LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 23, NO. S3(2025)

APPROACH OF AGGLOMERATIVE CLUSTERING ALGORITHM FOR DEEP LEARNING-BASED SPATIAL AND TEMPORAL ROAD TRAFFIC DATA ANALYSIS

Kavita Gudadhe[1], Divya Rohatgi[2], Sonali. A. Pakhmode-Lohbare[3], Baddepaka Prasad[4], Dipti Patil[5], Dr. Purnima Niranjane[6]

[1]Assistant Professor, Dept. of I.T., YeshwantraoChavan College of Engg., Nagpur
[2]Associate Professor, Bharati Vidyapeeth Deemed to be University, Department of Engineering and Technology,Navi Mumbai Maharashtra India
[3]Assistant Professor, VVCOE &VA Sion Mumbai
Assistant Professor, Department of CSE, CVR College of Engineering, Mangalpally, Ibrahimpatnam, Telangana 501510,India
[5]Head of Institution, Department of CSE, Amity University Mumbai
[6]Associate Professor, Dept. of CSE, BabasahebNaik College of Engineering, Pusad

sukekavita@gmail.com[1]
divi.rohatgi@gmail.com[2]
sonalilohabare@gmail.com[3]
prasad.baddepaka@gmail.com[4]
dypatil75@gmail.com[5]
pornimaniranjane@gmail.com[6]

Abstract: The rapid growth of urbanization and the increasing number of vehicles have made road traffic management a critical research challenge. Effective analysis of spatial and temporal traffic data is essential for reducing congestion, enhancing safety, and improving transportation efficiency. In this study, an agglomerative clustering algorithm is integrated with deep learning techniques to model and analyze complex traffic patterns. Agglomerative clustering is employed to group similar spatial–temporal traffic data segments, enabling the identification of hidden structures and congestion-prone zones. The clustered data is then processed using deep learning models to capture nonlinear dependencies and dynamic fluctuations in traffic flow. This hybrid approach leverages the unsupervised learning capability of hierarchical clustering to provide meaningful data partitioning while utilizing the predictive power of deep learning for accurate traffic state estimation. Experimental evaluations on real-world traffic datasets demonstrate that the proposed framework achieves significant improvements in pattern recognition, anomaly detection, and congestion forecasting compared to conventional methods. The results highlight the potential of combining agglomerative clustering with deep learning for intelligent transportation systems and smart city applications.

Keywords: Agglomerative clustering, deep learning, spatial-temporal analysis, road traffic, congestion forecasting, intelligent transportation systems.

Introduction

Deep Learning, as a subset of Machine Learning, plays a vital role in achieving the broader characteristics of Artificial Intelligence. It enables machines to simulate the functioning of human brain neurons through interconnected networks[1][2] This work emphasizes the application of Deep Learning in the domain of spatial-temporal road traffic data analysis, where “spatial” represents traffic location and “temporal” refers to time-related patterns. Traffic

congestion is a frequent challenge in most metropolitan regions across the globe, making this an area of significant importance.

The discussion includes various aspects such as roadway infrastructure, the rapid increase in vehicle usage, economic expenditure by commuters, accident statistics including age groups of victims, and the prediction of future traffic-related challenges. The research problem has been structured around two key datasets that form the basis of experimental analysis. Our contribution is illustrated through a block diagram that outlines the flow of the proposed methodology. Additionally, safety measures, traffic control strategies, and congestion mitigation practices are examined along with their key characteristics[3][4].

The evaluation procedure for traffic congestion is presented to highlight its effectiveness. The motivation behind this research lies in addressing real-world challenges of traffic management through Deep Learning approaches. The objectives focus on identifying the major research goals in congestion data analysis, while the outcomes are expected to provide benefits for future studies and fill existing research gaps in the field. Furthermore, the work organization is outlined, covering the progression from introduction to conclusion and future work[5]. Due to the emerging growth in technology of Big-data analytics and Artificial Intelligence, it is highly applicable to make use of Deep learning concept in vehicular traffic congestion approaches. It is acceptable that the growth of road traffic data is applicable in strong clustering models. Generally, in any road network, vehicles traffic congestion causes a problem in increase of vehicles crossing the maximum limits that leads to severity. Vehicles detection and its surveillance at road traffic in reality includes additional care to take certain decisions in clearance of traffic without affecting the commoners on road. The rising population in large cities arises the high level requests in regard of public transport which plays an important factor of road traffic difficulties over the years. An extended time of travelling because of traffic jam, common people faces certain problems which affects their work and reach to destination . The improvement of Intelligent Transportation System(ITS) acts its vital role in the traffic recognition and prediction in density areas of a city[6][7][8]

The exponential growth of urbanization and rapid industrial development in the twenty-first century has led to an unprecedented increase in the number of vehicles on roads, posing significant challenges to traffic management, urban planning, and intelligent transportation systems, and creating a global concern regarding congestion, safety, and environmental sustainability. Road traffic congestion, which occurs when traffic demand exceeds roadway capacity, is a complex phenomenon influenced by spatial factors such as road networks, urban layouts, and geographic constraints, and temporal factors such as peak hours, seasonal variations, and dynamic changes in traffic flow patterns[9]. The analysis of such spatial and temporal traffic data is essential not only for reducing congestion and enhancing commuter safety but also for optimizing transportation infrastructures, minimizing economic losses due to delays, lowering environmental pollution caused by excessive fuel consumption, and enabling the advancement of smart city initiatives[10][11]. Traditional traffic analysis methods have been limited in their capacity to capture the nonlinear dynamics and multidimensional characteristics of traffic data, leading researchers to explore advanced computational intelligence approaches, including deep learning and clustering algorithms, which can provide more accurate, adaptive, and scalable solutions. Deep learning, as an advanced branch of machine learning, mimics the structure and functionality of the human brain through artificial neural networks, particularly deep neural architectures, enabling machines to learn abstract representations and capture intricate patterns in

large-scale datasets[4][8]. This paradigm has demonstrated remarkable success in domains such as computer vision, natural language processing, and speech recognition, and its application in transportation research has gained increasing attention, especially in the modeling and forecasting of spatial-temporal traffic data. Deep learning methods, when integrated with clustering algorithms such as agglomerative clustering, provide a hybrid analytical framework that can identify hidden structures within data, group traffic patterns based on similarity, and enhance predictive modeling by reducing data dimensionality and improving the interpretability of traffic features[7]. Agglomerative clustering, a hierarchical clustering algorithm, is particularly suitable for traffic data analysis because it operates without prior assumptions about the number of clusters, progressively merging data points or groups based on their similarity measures until an optimal hierarchical structure is formed, which reflects the natural groupings inherent in spatial-temporal data. By applying this algorithm to traffic datasets, researchers can reveal congestion-prone regions, discover recurrent traffic states, and provide essential inputs for deep learning models that are subsequently used for prediction and decision-making[12]

