIJET.12.2.375_24

While useful, this approach fails to capture the behavioral patterns or temporal rhythms of crime. For instance, a neighborhood that experiences crime spikes only on weekends may have the same average crime rate as another area with consistent daily crime, yet both require entirely different policing strategies. To address this limitation, our project proposes a novel two-stage crime forecasting system.

In the first stage, we use the k-Shape clustering algorithm to group locations based on the shape of their daily crime patterns, enabling better identification of patterns such as "weekend spikes" and "stable" crime trends. In the second stage, we implement LightGBM to develop a set of specialized models, each trained on data from specific pattern groups identified by k-Shape.

II. Literature Review

To build our project, we first studied existing research to understand what works, what does not, and where the gaps exist. We focused on three main areas: forecasting models, clustering methods, and the role of feature engineering. Clustering is widely used in crime analysis to identify high-crime areas or "hotspots." Traditional approaches often rely on K-Means clustering, as highlighted in our initial proposal and supported by Howard et al. (2024), who showed that applying Time-Series K-Means before forecasting can improve prediction stability. However, the main limitation of K-Means is that it groups data based on average magnitude rather than behavioral patterns, making it unable to distinguish between areas with steady crime and those with periodic spikes, such as weekend surges. To overcome this, our project adopts the k-Shape algorithm, which is specifically designed for time-series data and clusters locations based on the shape of their crime patterns, enabling better identification of underlying behavioral rhythms.

The second focus area is forecasting models and feature engineering. Several studies have compared machine learning models to identify the most effective approach for crime prediction. Research by Alsubayhin et al. (2024) demonstrated that LightGBM outperformed models like Random Forest and Logistic Regression in terms of accuracy and F1 score.

Similarly, Ling et al. (2024) compared advanced models such as XGBoost and CatBoost and also found LightGBM to deliver superior performance. These studies further emphasized the importance of feature engineering, showing that incorporating features such as historical crime lags, holidays, and weather conditions significantly improves model accuracy. This evidence strongly influenced our decision to adopt LightGBM along with advanced feature engineering techniques in our system.

Finally, our review identified a key research gap. Many recent studies, such as Wang et al. (2024) and Fan et al. (2025), focus heavily on complex deep learning models like Graph Neural Networks (GNNs) and Long Short-Term Memory (LSTM) networks. While powerful, these models often act as "black boxes," making them difficult to interpret and computationally expensive to train. On the other hand, some approaches rely on a single generalized model for an entire city, which fails to capture localized crime patterns. To address this gap, our project proposes a hybrid approach that combines the strengths of multiple techniques. We use k-Shape clustering to capture behavioral patterns, LightGBM as a high-performance prediction engine, and robust feature engineering to enhance accuracy.

Additionally, we integrate SHAP analysis to provide interpretability, which is often missing in complex models. This approach enables us to create a "team of specialist models," where each LightGBM model is trained on a specific crime pattern group, resulting in forecasts that are not only highly accurate but also detailed, interpretable, and tailored to the unique characteristics of each patrol area.

III. Methodology

We carried out this research by following a systematic, five- stage procedure as evident in our flow diagram. Our design was to initially comprehend the embedded patterns of behavior in the data and then construct detailed, specialist models for every pattern.

A. Data Collection and Preprocessing

The dataset used in this study was obtained from the Chicago Data Portal, which provides publicly available crime records from 2001 to 2023. The dataset contains information such as crime type, date, and location.

Data preprocessing was performed to ensure data quality.

Records with missing or incomplete values, especially those lacking location identifiers such as "Police Beat," were removed. The cleaned dataset was then transformed into a structured time-series format by aggregating individual crime incidents into daily crime counts for each of the 274 police beats. This transformation enabled effective temporal analysis and modeling.

B. k-Shape Clustering

To capture distinct behavioral patterns in crime data, the k-Shape clustering algorithm was employed. Unlike traditional methods such as K-Means, k-Shape clusters time-series data based on similarity in shape rather than magnitude.

Before clustering, all time-series data were normalized to ensure fair comparison. The optimal number of clusters was determined using the Elbow Method, which indicated that six clusters provide the best balance between accuracy and complexity. The final model grouped all police beats into six clusters, each representing a unique crime pattern or temporal rhythm.

