AI-Driven Pattern DNA & GEO-Spatial DRIFT Analysis for Missing Person Cases in India

Dr. Satyawati S. Magar [1], Dr. Sandip R. Udawant [2]

[1] Associate Professor, E & TC, DVVP COE, Ahilyanagar
[2] Associate Professor, E & TC, DVVP COE, Ahilyanagar

Abstract

This study introduces an artificial intelligence- based analytical model created to identify patterns, anomalies, and leads in long-standing open missing persons cases in India. By utilizing publicly available demographic, geographic, and temporal data from the Missing People Dataset (Kaggle), NCRB data, and district reports, the system employs several analytical layers: Pattern DNA clustering to reveal hidden demographic trends, Geo-Spatial Drift Analysis to predict potential movement areas using spatial attributes and logistic regression, and a Semantic Case Matching module that uses BERT embedding to connect cases with similar narratives for cold case support. The pipeline also features a risk prediction module based on time-series models like Prophet to forecast upcoming peaks in missing persons cases by state and time intervals.

Extensive data preprocessing, exploratory data analysis, feature creation, and clustering were performed using Python. The resulting geo-risk heatmaps, similarity scores, and trend forecasts offer actionable insights to aid investigators in resource allocation, hotspot identification, and case prioritization. This framework demonstrates how AI can improve investigative efforts and highlights the potential for future integration with law enforcement databases and humanitarian programs, further emphasizing the value of technology-driven solutions for addressing key social issues.

Keywords: *Missing Person Analysis, Artificial Intelligence, Pattern Clustering, Geo-Spatial Analysis, Semantic Case Matching, Natural Language Processing, BERT embedding, Time Series Forecasting.*

1. Introduction

Missing persons cases in India have been consistently on the rise year after year, with several of these going unsolved for lengthy periods. These are actual people and families who are denied closure or resolution. With continued efforts by law enforcement bureaus, hurdles like disorganized data, disparate systems, and inadequate technological assistance prevent meaningful analysis and pattern identification.

To overcome these challenges, this paper introduces an AI- based analytical framework that incorporates a combination of techniques to provide better insights and investigation of missing person cases. The aim is to leverage accessible demographic, geographic, and narrative information more intelligently, revealing patterns that usual investigation procedures might miss.

The research starts with exploratory data analysis (EDA) with datasets from sources like Kaggle and the National Crime Records Bureau (NCRB) from 2016 to 2022 to determine trends based on age, gender, location, and time. To ensure data quality, significant cleaning and feature selection were performed. The Pattern DNA module then analyses eight key features from the dataset, transforming them into numerical vectors and clustering [5] them to reveal common case profiles that may guide investigative priorities and highlight vulnerable demographics.

To investigate potential movements, a Geo-Spatial Drift Model integrates last seen locations, distance to transport nodes or hotspots of crime, and urban–rural designation to forecast probable zones of movement through heat maps and logistic regression.

A Semantic Case Matching layer also utilizes natural language processing (NLP) methods with BERT embedding to match case stories and find comparable prior cases for the assistance of cold case linkages. A time-series model with Facebook Prophet also predicts possible future trends in missing person cases to inform early intervention measures.

Together, these modules form a flexible, AI-driven system designed to provide meaningful insights, support public safety efforts, and help families and authorities advance their search for missing persons. Given the scale and complexity of such cases, this framework demonstrates how data-driven tools can complement traditional workflows, improve investigative decision-making, and contribute to faster, more effective resolutions.

2. Methodology

The proposed methodology follows a step-by-step modular approach to process, analyze, and present data related to missing person cases in India.

A. Preprocessing and Data Collection

The research uses publicly available data, such as the Missing People Dataset (Kaggle), NCRB reports, and district-level records for missing person cases from various Indian states during the period 2017-2020. The datasets include demographic data, location coordinates, case descriptions, and temporal information. Duplicates were removed, and missing values were handled through data cleaning. Inconsistent entries were also normalized. Attributes such as age, gender, location, and time of disappearance were standardized.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to identify trends, seasonality, and demographic tendencies in the missing person database. The objective was to create insights that may help in selecting clustering features and enhance the accuracy of geospatial predictions.

