IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 1 October 2024, accepted 28 October 2024, date of publication 31 October 2024, date of current version 19 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3488784

RESEARCH ARTICLE

SafeRoutes: Charting a Secure Path-A Holistic Approach to Women's Safety Through Advanced Clustering and GPS Integration

KUSHAL AGRAWAL[1], AVIRAL SRIVASTAVA[1], (Member, IEEE), KANISHK SHARMA[1], SANDEEP KUMAR SATAPATHY[2], (Member, IEEE), SUNG-BAE CHO[2], (Senior Member, IEEE), SHRUTI MISHRA[3], (Member, IEEE), ABISHI CHOWDHURY[1], (Member, IEEE), AND AMRIT PAL[1], (Member, IEEE)

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India
[2]Department of Computer Science, Yonsei University, Seoul 03722, South Korea
[3]Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India

Corresponding author: Abishi Chowdhury (abishi.chowdhury@vit.ac.in)

This work was supported by the Vellore Institute of Technology, India.

ABSTRACT This research addresses the critical issue of women's safety in urban environments, emphasizing the need for innovative solutions to establish secure pathways. SafeRoutes presents a holistic approach, integrating advanced clustering methodologies and GPS technology, detailing its relevance, ideation, methodology, and anticipated results. During ideation, the team prioritized integrating cutting-edge technologies—artificial intelligence, data analytics, and cloud computing. Emphasizing the constraints of existing safety solutions, the focus was on crafting a sophisticated framework for detailed assessments and real-time risk detection during transit. SafeRoutes aims to redefine women's safety, providing actionable insights for urban planning and law enforcement. The methodology comprises three integral components. Firstly, a robust data ingestion pipeline connects to public and government data sources, ensuring near real-time models enriched with the latest data. The second component uses unsupervised machine learning models, comparing and employing various clustering algorithms. Parameters like crime rates, police presence, and infrastructure are utilized to cluster regions based on women's safety. Lastly, integration with map APIs and cab service vendors addresses the travel aspect, facilitating real-time alerts for deviations into unsafe areas. Results encompass a nuanced correlation matrix classifying regions based on safety clusters, offering valuable insights for urban planning and law enforcement. Integration with cab services ensures SafeRoutes not only identifies safe paths but actively contributes to enhancing women's safety during transit. The anticipated outcome positions SafeRoutes as a pioneering solution, contributing substantially to the discourse on urban safety and establishing a benchmark for future research.

INDEX TERMS Data ingestion, machine learning model, GPS integration, unsupervised learning, clustering, Gaussian mixture models, heatmaps, data lakes.

I. INTRODUCTION

Urban safety, particularly concerning women's well-being, remains a persistent and multifaceted challenge in contemporary urban design [1]. The "SafeRoutes" approach, presented here, represents a pioneering effort that tackles this complex issue by integrating cutting-edge technologies, data-driven analytics, and collaborative partnerships with mapping services and cab vendors. A comprehensive review of the literature reveals the historical evolution of safety measures within urban environments. Traditionally, urban safety initiatives have been anchored in law enforcement strategies and infrastructural development. However, the emergence of technology has triggered a paradigm shift, prompting

The associate editor coordinating the review of this manuscript and approving it for publication was Li Zhang.

researchers to explore innovative methodologies grounded in data analytics and artificial intelligence (AI). Studies like [2], [3] have emphasized the effectiveness of real-time data in crime analysis, highlighting its crucial role in guiding targeted interventions. Additionally, research in the realm of crime prediction and hotspot identification [4] has laid the groundwork for applying similar methodologies to address women's safety concerns.

Within the context of women's safety, existing studies have delved into the intricate interplay between socio-economic factors, infrastructural elements, and crime rates. Groundbreaking investigations [5] have underscored the influential role of these variables in shaping safety perceptions and outcomes. Nonetheless, the dynamic nature of safety concerns for women necessitates an adaptive approach, and this represents the unique contribution of the "SafeRoutes" approach. Capitalizing on insights gleaned from prior research, the "SafeRoutes" initiative introduces a holistic framework that seamlessly integrates data science, machine learning methodologies, and real-time interventions. Marking a distinct departure from conventional methods, which often grapple with subjective labelling and insufficient labelled data for women's safety, the approach leverages unsupervised machine learning models to unearth underlying patterns and correlations.

Previous innovations involving the application of clustering algorithms for crime analysis [6], [7] provide a foundational understanding. "SafeRoutes" not only incorporates but advances these methodologies by tailoring parameters specifically to women's safety. Employing statistical methods like the Pearson and Spearman indices, the approach meticulously identifies critical parameters that significantly contribute to the safety clustering algorithm. The nuanced selection of parameters, encompassing crime rates, police presence, and socio-economic indicators, reflects a comprehensive understanding of the multifaceted factors influencing women's safety. Furthermore, the study draws inspiration from successful endeavors in data integration aimed at enhancing public safety. Noteworthy studies, exemplified [8], [9], [10] have demonstrated the effectiveness of integrating data from diverse sources for comprehensive crime analysis. "SafeRoutes" extends this concept by seamlessly integrating with mapping services and cab vendors, recognizing the paramount importance of addressing the travel aspect of women's safety. By delineating safety zones along recommended paths and implementing an alert system for route deviations, the approach aspires to proactively safeguard women during their journeys.

This research demonstrates the potential of data-driven approaches to address complex societal challenges like women's safety in urban environments. By integrating cutting-edge technologies and collaborative partnerships, the "SafeRoutes" approach offers a promising way to fostering safer and more inclusive cities for all.

II. RELATED WORK

In the realm of women's safety, several noteworthy studies have been conducted recently. One such study [11] focussed on a female's safety system that is centered around AI, offering security to women at risk. Another significant contribution [12] that explores the role of IoT in women's safety through a systematic literature review. This study delves into research studies showcasing IoT devices for women's safety, detailing the main features, wearables, sensors used, and machine learning algorithms employed. Another study [13], discussed a women's safety system providing self-defense and incorporating a device with salient features. These studies provide valuable insights, setting a strong foundation for further research in the field of women's safety, emphasizing the importance of integrating advanced technologies like AI, IoT, and data analytics in developing effective safety solutions. A GPS based fly smart phone was proposed by Zhou et al. [14] which was involved in determining the speed, moving directions, distances using computer PR technologies and machine code which are data oriented.

Similarly, Pravin et al. [15] proposed neural network model for the purpose of classification and identity of the device owner for the purpose of safety. Ansari et al. [16] had suggested a GSM and GPS used scheme for the purposed of women safety and protection where the main objective was to provide a safe environment to face the societal issue. Another GPS and GSM used model was proposed [17] that would mark the women position and transmit emergency messages to control room for women protection. The only problem that the scheme faced that it could not provide immediate SMS service. To overcome this problem, another method was suggested by [18] where the photographs played a major role in classifying the people's location and tracking them.

