Predictive Risk Modeling for Safety Interventions in Transportation Networks Using Spatial Crime History

Nazmus Sakib Pallab
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
npallab2@illinois.edu

Jiarui Yu
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
jiaruiy9@illinois.edu

Favour Jack
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
frjack2@illinois.edu

Muhammad Fahad Ali
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
mali19@illinois.edu

***Abstract*—Integrating personal safety into transportation and pedestrian planning requires systematic use of crime data. Information on crime location, time, and type can be analyzed to identify unsafe streets, intersections, and transit hubs, uncovering vulnerable areas in the urban network. Such insights enable engineers to propose design interventions such as reducing dead-end streets, improving pedestrian connectivity, and strategically relocating public transit drop-off points to enhance safety.**

In this study, raw crime record data will be transformed into actionable hotspot maps and predictive risk models to optimize the allocation of traffic police and patrol routes, ensuring coverage in the areas of highest need. Using advanced machine learning techniques, the study predicts crime types based on factors such as location, time of day, victim profile, and premises description. These results provide Civil Engineers and Urban Planners with evidence-based tools to prioritize infrastructure improvements and safety investments, while also identifying specific locations likely to evolve into future hotspots for proactive deployment of patrols, surveillance, and safety infrastructure.

Together, these predictive and spatial approaches are expected to enhance response efficiency and guide long-term city planning initiatives—from upgrading street lighting and redesigning public spaces to improving transit accessibility and targeting community resources—thereby strengthening the overall resilience and safety of urban infrastructure.

***Index Terms*—Transportation safety, Crime data analysis, Predictive risk modeling, Hotspot mapping, Machine learning, Urban infrastructure planning, Pedestrian safety**

I. Description of Dataset

The dataset used for this project is the ***"Crime Data from 2020 to Present"*** dataset for the City of Los Angeles, which is publicly available on DATA.GOV [1]. It is maintained and released by the Los Angeles Police Department (LAPD) as part of the city's open-data initiative, based on official crime reports filed by law enforcement officers.

The dataset is provided in CSV format and contains over 1 million rows of crime incidents. The full dataset consists of 28 columns, while our project will focus on the following 12 key attributes:

- `DR_NO (Text)`: Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits.
- `Date Rptd (Floating and Timestamp)`: Date the incident was reported.
- `DATE OCC (Floating and Timestamp)`: Date the incident occurred.
- `TIME OCC (Text)`: Time the incident occurred.
- `AREA (Text)`: Area where the incident occurred.
- `AREA NAME (Text)`: ID of the area where the incident occurred.
- `Rpt Dist No (Text)`: A four-digit code that represents a sub-area within a Geographic Area.
- `Vict Age (Text)`: Age of the victim.
- `Vict Sex (Text)`: Sex of the victim.
- `Vict Descent (Text)`: Descent of the victim.
  - `LAT (Number)`: Latitude coordinate of the incident.
  - `LON (Number)`: Longitude coordinate of the incident.

This dataset provides both spatial (latitude/longitude, area, district) and socio-demographic (victim age, sex, descent) attributes, along with temporal information (date and time of crime occurrence), enabling spatial, temporal, and predictive risk modeling for transportation safety interventions.

II. Exploratory Data Analysis (EDA)

A. *Crime Type Distribution*

The Fig. 1 ranks the most common crimes reported across Los Angeles between 2020 and the present.

Vehicle-related crimes (particularly Vehicle Stolen, Burglary from Vehicle, and Theft from Motor Vehicle) dominate the dataset, together accounting for nearly half of all recorded incidents. These are followed by Battery – Simple Assault and Identity Theft, highlighting both property security and personal safety as key urban vulnerabilities.

To identify which areas experience the highest concentration of specific crimes, we generated a heatmap showing the top ten most frequent crime types across all police districts. Each cell represents the number of incidents for a given crime type within an area.

The heatmap reveals that crime intensity is not evenly distributed across the city. Property-related crimes such as Theft, Burglary from Vehicle, and Motor Vehicle Theft dominate the overall dataset but are heavily concentrated in a few areas—particularly Central, 77th Street, and Newton divisions. Conversely, violent offenses like Assault with a Deadly Weapon and Robbery cluster around Southeast and Southwest areas, indicating localized vulnerability.

The color gradients in the heatmap highlight that these high-frequency zones consistently report a wider mix of offenses than low-crime areas, suggesting persistent multi-type crime exposure. This uneven distribution underscores the importance of strategic resource allocation, as new or expanded police stations in these hotspots could improve response times and deter repeated offenses. These findings, combined with temporal and demographic analyses in later sections, provide quantitative evidence for prioritizing areas that face both high crime density and crime diversity, strengthening the case for targeted infrastructure and patrol expansion.

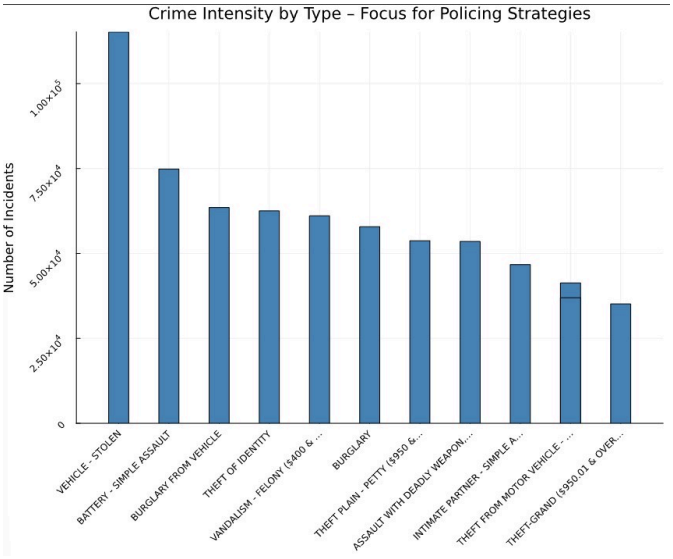

Fig. 1. Top 10 Crime Types in Los Angeles (2020–2024).

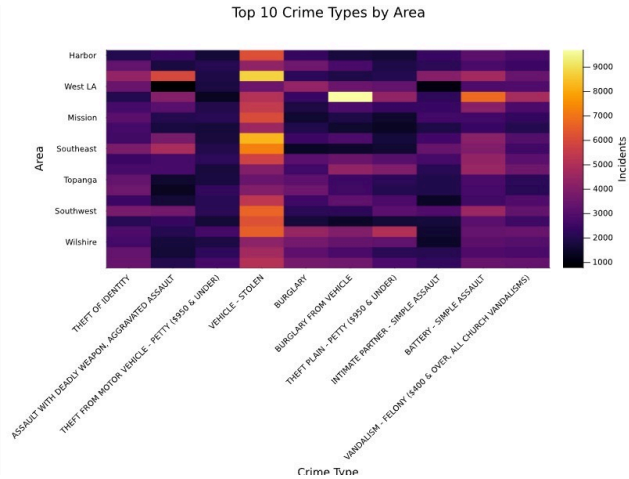

Fig. 2. Top 10 Crime Types in Los Angeles vs. Areas

B. *Temporal Analysis*

(i) Temporal Patterns of Different Crime Categories:

The temporal profile of generalized crime categories reveals that motor vehicle and bicycle theft consistently rank as the most frequent crime types, averaging 30,000–32,000 cases per year. These are followed by simple assault, personal or retail theft, and vandalism, which collectively account for a substantial share of the total crime volume. Aggravated assault and theft from vehicle form the next major tier, reflecting a stable yet diversified pattern of property and personal crimes. While total incident counts remained relatively stable from 2020–2023, a modest uptick was observed in 2022–2023, coinciding with post-pandemic normalization of urban activity. A slight decline in 2024 may partially reflect data latency or reporting lag rather than an actual reduction in crime rates.