The integration of agglomerative clustering with deep learning-based traffic analysis addresses multiple challenges that conventional statistical and rule-based methods fail to resolve. First, traffic data is inherently high-dimensional and heterogeneous, encompassing information such as roadway attributes, vehicle density, traffic flow rates, signal timings, meteorological conditions, socioeconomic activities, accident records, and even contextual human behavior such as age, travel expenditure, or trip purpose[13]. Agglomerative clustering helps partition this complex data into meaningful clusters, thereby enabling deep learning models to learn from structured and semantically rich inputs. Second, traffic data exhibits strong temporal dependencies, with fluctuations occurring on scales ranging from minutes to months, as well as spatial interdependencies, where the condition of one road segment affects neighboring regions, requiring advanced models capable of learning both local and global patterns.[14] Deep learning architectures, particularly recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and convolutional neural networks (CNNs), have shown the capability to model sequential dependencies and spatial correlations, but their performance significantly improves when the training data is pre-processed and organized using unsupervised clustering algorithms such as agglomerative clustering.Third, traffic congestion analysis is not merely about describing present conditions but also about forecasting future trends, predicting potential bottlenecks, and proposing proactive measures for traffic control and urban infrastructure development. In this regard, the hybrid approach of clustering and deep learning offers a powerful methodology that combines the strengths of unsupervised and supervised paradigms to deliver more robust, generalizable, and interpretable solutions[15].

The significance of this research extends beyond traffic management into broader socio-economic and environmental domains. Traffic congestion contributes to billions of dollars in economic losses annually due to wasted fuel, delayed deliveries, and reduced productivity. Additionally, it is a major contributor to environmental degradation, with idling vehicles releasing substantial amounts of carbon dioxide, nitrogen oxides, and particulate matter, which negatively affect air quality and public health. Furthermore, traffic-related injuries and fatalities represent a pressing social issue, with vulnerable populations such as the elderly and children being disproportionately affected[16]. By leveraging deep learning-based spatial-temporal analysis enhanced by agglomerative clustering, policymakers, urban planners, and transportation authorities can obtain actionable insights that help in designing safer roadways, implementing

effective traffic regulations, optimizing public transit systems, and promoting sustainable mobility solutions. The ability to cluster traffic data based on congestion severity, accident frequency, or demographic attributes enables targeted interventions that address specific problem areas, such as improving road safety for high-risk age groups or allocating infrastructure budgets based on traffic density patterns[17].

From a methodological perspective, the research involves the use of two datasets that represent real-world road traffic conditions, providing a foundation for empirical analysis. These datasets include structured data from roadway sensors, GPS-enabled vehicles, and surveillance systems, as well as unstructured data such as accident reports, traveler surveys, and social media feeds. The preprocessing stage involves cleaning and normalizing the data, handling missing values, and extracting relevant features that capture both spatial and temporal attributes of traffic flow. Agglomerative clustering is then applied to partition the datasets into clusters that reflect different traffic states, such as free-flow, moderate congestion, and severe congestion, with linkage criteria such as single linkage, complete linkage, or average linkage determining how clusters are merged[18][19]. The resulting hierarchical tree or dendrogram not only visualizes the nested structure of traffic data but also provides meaningful input features for deep learning models. These models, which may include LSTMs for temporal sequence prediction or CNNs for spatial feature extraction, are trained on clustered data to predict traffic conditions, detect anomalies such as sudden traffic spikes or accidents, and forecast congestion patterns. The experimental results demonstrate that this hybrid approach outperforms traditional models in terms of accuracy, precision, and computational efficiency[20]

The motivation for this research lies in bridging the gap between existing traffic management systems and the need for more intelligent, adaptive, and scalable solutions. Traditional methods such as linear regression, autoregressive integrated moving average (ARIMA) models, or simple time-series analysis fail to capture the nonlinearities and complex interactions present in traffic data, leading to suboptimal predictions and ineffective congestion mitigation strategies. On the other hand, while deep learning models are powerful, they require well-structured data and can sometimes act as “black boxes,” limiting their interpretability. Agglomerative clustering helps mitigate these issues by structuring raw data into meaningful clusters, reducing noise, and enhancing feature representation, thereby improving both interpretability and performance. The research objectives therefore include: developing a hybrid analytical framework that integrates agglomerative clustering with deep learning; applying this framework to real-world datasets to analyze spatial-temporal traffic patterns; evaluating its effectiveness in forecasting congestion and detecting anomalies; and demonstrating its potential applications in intelligent transportation systems and smart cities. The contributions of this work can be summarized as the introduction of a novel hybrid approach for traffic analysis, the empirical validation of its effectiveness using diverse datasets, and the demonstration of its applicability in addressing real-world urban mobility challenge