Another methodology to satisfy the purpose of women safety an android women protection app [19] was proposed to determine the current location of the women via GPS. This was further enhanced by [20] where a single button was used to determining the position of the location, also it registers the URL and calls the first identified connection to help in case of emergency. David and Bouldin [21] had presented a clustering algorithm for cluster separation measure for indicating the similarity of clusters that had a data density. This measure was also used to find the appropriateness of the data partitions. Similarly, Bholowalia and Kumar [22] had proposed a clustering technique with small and cost effection nodes for Wireless sensor nodes. They used Elbow method and K-means clustering algorithm for routing the traditional energy efficient protocol. Özarpacı et al. [23] had presented a GPS based velocity fields for clustering to create a block geometry. They used horizontal velocity fields to evaluate the effects of data dependences in finding the optimum number of clusters. MacQueen [24] had proposed a process for partitioning an n-dimensional population for classification and analysis of multivariate observations. The main objective was approximating the multivariate distributions and non-parametric tests for different variables. Many clustering techniques and ideas were considered for varied data for explaining the transformation from the prior to actual clustering. The data consisting of dissimilarities were considered between the objects [25]. Kılıç and Özarpacı [26] had proposed an ensemble clustering algorithm for presenting unique solutions for GPS velocities. They used block boundaries to identify and chance the clustering results for GPS velocities.

III. PROPOSED MODEL

The research study involves several key steps shown in the workflow diagram (as shown in figure 1). Initially, we focus on data ingestion, establishing a data pipeline connected to various public and government sources. This robust infrastructure maintains an updated near real-time model, with its efficacy tied to the volume and quality of the ingested data. In the Machine Learning Model phase, we use unsupervised machine learning models to address the challenge of labelling areas as safe. We delve into crucial parameters for women's safety, such as crime rates and infrastructure. Statistical methods help identify top parameters for our clustering algorithms. We used K Means Clustering and Gaussian Mixture Models, determining that 3-4 clusters align well with our dataset and parameters. We introduce dummy data from globally recognized safe cities as a reference point for scoring clusters, setting the stage for regional classification and heatmap generation. In the integration phase, we address the travel aspect of women's safety by leveraging Maps APIs and collaborating with cab service vendors. This allows us to visualize user rides, map recommended paths, and create a safety radius. An alert system notifies drivers of diversions and construction, ensuring adherence to recommended paths. Integration with the heatmap enables real-time alerts for drivers entering potentially unsafe areas, completing our methodology for enhancing women's safety.

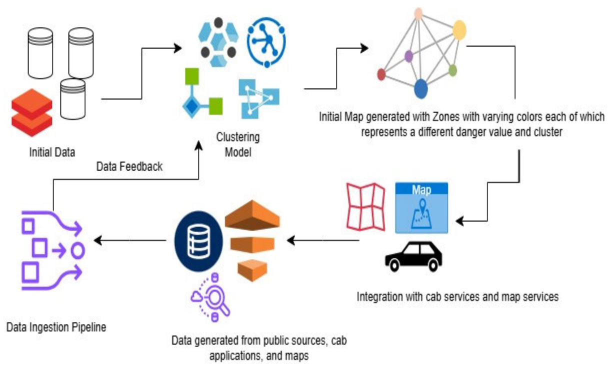


FIGURE 1. Data workflow of the complete approach.

IV. METHODOLOGY

A. DATA SOURCES

All data used in this study was retrieved from the Open Government Data Platform India (data.gov.in) [27]. The dataset collected focusses on diverse urban environment This platform ensures data reliability and consistency, making it suitable for research purposes. The specific datasets utilized are:

1. *district wise population for year* 2001 *and* 2011: This dataset provides district-level population information for two census periods, enabling insights into population growth and distribution.
2. *dstrIPC*_2013: This dataset contains information on crimes reported under the Indian Penal Code (IPC) across districts in 2013, specifically focusing on women-related crimes.
3. *literacydata*: This dataset provides literacy rates, potentially revealing correlations between educational attainment and women's safety.
4. *Per_Capita_Total_Expenditure*_1: This dataset offers insights into per capita expenditure, allowing for analysis of potential socio-economic factors influencing women's safety.
5. *polStr*_2013(Police Inspectors Data Per Region): This dataset comprises information on police inspector strength across regions, aiding in assessing the relationship between police presence and women's safety.
6. *populationDistrict*: This dataset contains additional population data, potentially enriching the analysis. This dataset contains additional population data like demographics breakdown by age, gender etc. and also social indicators like education, health, housing etc.
7. *Forest cover*: This dataset provides information on forest cover, which may be relevant for understanding geographical factors impacting women's safety.

B. DATA PRE-PROCESSING

Data preprocessing was conducted to ensure data quality and suitability for machine learning analysis. The steps involved are outlined below:

- *Missing Value Imputation*: Missing values were identified and addressed using appropriate techniques like mean/median imputation or k-Nearest Neighbors (KNN) imputation depending on the data type and distribution.
- *Feature Engineering*: New features were created based on existing data, such as population growth rate or literacy rate disparity. This step aimed to enrich the feature space and potentially improve model performance.
- *Feature Scaling*: StandardScaler or MinMaxScaler were employed to normalize features, ensuring all features contribute equally to the model during training.
- *Feature Selection*: A correlation matrix was generated to assess feature dependencies and identify potential multicollinearity. Subsequently, relevant features were selected using techniques like Principal Component Analysis (PCA) or feature importance scores from preliminary models.

While the study aims to enhance women's safety through data-driven methodologies, we acknowledge the potential for biases inherent in the data sources, which may affect marginalized groups or those in different socio-economic strata. To ensure a more equitable representation, several steps were taken:

- The datasets utilized, including crime rates and socio-economic indicators, may inherently reflect biases, particularly where marginalized communities face systemic neglect or under-reporting of crimes. Efforts were made to incorporate a range of socio-economic factors (such as literacy rates and per capita expenditure) to ensure that the analysis does not disproportionately favor affluent regions over those with fewer resources.
- SafeRoutes actively accounts for differences in urban, suburban, and rural regions by incorporating data on police presence, infrastructure, and reported incidents. However, we recognize that some areas may be under-represented due to a lack of reliable data. In future iterations, additional data sources, such as community-reported safety incidents and crowd-sourced information, will be integrated to provide a more inclusive safety model.
- The unsupervised learning algorithms employed in this study, such as K-Means Clustering and Gaussian Mixture Models, are sensitive to the quality and diversity of the input data. To mitigate biases, we implemented robust feature selection methods and correlation analyses that emphasized balanced data representation across various socio-economic groups. However, further refinements are necessary, particularly in expanding the dataset to include information specific to marginalized communities that may not be adequately represented in public crime statistics.
- The lack of data on vulnerable populations, including low-income communities, the outnumbered, and LGBTQ+ individuals, presents a challenge. As SafeRoutes evolves, we aim to address these ethical concerns by partnering with advocacy groups and local organizations to gather more inclusive and comprehensive data.

C. PROPOSED METHODOLOGY

The main steps involved in this research study include:

In the initial phase of our methodology, we prioritize data ingestion to build a robust foundation for our women's safety mechanism. This involves establishing a seamless data pipeline that connects to various public data sources and integrates with government databases. We leverage data lakes to ensure effective transportation insights and reliability. This robust infrastructure enables us to maintain an updated near real-time model, ensuring its efficacy is directly proportional to the volume and quality of the ingested data.