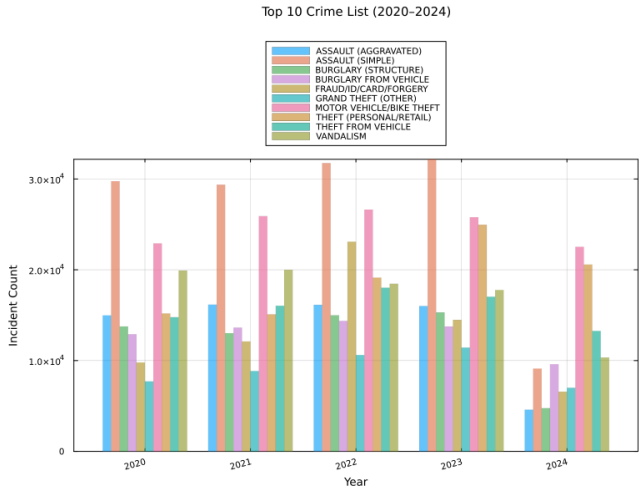

Fig. 3. Top ten generalized crime categories by year (2020–2024).

(ii) Incident Frequency Heatmap:

Incident frequency exhibits a strong diurnal and weekly rhythm, consistent with human activity cycles in an urban environment. As shown in the hourly heatmap, the lowest activity occurs during the early morning hours (03:00–

06:00), followed by a sharp increase after 08:00 that persists throughout the day. Evening hours remain active, peaking around typical commuting and social periods, and the highest overall frequencies are observed on Fridays and weekends. This pattern aligns with nightlife, leisure, and mobility trends, highlighting the influence of temporal human behavior on incident dynamics.

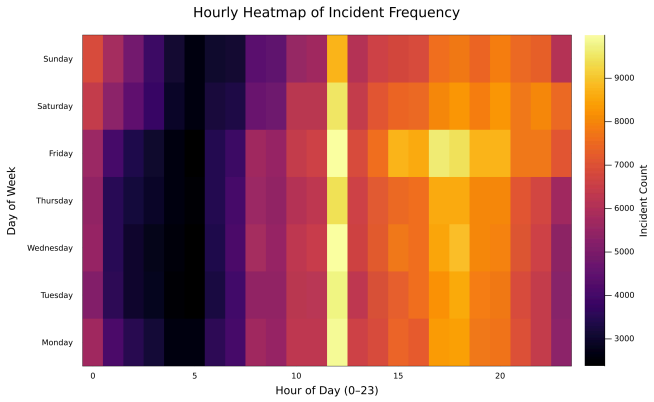

Fig. 4. Hourly heatmap showing diurnal and weekly variations in incident frequency.

(iii) Reporting Delay Distribution:

The analysis of reporting delay is defined as the difference between the date reported and the date of occurrence (Date Rptd − DATE OCC) — reveals a pronounced right-skewed distribution. Most incidents are reported either on the same day or within a few days of occurrence, with a mean delay of 2.96 days and a median delay of 1 day. Approximately 90% of all incidents are reported within five days, indicating generally prompt reporting behavior across most crime categories. The long upper tail in the delay distribution likely reflects crimes with delayed discovery or complex administrative workflows, such as fraud, forgery, or identity-theft–related offenses. This temporal asymmetry underscores the importance of considering both immediate and delayed reporting in operational planning and predictive modeling.

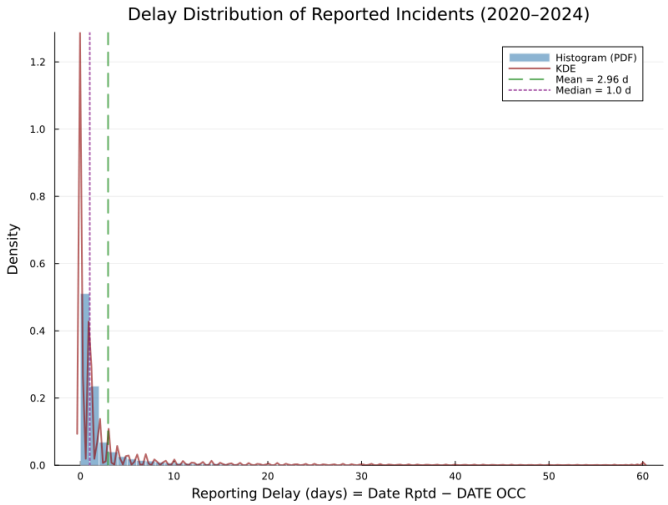

Fig. 5. Distribution of reporting delays (All recorded incidents, 2020–2024).

(iv) Spatial Distribution and Temporal Stability:

Spatially, incident clusters remain highly consistent across the five-year period, concentrating in Downtown, Hollywood, Westlake, and South Los Angeles. These areas exhibit persistent activity regardless of month or year, suggesting enduring socioeconomic and infrastructural factors driving higher incident density. The six representative panels display the months of peak activity for each year between 2020 and 2024. The scatter of points lies almost entirely within the official Los Angeles city boundary, confirming the spatial integrity and proper geocoding of the dataset. Such spatial persistence provides a reliable foundation for hotspot-based predictive modeling and targeted resource deployment.

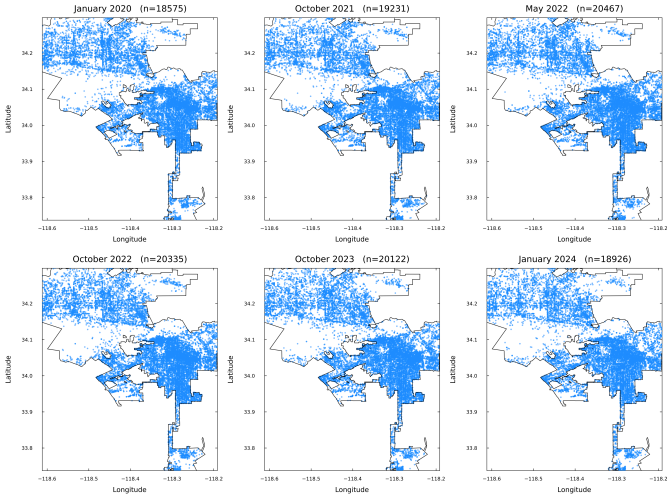

Fig. 6. Annual Peak Months of Incident Distributions (2020–2024)

C. *Spatial Analysis*

Another important aspect of our exploratory data analysis is the spatial distribution of crimes across Los Angeles. By mapping the latitude–longitude of each incident, we visualize hotspots and identify areas with high crime density.

We work with three spatial datasets: the set of crime points $C = \{C_i\}_{\{i=1\}}^{\{N_c\}}$, the set of street-lamp points $L = \{L_l\}_{\{l=1\}}^{\{N_l\}}$, and the city-boundary polygon B. We ensure longitudes and latitudes are numeric, finite, and within plausible ranges so that subsequent geometry remains meaningful. The street-light dataset is sourced from the City of Los Angeles GeoHub [2].