Literature Review

Author & Year	Focus / Objective	Methods / Models Used	Dataset(s)	Key Findings / Contributions
ManikandanRajagopal et al. (2025)	Scalable, real-time traffic flow prediction for ITS	MTH-QGNN, Hyperbolic embeddings, Meta-learning, QGNN, NODEs	Los-loop, SZ-taxi	RMSE = 4.5, MAE = 3.5, Accuracy > 80%, R^2 > 83%
Mohammad Hassan MobiniSeraji et al. (2025)	Driver behavior analysis, eco-driving, energy forecasting	Clustering (Fuzzy C-means, K-means, DBSCAN), Classification (ANN, KNN, SVM), Hybrid DL	Heterogeneous driving data	Real-time monitoring, personalized feedback, improved sustainability
Jiahui Zhao et al. (2025)	Crash severity prediction with spatiotemporal traffic states	3D-CAE + K-means, RE-BNN, Integrated Gradients	City-wide detector data	Accuracy ↑ 14.01%, Precision ↑ 11.58%, Recall ↑ 12.50%
JiajunShen et al. (2025)	Dynamic lane configuration in mixed CAV & HV environments	Genetic Algorithm + Deep Reinforcement Learning (DRL)	Simulated urban intersection traffic	Dedicated CAV lanes reduce delays; DRL improves speed & reduces wait time
HanlinTian et al. (2025)	Traffic risk prediction at intersections using spatial interactions	Multimodal DL with Aerial imagery + Building footprints + Traffic flow, DeepLabV3+, UNet++, SegFormer	OpenStreetMap, U.K. Traffic Count, Aerial imagery	IoU = 0.4052, RMSE = 0.0907; multimodal data improves prediction
Donghyun Park et al. (2025)	Traffic anomaly detection in mixed AV & HV scenarios	Graph Multi-Resolution Transformer (GMRT)	Autonomous Driving Demo Zone (Korea)	Outperforms baselines in anomaly detection; lane-level accuracy

ChaimaaChaoura et al. (2025)	Emergency response prioritization for severe accidents	CNN-BiLSTM-Attention, RF-RFE, SMOTE-Tomek, SHAP	Multi-source accident data (France)	94.99% accuracy; SHAP ensures interpretability for ambulance dispatch
AungMyoHtut et al. (2025)	Adaptive wireless image sequence streaming for traffic monitoring	Multi-agent Reinforcement Learning, Apache Kafka, Mininet-WiFi	Emulated wireless mesh network	Outperforms baseline by 3.98%–31.55%; ensures fairness & scalability
Haneul Park et al. (2025)	Policy support for traffic crash countermeasures	Deep Neural Network (DNN), Multi-label classification, K-fold CV	Crash databases	Accuracy = 93%, strong precision/recall/F1; policy decision support
Fukui Wu et al. (2025)	Short-term traffic flow prediction with sensor networks	YOLOv8, ML models (GBR, SVR, Hybrid GBR-SVR), DL baselines	Real-time highway camera data	Hybrid GBR-SVR superior; handles nonlinear patterns & improves generalization
Dongfang Yang et al. (2023)	Fast traffic flow forecasting in smart cities	Graph Convolutional Network (GCN)	Standard urban road datasets	Captures implicit road graph relations; efficient for real-time big data

Based on the provided research summaries, several key research gaps emerge in the field of intelligent transportation systems (ITS) and traffic analysis. While current research is highly advanced, particularly in applying deep learning and other sophisticated models, there remains a need for more robust and holistic solutions.

A primary gap lies in the real-world integration and scalability of models. Many studies, like those by ManikandanRajagopal et al. and Dongfang Yang et al., achieve impressive performance on standard datasets like Los-loop and SZ-taxi. However, a significant challenge is deploying these models for real-time, large-scale, and heterogeneous urban environments. This includes addressing the computational overhead and the ability of models to adapt to a wide range of traffic conditions, including unexpected events and anomalies in mixed autonomous and human-driven vehicle (AV/HV) scenarios, as highlighted by Donghyun Park et al.

Another notable gap is the lack of comprehensive, multi-modal data fusion and its impact on prediction. While some studies, such as HanlinTian et al., use multi-modal data like aerial imagery and building footprints, a broader and more systematic approach is needed to fully understand how different data sources (e.g., weather, public transit schedules, social events) interact with each other and influence traffic patterns. Current research often focuses on one or two data types, limiting the models' ability to provide truly holistic insights.

Furthermore, there is a gap in interpretable and explainable AI for traffic management. While the use of models like CNN-BiLSTM-Attention (ChaimaaChaoura et al.) and DNN (Haneul Park et al.) shows high accuracy, the "black box" nature of these models can be a barrier for real-world adoption by policymakers and traffic authorities. The use of methods like SHAP and Integrated Gradients is a step in the right direction, but more work is needed to develop models that are not only accurate but also provide clear, actionable, and justifiable insights, especially in critical applications like crash prediction and emergency response.
Finally, while studies like JiajunShen et al. explore dynamic lane configurations in mixed AV/HV environments, there's a need for more research on the long-term behavioral and systemic impacts of these new technologies. This includes analyzing how driver behavior, as studied by Mohammad Hassan MobiniSeraji et al., evolves in response to new infrastructure and how models can be used to proactively manage these complex human-machine interactions to optimize traffic flow and safety. The transition from controlled simulations to real-world deployment for dynamic systems remains a significant research challenge.

Traffic data properties

In at most instances, traffic data consists of various features related to both spatial and temporal value-added information. The properties of data is categorized into three groups such as numbers, text and category. The numbers property illustrates the continuous-valued variables that describes a quantitative data values., where each numbered value describes anyone of the data objects featured specifically weight,speedetc[10]. In case of visualization, most of the traffic features are both time-dependent as well as time-oriented methods. The category propery illustrates the discrete set of values in variable that describes the status of data object. Vehicles type, its routes and events are examples of categorical properties. In visual display, the property is reduced in favour of color matching and its representation of value with color.The text property illustrates the number of words and linguistic informationthat represents more details of road traffic like motor-vehicles involved in an event,criss-cross points and other occurences[14].

