**Optimal Ambulance Positioning for Road  
Accidents with Deep Embedded Clustering**

**Allen Mathews BK (23BCE11302), Gokul Santhosh (23BCE11291), Shivam Sharma (23BCE11292)**School of computer science and engineering, VIT Bhopal University, School of computer science and engineering, VIT Bhopal University, School of computer science and engineering, VIT Bhopal University

**INDEX TERMS**Deep embedded clustering (DEC), Cat2Vec, K-means, ambulance positioning, accident hotspots. **ABSTRACT:**  
In today’s world road accident is one of the major issues across worldwide, causing countless injuries and lives lost every year. One of the ways to reduce this is by reducing the ambulance response time. Many at times the major cause of death causalities of accident is due to longer time of ambulance reaching the accident spots. One way to prevent this scenario is pre positioning of ambulance in the right location to make a big difference by ensuring faster medical help when it’s needed the most, instead of waiting to dispatch the ambulance.   
This study focuses on using deep learning to predict the best spots for placing ambulances. By studying the patterns and factors that lead to road accidents in specific areas, the approach ensures these insights are preserved while building the model. A deep-embedded clustering approach is used, supported by another deep-learning model, Cat2Vec, to capture these patterns effectively.

The proposed framework is tested against traditional clustering methods like K-means and GMM. The system demonstrated outstanding performance, achieving 95% accuracy through k-fold cross-validation and a distance score of 7.581, outperforming traditional methods. This approach shows significant promise in improving emergency healthcare services.

**INTRODUCTION**Emergency medical services (EMS) are a crucial lifesaving component because they provide timely medical assistance during emergencies. In the event of road accidents or any other emergency, the time taken for an ambulance to arrive at the incident location is what determines how many people will survive such incidents. The idea behind optimal ambulance positioning is that the ambulances will be placed in locations that maximize coverage and,   
hence, minimize distance travelled to incident sites. In recent times, the emergence of deep learning and advanced techniques in clustering has transformed this field, providing data-driven solutions to this complex problem.

The techniques of clustering are K-Means, Gaussian Mixture Models (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Such research bases its determination of the optimum positions of ambulances with historical data on all of these kinds of methods of clustering because they allow an in-depth spatial examination and the establishment of patterns or hotspots.

K-Means is a very popular algorithm that can divide the data set into a fixed number of clusters. The algorithm follows the concept of minimizing the sum of the squared distances of data points to their closest cluster centroids. It works best in determining central points that may act as good locations for an ambulance to be placed.

K-Means is very simple, easy to implement, efficient, and scalable for large datasets. It has clear and interpretable cluster centroids. However, it is not robust because the number of clusters has to be predefined. It is sensitive to outliers and noise and assumes spherical cluster shapes, which are not usually the case with real-world spatial distributions.

Gaussian Mixture Models is another probabilistic clustering approach, which models data in the form of a mixture of multiple Gaussian distributions. It estimates the probability that a data point belongs to every cluster and therefore becomes more flexible in capturing a variety of data patterns.

The advantages of GMM are that it can model complex shapes of clusters, overlapping clusters, and provide probabilistic assignments for nuance interpretation, and adapts to various distributions. However, it is much more computationally intensive compared to K-Means. It requires careful initialization to avoid suboptimal solutions, and it assumes that data points follow a Gaussian distribution, which is not necessarily the case in all distributions.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identifies clusters based on density and groups points that are closely packed into a cluster while marking the outliers as noise. DBSCAN is appropriate for data whose clusters are irregular in shape and have varying densities.

DBSCAN does not require the number of clusters to be predefined, is robust to noise and outliers, and identifies clusters of arbitrary shapes. However, its performance depends on the choice of parameters, such as epsilon and minimum points, it struggles with datasets of varying densities, and it is computationally intensive for large datasets.

Ambulance positioning is an essential activity that can enhance the efficiency of EMS, reduce response times, and thus improve patient outcomes. With the integration of historical accident data and population density metrics, this project offers actionable insights to EMS planners on the best places to position ambulances at strategic locations identified through clustering methods, thereby ensuring high-risk areas receive adequate coverage that will save lives.