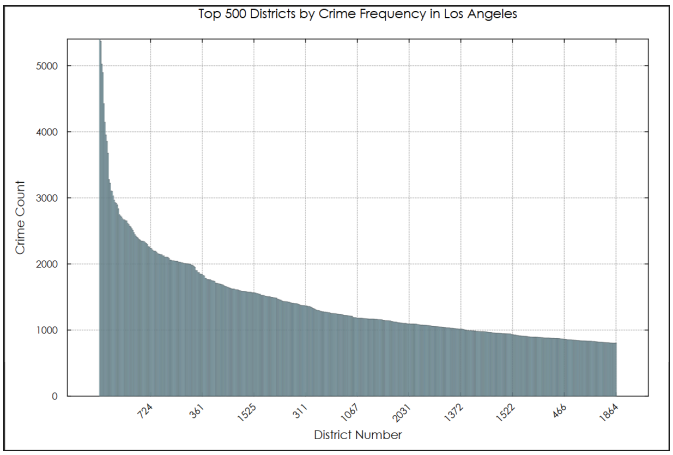

Fig. 7. Top 500 crime count in each designated LAPD district.

Figure Fig. 7 ranks Los Angeles (city) LAPD reporting districts by total reported incidents (2020–present). The x-axis lists districts in descending order (leftmost = highest), and the y-axis shows incident counts. The distribution is heavy-tailed: a few districts concentrate many incidents, followed by a long tail of moderate activity.

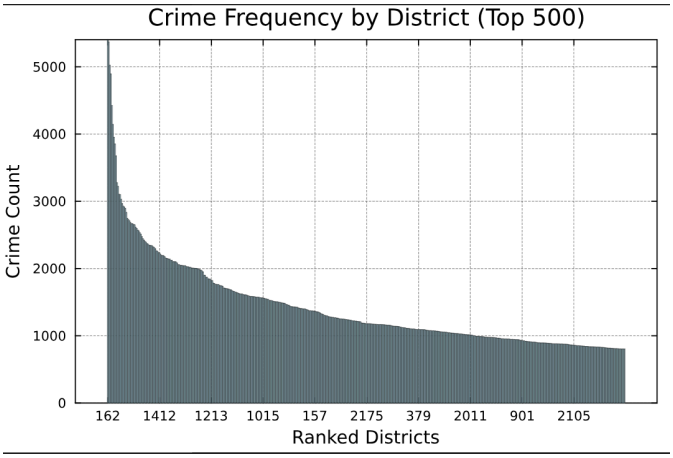

Fig. 8. LA top-500 crime counts (ranked).

Figure Fig. 8 shows the top 500 LAPD reporting districts sorted by total incidents, with the x-axis as rank (left = highest). The y-axis is the incident count. The curve is heavy-tailed: a few districts account for many incidents, followed by a long taper of moderate counts across the remaining ranked districts.

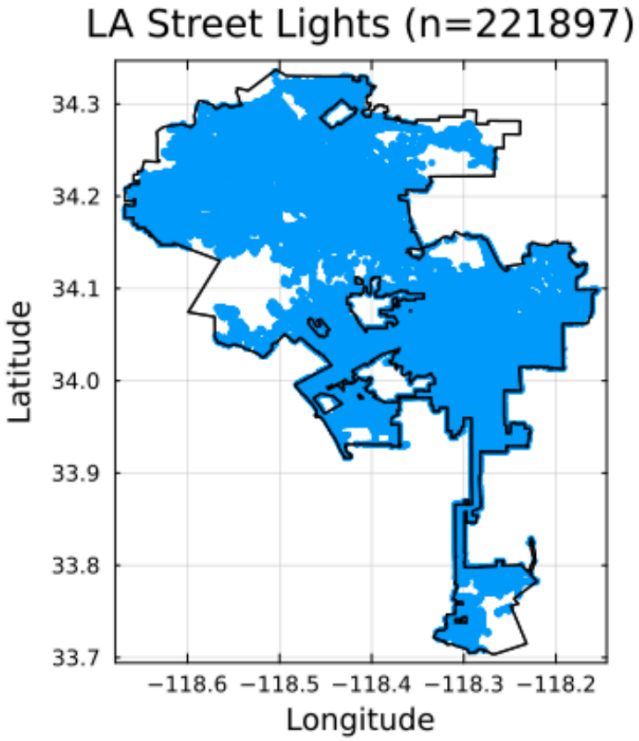

Fig. 9. LA street lights within the city boundary (n=221,897).

This map Fig. 9 plots the reported street-light point locations for Los Angeles in geographic coordinates (longitude/latitude). The high point density makes many neighborhoods appear as solid filled regions, revealing broad coverage across the city and sparser coverage near edges and open spaces. Axes are in degrees to match the source data.

a) *Finding relationship between crime locations and street lights:*

We project all coordinates into a single metric CRS so distances are measured in meters:

$((x_i, y_i) = \Phi(\lambda_i, \varphi_i),\ (u_l, v_l) = \Phi(\lambda'_l, \varphi'_l))$

where:

- (λ_i, φ_i) and (λ'_l, φ'_l) are geographic (lon/lat, degrees) for crime C_i and lamp L_l,
- Φ denotes the projection to a local metric CRS,
- (x_i, y_i) and (u_l, v_l) are projected (planar) coordinates in meters.

A local projected CRS is used because Euclidean lengths in degrees are not physically meaningful, and a single metric CRS avoids unit mismatches.

We restrict the analysis to the jurisdiction by clipping points to the city polygon B, keeping only crimes and lamps whose projected coordinates fall inside B. This removes out-of-area points and reduces boundary artifacts that would otherwise inflate nearest-distance values.

We define planar Euclidean distance between a crime and a lamp:

$$d(C_i, L_l) = \sqrt{(x_i - u_l)^2 + (y_i - v_l)^2} \qquad (1)$$

For each crime, we keep its nearest-lamp distance: $d_i = m \underset{\{1 l e l l e N_l\}}{\in} d(C_i, L_l), q \quad i = 1, ..., n$, where n is the number

of crimes inside B. This yields a single, interpretable proximity value per incident. A spatial index keeps these queries fast.

From $\{d_i\}_{\{i=1\}}^{n}$ we build the empirical cumulative distribution function (ECDF): $F(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{d_i \le r\}, \quad r \ge 0$, so the coverage at an operational radius R is simply: $\text{Coverage}(R) = F(R)$.

We report robust summaries of proximity (median) and tail behavior (p90, p99) along with the maximum distance. Quantiles are preferred over the mean because the distribution is typically right-skewed and outlier-sensitive.

To guard against faulty records, we flag implausible distances above a conservative cap and, if needed, compute robust z-scores using the median and MAD to identify unusual d_i values.

Finally, to see how coverage accumulates with distance, we partition r into analyst-chosen bands (e.g., 0–50–100–250 m, …) and tabulate the share of crimes whose nearest-lamp distance falls into each band. This "coverage by band" view highlights where most gains occur (very small radii) and where diminishing returns set in as the radius grows.

Table II summarizes the distance from each reported crime in Los Angeles to its nearest street light. It shows total valid records, distribution percentiles (p25–p99), the maximum, and how many / what share fall within the working radius **R = 100 m**. Most crimes are close to a light, with a long right tail driven by a small set of far-out points. More such comparison will be carried out based upon availability of the data.