Moving to the core of our approach, the Machine Learning Model phase addresses the challenge of labelling areas as completely safe, given the absence of labelled data. To overcome this, we turn to unsupervised machine learning models, delving into parameters crucial for women's safety. Parameters such as rape cases, kidnapping cases, police presence, infrastructure, and more are meticulously chosen based on domain knowledge. Applying statistical methods such as the Pearson Index and Spearman Index, we identify the top parameters that feed into our clustering algorithms. Utilizing K Means Clustering and Gaussian Mixture Models as our primary machine learning models, we determine that 3-4 clusters align well with our extensive dataset and parameters. To establish a baseline for comparison, we introduce dummy data representing cities globally recognized as the safest. This becomes a reference point for scoring clusters, with the highest-scoring cluster deemed the safest, setting the stage for regional classification and heatmap generation.

Transitioning to the integration phase, we address the critical travel aspect of women's safety by leveraging Maps APIs and collaborating with major cab service vendors. This strategic integration allows us to visualize user rides, crucial for mapping recommended paths and creating a safety radius around them. An alert system is implemented to notify drivers of diversions and construction, ensuring adherence to recommended paths. Additionally, going beyond regional boundaries triggers alerts, linking back to the data ingestion layer for enhanced safety insights. Integration with the heatmap enables real-time alerts for drivers entering potentially unsafe areas, thus completing our comprehensive methodology for enhancing women's safety.

Addressing real time deviations, figure 2 depicts a flow chart to represent the same. There is a ride request process in which the taxi service requests the safe path based on SafeRoutes data, including crime clusters and police presence. The 1.96x safety buffer (confidence interval) is calculated to allow for deviations due to construction or temporary roadblocks. It also consists of the real-time monitoring, the system continuously checks if the taxi is following the safe path or within the predefined buffer zone. If the taxi deviates, it checks whether this leads to an unsafe zone (red area). Another level is used for handling unsafe zones, if a red zone is entered, immediate alerts are triggered, and rerouting is initiated. If the driver does not follow the new safe route, the system escalates monitoring to critical status, potentially notifying authorities or emergency services. The temporary deviation handling, is concerned if the deviation is due to legitimate road conditions like construction, the system temporarily allows the deviation while ensuring the taxi remains within the safe buffer. If the taxi doesn't return to the path, a new safe route is calculated using real-time data. At the end of ride, the system checks if the ride is completed safely, logging the data for further analysis.

V. RESULTS AND INFERENCES

The study commences with a meticulous analysis of forest cover data at the state level, exploring parameters such as very dense, moderately dense, and open forest areas. This foundational environmental insight establishes a backdrop for

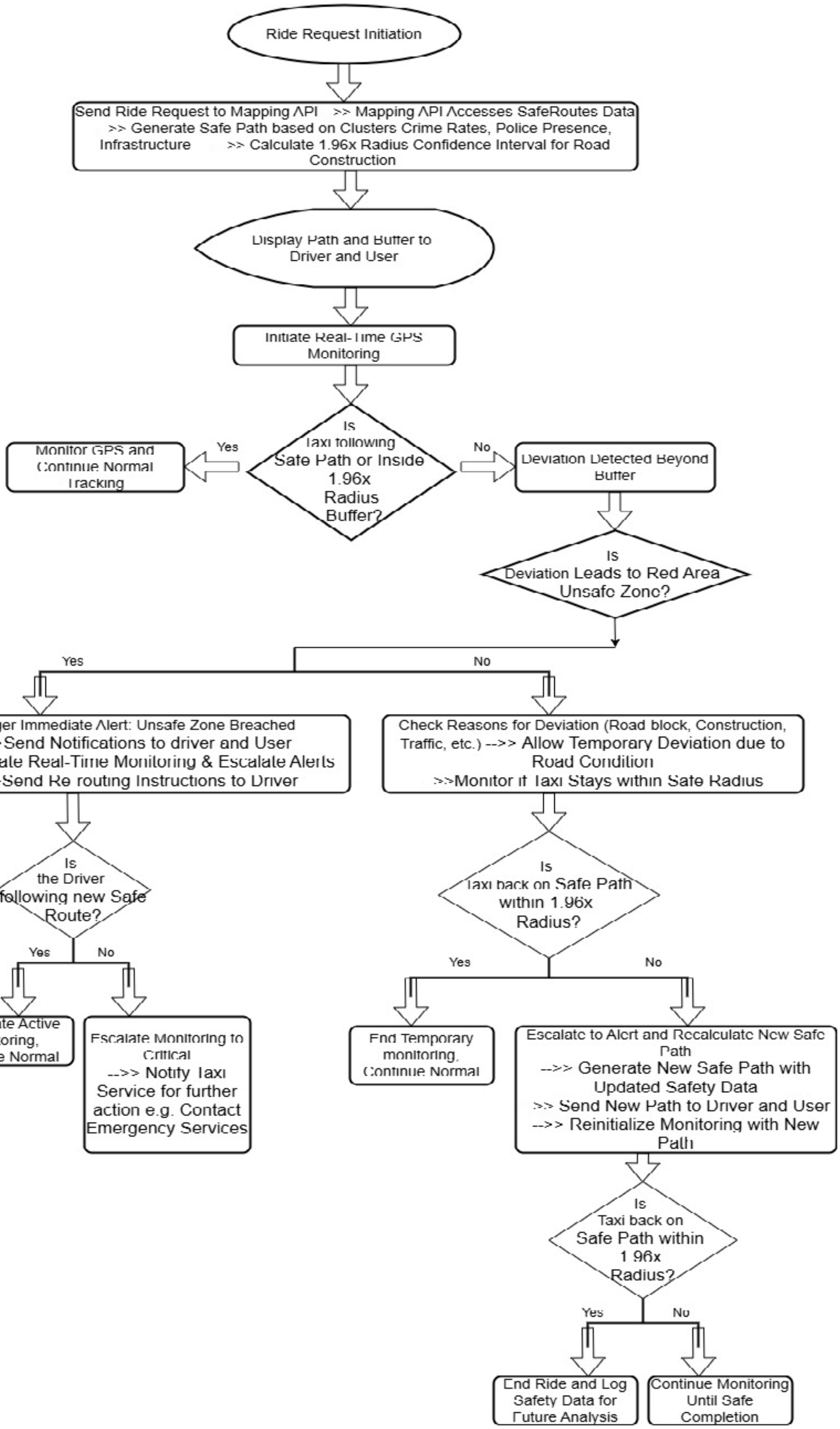


FIGURE 2. **A detailed flowchart for depicting real-time deviations.**

the subsequent examination of women's safety. Moving to the crime and demographic analysis, the study provides a district-wise breakdown of crimes against women, encompassing rape, custodial rape, and other related incidents. The inclusion of demographic elements like population, total expenditure, and per capita expenditure offers a comprehensive socio-economic context. Categorical data encoding ensures the preparedness of the dataset for subsequent machine learning analyses.