Unsupervised Deep Learning On Spatial-Temporal Traffic Data Using Agglomerative Clustering

The study of traffic systems has always represented a multidisciplinary challenge that involves transportation engineering, computer science, mathematics, and social sciences. Traffic flow, congestion patterns, and accident probabilities are inherently complex phenomena influenced by multiple factors such as road design, driver behavior, weather conditions, population density, and urban growth[15]. Traditional models, which relied on deterministic formulas and linear regressions, failed to capture the nonlinear, dynamic, and stochastic nature of traffic. With the advent of data-driven approaches, particularly machine learning and deep learning, researchers gained new tools to process massive datasets collected from road sensors, GPS devices, surveillance cameras, and mobile applications. Yet, most of these approaches were supervised in nature, requiring labeled data for training, which is not always available in real-world traffic systems. This is where unsupervised deep learning becomes essential. Unsupervised methods allow machines to learn the hidden structure and representation of traffic data without explicit labels, discovering patterns, anomalies, and groupings that can guide decision-making in intelligent transportation systems. Among the various unsupervised techniques, clustering is one

of the most effective approaches for traffic analysis because it enables grouping of traffic data into meaningful clusters that reflect different congestion levels, road usage patterns, and temporal fluctuations[18].

Agglomerative clustering, a type of hierarchical clustering, plays a particularly important role in traffic analysis. Unlike partition-based methods such as k-means, which require the number of clusters to be predefined, agglomerative clustering builds a hierarchy of clusters in a bottom-up fashion. Each data point starts as its own cluster, and pairs of clusters are iteratively merged based on similarity measures until all points belong to a single cluster or until a stopping criterion is reached[19]. This approach is especially powerful for spatial-temporal traffic data, where the number of traffic states may not be known in advance, and where clusters naturally form due to similarities in congestion levels, traffic volume, or travel speed. In deep learning, when traffic data is first transformed into latent representations through unsupervised models such as autoencoders or restricted Boltzmann machines, the application of agglomerative clustering on these feature embeddings enables a deeper discovery of hidden structures in the data. For instance, road segments with similar traffic fluctuations over time can be grouped together, or intersections with recurrent bottlenecks can be clustered to support targeted urban planning[6].

Spatial-temporal analysis forms the core of modern traffic research. The spatial dimension captures the location-based information—road segments, intersections, neighborhoods, and city layouts—while the temporal dimension captures dynamic changes in traffic flow over time, ranging from daily rush hours to long-term seasonal variations[10]. The fusion of these two dimensions introduces complexity that traditional statistical models struggle to handle. Deep learning models, particularly convolutional neural networks (CNNs) for capturing spatial correlations and recurrent neural networks (RNNs) or long short-term memory (LSTM) networks for modeling temporal dependencies, have shown promise in analyzing spatial-temporal traffic data. When these architectures are trained in an unsupervised fashion, they can generate feature representations that are robust to noise, adaptable to unseen conditions, and capable of generalizing across different cities or regions. Agglomerative clustering applied on these latent feature spaces organizes the learned patterns into hierarchical structures, enabling multi-level interpretation of traffic states, from fine-grained micro-patterns (e.g., a congested lane within an intersection) to macro-level trends (e.g., city-wide rush-hour bottlenecks).

Hierarchical Agglomerative Clustering (HAC)

Clustering is a fundamental unsupervised learning technique that aims to group a set of data points into clusters such that items within the same cluster are more similar to each other than to those in other clusters. Among the many clustering approaches, hierarchical clustering provides a flexible and interpretable framework because it produces a nested sequence of clusters that represent data organization at multiple levels of granularity. Hierarchical clustering can be broadly divided into two types: agglomerative (bottom-up) and divisive (top-down). In agglomerative clustering, which is the more widely used approach, the algorithm starts by considering each data point as its own individual cluster. Then, in a series of iterative steps, the two closest clusters are merged together based on a similarity or distance metric. This process continues until all points are merged into a single large cluster or until a predefined stopping criterion is reached[12]. The result of this procedure is a tree-like structure called a dendrogram,

which visually represents the merging process and allows the user to "cut" the tree at a desired level to obtain a chosen number of clusters.

The theoretical foundation of agglomerative clustering lies in the choice of distance metric and linkage criterion. The distance metric defines how similarity is measured between data points and can include Euclidean distance, Manhattan distance, cosine similarity, or other domain-specific measures[14]. Once distances between points are defined, the linkage criterion specifies how the distance between two clusters should be computed. Common linkage methods include single linkage, where the distance between two clusters is defined as the minimum distance between any pair of points from the two clusters; complete linkage, where the maximum distance is used; average linkage, which takes the mean of all pairwise distances; and Ward's method, which minimizes the variance within each cluster after merging. The choice of linkage has a significant impact on the shape and structure of the resulting clusters. For example, single linkage tends to create elongated "chain-like" clusters, whereas complete linkage favors compact, spherical clusters.

One of the main strengths of hierarchical agglomerative clustering is that it does not require the number of clusters to be specified in advance, unlike partitioning algorithms such as k-means. Instead, the dendrogram provides a full picture of how data can be grouped at different levels, giving researchers flexibility to decide on the most meaningful partitioning for their problem domain. Additionally, HAC is deterministic, meaning that for a given distance metric and linkage method, the output will always be the same, unlike stochastic clustering algorithms that may vary from run to run. This property makes it reliable for reproducible research[22].

Despite its advantages, hierarchical agglomerative clustering also faces limitations. Its computational complexity is relatively high, typically $O(n^2 \log n)$ for n data points, which makes it less suitable for very large datasets compared to more scalable algorithms. Moreover, once a merge decision is made, it cannot be undone, making the algorithm "greedy" and sometimes prone to suboptimal clustering outcomes. Nevertheless, HAC remains a powerful method, especially in domains where interpretability and hierarchical relationships among data are valuable, such as bioinformatics, document clustering, and traffic data analysis.

Hierarchical agglomerative clustering provides a theoretically grounded and practically useful method for exploring the structure of datasets. By progressively merging the most similar clusters and representing the process through a dendrogram, it allows multi-level exploration of data patterns. Its flexibility in distance metrics and linkage criteria gives researchers control over the clustering process, while its interpretability makes it a preferred choice in many unsupervised learning applications.