- District-wise Mapping: A choropleth map was created to map missing cases by district, indicating geographical hotspots and areas with repeated high frequencies of reports.
- Seasonal Trend Analysis: Monthly trend lines indicated peaks in missing cases within some months, suggesting potential connections with festivals, migration, or seasonal activities.
- Demographic Distributions: Age and gender distribution histograms validated that some age ranges and sexes were disproportionately represented, justifying the choice of demographic attributes for Pattern DNA Clustering [5].
- Outlier detection and standardization: Numerical columns like age were screened for extreme outliers via the Z-score technique.

$$Z = \frac{(X - \mu)}{\sigma}$$

- Scaling for Clustering: After outlier removal, key features were normalized using Min-Max scaling.

C. Pattern DNA

In this module, the system reveals latent demographic signatures by converting raw categorical counts into normalized numerical feature vectors. From the district-wise dataset, primary ratios were obtained — like Male Ratio, Female Ratio, and Transgender Ratio — each of which was computed as:

$$Ratio = \frac{Category\ Count}{Total\ Missing\ Cases\ in\ District}$$

In addition, age-based ratios for below 18, 18–30, 30–45, 45–60, and 60+ years were calculated, leading to an eight-dimensional Pattern DNA vector per district. This conversion facilitates comparison of demographic profiles among areas, regardless of raw population size. Following numerical encoding and normalization, the resultant feature matrix was clustered with the K-Means algorithm (k=4).

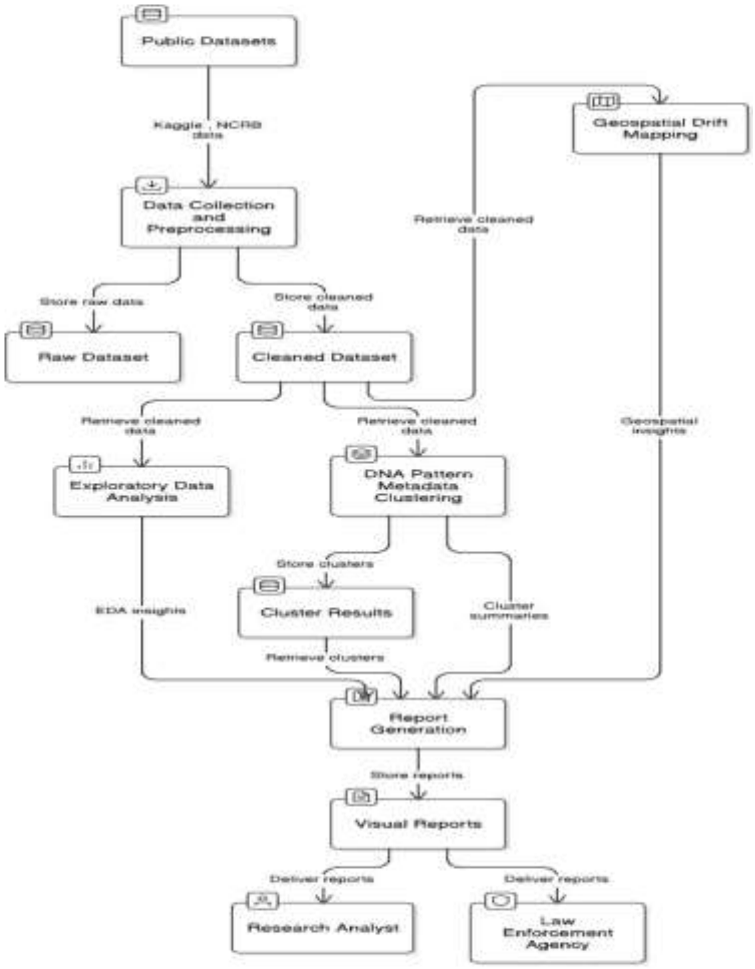

Figure 1 Methodology Block Diagram

This yielded four unique hidden profiles of missing person trends:

- Cluster 0: Under-18 Female Heavy
- Cluster 1: Adult Male Migration
- Cluster 2: Balanced Profile
- Cluster 3: Young Urban Male Drift

Dimensionality Reduction and Visualization:

In order to facilitate easy interpretation of the eight- dimensional Pattern DNA vectors, Principal Component Analysis (PCA) was used to project the data into a two- dimensional space without sacrificing maximum variance. This is mathematically represented as:

$$Z = XW$$

Where,

- X: X is the scaled feature matrix,
- W: W holds the eigenvectors for the two largest principal components,
- Z: Z is the estimated dataset with new dimensions PCA1 and PCA2.