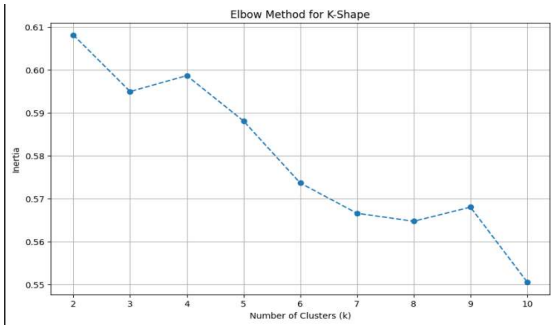

Fig 1. Graph-1: Showing optimal number of clusters

Pattern Identification: We trained the final k-Shape model to cluster all 274 beats into one of these six groups. This provided us our patterns (e.g., "High-Crime, Busy Commercial," "Low- Crime, Quiet Residential," etc.).

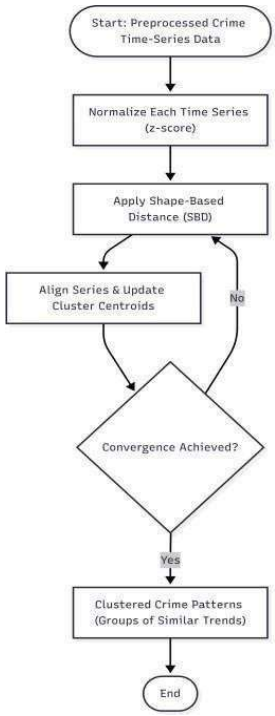

Fig 2. K-Shape Flow Chart

C. Feature Engineering

- Time-Based: Day of week, month, year, day of year, etc.
- Temporal Lags: We developed features for the "7 -day prior crime count" and the "7-day rolling average".
- External Data: We joined in two external data sources: a U.S. holiday list and a historical daily weather data set (maximum temperature and precipitation).

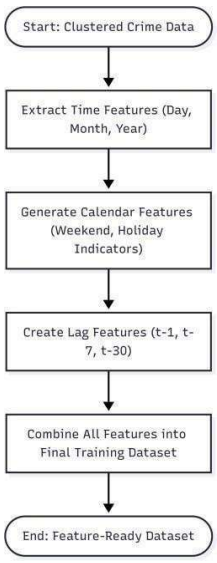

Fig 3: Feature Engineering Flow Chart

D. LightGBM Forecasting Models

Instead of one overall model, we used a "team of experts" using LightGBM.
Data Split: The data was split on a chronological basis. We used all data from 2001-2022 as our training sample and reserved the entire year of 2023 as our test sample to validate the models.
Multi-Target Structure: We specified 10 new columns, one for each of our top 10 crime types, for the daily count (THEFT, BATTERY, etc.).
Training: We trained 66 expert models:
6 models (one per cluster) to predict the Overall Crime Total.
60 models (6 clusters x 10 crime types) predict Specific Crime Types.

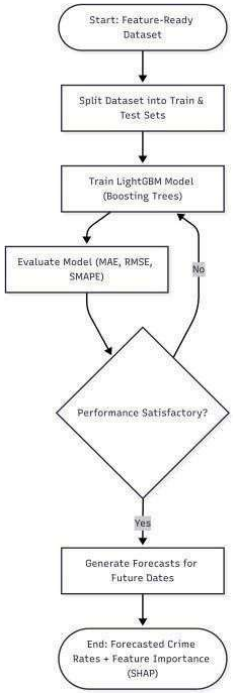

Fig 4: LightGBM Flow Chart

IV. RESULTS AND DISCUSSION

We introduce the practical realization of our felony predicting system. The application of the system was done through a web application to make it easier for the police officers to work together with the expert models through a user-friendly interface.

The Crime Forecaster user interface was designed to be extremely operationally efficient, allowing even non-tech users to benefit from complex machine learning insights. The dashboard is laid out in two main sections: the Control Sidebar and the System-Wide Overview Dashboard

Fig 5: System Dashboard and Control Interface

The dashboard transitions from general historical context to active intelligence by generating a forecast for a specific geographical unit. For the date of January 22, 2026, the system thinks there will be 2 crimes in Patrol Zone #833. The system uses a computer program called LightGBM to make this prediction. This program looks at what people in the Ashburn region do and then makes a guess. The number 2.10 is what the LightGBM program comes up with for Patrol Zone #833. This number is like a score that shows what the system thinks will happen with crime in that area. The system uses this score to help figure out what might happen with crime, in Patrol Zone #833 on January 22, 2026.

The 1-Day Forecast plot is helpful because it shows what we think will happen on a timeline. This means officers can see what to expect for the volume compared to what happens. The 1 - Day Forecast plot marks the predicted value on a timeline, which's useful for officers to see the expected volume relative to the baseline of the 1-Day Forecast.