The heart of the research lies in the machine learning model, where unsupervised learning techniques, including K Means Clustering and Gaussian Mixture Models, are applied. Parameters like crime rates, demographic factors, and forest cover intricacies contribute to the formation of safety clusters, similar to previous works in crime analysis and hotspot identification. Dummy data from globally recognized safest areas serves as a benchmark, enabling the assignment of scores to relative safety levels within each cluster, drawing inspiration from established data integration practices for public safety enhancement. To address the travel aspect of women's safety, the study integrates with major cab service vendors, omitting the API-specific details. It establishes recommended paths, incorporating a real-time alert system for drivers entering unsafe areas. This integration aims not only to recommend secure routes but actively enhances women's safety during transit.

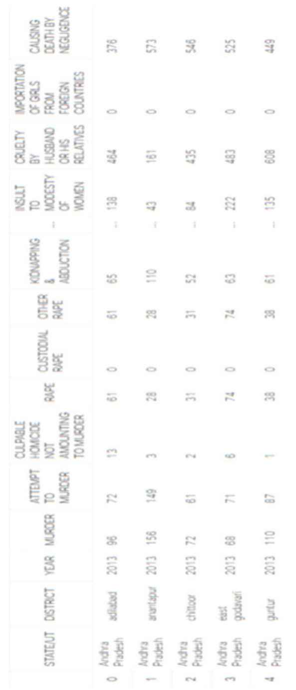
Correlation analysis (as shown in figure 3-5) becomes pivotal, examining relationships among variables such as crime rates, demographic factors, and total expenditure. The Pearson correlation matrix unfolds intricate connections, providing a deeper understanding of the multifaceted influences on women's safety.

In the data pre-processing phase, duplicates are handled, and categorical data is transformed using Label Encoding (as shown in figure 6). This meticulous preparation ensures data compatibility and sets the stage for subsequent analytical steps.

Machine learning model evaluation involves techniques like Label Encoding and correlation analysis. The resulting outputs are scrutinized for their potential to offer actionable insights to policymakers, urban planners, and law enforcement agencies. In conclusion, the fusion of forest cover, crime, demographic, and machine learning analyses positions SafeRoutes as a holistic and pioneering solution for enhancing women's safety in urban environments. The research contributes significantly to the urban safety discourse, providing a comprehensive framework for future studies in this critical domain.

A. UNDERSTANDING URBAN SAFETY DYNAMICS

To comprehend the dynamics of urban safety, we turn our attention to critical case studies that form the bedrock of our analytical framework [28]. The metropolis, with its ever-changing landscape and diverse demographics, poses a myriad of challenges. Case study analyses of urban areas offer invaluable insights into the nuanced relationship



	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT TO MURDER	CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	RAPE	CUSTODIAL RAPE	OTHER RAPE	KIDNAPPING & ABDUCTION	...	INSULT TO MODESTY OF WOMEN	CRUELTY BY HUSBAND OR HIS RELATIVES	IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES	CAUSING DEATH BY NEGLIGENCE
0	Andhra Pradesh	adilabad	2013	96	72	13	61	0	61	65	...	138	464	0	376
1	Andhra Pradesh	anantapur	2013	156	149	3	28	0	28	110	...	43	161	0	573
2	Andhra Pradesh	chittoor	2013	72	61	2	31	0	31	52	...	84	435	0	546
3	Andhra Pradesh	east godavari	2013	68	71	6	74	0	74	63	...	222	483	0	525
4	Andhra Pradesh	guntur	2013	110	87	1	38	0	38	61	...	135	608	0	449

FIGURE 3. A sample correlation analysis.

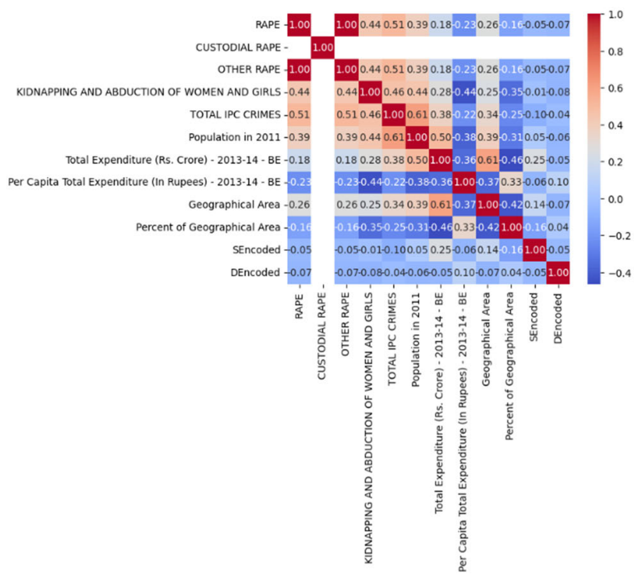


FIGURE 4. Spearman correlation matrix.

between crime patterns, socio-economic variables, and safety perceptions [8]. These case studies serve as benchmarks, guiding the selection and refinement of parameters for our clustering algorithms [6], [7]. Machine learning models, particularly K Means Clustering and Gaussian Mixture Models, reveal their prowess in discerning subtle patterns within these urban landscapes [6], [7]. Case studies involving these models illuminate their ability to categorize regions based on safety, offering a tangible representation of the effectiveness of our approach [4]. By applying statistical methods, such as

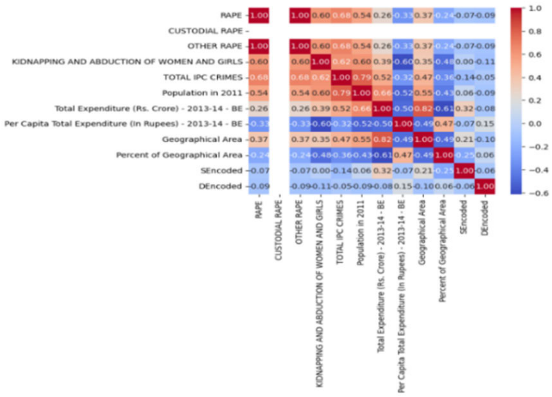


FIGURE 5. Pearson correlation matrix.

	RAPE	CUSTODIAL RAPE	OTHER RAPE	KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS	TOTAL IPC CRIMES	Population in 2011	Total Expenditure (Rs. Crore) - 2013-14 - BE	Per Capita Total Expenditure (In Rupees) - 2013-14 - BE	Geographical Area	Percent of Geographical Area	SEncoded	DEncoded	logpop
0	61	0	61	47	6381	2741239	152722.0	17622.51	275069.0	16.77	0	0	46.964634
1	28	0	28	84	6913	4081148	152722.0	17622.51	275069.0	16.77	0	12	69.497795
2	31	0	31	27	5610	4174064	152722.0	17622.51	275069.0	16.77	0	71	71.060358
3	74	0	74	33	6779	5154296	152722.0	17622.51	275069.0	16.77	0	99	87.544855
4	38	0	38	54	7026	4887813	152722.0	17622.51	275069.0	16.77	0	128	83.063428

FIGURE 6. Encoding data representation.

the Pearson and Spearman indices, we ascertain the relevance and significance of various parameters, ensuring a judicious selection that aligns with the nuances of women's safety concerns [6].