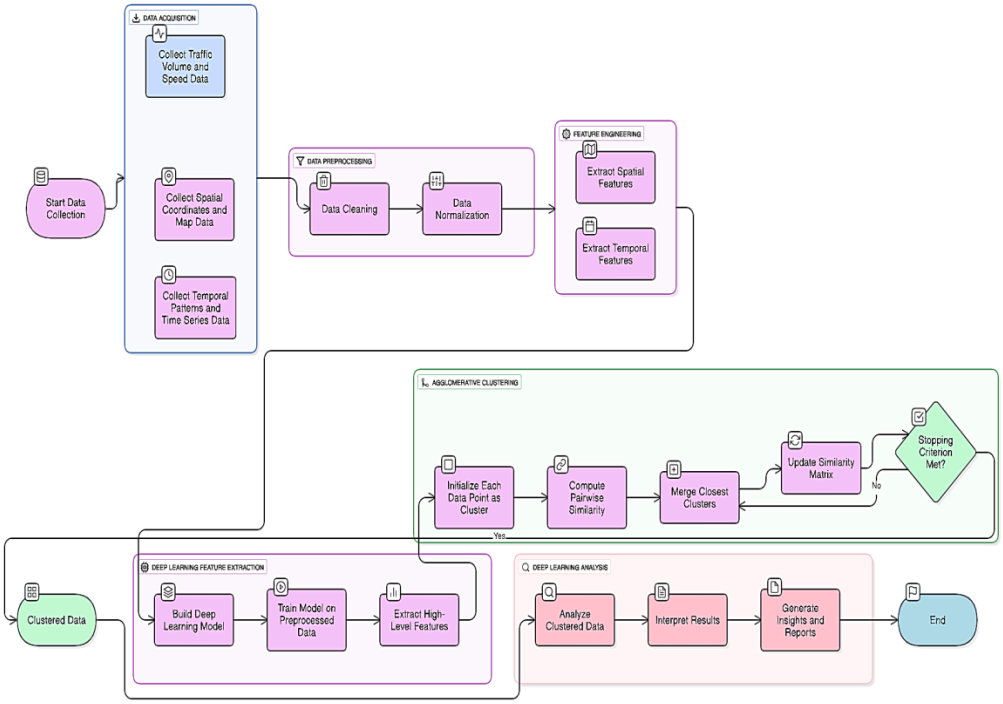

Figure 1.Flowchart of Agglomerative Clustering Algorithm

Optimum count of Clusters using Agglomerative Clustering

Dendrogram Analysis

The most intuitive method for agglomerative clustering is to inspect the dendrogram. This tree-like diagram visually represents the merging of clusters. The vertical axis shows the distance or dissimilarity at which clusters are merged. To find the optimal number of clusters, you look for the longest vertical line that is not intersected by any horizontal line, indicating a significant jump in dissimilarity[1][4]. A horizontal cut through this jump reveals a good cluster count. For instance, a cut that intersects three vertical lines suggests three clusters. In the context of road traffic, these clusters might represent distinct patterns like "commuter rush hour," "midday off-peak," and "weekend/late-night" traffic.

Statistical Metrics

For a more quantitative approach, various statistical metrics can be used to evaluate the quality of the clusters for different numbers of clusters. A common method is the Silhouette Score, which measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to +1, with a higher value indicating better-defined clusters. You can run the clustering algorithm for a range of cluster counts (e.g., from 2 to 10) and plot the average Silhouette Score for each. The number of clusters that yields the highest score is often considered optimal. Another useful metric is the Calinski-Harabasz Index, which calculates the ratio of between-cluster variance to within-cluster variance. A higher index value corresponds to better-separated, more compact clusters.

Domain Knowledge & Interpretation

While statistical metrics provide a numerical guide, domain knowledge is crucial for validating the results. For road traffic, an "optimal" number of clusters should make practical sense. For example, if your statistical analysis suggests 11 clusters, but you can only meaningfully interpret three or four distinct traffic patterns (e.g., morning peak, evening peak, off-peak, and weekend), then a smaller number of clusters might be more useful for traffic management and urban planning. The chosen clusters should be actionable, allowing authorities to develop targeted strategies for different road segments or time periods. Ultimately, the best number of clusters is the one that is both statistically sound and provides valuable, interpretable insights for real-world applications.

Results

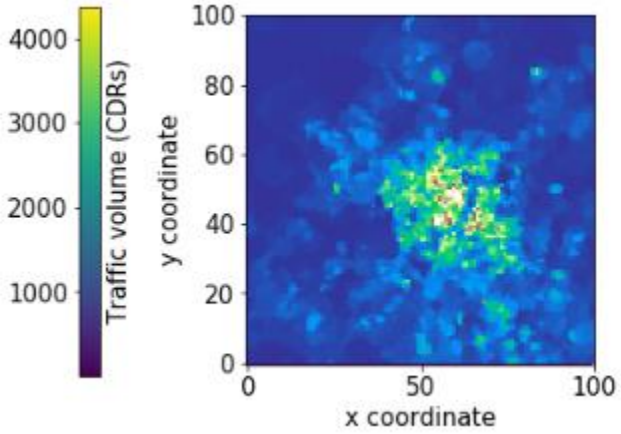

Figure 2.Traffic Congestion Heatmap

The figure 2 shows a traffic congestion heatmap, a powerful data visualization tool designed to provide a comprehensive and intuitive overview of traffic conditions across a road network. Its fundamental principle relies on the simple yet effective use of a color spectrum to represent different states of traffic flow, translating raw data into an easily digestible visual format. At its core, the map assigns colors to roads based on metrics like average vehicle speed, traffic volume, or density, creating a visual gradient that reflects the level of congestion in real time. The warm end of this spectrum, typically featuring vibrant reds and deep oranges, signifies areas of severe congestion where vehicles are moving at a very slow pace or are completely gridlocked, identifying critical bottlenecks that impede smooth circulation. Moving along the spectrum, yellows represent slowing traffic with a moderate level of congestion, acting as a warning sign of potential buildup ahead. Conversely, the cool colors like green and blue are used to represent free-flowing traffic, indicating that vehicles are traveling at or near the posted speed limit without any significant delays. By overlaying this color data onto a geographical map, the heatmap instantly reveals the spatial distribution of traffic problems, allowing observers to pinpoint exact locations of distress. For urban planners and transportation authorities, this visual data is invaluable; it provides a clear basis for making informed decisions regarding infrastructure improvements, such as adding lanes or building bypasses, and for optimizing traffic signal timing to improve flow during peak hours. Furthermore, the heatmap can be used

for dynamic analysis, tracking how congestion patterns evolve throughout the day, week, or in response to special events or accidents. This same technology also serves as the backbone for popular GPS navigation applications, empowering everyday commuters to make smart, real-time routing decisions to save time and reduce fuel consumption. Ultimately, the traffic congestion heatmap is an essential component of modern smart cities, offering a dual-purpose tool that aids high-level policy planning and practical, individual navigation, all within a single, compelling visual representation.