Collects historical aggregated accident information along roads, spacial coordinates, time, and severity of incidents. The data pre-processed to remove noise, further made standard. Then, clustering like K-Means, GMM, DBSCAN is used to analyse hotspots and optimal placement of ambulances for emergency services.

The results of clustering are validated using silhouette score, Davies-Bouldin index, and practical evaluation through simulated response times. The most appropriate technique of clustering is chosen based on the comparison of the advantages and disadvantages of each method. The final model integrates real-time data streams for the dynamic update of ambulance positions according to the changing patterns.

The project starts with collecting historical road accident data that contains spatial coordinates, time, and severity of incidents. The data is pre-processed to remove noise and standardize features. Clustering algorithms such as K-Means, GMM, and DBSCAN are applied to identify hotspots and optimal ambulance positions.

The main aim of this project is to create a robust, efficient system for determining optimal ambulance placement using advanced techniques in clustering. This, through the analysis of historical accident data and the identification of high-risk zones, would minimize response time and ensure maximum coverage of the critical zones. This not only enhances the operational efficiency of emergency medical services but saves lives as well by enabling faster medical assistance during emergencies.

This research shows the potential of clustering algorithms in optimizing ambulance positioning. Combining deep learning with advanced clustering techniques, the project is expected to contribute to more efficient and responsive EMS systems, ultimately improving public safety and health outcomes.

**LITERATURE REVIEW**

Clustering Techniques for Optimal Ambulance Placement

Researchers worldwide have extensively explored methods for predicting crash sites, identifying factors contributing to accidents, and determining ideal locations for paramedic teams. This section summarizes key studies employing exploratory data analysis, machine learning, and deep learning for clustering-based ambulance positioning.

Crash Site Prediction and Severity Analysis

Assi et al. and Xiong et al. used machine learning models like Gaussian Mixture Models and SVMs to predict accident vs. non-accident patterns. They also clustered crashes using fuzzy c-means, Feed Forward Neural Networks, and SVM to predict injury severity. Among these, fuzzy c-means outperformed k-means and SVM in terms of accuracy.

Identifying Risk Factors in Road Accidents

Ghandour et al. and Tiwari et al. developed a hybrid ensemble classifier combining decision trees with the MSO algorithm to analyse risk factors contributing to fatal accidents. Using the Lebanese Road Accident Platform (LARP) dataset, they identified seven significant variables linked to casualties. Their models, evaluated using F1 score, precision, and AUC-PR curves, showed high performance.

Simulation for Ambulance Demand Prediction

Granberg et al. employed a multivariate regression model to predict emergency ambulance demand. They used census data and a distance matrix for clustering 35 probable ambulance locations. Their genetic algorithm achieved an R² value of 0.71, outperforming traditional forecasting techniques.

Clustering Techniques and Applications

Several studies explored clustering for crash location analysis. Cao et al and Moriya et al. applied fuzzy c-means, k-means, and batch clustering methods to group crash locations based on factors influencing accidents. Alkheder et al. used decision trees, MLP, and Naïve Bayes to identify key attributes predicting accident severity, with decision trees providing the highest accuracy.

Genetic Algorithms and Bayesian Networks

Hashmienejad et al. combined genetic algorithms with decision tree models like CART and C4.5 for accident severity prediction, achieving an accuracy of 88.2% and outperforming ANN, SVM, KNN, and Naïve Bayes. Ghosh et al. and Sasaki et al. employed Bayesian Networks to model relationships between attributes, predict accident severity, and evaluate performance using sensitivity, specificity, MAE, and RMSE.

Dizaji et al. and Tian et al explored the use of autoencoders to reduce data dimensionality and highlight the features most relevant to clustering. After dimensionality reduction, K-means clustering was applied to group these features effectively. Their method involved using autoencoders to create representations of accident locations, skipping the decoder step for simplicity, and adding a K-means layer to refine the clusters. However, this approach faced a limitation: it did not fully optimize the two separate processes of feature mapping and cluster formation together.