B. PARAMETRIC SIGNIFICANCE AND CLUSTERING ALGORITHMS

The significance of individual parameters within our clustering algorithms is exemplified through case studies focusing on specific variables. For instance, we examine case studies where the prevalence of crime rates is high but mitigated by a robust police presence, leading to the identification of safe clusters. Contrarily, instances where socio-economic indicators heavily influence safety clustering highlight the socio-economic landscape's pivotal role in shaping safety perceptions. Through the lens of parametric analysis, we uncover the strength of our clustering algorithms in identifying key contributors to safety classifications. The statistical indices guide us in discerning the most influential variables, refining our model to ensure a more accurate representation of women's safety concerns. Case studies allow us to fine-tune the algorithms, ensuring their adaptability to diverse urban environments.

C. INTEGRATION WITH MAPS API AND CAB SERVICES

The analytical focus extends beyond static safety categorizations to dynamic aspects of women's safety during travel. Case studies involving the integration with Maps API and cab services provide an intricate understanding of how real-time interventions influence the safety landscape. The creation of safety zones around recommended paths, coupled with alerts for deviations into potentially unsafe areas, emerges as a proactive measure with tangible implications. We delve into specific instances where the integration with cab services has not only enhanced the reliability of women's travel but has also served as a valuable data source for refining our safety models. Insights gained from these case studies underscore the proposed approach's potential to influence and shape the behaviour of transportation services to prioritize women safety.

D. SCORING SYSTEM AND COMPARATIVE ANALYSES

Our analytical journey culminates in the examination of the scoring system derived from case studies of 'perfect cities.' By assigning an arbitrary score of 100 to the cluster with the highest number of these idealized cities, we establish a comparative framework (similar to [8]). Case studies of cities considered the safest serve as benchmarks, allowing for a robust comparative analysis of other clusters, following established data integration practices for public safety enhancement [28] Through these comparative analyses, we discern the relative safety levels of diverse urban regions, building upon previous works in crime analysis and hotspot identification [4]. The case studies not only validate the effectiveness of our scoring system but also provide a nuanced understanding of the contextual factors influencing safety clusters, similar to existing studies that delve into the intricate interplay between socio-economic factors, infrastructural elements, and crime rates [16]. Insights derived from these analyses contribute to the refinement of our scoring system, ensuring its applicability across diverse urban landscapes, extending the concept of data-driven interventions beyond static predictions. By unravelling the intricacies of urban safety dynamics [15], parametric significance [6], [7], integration with dynamic travel components, and the establishment of a scoring system, we illuminate the multifaceted nature of our approach. These analyses not only validate the robustness of our methodologies but also furnish actionable insights for the advancement of women's safety in urban environments, aligning with the goals of SafeRoutes strategy as a pioneering effort to address this complex issue.

VI. ABLATION STUDY

In this study, various clustering methodologies are meticulously examined to discern the optimal configurations for the SafeRoutes approach. Through detailed analyses of explained variance, 3D visualizations, Bayesian Information Criterion scores, hierarchical clustering dendrograms, and validation metrics, this section elucidates the strengths and limitations of each technique.

The explained variance used here is to determine how much of the total variance in the data is accounted for by a statistical model i.e. how well the clustering algorithm has partitioned the data into groups. It indicates the proportion of the total variance that is explained by the differences between clusters. The explained variance in the context of clustering is the ratio of the between-cluster variance to the total variance.

A. DETERMINATION OF OPTIMAL CLUSTERS

The "Explained Variance vs. Number of Clusters" graph (as shown in figure 7 and table 1) demonstrates that as the number of clusters increases, the explained variance tends to decrease. An optimal point is reached where increasing clusters no longer significantly improves variance explanation. This analysis aids in selecting an appropriate balance between capturing underlying patterns and avoiding overfitting.

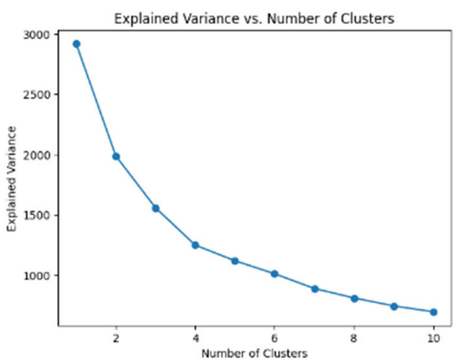


FIGURE 7. Optimal clusters determination.

TABLE 1. Number of clusters vs explained variance.

No. of Clusters	Explained Variance
1	41.682293
2	22.303359
3	17.203420
4	13.385043
5	11.645606
6	10.093617
7	9.265175
8	8.353212
9	7.909966
10	7.180986

B. K-MEANS CLUSTERING VISUALIZATION IN 3D

The 3D scatter plot (as shown in figure 8) vividly displays the outcome of the K-Means clustering algorithm (as shown in eq 1). Clusters are visually distinguishable, with each data point plotted in a 3D space based on its assigned cluster. The black 'x' markers represent the cluster centers, showcasing the algorithm's ability to identify central points within each cluster.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

where, WCSS is within -cluster sum of squares, C_i is the i-th cluster and μ_i is the centroid of the i-th cluster.

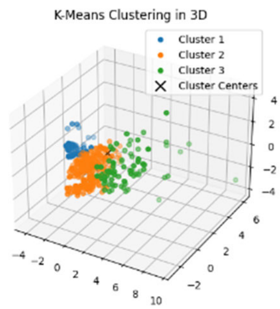


FIGURE 8. 3D visualization of K-Means clustering.

C. GAUSSIAN MIXTURE MODEL (GMM) EVALUATION

The *BIC Score* vs. *Number of Components* graph (as shown in figure 9) and table (as shown in table 2) facilitates the identification of the optimal number of components for GMM. The curve demonstrates an elbow point, indicating the most suitable complexity level. A lower BIC score signifies a better balance between model fit and complexity, guiding the selection of the optimal GMM configuration.

TABLE 2. Components vs BIC score table.

No. of Components	BIC Score
1	-1758.58
2	-2346.56
3	-2419.12
4	-2528.44
5	-2550.13
6	-2653.48
7	-2462.27
8	-2611.79
9	-2560.50
10	-2502.21

D. HIERARCHICAL CLUSTERING DENDROGRAM

The hierarchical clustering dendrogram (as shown in figure 10), constructed using the ward linkage method, provides a visual representation of the relationships between

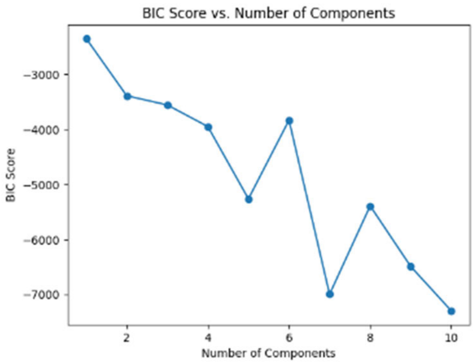


FIGURE 9. Gaussian mixture model.