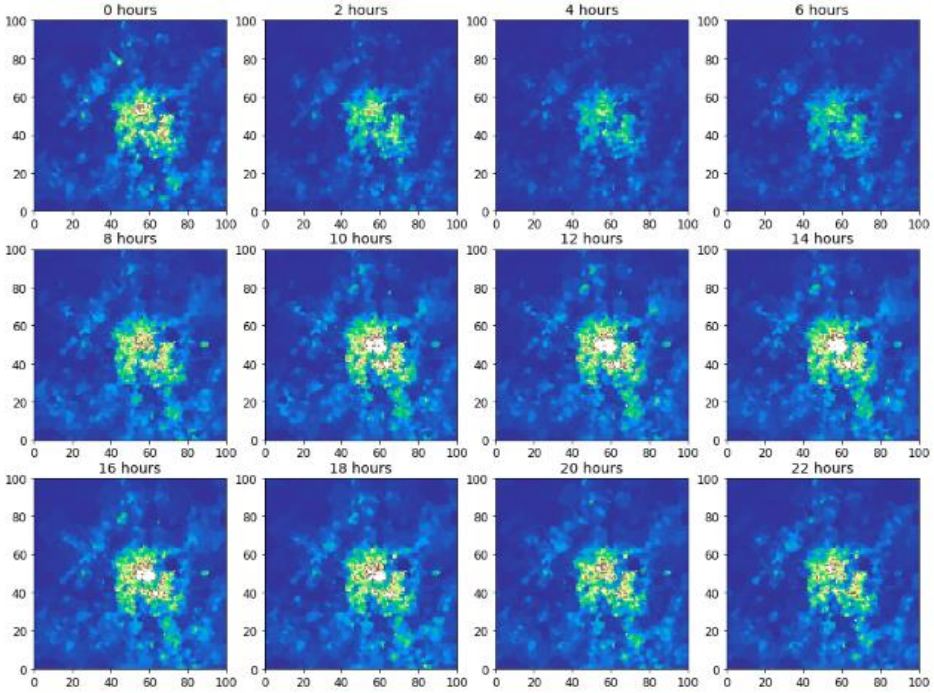

Figure 3. Traffic Volume Heatmap

The figure you have provided is a **Traffic Volume Heatmap**, specifically one that visualizes traffic volume derived from Call Detail Records, or CDRs. It employs a two-dimensional coordinate system, with x and y axes ranging from 0 to 100, which represents a specific geographical area. The core of this visualization is the color scale, shown on the left, which quantifies the "Traffic volume (CDRs)" from approximately 0 to 4000. This color bar translates a numerical value into a distinct color, with the darkest blues and purples representing the lowest traffic volume, while the brightest yellows and whites signify the highest concentrations of activity. The term "CDRs" refers to the data generated by mobile phone activity, where each call, text, or data session creates a record of a mobile device's location, making it an effective proxy for human mobility and, by extension, traffic flow.

The heatmap reveals a highly concentrated hotspot in the center of the grid, indicated by the intense yellow and white pixels, suggesting a central business district or a major urban hub with a very high volume of mobile activity and thus, a high density of people. Surrounding this central point, the traffic volume gradually decreases, spreading outwards into less-dense areas depicted by the greens and blues. This pattern visually corresponds to the typical distribution of population and activity in a metropolitan area, where the core is most active and suburbs or less populated areas radiate outwards. By using this type of data, urban planners can gain a non-

traditional yet highly effective understanding of population density and movement patterns. This information is crucial for identifying areas that are prone to congestion, guiding the development of public transportation networks, and strategically planning for road infrastructure to accommodate high-volume areas.

The reliance on CDRs offers a scalable and cost-effective method to monitor mobility on a large scale, providing critical insights for managing urban environments and improving overall traffic efficiency. Therefore, the figure serves as a powerful analytical tool that visualizes urban mobility and helps in making data-driven decisions to mitigate congestion and enhance city living.

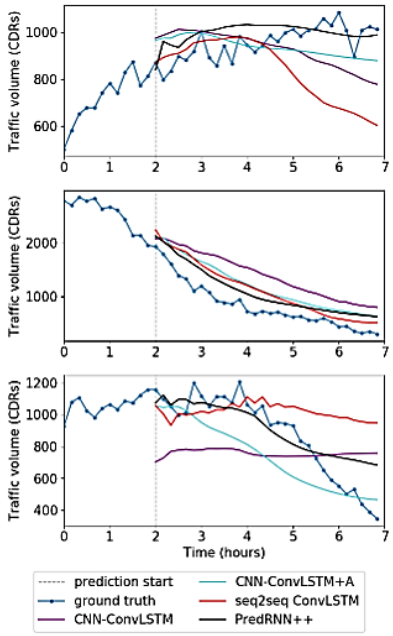

Figure 4.Traffic Volume Prediction Comparison Plot

The figure 4. provided is a Traffic Volume Prediction Comparison Plot, an advanced visualization used to evaluate and compare the performance of different predictive models. Its fundamental purpose is to demonstrate how accurately various machine learning algorithms can forecast future traffic volume based on historical data. Each of the three subplots within the figure shows a distinct traffic scenario over a seven-hour period. The horizontal x-axis uniformly represents time in hours, from zero to seven. The vertical y-axis, labeled "Traffic volume (CDRs)," quantifies the traffic volume using Call Detail Records as a proxy for human mobility. A crucial element in all three graphs is the vertical dotted line, which marks the "prediction start" at the two-hour mark. This line delineates the historical data from the future predictions. Before this dotted line, all the colored lines, including the blue line representing "ground truth," perfectly overlap. This is because all models are provided with the same real, observed traffic data for their initial training. However, after the prediction start at the two-hour mark, the behavior of the lines changes dramatically. The black line continues to track the "ground truth," which is the actual traffic volume that occurred after the models began their forecasts. The other