III. Spatial Crime-Risk Modelling with Random Forests and Spatial Accessibility

Point-level crime incidents come from the LAPD "Crime Data from 2020 to Present" dataset [1], combined with a detailed inventory of street-light locations from the Bureau of Street Lighting [2]. Accessibility features are constructed using countywide layers on mental health centers [3], food assistance providers [4], and public libraries [5], together with parks and open space data at both the county [6] and city [7] levels. School district boundaries [8] and LAPD community police station locations [9] provide additional institutional context. Following analysis was performed.

A. 1. Data, Notation, and Study Region

a) *1.1 Crime and Non-Crime Points:*

Let each location be represented by geographic coordinates:

$$p_i = (\text{lat}_i, \text{lon}_i) \in \mathbb{R}^2, i = 1, ..., N. \tag{2}$$

Each point has a binary crime label $y_i \in \{0, 1\}$, where

- $y_i = 1$ if a crime is observed at p_i,
- $y_i = 0$ if the point is treated as a non-crime location.

The final analysis dataset contains:

- N = 2,009,982 total records,
- 1,004,991 crime locations ($y_i = 1$),
- 1,004,991 non-crime locations ($y_i = 0$),

so that the crime proportion is

$$\frac{1}{N} * \sum_{i=1}^{N} y_i = 0.50 \quad (50\% \text{ crimes, } 50\% \text{ non-crimes}). \tag{3}$$

b) *1.2 Generating Non-Crime Locations:*

The original crime dataset provides only locations where crime occurred. To train a binary classifier, we need contrasting "non-crime" points. The notebook generates these using a mixture of two strategies:

1) **Nearby perturbations** around crime points (simulating "similar" areas where crime did not happen).
2) **Random points** uniformly spread across the city's bounding box.

Formally, let N_c be the number of original crime locations. We create:

- $N_{\text{near}} = \lfloor 0.7 N_c \rfloor$ nearby points, and
- $N_{\text{rand}} = N_c - N_{\text{near}}$ random points.

For a subset of crime locations p_i, nearby non-crime points are generated as

$$p_i^{(\text{near})} = p_i + \Delta p_i, \tag{4}$$

where Δp_i is a random perturbation with typical magnitude

$$\|\Delta p_i\| \approx 0.01°, \tag{5}$$

corresponding to roughly 1.1 km in latitude.

Random points are sampled uniformly inside the geographic bounds:

$$\text{lat}_{\min} \le \text{lat} \le \text{lat}_{\max}, \quad \text{lon}_{\min} \le \text{lon} \le \text{lon}_{\max}. \tag{6}$$

c) *1.3 Geographic Coverage:*

The combined dataset spans:

- Latitude range: 0.0° to 34.3343°,
- Longitude range: −118.6676° to 0.0°.

The approximate spatial extent is computed using the standard "111 km per degree" approximation:

- North–south size:

$$L_{\text{NS}} \approx (\text{lat}_{\max} - \text{lat}_{\min}) * 111 \ \text{ km} \approx 3811.1 \ \text{ km} \tag{7}$$

- East–west size:

$$L_{\text{EW}} \approx (\text{lon}_{\max} - \text{lon}_{\min}) * 111 \cos(|(\text{lat})) \ \text{ km} \approx 12585.8 \text{ km}, \tag{8}$$

where $|(\text{lat})$ is the mean latitude of the area.

This is a conceptual bounding box; the actual points of interest are concentrated near the Los Angeles region.

B. 2. External Spatial Datasets and KD-Trees

The notebook loads several facility layers, each represented as a set of points:

- Mental health centers,
- Food assistance providers,
- Public libraries,
- County parks,
- City parks,
- Metro lines / stations (where coordinates are available),
- Police stations.

For each facility type k, we denote its locations as

$$F^k = \left\{ f_j^k \in \mathbb{R}^2 : j = 1, ..., M_k \right\}. \tag{9}$$

To efficiently query distances and neighbours, each F^k is indexed with a KD-tree (using the

`NearestNeighbors.jl` package). A KD-tree enables fast queries of the form:

- **Nearest neighbour**:

$$f_{\text{nn}}^k(p) = \text{argmin}_{f \in F^k} \| \, p - f \, \|_2, \tag{10}$$

- **Radius search** (all facilities within distance r):

$$\{f \in F^k : \| \, p - f \, \|_2 \leq r\}. \tag{11}$$

The notebook's `build_kdtree` function converts latitude/longitude columns into a 2×M_k matrix and builds `KDTree(coords, Euclidean())`.

C. 3. Spatial Accessibility Features

For each crime or non-crime point p_i, and each facility type k, the notebook computes two families of features:

1) **Nearest distance**:

$$d_i^k = \min_{1 \leq j \leq M_k} \| \, p_i - f_j^k \, \|_2. \tag{12}$$

2) **Count within a radius r**:

$$c_i^k(r) = \sum_{j=1}^{M_k} \mathbf{1}\big(\| \, p_i - f_j^k \, \|_2 \leq r\big), \tag{13}$$

where $\mathbf{1}(\cdot)$ acts as an indicator function (1 if the condition holds, 0 otherwise).

The code implements this via:

- `knn(tree, query_coords, 1, true)` for nearest distances,
- `inrange(tree, query_point, radius, false)` to count facilities within radius.

The chosen radius is

$$r = 0.01° \approx 1.1 \ \text{km}, \tag{14}$$

interpreted as a local neighbourhood scale.

For seven facility types, the total number of engineered features is:

- 7 nearest-distance features, and
- 7 radius-count features,

for a total of 14 spatial features:

$$p = 14. \tag{15}$$

In Typst-style notation, the feature vector for point p_i can be written as:

$$x_i = \left(d_i^{\text{mh}}, c_i^{\text{mh}}, d_i^{\text{food}}, c_i^{\text{food}}, ..., d_i^{\text{police}}, c_i^{\text{police}}\right) \in \mathbb{R}^{14}. \tag{16}$$

The `augment_with_spatial_features!` function takes a DataFrame with `latitude`, `longitude`, and `crime`, and **mutates** it in-place by adding these 14 columns.

D. 4. Correlation Analysis and Descriptive Statistics

Before modelling, the notebook computes the Pearson correlation between each spatial feature and the crime label.

Given a numeric feature vector $X = (X_1, ..., X_N)$ and labels $Y = (Y_1, ..., Y_N)$ with $Y_i \in \{0, 1\}$, the Pearson correlation coefficient is

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{17}$$

The function `analyze_correlations`:

- Selects all columns whose names contain `_nearest_dist` or `_count_radius`,
- Replaces `Inf` values (e.g., if no facility is found) with a large constant 10^{10},
- Computes ρ between each feature and the crime indicator,
- Sorts features by absolute correlation.

Additionally, the function `summary_by_crime` compares **mean feature values** between crime and non-crime groups:

- Mean among crime points:

$$\mu_{\text{crime}}(X) = \frac{1}{N_1} \sum_{i : y_i = 1} X_i, \tag{18}$$

- Mean among non-crime points:

$$\mu_{\text{no-crime}}(X) = \frac{1}{N_0} \sum_{i : y_i = 0} X_i. \tag{19}$$

This helps interpret whether, for example, crime locations tend to be closer or farther from certain facilities than non-crime locations.

E. 5. Machine-Learning Dataset Construction

We now build the design matrix X and label vector y for supervised learning.

Let

$$X \in \mathbb{R}^{N \times p}, \quad y \in \{0, 1\}^N. \tag{20}$$

The `prepare_ml_data` function:

1) Selects all spatial features and any existing lighting features (columns containing `"light"`).
2) Replaces infinite values with a large constant (e.g. 10^{10}).
3) Imputes any NaNs with the feature median.
4) Encodes the target as string labels `"0"` or `"1"` for use with `DecisionTree.jl`.
5) Creates a random train–test split with test ratio $\alpha = 0.2$.