We employed the Python implementation using *sklearn decomposition*. PCA with *n. components=2* to derive these top components. The derived PCA1 and PCA2 features were appended to the dataset and displayed on a scatter plot with Seaborn, and it revealed clear cluster groupings and confirmed the occurrence of demographic drift patterns between districts.

D. Geo Spatial Drift Forecasting

The Geo-Spatial Drift Forecasting module takes the Pattern DNA outcomes further by incorporating a spatial intelligence layer to forecast potential paths of movement for missing individuals. This assists in determining where a missing individual may next travel from their last known location, local transport facilities, and high-risk crime hotspots. Salient steps in this module:

- Last Known Location Tagging: Every record for a missing individual contains accurate latitude and longitude coordinates of the last confirmed sighting.
- Proximity Analysis: Based on geodesic distance calculations (Haversine formula through Ball Tree) , the system calculates distances from the recent point to: Closest airport, Closest bus stop, Closest railway station, Closest known crime/trafficking destination
- Normalization: Distances are normalized through Min-Max scaling to make them comparable.

$D_{normalized} = (D_{actual} - D_{min})/(D_{max} - D_{min})$

This converts all distances to a common 0–1 range.

- 4. Drift Risk Score Computation: A composite Drift Risk Score is computed by summing the normalized distances with increased risk to cases near transport points and hotspots.

$Drift\ Risk\ Score = w1\ x\ )1\text{-}D_{airpoirt}) + w2\ x\ (1\text{-}D_{bus}) + w3\ x\ (1\text{-}D_{train}) + w4\ x\ (1\text{-}D_{hotspot})$

Here, w1–w4 are variable weights depending on expert opinion or historical trends.

- Directional Bearing Calculation: To make predictions of the probable direction of movement, the system calculates the bearing angle between the most recent known location and the nearest hotspot.

$$Bearing = atan2(sin(\Delta lon)\ x\ cos(lat2),$$
$$Cos(lat1)\ x\ sin(lat2) - sin(lat1)\ x\ cos(lat2)\ x\ cos(\Delta lon))$$

- **Spatial Visualization:** The results of drift risk appear on an interactive Folium map with color- coded, size-by-case markers by Drift Risk Score. Directional arrows show potential movement trajectories towards proximal transport hubs and crime hotspots, with shaded risk areas indicating probable drift areas. Each marker has a popup with important information such as the individual's age, risk score, proximal airport, bus or rail station distances, nearest city or hotspot, and computed drift bearing.

E. Semantic Case Matching:

To augment the geographic shift and demographic aggregation modules, a Semantic Case Matching layer is added to identify narrative commonalities among missing person reports. Most long-standing unsolved cases have unstructured text descriptions that have minor contextual patterns that structured information might overlook.

Dataset: The semantic matching module takes advantage of the cases_with_descriptions.csv dataset, which contains comprehensive free-text descriptions for every missing person case, as well as the supporting features of age, location, and date of disappearance.

Text Embedding: Every description is then mapped to a high-density numerical vector through a pre-trained Sentence-BERT model (all-MiniLM-L6-v2). The embedding process translates variable-length text into fixed- sized 384-dimensional vectors that maintain semantic meaning beyond literal keywords. For instance, a case notes like "Teen girl missing near bus terminal after festival" is translated into an embedding vector that can be compared in similarity. [7,8]

To determine narrative connections between cases, the cosine similarity is calculated between the embedding vector of a query case and all historical case embedding vectors. The similarity score is determined as:

$$\text{Similarity}\ (q,di) = (q.di)/\ (\|q\|\ x\ \|di\|)$$

After embedding case descriptions with the Sentence-BERT pre-trained model, the system calculates cosine similarity scores between new input queries and stored records to find the most semantically similar matches. This allows for automatic connecting of cases that have common patterns of narrative, like disappearances around transport centers, suspected kidnappings, or recurring situations, irrespective of time or locational differences. By bringing to light these subtle similarities, the method assists investigators in creating new leads, linking outstanding files, and identifying common patterns of occurrence that could signal organized crime or trafficking circuits. To further maximize its usefulness, the module could be supplemented by refining the model through training with local case reports and grouping matched cases into thematic categories, adding a useful layer of semantic insight to spatial and demographic examination.