A. The Forecasting Control Sidebar

The sidebar serves as the main input interface for the entire application. With its help, the user can transform the enormous stigma Chicago crime dataset into an extremely accurate way.

Area and Zone Selection: Users by initially selecting a broad area (like "Albany Park") can further specify a certain Patrol Zone or Police Beat (for example, "1712"). The corresponding specialist model for the area of that zone is then loaded. Time Parameters: The Forecast Start Date and the Number of Days to Forecast slider allow the users to define the length of the prediction. This is especially important for scheduling the police shifts on a weekly basis. Prediction Type: The users are given an option to switch between Overall Crime Total and Specific Crime Types and benefit from the 60+ special models created by us. Execution: The moment the "Generate Forecast" button gets clicked, it is interpreted as the initiation of the backend LightGBM model processing and SHAP explainer calling.

B. The Home Page: System-Wide Overview

The Home Page shows an overview of the old data that the system has stored before you make a specific forecast. This gives you an idea of what to expect from the prediction. The Home Page is helpful because it gives you an overview of the data.

Key Performance Indicators (KPIs): At the top of the dashboard, you will see three things that tell you what is going on right now with the system. These three things are very important. They help the user understand what is happening immediately. The three critical insights are shown to give the user a sense of what's happening with the system. The place with the crime in Austin is where the police are always busy. This is the area that has the most crime happening all the time, so the police need to be a lot. The Most Active Area, in Austin, is where the crime rate is high, and that means the police must work in this area

C. Granular Crime Type Decomposition

This project does something cool. It can break down a forecast of Total Crime into parts. This gives the police information that they can use. The different police units that specialize in things can use this information to do their jobs better. The Total Crime forecast is basically taken apart so that the police can see what is really going on and make decisions. This helps the police units because they get information about the Total Crime things that they care about. Component Analysis: If we look at the Crime Type Breakdown chart, we can see that even though there are about 2 crimes predicted the Crime Type Breakdown does not show that the risk of Crime Type Breakdown is the same, for all types of Crime Type Breakdown. The main things that can go wrong in Ashburn are Motor Vehicle Theft and Theft. These are the crimes that're most likely to happen. Motor Vehicle Theft and Theft are the risks we see. Risks: There are other things to think about like Battery and Deceptive Practice. These things are not as likely to happen. They could still happen. On the hand things like Narcotics are not expected to be a problem at all, during this time. Operational Impact: By providing this breakdown, the system enables the police department to deploy specific resources—such as auto-theft task forces—rather than just increasing general patrols, thereby optimizing resource efficiency.

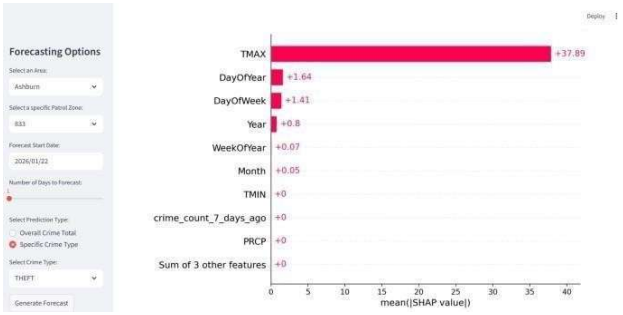

Fig 7: SHAP Feature

D. ExplainableAI (XAI)Integration

Our system is trying to do something with machine learning. We want to make it clearer how it works. So, we use something called SHAP values. These values help us understand each forecast that the system makes. This is important for people like police officers who use our system. They need to know why it says something is a risk. The police officers must trust the system. They can only do that if they understand what is behind each forecast. We use SHAP values for every forecast to make sure the police officers know what is going on with the machine learning system. Analysis of Feature Contributions In this case when we are trying to predict Theft in Ashburn, the SHAP plot shows us some important things about how the Theft prediction model works. The Theft prediction model is trying to figure out what is going on with Theft in Ashburn. The SHAP plot for Theft prediction in Ashburn gives us an understanding of the Theft prediction model. The weather plays a role in this. The maximum temperature is the thing that affects the outcome. It adds a lot of points to the score of thirty-seven points eight nine. This makes sense because when it is hot outside people are more likely to be outside and that means there are more chances for theft to happen. The maximum temperature is what really drives this. It is important to consider when looking at the weather and how it affects theft. Temporal Seasonality is important here. The Temporal Seasonality features like DayOfYear and DayOfWeek are very useful. DayOfYear has a score of +1.64 and DayOfWeek has a score of +1.41. These scores show that Temporal Seasonality features are helping a lot. This means that the model is good at understanding the weekly patterns of crime in this group of people. The model has learned the Temporal Seasonality rhythms of crime that happen at times of the year and week. When we look at things that happened in the past and the amount of rain, something stands out with this prediction. Things like the number of crimes that happened seven days ago, and the amount of Precipitation show that they do not really matter. This means that the model can change, and it ignores things that're not important, for what we are trying to predict right now. It only looks at the things that will make a difference in prediction modellooking at Historical Lags and Precipitation like crime_count_7_days_ago and Precipitation. It is ignoring them because they are not important.