Given $N = 2,009,982$ and $\alpha = 0.2$,

- Training size:

$$N_{\text{train}} = (1 - \alpha)N = 0.8N = 1,607,986, \tag{21}$$

- Test size:

$$N_{\text{test}} = \alpha N = 401,996. \tag{22}$$

Class counts:

- Train: 803,826 crimes, 804,160 non-crimes,
- Test: 201,165 crimes, 200,831 non-crimes.

The dataset is therefore close to perfectly balanced in both train and test splits.

F. 6. Random Forest Classifier

The model is a Random Forest classifier trained on the feature matrix X to predict the binary label y.

a) *6.1 Single Decision Tree:*

A single decision tree recursively partitions the feature space $\mathbb{R}^p$ into axis-aligned regions by splitting on conditions of the form

$$x_j \leq \tau, \tag{23}$$

where x_j is feature j and τ is a threshold. At each node, the algorithm chooses a feature and threshold to reduce impurity (e.g. Gini impurity or entropy).

For a node with class probabilities (p_0, p_1), the Gini impurity is:

$$I_{\text{Gini}} = 1 - p_0^2 - p_1^2. \tag{24}$$

b) *6.2 Random Forest Ensemble:*

A Random Forest builds an ensemble of T trees $\left\{h_{t(x)}\right\}_{t=1}^{T}$ using:

- Bootstrap samples of the training data,
- A random subset of features (of size $m_{\text{sub}} \approx \sqrt{p}$) considered at each split.

Given the notebook's configuration:

- Algorithm: Random Forest,
- Number of trees: $T = 100$,
- Maximum depth per tree: $\max\_\text{depth} = 10$,
- Number of features: $p = 14$,
- Subset size at each split: $m_{\text{sub}} = \max(1, \lfloor\sqrt{p}\rfloor) = 3$.

For a new feature vector x, each tree outputs a class prediction $h_{t(x)} \in \{0, 1\}$, and the forest prediction is the majority vote:

$$\hat{y}(x) = \text{mode}\left\{h_{t(x)} : t = 1, ..., T\right\}. \tag{25}$$

The probabilistic prediction used for ROC analysis is

$$\hat{p}(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left(h_{t(x)} = 1\right), \tag{26}$$

i.e. the fraction of trees voting for class 1 (crime).

c) *6.3 Feature Importance:*

`DecisionTree.impurity_importance(model)` returns the **impurity-based feature importance**. For feature j, the importance score I_j is the sum over all splits that use feature j of the decrease in impurity, weighted by the number of samples that pass through the split.

The importances are normalised so that

$$\sum_{j=1}^{p} I_j = 1. \tag{27}$$

The top 5 most important features in the final model are:

1) `mental_health_nearest_dist`: 36.01%
2) `food_assistance_nearest_dist`: 17.31%
3) `county_park_nearest_dist`: 16.45%
4) `police_nearest_dist`: 14.66%
5) `library_nearest_dist`: 10.56%

Six of the 14 features have zero importance (never used in any split), suggesting they do not contribute to the forest's decisions.

G. 7. Evaluation Metrics: Accuracy, Precision, Recall, ROC

Let the test set contain N_{test} instances with true labels $y_i \in \{0, 1\}$ and predicted labels $\hat{y}_i$.

We define:

- **True Positives (TP)**: $y_i = 1$ and $\hat{y}_i = 1$,
- **True Negatives (TN)**: $y_i = 0$ and $\hat{y}_i = 0$,
- **False Positives (FP)**: $y_i = 0$ but $\hat{y}_i = 1$,
- **False Negatives (FN)**: $y_i = 1$ but $\hat{y}_i = 0$.

From the notebook's final confusion matrix on the test set:

- TP = 195,396
- TN = 70,697
- FP = 130,134
- FN = 5,769

The key performance metrics are:

1) **Accuracy**:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \approx 66.19\% \tag{28}$$

2) **Precision** (positive predictive value):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \approx 60.02\%. \tag{29}$$

3) **Recall** (true positive rate, TPR):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \approx 97.13\%. \tag{30}$$

Thus, the model catches most crime locations (very high recall) but at the cost of a substantial number of false positives (only moderate precision).

a) *7.1 ROC Curve and AUC:*

The ROC (Receiver Operating Characteristic) curve is constructed by varying the decision threshold θ on the predicted probability $\hat{p}(x)$ and, for each threshold, computing:

- $\text{TPR}(\theta) = \text{Recall}$,
- $\text{FPR}(\theta) = \frac{\text{FP}(\theta)}{\text{FP}(\theta) + \text{TN}(\theta)}$.

The Area Under the Curve (AUC) is approximated numerically via the trapezoidal rule:

$$\text{AUC} \approx \sum_{k=2}^{K} [\text{FPR}_k - \text{FPR}_{k-1}] * \frac{\text{TPR}_k + \text{TPR}_{k-1}}{2}. \tag{31}$$

The function `compute_roc` implements this calculation explicitly.

H. 8. Spatial Prediction and Crime Risk Heatmaps

To evaluate the learned model at an arbitrary location (lat, lon), the function `predict_at_location`:

1) Constructs a query point $q = (\text{lon}, \text{lat})$.
2) For each facility type k, computes:
   - the nearest distance $d^k(q)$ via `knn`,
   - the count $c^k(q; r)$ via `inrange`.
3) Builds a feature dictionary and aligns it with the trained model's `feature_names`.
4) Replaces any `Inf` distances with 10^{10}.
5) Calls `apply_forest_proba(model, feature_vec, classes)` to obtain class probabilities.

The returned value is the probability

$$\hat{p}(\text{crime} \mid q) = P(y = 1 \mid \text{features at } \ q). \tag{32}$$

The `predict_grid` function evaluates this probability over a regular grid:

$$\text{lat}_1, ..., \text{lat}_R; \quad \text{lon}_1, ..., \text{lon}_C, \tag{33}$$

and stores probabilities in a matrix

$$P_{ij} = \hat{p}\big(\text{crime} \mid \text{lat}_i, \text{lon}_j\big), \tag{34}$$

which is then visualised with a heatmap in `plot_heatmap` as a **crime risk map**.