Building on this, Alqahtani et al. introduced a more advanced technique by embedding a clustering layer directly within deep autoencoders. Unlike traditional methods, this approach simultaneously learned feature representations and formed clusters. During optimization, cluster centres were recalibrated based on accident locations, iteratively updating until achieving stable and well-defined clusters. This improved performance compared to older methods.

Challenges and Gaps in Existing Research

An analysis of existing studies reveals several gaps and limitations:

Loss of Information in Data Representation:

Many methods rely on traditional ways to encode categorical data, like one-hot or numerical encoding. These fail to capture relationships between categories, leading to a loss of valuable information and reducing model accuracy.

Limited Real-Time Data Integration:

Current research often overlooks the use of dynamic, real-time data like live traffic updates, weather conditions, or recent accident reports. Incorporating such data could greatly enhance model accuracy and responsiveness in practical scenarios.

Focus on Cluster Characteristics Over Practical Effectiveness:

Metrics like point distance, inter/intra-cluster similarity, and cluster dispersion are commonly used to evaluate clustering algorithms. While useful, these focus on geometric properties and do not account for the real-time applicability of models. This gap often results in an incomplete assessment of a model's effectiveness in live situations.

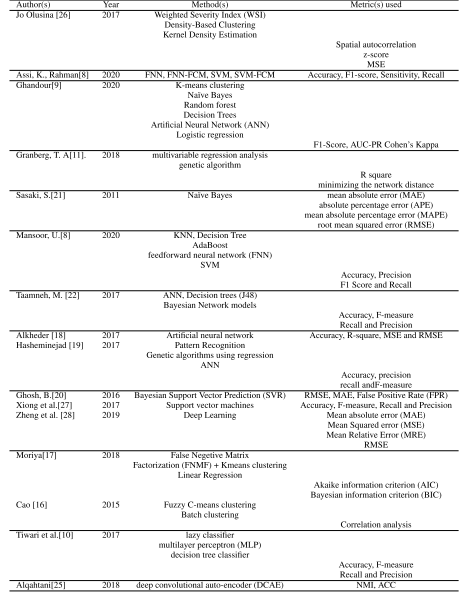
Lack of Optimization in Combined Processes:

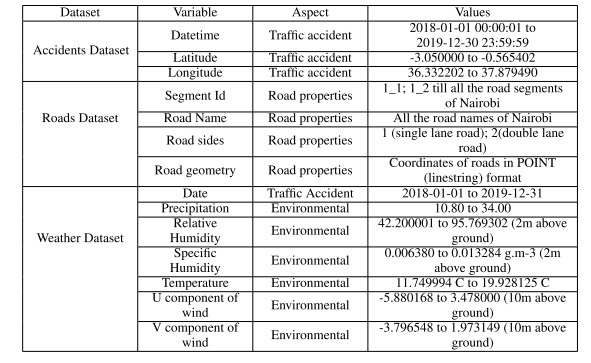
Existing methodologies often treat feature selection, dimensionality reduction, and clustering as separate processes, rather than optimizing them holistically.

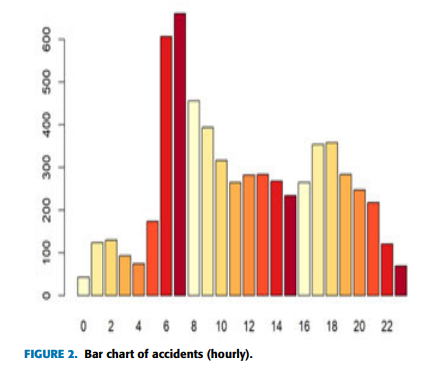
These gaps underscore the need for new methods that address the limitations of current research. Future approaches should focus on capturing relationships between categories, integrating real-time data streams, and

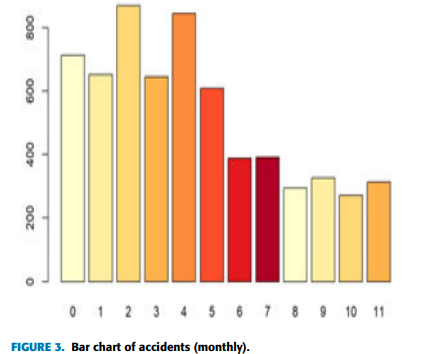
optimizing processes to ensure robust and responsive models for real-world applications.