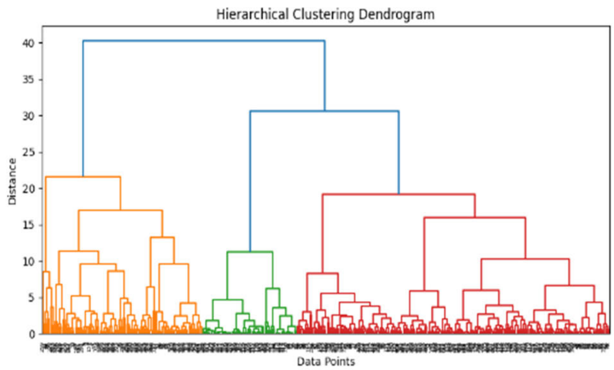


FIGURE 10. Hierarchical clustering dendrogram.

data points. The vertical lines showcase the merging process, with shorter lines indicating closer data point relationships. This dendrogram aids in identifying natural clusters and their hierarchical arrangement.

E. SILHOUETTE SCORE

The *Silhouette Score* vs. *Number of Clusters* graph (as shown in figure 11) reveals the highest silhouette score at a specific cluster count (as shown in table 3). This score (as shown in eq 2) signifies the degree of cohesion within clusters and separation between clusters, helping to identify the optimal number of clusters for hierarchical clustering [29].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2}$$

where, $a(i)$ is the average distance from point i to other point in the same cluster and $b(i)$ is the minimum average distance from point i to point in a different cluster.

F. CALINSKI-HARABASZ INDEX

The *Calinski-Harabasz Index* vs. *Number of Clusters* graph (as shown in figure 12) illustrates peaks, indicating optimal cluster configurations (as shown in table 4). It's basically a method for identifying clusters (as shown in eq 3) of points in a multidimensional Euclidean space is described, along with its application to taxonomy [30]. Higher Calinski-Harabasz scores denote well-defined, distinct clusters. The graph

TABLE 3. Silhoutte score vs clusters.

No. of Clusters	Explained Variance
2	0.440273
3	0.415679
4	0.292236
5	0.292553
6	0.297964
7	0.284529
8	0.285107
9	0.240554
10	0.242141

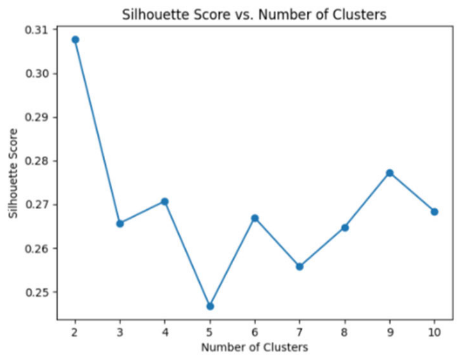


FIGURE 11. Silhouette score.

TABLE 4. Calinski Index vs clusters.

No. of Clusters	Explained Variance
2	264.676082
3	220.956094
4	222.705943
5	207.773331
6	199.884824
7	187.941316
8	180.395238
9	172.077039
10	165.649681

assists in determining the number of clusters that maximizes inter-cluster variance and minimizes intra-cluster variance.

$$CH = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n-k}{k-1} \tag{3}$$

where, $tr(B_k)$ is the trace of the between-cluster dispersion matrix, $tr(W_k)$ is the trace of within-cluster dispersion matrix, n is the number of points and k is the number of clusters.

G. DAVIES-BOULDIN INDEX

The *Davies-Bouldin Index* vs. *Number of Clusters* graph (as shown in figure 13) showcases valleys representing optimal configurations (as shown in table 5). Lower Davies-Bouldin

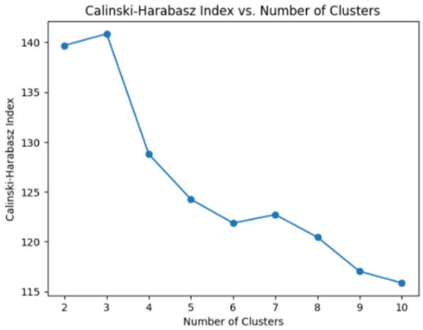


FIGURE 12. Calinski-Harabasz index.

TABLE 5. Davies-Bouldin index.

No. of Clusters	Explained Variance
2	1.020654
3	0.970371
4	1.011432
5	1.094263
6	1.152882
7	1.102814
8	1.142780
9	1.199943
10	1.147872

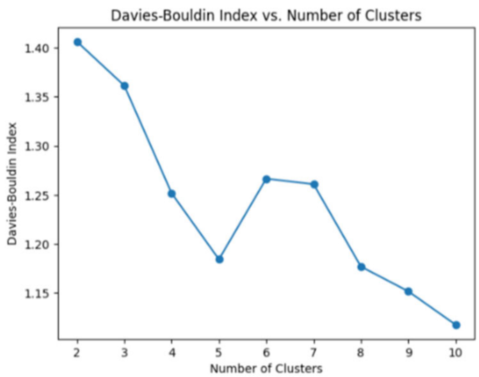


FIGURE 13. Davies-Bouldin index.

scores indicate better cluster separation. This metric (as shown in eq 4) assists in selecting the number of clusters that results in the most balanced and well-separated clusters.

$$DB = \frac{1}{k}\sum\nolimits_{i=1}^{k} \max_{j\neq 1}\left(\frac{s_i + s_j}{d_{ij}}\right) \tag{4}$$

where, s_i is the average distance between each point in the i-th cluster and the centroid of that cluster and d_{ij} is the distance between the centroids of clusters i and j.

These specific insights offer a nuanced understanding of each clustering method's performance, enabling informed decisions on the optimal configurations for subsequent analyses.

VII. DISCUSSION

The innovative solution presented in this study, SafeRoutes, strategically addresses the complex challenge of ensuring women's safety in urban environments. By integrating advanced clustering methodologies and GPS technology, the system offers a comprehensive approach that surpasses traditional safety measures. During the ideation phase, a careful emphasis was placed on integrating cutting-edge technologies, including artificial intelligence, data analytics, and cloud computing. This strategic integration enables SafeRoutes to overcome the limitations of existing safety solutions, resulting in a sophisticated framework capable of granular assessments and real-time risk identification during transit.

At the core of SafeRoutes lies a robust data ingestion pipeline intricately connected to diverse public and government data sources. By leveraging data lakes for effective transportation and reliability, this pipeline ensures the availability of near real-time models enriched with the latest data. The efficacy of the clustering model directly depends on the richness and timeliness of the ingested data. Recognizing the absence of labelled data for explicitly categorizing safety levels, SafeRoutes employs unsupervised machine learning models. Parameters correlated with women's safety, such as crime rates, police presence, and infrastructure, are carefully considered. Statistical methods, including Pearson Index and Spearman Index, are used to identify the most influential factors. Clustering algorithms like K Means and Gaussian Mixture Models ensure nuanced safety assessments and the generation of informative heatmaps. Acknowledging the integral role of travel dynamics in women's safety, SafeRoutes integrates with maps APIs and cab service vendors. By visualizing ride data, the system generates an approximate safe zone for drivers, triggering alerts if deviations into unsafe areas occur. This integration not only fortifies travel safety but also enhances the overall safety data available to the system.