I. 9. Comprehensive Summary and Result Table

The notebook prints a comprehensive summary of the data, features, model configuration, and performance. Key points:

- **Data overview** – Total records: 2,009,982
  - Crime: 1,004,991 (50.0%)
  - Non-crime: 1,004,991 (50.0%)
- **Feature engineering** – Total engineered features: 14
  - Distance features: 7
  - Count features: 7
  - Search radius: 0.01° ≈ 1.1 km.
- **Train/test split** – Training: 1,607,986 records (80%)
  - Crime: 803,826
  - Non-crime: 804,160
  - Test: 401,996 records (20%)
    - Crime: 201,165
    - Non-crime: 200,831
- **Model configuration** – Algorithm: Random Forest
  - Trees: 100
  - Max depth: 10
  - Features used: 14
- **Performance** – Training accuracy: 66.15%
  - Test accuracy: 66.19%
  - Precision: 60.02%
  - Recall: 97.13%
  - 6 features with zero importance.

a) *9.1 Results Summary Table:*

Category	Metric	Value / Comment
Data	Total records	2,009,982
Data	Crime locations	1,004,991 (50.0%)
Data	Non-crime locations	1,004,991 (50.0%)
Geography	Latitude range	0.0° to 34.3343°
Geography	Longitude range	−118.6676° to 0.0°
Geography	Approx. coverage	≈ 3811.1 km × 12585.3 km
Features	Total features	14 (7 distances, 7 counts)
Features	Search radius	0.01° ≈ 1.1 km
Split	Training set size	1,607,986 (80.0%)
Split	Test set size	401,996 (20.0%)
Model	Algorithm	Random Forest (100 trees, depth 10)
Model	Features used	14, with 6 zero-importance
Performance	Train accuracy	66.15%
Performance	Test accuracy	66.19%
Performance	Precision (test)	60.02%
Performance	Recall (test)	97.13%
Importance	1st: mental_health	36.01%
Importance	2nd: food_assistance	17.31%
Importance	3rd: county_park	16.45%
Importance	4th: policet	14.66%
Importance	5th: library	10.56%

J. 10. Conclusions and Insights

- The model trades precision for recall: it correctly flags most crime locations (recall ≈ 97%) but produces many false positives (precision ≈ 60%).
- Distance-based accessibility to mental health centers, food assistance, county parks, police stations, and libraries carries the strongest signal for distinguishing crime vs non-crime points in this setup.
- Several derived features are unused, suggesting opportunities for feature reduction, alternative feature engineering (e.g., non-linear transformations, interaction terms), or different models.

The ROC curve in Fig. 11 shows that the model has a reasonably strong ability to distinguish crime from non-crime locations: the blue ROC line sits well above the grey random-guess baseline and yields an AUC of about 0.76, meaning that in roughly three out of four randomly chosen crime/non-crime pairs the model assigns a higher risk to the crime point. The crime risk map in Fig. 10 translates these probabilities into space, displaying predicted crime risk over a latitude–longitude grid: each cell's colour encodes the model's estimated probability of crime, with greener cells indicating higher risk and red–orange cells indicating lower risk. Together, Figures Fig. 11 and Fig. 10 show that the classifier is meaningfully better than random and that it identifies coherent high-risk clusters and lower-risk areas across the Los Angeles region rather than uniform risk everywhere.

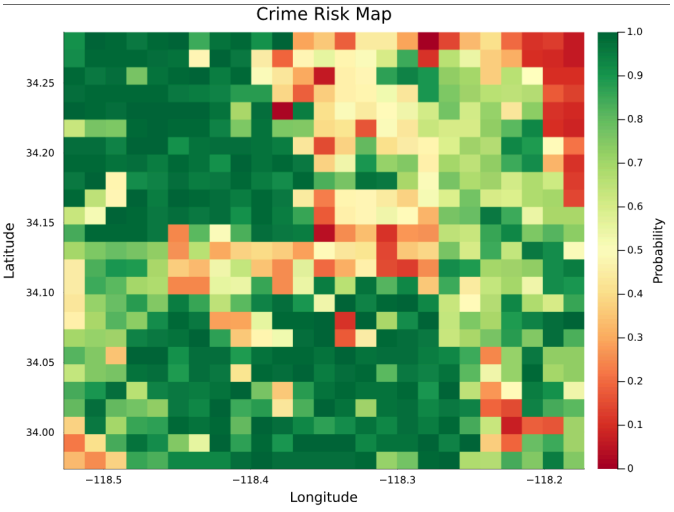

Fig. 10. Crime risk map over Los Angeles generated from Random Forest predictions. Colours show the estimated crime probability on a longitude–latitude grid (green = higher risk, red = lower risk).

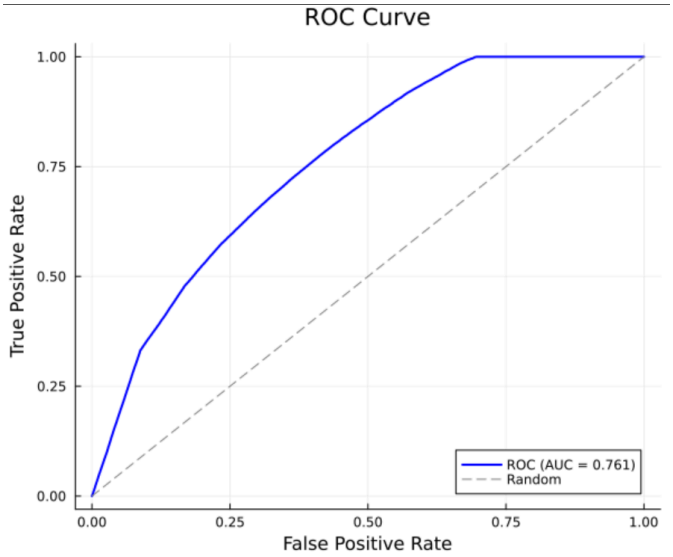

Fig. 11. Receiver operating characteristic (ROC) curve for the Random Forest crime classifier on the test set. The blue line shows the trade-off between true positive rate and false positive rate, and the dashed line is the random-guess baseline. The area under the curve (AUC) is approximately 0.76.

K. *Demographic Anlysis*

Besides analyzing crime types and its patterns over time and space, exmining the demographic characteristics of crime victims might provide some useful sights. We analyzed age, sex, and descent compositions of victims. Overall, the victim population is 40.19% male, 35.68% female, and 24.13% unknown or missing. The age distribution of victims is shown in Fig. 12. It shows that the age group of 30-34 has the highest number of victims, followed by the age group of 25-29. The descent distribution of victims is shown in Fig. 13. It shows that the major victim descent groups are Hispanic/ Latin/Mexican, White, and Black, with the percentages of 34.45%, 23.41%, and 15.79%, respectively.

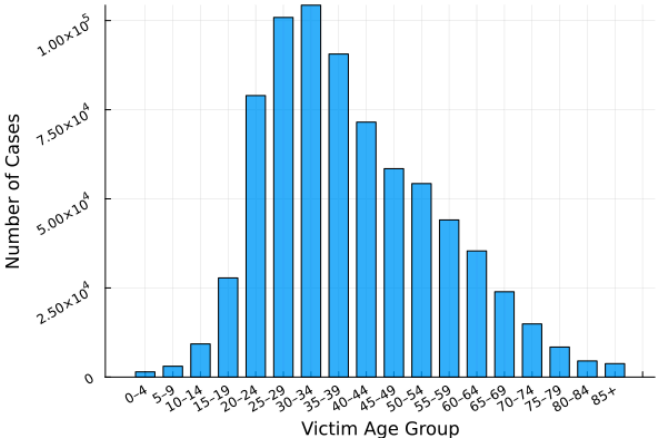

Fig. 12. Age Distribution of Crime Victims

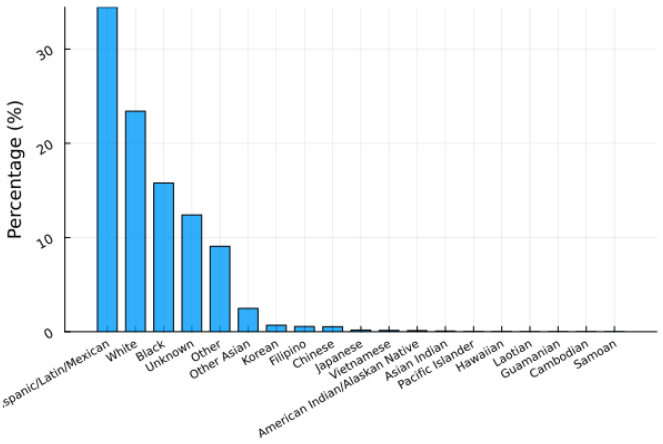

Fig. 13. Descent Distribution of Crime Victims

To see the correlation between age and descent, we created a heatmap shown in Fig. 14. The pattern of age among all descent groups is similar, with the age group of 25-34 having the highest number of victims across all descent groups.