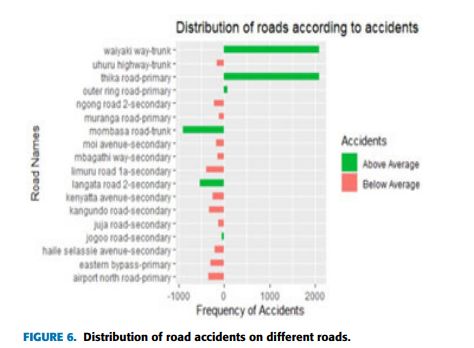
**DATA SETS**

****

****

****

****

****

**METHODOLOGY**

The methodology used here are as follows

**1. K-Means Clustering**  
  
K-Means is a popular clustering algorithm used to group data points into k clusters. The aim is to partition data such that points within a cluster are as similar as possible while being distinct from other clusters.   
  
**Choose the Number of Clusters (k)**:  
Decide how many clusters you want to create. This can be based on domain knowledge or trial and error.

**Initialize Centroids**:  
Randomly place k centroids (points that represent the center of clusters) in the dataset.

**Assign Points to the Nearest Centroid**:  
Each data point is assigned to the cluster of the nearest centroid based on a distance measure (usually Euclidean distance).

**Recalculate Centroids**:  
After all points are assigned to clusters, calculate the new centroids as the mean of all points in each cluster.

**Repeat**:

* Reassign points to the nearest centroid.
* Recalculate centroids.
* Continue this process until the centroids no longer change significantly or a set number of iterations is reached.

**Output**:  
The algorithm outputs the final cluster assignments and the positions of the centroids.

**How it Works on Our Model:**  
It follows these steps:  
  
1. Initialization: Randomly places k cluster centroids (starting points).  
2. Assignment: Each data point is assigned to the nearest centroid, based on a distance measure like Euclidean distance.  
3. Update: The centroid of each cluster is recalculated as the mean of all points within that cluster.  
Repeat: Steps 2 and 3 are repeated until the centroids stabilize, meaning the clusters no longer change significantly.  
  
Data Collection / Simulation

Input Data: In real-world applications, data might come from historical accident records, population density, or emergency calls. Each data point represents the location (e.g., latitude and longitude) of an event where an ambulance may be needed.

Simulated Data: For demonstration purposes, we generate random 2D data points to represent accident locations. In practice, this would be replaced with actual accident or emergency data.

The Accidents Array contains 100 points each with a latitude and longitude.  
  
**Advantages:**

Simple to implement and computationally efficient.  
Works well when clusters are spherical and of similar size.  
For Ambulance Positioning:

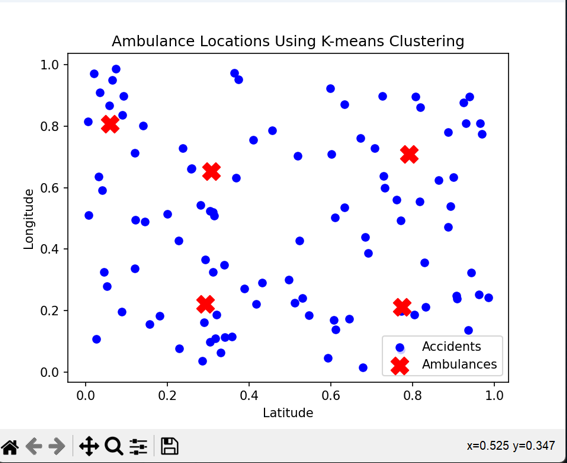
Historical accident data (latitude, longitude, time, and frequency) can be fed into K-Means to identify clusters of accident-prone areas.

The centroids of these clusters suggest optimal locations for ambulances to reduce response times.

For example, in a city with 5 major hotspots, K-Means can recommend 5 key ambulance locations, ensuring quick access to each cluster.

**Limitations:**

Assumes all clusters are circular and equally sized, which may not align with real-world accident patterns.  
Sensitive to the choice of k (number of clusters), which must be determined carefully.