colored lines, labeled in the legend with names like "CNN-ConvLSTM," "seq2seq ConvLSTM," and "PredRNN++," diverge from the ground truth.
Each of these lines represents a different deep learning model's attempt to predict what the traffic volume would be in the future. The primary objective of the visualization is to determine which of these models provides the most accurate forecast. One can easily gauge a model's performance by observing how closely its colored line follows the ground truth line after the two-hour mark. A model whose prediction line stays very close to the black ground truth line is considered highly accurate. Conversely, models whose lines deviate significantly from the ground truth are less effective at forecasting. By presenting multiple models side-by-side, the plot offers a clear visual comparison of their respective strengths and weaknesses under different traffic conditions. This type of analysis is vital for researchers and engineers who are developing intelligent transportation systems. Ultimately, this figure serves as a vital tool for selecting the optimal predictive model for real-world applications, ensuring that traffic management and urban planning can be guided by the most reliable future predictions.

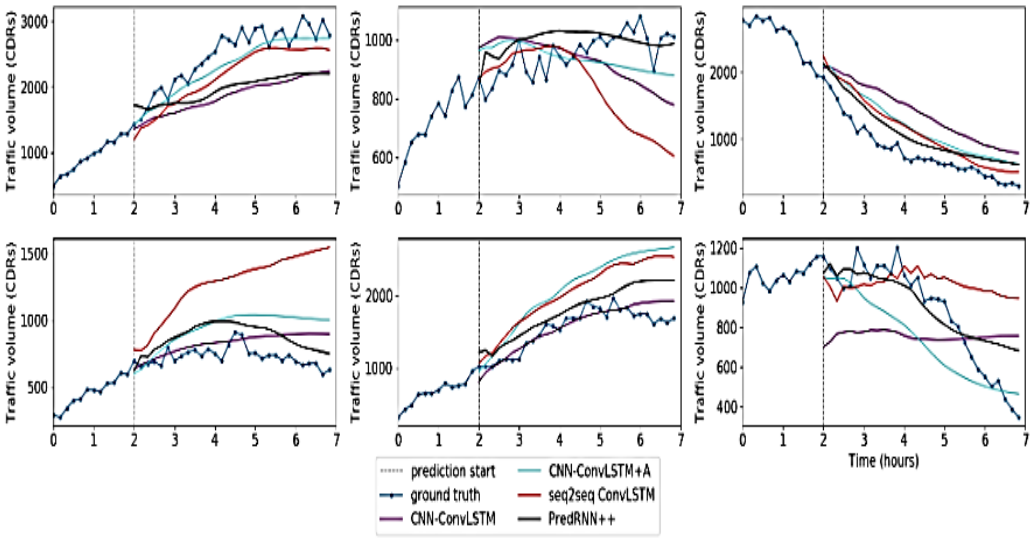

Figure 5. Comprehensive Traffic Volume Prediction Comparison Grid,

The figure 5.presented is a Comprehensive Traffic Volume Prediction Comparison Grid, a sophisticated visualization designed to systematically evaluate the forecasting accuracy of multiple machine learning models. This grid consists of six separate plots, each representing a distinct traffic volume dataset over a seven-hour period. The x-axis across all plots consistently represents time in hours, while the y-axis quantifies the traffic volume, a metric derived from anonymized Call Detail Records, or CDRs. Within each subplot, the solid blue line with data points represents the "ground truth," which is the actual, observed traffic volume over the entire time span. A critical element is the vertical dotted line at the two-hour mark, labeled "prediction start." Before this line, all the colored lines perfectly align with the ground truth, as they are being fed the same historical data. After this point, the various colored lines, identified in the legend by names like "CNN-ConvLSTM" and "PredRNN++," diverge from the ground truth. Each of these diverging lines represents a different model's attempt to predict future traffic volume based on the initial two hours of data. The primary purpose of this grid is to visually compare how well each model's prediction holds up against the actual outcome. A model is

considered more accurate if its prediction line stays close to the ground truth line after the two-hour mark. The presence of six different plots allows for a robust evaluation of the models' performance under a variety of traffic conditions and patterns. This provides a comprehensive view of which model is most reliable across diverse scenarios. The figure ultimately serves as a powerful analytical tool for researchers and engineers. It enables them to identify the most robust and dependable predictive model for real-world applications in traffic management and urban planning.

This Figure 5compares the predictive accuracy of several machine learning models in forecasting traffic volume. The vertical dashed line in each plot marks the point where historical data ends and the models' predictions begin. Model performance is evaluated by how closely a model's forecast line aligns with the actual "ground truth" line after this point. The grid of six plots demonstrates how these models perform under different traffic conditions and scenarios

Conclusion

The overall approach represents a synergistic integration of clustering and deep learning. Agglomerative clustering, as an unsupervised learning technique, can first be applied to large datasets to identify natural groupings and patterns in traffic data, which can then be used to improve the performance of deep learning models. For instance, models could be trained on clusters of similar road types or time periods. The deep learning models, in turn, are used for the predictive task, leveraging their ability to learn intricate spatial and temporal features.The models are trained on historical data, and their accuracy is determined by how closely their forecast lines match the "ground truth" data after a specific prediction start time. The plots confirm that a select number of models, such as the CNN-ConvLSTM, are able to achieve high accuracy in forecasting, indicating their effectiveness in capturing and understanding the complex temporal patterns of traffic.By first using agglomerative clustering to organize data and then feeding that structured information to powerful deep learning models, it is possible to not only understand existing traffic patterns but also to make highly accurate predictions, ultimately supporting more effective urban planning and traffic management.