The anticipated results include a detailed correlation matrix classifying regions based on safety clusters, providing invaluable insights for urban planning and law enforcement. By actively contributing to the discourse on urban safety, SafeRoutes stands as a pioneering solution, setting a benchmark for future research in this domain. As the system evolves, continuous enhancements and refinements promise to elevate women's safety to new heights, catalysing positive changes in urban landscapes. Policymakers can adopt SafeRoutes into existing urban infrastructure through a series of actionable steps that extend beyond theoretical frameworks:

- Establish collaborations with local governments and law enforcement agencies to ensure SafeRoutes is integrated with existing data sources, such as crime statistics and traffic monitoring systems. This requires standardization and alignment of SafeRoutes data with current urban databases, ensuring seamless integration and real-time updates.

- Engage with public and private transportation providers, including cab services and public transit systems, to embed SafeRoutes into route planning and navigation systems. This can involve integration with maps and navigation apps, where real-time risk alerts can guide commuters.
- Utilize the data-driven insights from SafeRoutes to influence urban planning decisions. By identifying high-risk areas, urban planners can prioritize infrastructure improvements, such as better lighting, increased police presence, and CCTV coverage in vulnerable zones. This ensures that urban spaces are designed with safety as a core principle.
- Develop public awareness campaigns to educate citizens on using SafeRoutes. Additionally, community participation can provide critical feedback to enhance the system's efficacy, encouraging public reporting of unsafe zones and incidents.
- Governments must legislate policies that support the adoption and funding of SafeRoutes in cities. This includes earmarking funds for its continuous development, ensuring that the system is scalable, adaptable, and sustainable over time.

VIII. CONCLUSION

SafeRoutes represents a transformative leap in addressing the critical issue of women's safety within urban landscapes. By integrating state-of-the-art technologies, advanced clustering methodologies, and real-time data analytics, the system delivers a holistic and proactive approach to ensuring secure pathways. The ideation phase strategically harnessed artificial intelligence, data analytics, and cloud computing, forming the backbone of a sophisticated framework capable of granular risk assessments during transit. The success of SafeRoutes relies on a resilient data ingestion pipeline, ensuring the availability of up-to-date information from diverse public and government sources. The utilization of unsupervised machine learning models, guided by statistical methods, allows SafeRoutes to navigate the absence of labelled data for safety categorization. Clustering algorithms, including K-Means and Gaussian Mixture Models, facilitate the creation of insightful heatmaps, offering a nuanced understanding of safety clusters based on various parameters.

Integral to the system is its seamless integration with maps APIs and cab service vendors, addressing the dynamic nature of travel safety. By visualizing ride data and establishing safe zones for drivers, coupled with real-time alerts for potential deviations into unsafe areas, SafeRoutes actively fortifies women's safety during transit. Anticipated outcomes include detailed safety heatmaps, providing actionable insights for urban planning and law enforcement. SafeRoutes stands as a pioneering solution, contributing significantly to the discourse on urban safety. As the system evolves, continuous refinements promise to set new standards, fostering positive transformations in urban landscapes and exemplifying a commitment to creating safer and more inclusive environments for women globally. SafeRoutes not only signifies a technological milestone but also represents a catalyst for positive societal change.

ACKNOWLEDGMENT

The authors declare that they have no competing interests or personal relationships that could have appeared to influence the work reported in this article.

REFERENCES

[1] H. Fang, "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem," in *Proc. IEEE Int. Conf. Cyber Technol. Autom., Control, Intell. Syst. (CYBER)*, Jun. 2015, pp. 820–824.

[2] T. Hlupic and J. Puniš, "An overview of current trends in data ingestion and integration," in *Proc. 44th Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, Sep. 2021, pp. 1265–1270.

[3] M. Roche and M. Teisseire, "Integrating textual data into heterogeneous data ingestion processing," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 6008–6010.

[4] H. Isah and F. Zulkernine, "A scalable and robust framework for data stream ingestion," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2900–2905.

[5] G. Esppenchutz, *Data Ingestion With Python Cookbook: A Practical Guide to Ingesting, Monitoring, and Identifying Errors in the Data Ingestion Process*. Birmingham, U.K.: Packt Publishing, 2023.

[6] S. Juned Ali, A. Gavric, H. Proper, and D. Bork, "Encoding conceptual models for machine learning: A systematic review," in *Proc. ACM/IEEE Int. Conf. Model Driven Eng. Lang. Syst. Companion (MODELS-C)*, Oct. 2023, pp. 562–570.

[7] K. Virupakshappa and E. Oruklu, "Unsupervised machine learning for ultrasonic flaw detection using Gaussian mixture modeling, K-Means clustering and mean shift clustering," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2019, pp. 647–649.

[8] R. Pryss, D. John, M. Reichert, B. Hoppenstedt, L. Schmid, W. Schlee, M. Spiliopoulou, J. Schobel, R. Kraft, M. Schickler, B. Langguth, and T. Probst, "Machine learning findings on geospatial data of users from the TrackYourStress mHealth crowdsensing platform," in *Proc. IEEE 20th Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Jul. 2019, pp. 350–355.

[9] M. Hashemi and H. A. Karimi, "A machine learning approach to improve the accuracy of GPS-based map-matching algorithms (invited paper)," in *Proc. IEEE 17th Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2016, pp. 77–86.

[10] I. Agarwal, A. Singh, A. Agarwal, S. Mishra, S. K. Satapathy, S.-B. Cho, M. R. Prusty, and S. N. Mohanty, "Enhancing road safety and cybersecurity in traffic management systems: Leveraging the potential of reinforcement learning," *IEEE Access*, vol. 12, pp. 9963–9975, 2024.

[11] M. Naved, A. Fakih, A. N. Venkatesh, A. Vani, P. Vijayakumar, and P. R. Kshirsagar, "Artificial intelligence-based women security and safety measure system," *Nucleation Atmos. Aerosols*, vol. 2393, pp. 1–7, 2022.

[12] D. Devi, M. Pavithra, K. Monalisha, T. S. Kirthana, and S. Pooja, "IoT based safety system for women," in *Proc. 6th Int. Conf. Commun. Electron. Syst. (ICCES)*, Coimbatre, India, Jul. 2021, pp. 731–736.

[13] D. Chanda, V. Reddy, and S. Aithal, "Tracking of mobile phones for piracy detection," *Int. Adv. Res. J. Sci.*, vol. 10, no. 1, pp. 16–25, 2023.

[14] C. Zhou, H. Jia, Z. Juan, X. Fu, and G. Xiao, "A data-driven method for trip ends identification using large-scale smartphone-based GPS tracking data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 2096–2110, Aug. 2017.

[15] K. Pravin and S. Akojwar, "Prediction of neurological disorders using PSO with GRNN," in *Proc. IEEE SCOPES Int. Conf.*, Jul. 2016, pp. 1–10.