### 

### 2. Gaussian Mixture Model (GMM) The Gaussian Mixture Model (GMM) is a probabilistic clustering algorithm that models data as a mixture of multiple Gaussian distributions (bell-shaped curves). Unlike K-Means, which assigns each data point to a single cluster, GMM provides probabilistic cluster memberships, meaning a point can belong to multiple clusters with varying probabilities.

**How it Works**:  
GMM assumes the data comes from a mixture of multiple Gaussian distributions (bell-shaped curves). Each distribution is defined by:

* **Mean** (centre of the distribution).
* **Covariance** (spread and orientation of the distribution).
* **Weight** (the proportion of data points belonging to that distribution).

The algorithm uses the **Expectation-Maximization (EM)** method:

1. **E-step**: Estimates the probability that each point belongs to each Gaussian component.
2. **M-step**: Updates the parameters (mean, covariance, and weight) to maximize the likelihood of the data fitting the model.

**Advantages**:

* Handles overlapping clusters by assigning probabilities rather than fixed labels.
* Flexible in modelling clusters of different shapes and sizes.

**For Ambulance Positioning**:

* GMM can model accident data where hotspots overlap, such as urban areas with high traffic density.
* Probabilistic assignments allow for flexible ambulance deployment, especially in areas with multiple high-risk zones.
* For instance, if an accident hotspot overlaps two neighbourhoods, GMM helps decide how to split ambulance coverage efficiently.

**Limitations**:

* Computationally intensive, especially for large datasets.
* Assumes data fits Gaussian distributions, which may not always be true.

**K-Means Limitation Solved**:

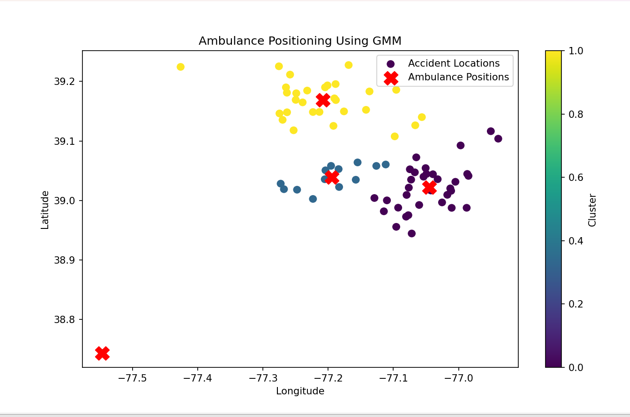
* **Fixed Cluster Shapes**: K-Means assumes spherical clusters with equal variance, which limits its effectiveness with complex or overlapping clusters.
* **Hard Assignments**: K-Means assigns each point to a single cluster rigidly, while GMM uses probabilities, allowing for soft assignments.

**Why GMM is Better**:

* Model clusters with varying shapes, sizes, and densities using Gaussian distributions.
* Handles overlapping clusters and provides probabilistic interpretations of cluster membership.

**When to Prefer GMM**:

* If your data contains non-spherical clusters or overlaps, GMM typically performs better than K-Means.



**3.DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**DBSCAN is a density-based clustering algorithm that identifies clusters as regions of high data density separated by regions of low data density. Unlike K-Means and Gaussian Mixture Models, DBSCAN does not require you to specify the number of clusters in advance, and it can identify clusters of arbitrary shape while handling outliers effectively.

**How it Works**:  
DBSCAN groups data points based on density rather than predefined cluster shapes or numbers. Key concepts include:

* **Core Points**: Data points with a minimum number of neighbours (minPts) within a given radius (epsilon).
* **Border Points**: Points that are close to core points but don’t meet the density criteria themselves.
* **Noise**: Outliers or points that do not fit into any cluster.

The algorithm forms clusters by connecting dense regions, allowing for clusters of arbitrary shapes.

**Advantages**:

* Identifies clusters of varying shapes and sizes.
* Automatically detects outliers (isolated accidents).
* Doesn’t require specifying the number of clusters in advance.