F. Risk Forecasting

The Risk Forecasting module brings a vital temporal forecasting function to the overall analysis pipeline by predicting the future evolution of missing person case trends. By using sophisticated time-series forecasting algorithms, this module produces month-ahead forecasts that uncover patterns, seasonal spikes, and long-term drifts that may otherwise be masked in raw data. Such anticipatory insight is priceless for investigative authorities, law enforcement, and NGOs since it equips them to prepare for targeted interventions well in advance, assign resources more strategically during periods of high risk, and create preventive strategies specific to expected spikes in cases. Finally, this module converts static historical records into actionable intelligence to inform proactive decision-making and policy development.

- **Input Dataset:**

The module takes the cases_with_descriptions.csv dataset, to guarantee accurate forecasting, the 'Date' column is then converted into a standard date time format. Any entry with missing or erroneous date entries is deleted to preserve the integrity and continuity of the time series. This cleaning process guarantees that the final dataset correctly represents when every case was reported, which is absolutely important for accurate trend analysis.

- **Monthly Aggregation:**

After cleaning, the dataset is grouped into monthly intervals via the Period Index tool of the panda’s library. This converts single daily records into aggregated monthly totals of missing persons reports. The data is then reformatted into a format well-suited to Prophet's demands, into two columns: ds (standing for the timestamp for each month) and y (the cases reported that month). This format constitutes the input to the forecasting model.

- **Time-Series Forecasting with Prophet:**

The forecasting is done with Facebook's Prophet Library, which is suited for time-series data with several seasonal cycles and irregular trends. Prophet breaks down the time series according to the following model:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

Where:

- $g(t)$ is the trend component that captures non-periodic changes over time,
- $s(t)$ represents the seasonal effects that recur annually or monthly,
- $h(t)$ compensates for holiday effects or other unique events if designated,
- $\varepsilon(t)$ is an error term for unforeseen variations.

A future data frame is then created, which extends the timeline 72 months (6 years) after the historical data available, allowing forecasts through the year 2030. The forward projection shows possible future spikes or dips in case volumes, which provide stakeholders with a rich, advanced look at probable trends. [6,9]

Results & Conclusion

The suggested multi-stage pipeline was rigorously tested using real-world missing persons datasets gathered from public sources like NCRB and Kaggle. The pipeline combines four synergistic analytical modules: Pattern DNA Clustering, which exposes latent clusters and repeated patterns in case metadata; Geo-Spatial Drift Forecasting, which forecasts likely directions of movement and danger areas from geographic proximity analysis; Semantic Case Matching, which connects cases with comparable narrative information based on advanced NLP; and Temporal Risk Forecasting, which models longer-term trends to emphasize seasonal or periodic spikes in missing persons reports.

Taken together, these findings demonstrate how raw, unstructured case data can be re-fashioned into actionable, multi-dimensional intelligence. Through the blending of spatial, temporal, and semantic layers, the system empowers law enforcement, NGOs, and analysts with richer situational awareness, which they can use to set priorities among leads, deploy resources more effectively, and build targeted interventions. The next sections describe visual outputs and expound upon practical insights gained by each module.

A. Pattern DNA :

The Pattern DNA module generated effective clusters based on parameters like age, gender, duration of disappearance, and state-wise occurrence. Figure 1 shows the state-wise distribution of these clusters, indicating how the missing person cases cluster differently across states.

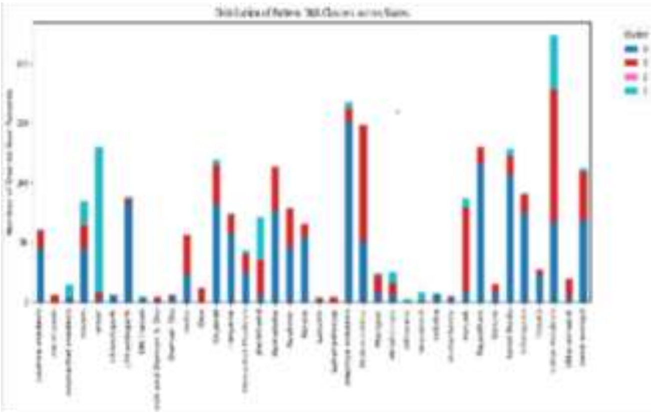

Figure 2 State-wise distribution , Data source

Figure 2 shows the distribution of Pattern DNA clusters across various Indian states. The stacked bar chart represents four distinct clusters identified by the unsupervised learning process. It can be observed that states like Uttar Pradesh, Maharashtra, Bihar, and Madhya Pradesh exhibit significantly larger counts spread across multiple clusters. This indicates that different demographic segments or local conditions may influence how and why disappearances occur in these states.