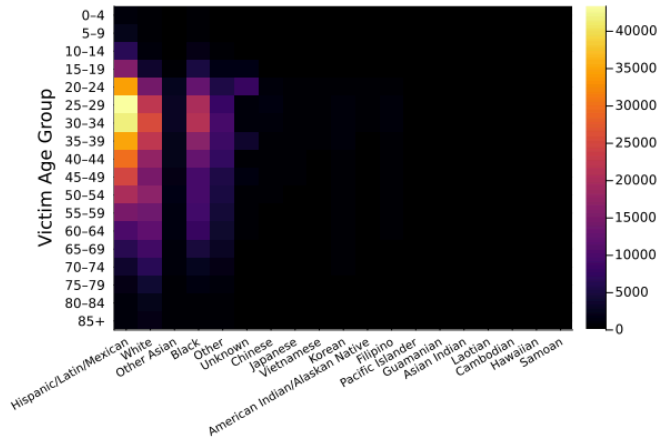

Fig. 14. Heatmap of Victim Age vs. Descent

L. *Summary Statistics of Dataset*

TABLE I
OVERVIEW OF KEY STATISTICS OF CRIME DATA FROM 2020 TO 2024

Metric	Value
Total records (2020–2024)	≈ 1,004,991 incidents
Average monthly incidents	≈ 20,000–23,000
Distinct crime types	≈ 140
Generalized crime types	37 aggregated categories
Earliest date of Dataset	2020-01-01
Latest date of Dataset	2024-03-31
Mean Hotspot for Crime (lat / lon)	34.05 / −118.32 (≈ Downtown LA)
Mean reporting delay after incident	2.96 days (mean), 1 day (median)
Victim Sex Composition	40.19% male, 35.68% female
Largest Victim Age Group	30–34 years
Top Victim Descent	Hispanic/Latin/Mexican

TABLE II
LA DISTANCE-TO-LIGHT SUMMARY.

Metric	Value
N (valid)	1,004,996
min (m)	0.0852
mean (m)	25,702.8
std (m)	5.43e3
p25 (m)	10.8821
median (m)	14.2868
p75 (m)	20.8421
p90 (m)	81.4454
p95 (m)	152.942
p99 (m)	342.602
max (m)	1.15167e7
≤R (count)	922,403
≤R (%)	91.7822

IV. PREDICTIVE MODELING

A. *Crime Type Prediction with Decision Tree Method*

*i. **Crime Type Prediction with Demographic Features**:*

We first examined the relationship between demographic features of victims and the crime types happened on them. The characteristics of victims includes age, gender, and descent. After extracted necessary data from the original dataset, we set victim age, victim gender, and victim descent as features, and crime type as labels to train a decision tree classifier model. After organizing and encode data, the gender feature includes male and female; the age feature includes 9 age groups; the descent feature keeps the original categories; and the crime type includes top 20 types. 80% of the data was used for training and the remaining 20% was used for testing. After apply the standard decision tree classifier from DecisionTree.jl package, the accuracy of the model on the test data is 16.14%, which indicates low correlation between demographic features and crime types. Besides, only two features were used as features in the model in pairs (i.e. age and gender; age and descent; gender and descent), the accuracy of the model are 12.48%, 15.06%, and 13.45%, respectively. The results shows that either the decision tree method is not suitable for predicting crime types, or the demographic features of victims are not strongly related to the crime types happened on them.

*ii. **Crime Type Prediction with Temporal-Spatial Features**:*

Then, we tried to figure out if crime types are relevant to the crime time and location. In the dataset, the region is devided into 21 areas, and the exact hour of the crime are recorded. We set area index and rounded the exact time to 24 hours as feartures for the decision tree, tope 20 crime types as labels. Similarly, 80% of the data was used for training and the remaining 20% was used for testing. Under the settings that the maximum depth is 10, the minimum sample leaf is 50, and the minimum sample split is 100, the minimum purity increase is 0.001 for the decision tree, the accuracy of the model on the test data is 16.75%, which doesn't show significant relationship between crime types and temporal-spatial features either. Besides, the parameters of the decision tree model were tuned to find the best accuracy. However, the accuracy is always lower than 17%. Therefore, we can conclude that either the decision tree method is not ideal for predicting crime types for our dataset, or the crime types are not strongly relevant to both temporal-spatial features and demographic features of victims.

B. *Temporal–Spatial Crime Hotspot Prediction Model*

Motivated by urban safety and resource allocation in Los Angeles, we focused on the following predictive question:

Can we predict which spatial crime hotspot an incident will occur in using time of occurrence (hour of day, day of week, and month)?

This question is very important from a civil and environmental engineering perspective because police, EMS, and traffic management agencies often require time-of-day deployment strategies without knowing the exact location of future incidents. If reliable temporal–spatial patterns exist, agencies could proactively position resources within a select number of hotspot regions during specific time windows.

*i. **Spatial Clustering of Crime Locations (Unsupervised)**:*

To forecast crime hotspots in Los Angeles, we developed a predictive modeling framework that integrates unsupervised

spatial clustering with supervised softmax classification. The objective is to predict which hotspot (cluster) is most likely to experience crime under specific temporal and categorical conditions such as month, day of week, hour, crime category, and LAPD geographic area.

We applied k-means clustering to three years of historical crime data (2020–2022) using only geographic coordinates (latitude and longitude). After experimentation, we selected $K = 8$ clusters, which produced meaningful and well-distributed hotspot regions across Los Angeles.

For crimes occurring in the test years (2023–2024), each incident was assigned to the nearest cluster centroid using Euclidean distance. This ensures generalization and avoids information leakage, since test points were never used to form the clusters.

This stage is fully unsupervised because the hotspot labels emerge solely from underlying spatial density patterns.

ii. ***Predicting Hotspot (Supervised)****:*

After generating cluster IDs, I treated them as labels for a supervised classifier. The following predictive features were used:

- hour of day
- day of week
- month
- type of crime
- LAPD reporting area

Numeric features were normalized, and categorical features were consistently encoded using training-set levels. A softmax neural network classifier was trained using cross-entropy loss and gradient descent, with accuracy and loss tracked across iterations.

iii. ***Evaluation and Performance****:*

The dataset was split chronologically:

- Training: 2020–2022
- Testing: 2023–2024

Model performance:

- Training accuracy: 87%
- Test accuracy: 88%
- Cross-entropy loss: steadily decreased and stabilized

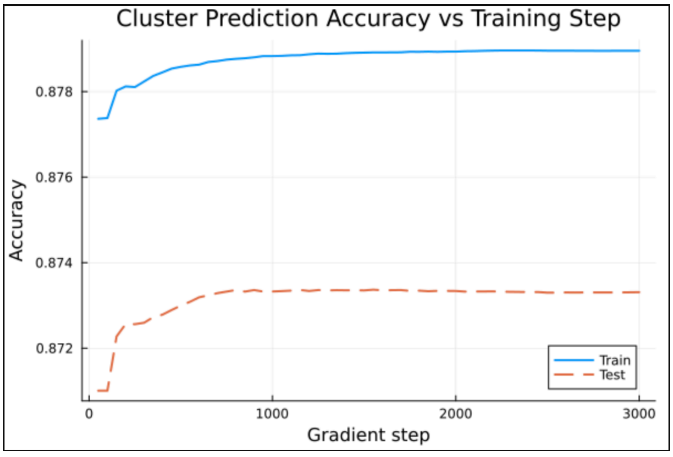

Fig. 15. Training and testing accuracy over gradient steps.)