**For Ambulance Positioning**:

* DBSCAN can handle real-world accident patterns, such as clusters along highways, which may be elongated or uneven.
* Outliers (e.g., rare but severe accidents in remote areas) can be identified separately, allowing for contingency planning.
* For example, accident data near winding mountain roads might form elongated clusters, which DBSCAN handles well compared to K-Means.

**Limitations**:

* Requires careful tuning of epsilon and minPts parameters.
* Struggles with datasets of varying density, where clusters aren’t well-separated.

**K-Means Limitation Solved**:

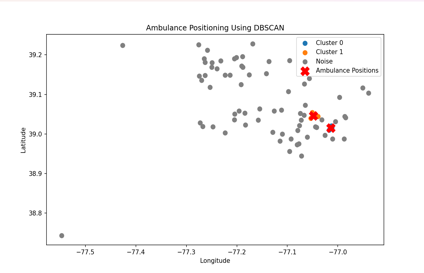
* **Predefined Number of Clusters**: K-Means requires the user to specify the number of clusters (k) in advance, which is not always practical.
* **Sensitivity to Outliers**: K-Means is highly sensitive to noise and outliers, which can skew the cluster centroids.

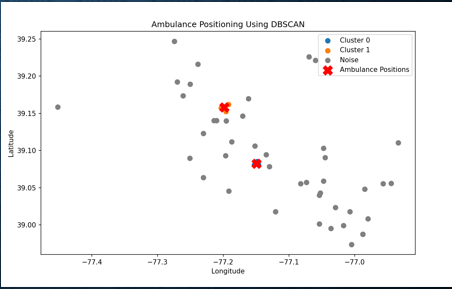
**Why DBSCAN is Better**:

* Does not require the number of clusters as input.
* Automatically identifies outliers and treats them as noise.
* Effectively clusters data with irregular shapes or density variations.

**When to Prefer DBSCAN**:

* If your data contains noise or irregularly shaped clusters, DBSCAN provides a more robust alternative.





**4. Cat2Vec (Category-to-Vector Embedding)**

**How it Works:**

Cat2Vec converts non-numerical (categorical) data, like road type or weather, into meaningful numbers. Unlike simple methods (like assigning 0s and 1s), Cat2Vec captures relationships between categories. For example, “rainy” and “foggy” conditions might be closer in the output than “sunny.”