Technical Validation of Specialization: The high SHAP value for TMAX proves that the Specialist Model for this cluster has correctly identified weather as a primary catalyst for crime in this zone. By quantifying these impacts, the system provides why behind what, allowing police departments to justify their deployment strategies based on objective environmental data rather than intuition alone.

V. CONCLUSION

This research has been extremely positive in developing a complex combination of Forecasting Crime Models and has developed "Crime Pattern Clusters using Time Series Forecasting and Geographic Intelligence," and demonstrated clearly that combining techniques significantly improves crime prediction accuracy. By employing the Combination of k-Shape Clustering techniques into our Model Designs, we have been able to identify several distinct patterns of Criminal Behaviour within all six patrolled zones that were S. Alagarsamy, K. Selvaraj, V. Govindaraj, A. A. Kumar, S. Hari Shankar, and G.L. Narasimman, "Automated data analytics approach for examining the background economy of cybercrime," in Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore,India,2021,pp.332–336, doi:10.1109/ICIRCA51532.2021.9544845.created. The cluster designs identified each of the six zones was attained through the development of specialized LightGBMs for predicting specific types of criminal behaviour that correlate with similar types of Criminal Behaviour. Using the results from our trained models allows for accurate and precise estimates of Criminal Behaviour for future analysis based on those models. Using Explainable AI (SHAP) in conjunction with our models allows our trained models to provide easily understood justification for their predictions, thus increasing user confidence

REFERENCES

[1] Ling, H. G., Jian, T. W., Mohanan, V., Yeo, S. F., & Jothi, N. (2024). "Crime Prediction using Machine Learning," Chapter in International Conference on Emerging Technologies and Applications . DOI: 10.1007/978-3-031-62971-9-8.

[2] Paparrizos, J.,& Gravano, L. (2015)."k-Shape: Efficient and Accurate Clustering of Time Series." In Proc. of the 2015 ACM SIGMOD International Conference on Management of Data.

[3] Lundberg, S. M., & Lee, S. I. (2017). "A Unified Approach to Interpreting Model Predictions." In Advances in Neural Information Processing Systems.

[4] Howard, D. A., Jørgensen, B. N., & Ma, Z. G. (2024). "Adata driven approach using time series K-means clustering and forecasting." arXiv, arXiv:2401.04751.

[5] Alsubayhin, A., Ramzan, M. S., & Alzahrani, B. (2024). "Crime prediction modelusing three classification techniques." International Journal of Advanced Computer Science and Applications, 15(1) .

[6] Ke, G., Meng, Q., Wang, T., et al. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." In Advances in Neural Information Processing Systems.

[7] Zhang, B., Panagiotelis, A., & Li, H. (2024)."Constructing hierarchical time series through clustering." arXiv, arXiv:2404.06064.

[8] Fan, Y., Hu, X., & Hu, J. (2025). "Research on a crime spatiotemporal prediction method integrating Informer and ST GCN." Big Data and Cognitive Computing, 9(7), 179 .

[9] Wang, Z., Ma, X., Yang, H., et al. (2024). "Uncertainty aware crime prediction with spatial-temporal multivariate GNNs." arXiv, arXiv:2408.04193.

[10] Iqbal, M. F. B., Ullah, A., Alhomoud, A., et al. (2024). "A 2D clustering based hotspot identification approach." International Journal of Interactive Multimedia and Artificial Intelligence .

[11] Alagarsamy, S., Selvaraj, K., Govindaraj, V., Kumar, A. A., HariShankar, S., & Narasimman, G. L. (2021, September). Automated Data analytics approach for examining the background economy of Cybercrime. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 332-336). IEEE.

[12] Alagarsamy, S., Selvaraj, K., Govindaraj, V., Kumar, A. A., HariShankar, S., & Narasimman, G. L. (2021, September). Automated Data analytics approach for examining the background economy of Cybercrime. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 332-336). IEEE.