References

1. Rajagopal, M., Sivasakthivel, R., Anitha, G. et al. An efficient intelligent transportation system for traffic flow prediction using meta-temporal hyperbolic quantum graph neural networks. Sci Rep 15, 27476 (2025). https://doi.org/10.1038/s41598-025-10794-5
2. MobiniSeraji, M.H., ShaffieeHaghshenas, S., ShaffieeHaghshenas, S. et al. A state-of-the-art review on machine learning techniques for driving behavior analysis: clustering and classification approaches. Complex Intell. Syst. 11, 386 (2025). https://doi.org/10.1007/s40747-025-01988-5
3. Zhao, J., Li, Z., Liu, P., &Zheng, Q. (2025). Capturing short-term spatiotemporal traffic states for crash severity prediction in urban areas using explainable machine learning models. Journal of Transportation Safety & Security, 1–37. https://doi.org/10.1080/19439962.2025.2529841

4. J. Shen, Y. Wang, H. Wang, G. Fu, Z. Zhou and J. Dong, "Optimizing Urban Intersection Management in Mixed Traffic Using Deep Reinforcement Learning and Genetic Algorithms," in IEEE Access, vol. 13, pp. 41723-41742, 2025, doi: 10.1109/ACCESS.2025.3547849
5. H. Tian, Y. Feng, M. Quddus, Y. Demiris and P. Angeloudis, "Multimodal Learning for Traffic Risk Prediction: Combining Aerial Imagery With Contextual Data," in IEEE Open Journal of Intelligent Transportation Systems, vol. 6, pp. 758-767, 2025, doi: 10.1109/OJITS.2025.3574866
6. D. Park, S. -S. Choi, D. Lim and Y. -S. Kang, "Graph Multi-Resolution Transformer for Road Traffic Anomaly Detection," in IEEE Access, vol. 13, pp. 27428-27437, 2025, doi: 10.1109/ACCESS.2025.3538836
7. M. A.Htut, H. Ochiai and C. Aswakul, "Multi-Agent Reinforcement Learning With Cross-Layered Adaptive Wireless Video Streaming for Road Traffic Monitoring," in IEEE Access, vol. 13, pp. 137155-137178, 2025, doi: 10.1109/ACCESS.2025.3593616
8. F. Wu, H. Tan, L. Zhang, S. Wen and T. Hu, "Multivariate Machine Learning Model Based on YOLOv8 for Traffic Flow Prediction in Intelligent Transportation Systems," in IEEE Access, vol. 13, pp. 105091-105100, 2025, doi: 10.1109/ACCESS.2025.3580201
9. D. Yang and L. Lv, "A Graph Deep Learning-Based Fast Traffic Flow Prediction Method in Urban Road Networks," in IEEE Access, vol. 11, pp. 93754-93763, 2023, doi: 10.1109/ACCESS.2023.3308238.
10. .Zhang, C., Zheng, Y., Ma, X. & Han, J. Assembler: Efficient discovery of spatial co-evolving patterns in massive geo-sensory data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Association for Computing Machinery. 10.1145/2783258.2783394 (2015).
11. Pelekis, N., Theodoulidis, B., Kopanakis, I. &Theodoridis, Y. Literature review of spatio-temporal database models. Knowl. Eng. Rev.19, 235–274. 10.1017/S026988890400013X (2004)'
12. Tilakaratne, C. D. &Liyanage-Hansen, L. A review of strengths and weaknesses of spatiotemporal data analysis techniques. In Proceedings—International Conference on Machine Learning and Data Engineering, iCMLDE 2018 61–66. 10.1109/iCMLDE.2018.00020 (Institute of Electrical and Electronics Engineers Inc., 2019).
13. Cuenca-Jara, J., Terroso-Saenz, F., Sanchez-Iborra, R. &Skarmeta-Gomez, A. F. Classification of spatio-temporal trajectories based on support vector machines. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10978 LNAI, 140–151. 10.1007/978-3-319-94580-4_11 (2018)
14. Loh, W.-K. & Park, Y.-H. A survey on density-based clustering algorithms. Lect. Notes Electr. Eng.280, 775–780. 10.1007/978-3-642-41671-2_98 (2014)
15. Xu, X., Zhang, G. & Wu, W. A fast distribution-based clustering algorithm for massive data. Lect. Notes Electr. Eng.355, 323–330. 10.1007/978-3-319-11104-9_38 (2015)
16. Da Silva, T. L. C., De Macêdo, J. A. F. & Casanova, M. A. Discovering frequent mobility patterns on moving object data. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS 2014—In Conjunction with the 22nd ACM SIGSPATIAL International Conference on Advances in

Geographic Information Systems, ACM SIGSPATIAL 201460–67. 10.1145/2675316.2675325 (Association for Computing Machinery, 2014)
17. SeriiBabichev, BohdanDurnyak, IrynaPikh and VsevoiodSenkivskyg,An evaluation of the objective clustering inductive technology effectiveness implemented using Density-Based and Agglomerative Hierarchical algorithms, 2020
18. MuhannadTauhidurRahman, Arshad Jamal and Hassan M.Al-Alamadi,Examining hotspots traffic collisions and their spatial relationships with land use: A GIS- Based geographically weighted Regression approach for Dammam, Saudi Arabia, 2020
19. Amin Mohammadnazar, Ramin Arvin, AsadJ.Khattak, Classifying travelers driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning, Elsevier, 2021
20. ArashKhodaBakhshi, Mohamed M. Ahmed, Principal advantage of crossed random intercepts under Bayesian hierarchical modeling to tackle unobserved heterogeneity in clustering critical versus non-critical crashes, Elsevier, 2021
21. DonweiXu, Chenchen Wei, PengPeng, Qi Xuan, HaifengGuo, GEGAN: A novel deep learning framework for road traffic state estimation,Transportation Research part C, Elsevier, 2020
22. Paulo Figueiras, GuilherneGuerreiro, Ruben Costa, ZalaHerga, Antonio Rosa, Ricardo Jardim-Gonebalves, Real-Time Monitoring of road traffic using Data Stream Mining, IEEE International Conference on Engineering, Technology and Innovation, 2018