[16] A. H. Ansari, B. Pratiksha, M. Tejal, and M. Y. Snehal, "Virtual friendly device for women security," in *Proc. Int. Conf. Phys. Photon. Processes Nano Sci.*, 2017, vol. 13, no. 3, p. 1.

[17] M. Deshpande, K. Kalita, and M. Ramachandran, "Artificial Intelligence based algorithm to support disable person," *Int. J. Appl. Eng. Res.*, vol. 9, no. 23, pp. 21975–21992, 2014.

[18] P. Pradeep, J. E. R. Dhas, and M. Ramachandran, "Comparative analysis of bamboo using jute and coir fiber reinforced polymeric composites," *Int. J. Appl. Eng. Res.*, vol. 10, no. 11, pp. 10392–10396, 2015.

[19] B. Vijaylashmi, S. Renuka, P. Chennur, and S. Patil, "Female safety gadget system using GPS & GSM module," *JRET Int. J. Res. Eng. Technol.*, vol. 6, no. 5, pp. 411–415, 2019.

[20] S. Akojwar and P. Kshirsagar, "A novel probabilistic-PSO based learning algorithm for optimization of neural networks for benchmark problems," in *Proc. WSEAS Int. Conf. Neural Netw.*, 2016, pp. 1–10.

[21] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[22] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and K-means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 17–24, 2014.

[23] S. Özarpacı, B. Kılıç, O. C. Bayrak, A. Özdemir, Y. Yılmaz, and M. Floyd, "Comparative analysis of the optimum cluster number determination algorithms in clustering GPS velocities," *Geophys. J. Int.*, vol. 232, no. 1, pp. 70–80, Sep. 2022.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, vol. 1, no. 14, pp. 281–297.

[25] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: Wiley, 1990.

[26] B. Kılıç and S. Özarpacı, "Ensemble clustering in GPS velocities: A case study of Turkey," *Appl. Sci.*, vol. 12, no. 24, p. 12636, Dec. 2022.

[27] *Open Government Data Platform*, India. Accessed: Feb. 4, 2023. [Online]. Available: https://data.gov.in/

[28] T. Hlupic, D. Oreščanin, D. Ružak, and M. Baranovic, "An overview of current data lake architecture models," in *Proc. 45th Jubilee Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, May 2022, pp. 1082–1087.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[30] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist. Simul. Comput.*, vol. 3, no. 1, pp. 1–27, 1974.

KUSHAL AGRAWAL is currently pursuing the B.Tech. degree in computer science engineering with Vellore Institute of Technology, Chennai, India. His research interests include cyber physical systems, data analytics, and deep learning and its applications.

AVIRAL SRIVASTAVA (Member, IEEE) is currently pursuing the B.Tech. degree in computer science engineering with Vellore Institute of Technology, Chennai, India. His research interests include artificial intelligence, machine learning, data analytics, and deep learning and its applications.

KANISHK SHARMA is currently pursuing the B.Tech. degree in computer science engineering with Vellore Institute of Technology, Chennai, India. His research interests include artificial intelligence, machine learning, data analytics, and deep learning and its applications.

SANDEEP KUMAR SATAPATHY (Member, IEEE) received the Ph.D. degree in data mining and machine learning. His Ph.D. thesis includes a detailed classification of brain EEG signals using machine learning techniques. He was an Associate Professor with the Department of Computer Science and Engineering and the Head of the Department of Information Technology, Vignana Bharathi Institute of Technology, Hyderabad. He was an Associate Professor with the Centre for Advanced Data Science, VIT University, Chennai. He is currently a Postdoctoral Fellow with Yonsei University, Seoul, South Korea. He is highly engrossed in the areas of deep learning, image processing, and machine learning. He has many research publications to his credit, such as more than 40 research articles, three books, and many book chapters in various peer-reviewed journals. He has guided more than 15 master's thesis. He has also authored two books, such as Frequent Pattern Discovery from Gene Expression Data: An Experimental Approach (Elsevier) and EEG Brain Signal Classification for Epileptic Seizure Disorder Detection (Elsevier). He has been a member of various academic committees within the institution. He is a member of many professional organizations and societies. He has been an active reviewer of various peer-reviewed journals and prestigious conferences. He has also reviewed many research articles and books in Elsevier for possible publication.

SUNG-BAE CHO (Senior Member, IEEE) received the B.S. degree in computer science from Yonsei University, Seoul, South Korea, and the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He was an Invited Researcher with the Human Information Processing Research Laboratories, Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, from 1993 to 1995; and a Visiting Scholar with The University of New South Wales, Canberra, Australia, in 1998. He was a Visiting Professor with The University of British Columbia, Vancouver, Canada, from 2005 to 2006; and the King Mongkut's University of Technology, Thonburi, Bangkok, Thailand, in 2013. Since 1995, he has been a Professor with the Department of Computer Science, Yonsei University, and an Underwood Distinguished Professor, since 2021. He has published more than 230 journal articles and more than 680 conference papers. His research interests include neural networks, pattern recognition, intelligent man–machine interfaces, evolutionary computation, and artificial life. He was a recipient of the Richard E. Merwin Prize from the IEEE Computer Society, in 1993. He received several distinguished investigators awards from Korea Information Science Society, in 2005, and the GaheonSindoricoh, in 2017. He was a recipient of the Service Merit Medal from Korean Government, in 2022.

SHRUTI MISHRA (Member, IEEE) received the Ph.D. degree in computer science and engineering from Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India. She had been an Associate Professor with the Department of Computer Science and Engineering, Vignana Bharathi Institute of Technology, Hyderabad. She is currently an Assistant Professor (Senior) with the Centre of Advanced Data Science, Vellore Institute of Technology, Chennai. She has more than 30 articles in both national and international to her credit along with three books. She has guided more than 40 postgraduate and undergraduate students. She is the guest editor of many reputed publishers, such as Elsevier. She has also served as a reviewer for many reputed journals and conferences.

ABISHI CHOWDHURY (Member, IEEE) received the B.E. degree from the University Institute of Technology, West Bengal, India, the M.Tech. degree from the National Institute of Technical Teachers' Training and Research, Bhopal, Madhya Pradesh, India, and the Ph.D. degree from the Visvesvaraya National Institute of Technology, Nagpur, India. She is currently an Assistant Professor with Vellore Institute of Technology, Chennai Campus, Chennai, India. Her research interests include cloud computing, cloud resource scheduling, machine learning, and the Internet of Things.

AMRIT PAL (Member, IEEE) received the B.Tech. degree from Kurukshetra University, Kurukshetra, India, in 2011, the M.Tech. degree from the National Institute of Technical Teachers' Training and Research, Bhopal, India, in 2014, and the Ph.D. degree from Indian Institute of Information Technology Allahabad, India, in 2020. He was an Assistant Professor with the Centre for Advanced Studies, AKTU, Lucknow, India. He is currently an Assistant Professor with Vellore Institute of Technology, Chennai, India. His research interests include big data analytics, cloud computing, machine learning, and the Internet of Things.