The presence of multiple clusters within the same state suggests underlying socio-economic, cultural, or migratory factors that create unique risk profiles. For example, certain clusters may capture patterns such as young females missing near transit points, while others highlight long-term disappearance trends in specific age brackets. By visualizing these hidden structures, the Pattern DNA Clustering output validates the system's ability to extract actionable patterns that are not evident from raw tabular records alone. [5,7]

B. Geo-Spatial Drift Forecasting:

The Geospatial Hotspot Analysis module introduces an important spatial intelligence layer into the pipeline, which helps identify and visualize the most densely populated areas of missing person cases within India. By grouping historical reports with accurate geographic coordinates, the system produces rich heat maps — such as the one above — where darker colors indicate areas with greater densities of missing persons incidents.

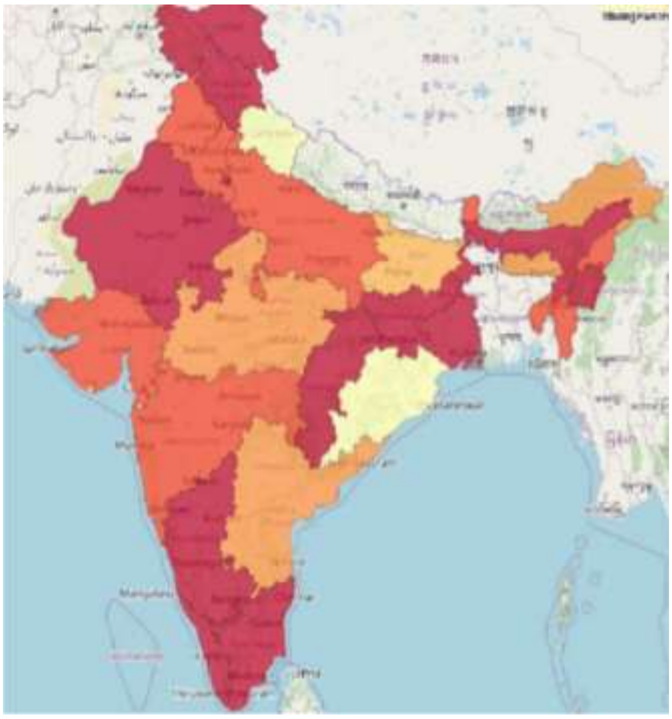

Figure 3: Heatmap showing regional concentration of missing person cases across India. Data Source: [1,10]

C. *Semantic Case Matching:*

The Semantic Case Matching module solves the problem of linking missing person cases that might have latent narrative similarities but no obvious geographic or demographic similarity. With a transformer-based language model such as Sentence-BERT [7], each free-text case summary is reduced to a high-dimensional semantic embedding. When a fresh report is submitted, the system converts the query to the same vector space and computes cosine similarity scores against all previous case embeddings. This will make reports with similar context, wording, or situational hints surface as potential matches.

Figure 4: Top 5 similar missing person cases retrieved for the query, ranked by cosine similarly scores, Data Source: [1,11]

For example, the test query — "young girl went missing near Pune railway station, could be abducted" — returned the top five most relevant responses, ranked by cosine similarity scores between 0 and 1. Higher scores indicate a closer match to the narrative. The highest-ranked match was a 21-year-old woman missing from a bus stand in Kadapa, with a similarity score of 0.5975, demonstrating the system's ability to identify contextual links such as transport stations and suspected kidnappings. Other retrieved matches mentioned disappearances near stations or terminals involving victims of similar age, showing how semantic similarity matching can turn weak or incomplete reports into useful investigative leads that would be difficult to find manually.