Compared to the baseline accuracy of 1/8 = 12.5% (random guessing), the classifier shows strong predictive capability. Learning curves indicate stable convergence with no signs of overfitting.

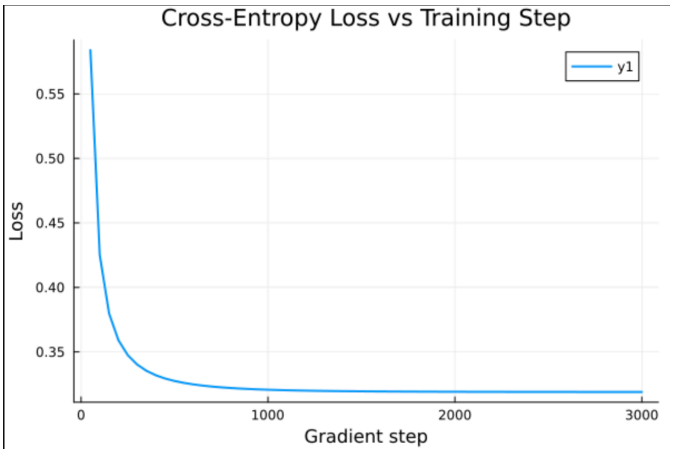

Fig. 16. Cross-entropy loss during training

We also observed loss decreases sharply during the first 300 gradient updates, indicating rapid learning of the major temporal–spatial decision boundaries from the training data. After this initial phase, the loss continues to decline more slowly and eventually stabilizes around 0.32, demonstrating smooth and monotonic convergence without instability or oscillation. This behavior confirms that the learning rate and optimization setup are well-chosen, and that the model successfully minimizes classification error as it adapts to the training patterns. The absence of sudden spikes further suggests that the model is not overfitting to rare or noisy samples.

iv. ***Hotspot Visualization****:* To improve interpretability, I generated multiple geographic visualizations that overlay model predictions on the official Los Angeles boundary shapefile. These figures demonstrate how predicted hotspot regions shift depending on the temporal query.

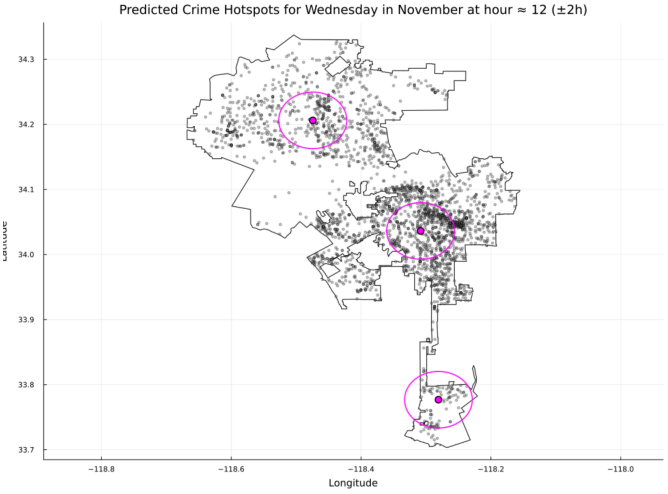

Fig. 17. Spatial distribution of predicted crime hotspots for a sample temporal query (Wednesday, November, 12:00 ±2h)

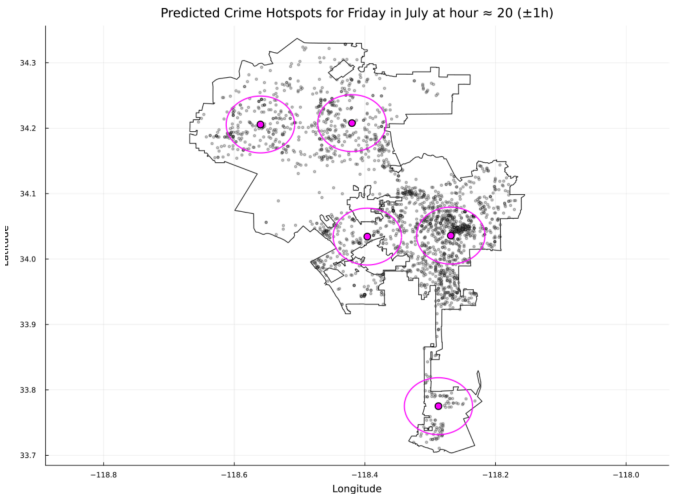

Fig. 18. Spatial distribution of predicted crime hotspots for a sample temporal query (Friday, July, 20:00 ±1h)

For each user-defined query (e.g., "Fridays in July at 20:00 ± 1 hour" or "Wednesdays in November at 12:00 ± 2 hours"), the visualizations display:

- historical crime points within the specified temporal window (gray)
- the LA city boundary outline
- the model's top predicted hotspot centroids (magenta points)
- approximately 1-mile radius highlight zones around each centroid (magenta circles)

The first visualization corresponds to Friday in July around 20:00 (±1 hour) and shows multiple active hotspots distributed across central and southern Los Angeles.

The second visualization corresponds to Wednesday in November around 12:00 (±2 hours) and reveals a different spatial pattern with fewer but more concentrated predicted hotspots.

Together, these visualizations illustrate how the model adapts hotspot predictions to specific time-of-day and seasonal contexts, providing interpretable, actionable spatial insights for operational decision-making.

*v. **Future Works**:*

Future improvements may include incorporating additional features such as weather, special events, socioeconomic indicators, or crime-type interactions. Alternative clustering approaches (e.g., DBSCAN, Gaussian Mixture Models) could capture more flexible hotspot shapes. In the supervised stage, deeper neural networks or ensemble models may further enhance accuracy. Deploying this framework as an interactive real-time tool would expand its usefulness for operational planning and situational awareness.

References

[1] Los Angeles Police Department / LAPD OpenData, "Crime Data from 2020 to Present." data.lacity.org / Data.gov, 2025.
[2] City of Los Angeles Bureau of Street Lighting (BSL), "Street Lights (Feature Layer)." City of Los Angeles GeoHub, 2025.
[3] County of Los Angeles Department of Mental Health, "Mental Health Centers." County of Los Angeles Open Data, 2025.
[4] County of Los Angeles / Food Oasis LA, "Food Assistance." County of Los Angeles Open Data, 2025.
[5] County of Los Angeles Public Library, "Library Buildings 2023." County of Los Angeles Open Data, 2025.
[6] Los Angeles County Department of Parks and Recreation, "Countywide Parks and Open Space (Public – Hosted)." LA County eGIS / Open Data, 2025.
[7] City of Los Angeles Department of Recreation and Parks, "Los Angeles City Parks Boundaries." City of Los Angeles GeoHub, 2025.
[8] County of Los Angeles, "School District Boundaries." LA County eGIS / ArcGIS Hub, 2025.
[9] City of Los Angeles Police Department, "LAPD Police Stations." City of Los Angeles GeoHub, 2025.