D. Temporal risk forecasting :

The Temporal Risk Forecasting module introduces a time- series analysis layer into the pipeline, which assists in predicting long-term trends in missing persons reports. From the Facebook Prophet model [6], the system was learned using past date-stamped case data to produce a stable monthly prediction up to the year 2030. The resulting prediction is shown in Figure 5: the blue solid line is the forecasted trend, the shaded area is the model's confidence band, and the black dots are the model-fitting and validation actual historical observations.

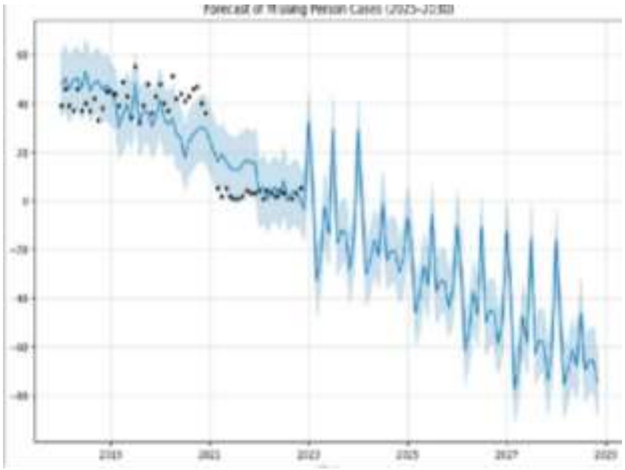

Figure 5: Forecast of monthly missing person cases (2025–2030) using
Facebook Prophet. Data Source: [1,12]

The narrative picks out obvious seasonality and cyclical spikes in the reported cases, whereas the trend line as a whole shows how much the number of missing persons reports is likely to go up or down in the next few years. By measuring these patterns over time, the model helps authorities to better gauge when caseload surges are likely to happen, for example, at specific months or holiday periods when vulnerable groups might be particularly at risk, and equips them to prepare carefully targeted awareness campaigns, staff up more people during peak periods, and develop forward-looking interventions to avoid disappearances in the first place. By deriving both underlying trends and recurring seasonal variation, this module turns raw historical data into actionable information that can support risk mitigation and resource planning for law enforcement, NGOs, and policymakers to transition from reactive response to data-driven prevention, ultimately fortifying the safety net for vulnerable people.

CONCLUSION

This study introduces an integrated, multi-module pipeline that meets the intricate and urgent problem of missing persons in India using the aggregated application of Geo- Spatial Hotspot Analysis, Semantic Case Matching, and Temporal Risk Forecasting. These modules collectively convert raw and fragmented case records into actionable intelligence that has the potential to significantly improve investigative procedures and strategic planning. The Geo-Spatial Hotspot Analysis identifies areas of high risk, allowing for focused resource deployment and proactive preventive action. The Semantic Case Matching module harnesses sophisticated language models to reveal latent connections in narrative case information, helping investigators make meaningful associations between disparate-seeming reports. The Temporal Risk Forecasting module, applying time-series analysis, offers foresight into emerging trends, enabling agencies to plan intervention and deploy resources in advance during expected surge periods.

Although the Pattern DNA Clustering module was discovered on an independent dataset to investigate demographic clustering, it extends the broader vision of employing unsupervised methods to identify embedded correlations.

The practical applicability of these modules by way of an interactive Streamlit dashboard also provides real- world viability for law enforcement and NGOs. Future work will extend the route prediction module to incorporate rail and highway networks in addition to bus networks so that possible exit routes can be mapped out even more comprehensively. Together, this framework lays the basis for scalable, data-driven, and preventative action to protect vulnerable people and communities.

References

[1] Arjoonn, “Missing People,” *Kaggle*, n.d.

[2] PowerDrill, “Districtwise Missing Persons,” *PowerDrill*, n.d. ericsiq, “India 5 Years District-wise Missing Persons Dataset,” *Kaggle*, n.d.

[3] Government of India, “State/UT-wise Details of Total Missing and Traced Persons during 2022,” *Data.gov.in*, 2022.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[5] S. J. Taylor and B. Letham, “Forecasting at scale,” *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[6] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings,” in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2019.

[7] NCRB, “NCRB Statistics – Missing Persons,” n.d.

[8] A. S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.

[10] L. Rokach and O. Maimon, “Clustering methods,” in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 321–352.

[11] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017.

[12] Y. Bengio, I. Goodfellow, and A. Courville, *Deep Learning*, MIT Press, 2016

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.