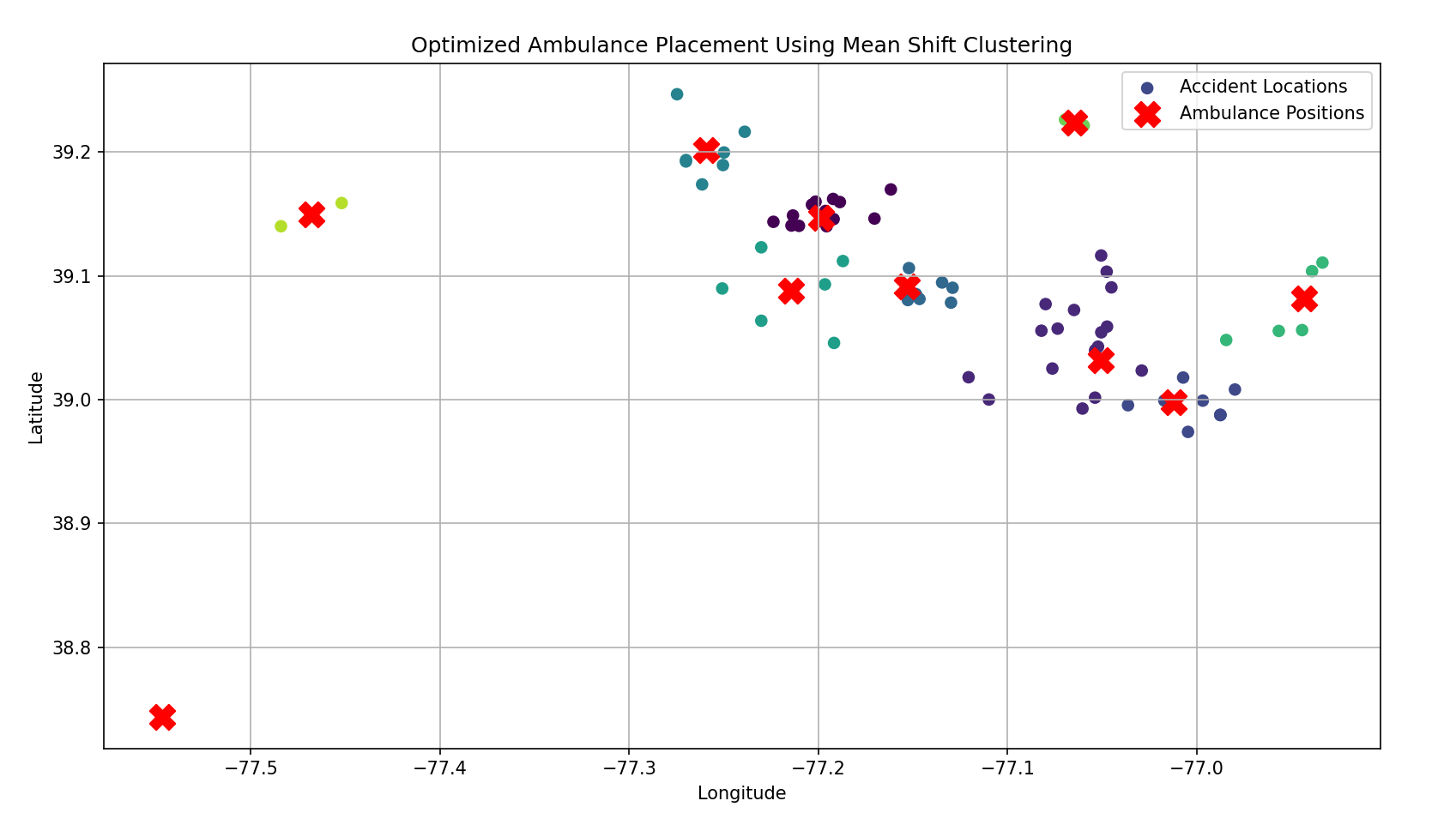
**For Ambulance Positioning:**

Using Cat2Vec allows clustering models to better handle variables like time of day, weather, or road types. This leads to more accurate identification of high-risk zone

**5. MEAN SHIFT CLUSTERING**  
The Mean Shift algorithm is a clustering technique commonly used in data analysis and computer vision tasks such as image segmentation and object tracking. Unlike other clustering algorithms like K-Means, Mean Shift does not require the number of clusters to be specified in advance. Instead, it discovers clusters by identifying dense regions in the feature space.

Advantages of Mean Shift Clustering

1. Adaptive Number of Clusters:
   1. Unlike K-Means, GMM, and AHC, Mean Shift does not require you to specify the number of clusters in advance. It automatically finds the number of clusters based on the data distribution.
2. Identifies Dense Regions:
   1. Mean Shift works by iteratively shifting points toward regions of higher density. This makes it excellent for locating optimal positions for ambulances in areas with high accident frequency.
3. Robust to Outliers:
   1. Mean Shift focuses on dense regions, so outliers or sparsely located accidents have minimal impact on the final cluster centers.
4. No Assumptions About Cluster Shape:
   1. Unlike K-Means (which assumes spherical clusters) and GMM (which assumes Gaussian distributions), Mean Shift can handle arbitrarily shaped clusters.
5. Real Data Points as Centers:
   1. Cluster centers found by Mean Shift correspond to dense regions, making them practical for positioning ambulances close to real accident location.



**CONCLUSION**   
The optimal positioning of ambulances through deep learning and clustering techniques is a transformational approach to enhancing EMS by minimizing the response time and maximizing the coverage of high-risk areas. Methods such as K-Means, Gaussian Mixture Models (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identify critical hotspots based on historical accident data and position ambulances strategically to achieve timely medical assistance. Each method offers unique advantages and challenges, making their combined application highly effective for diverse datasets.

K-Means is one of the most widely used algorithms. It is efficient, simple, and scalable, making it very useful for large datasets. It finds central locations by minimizing the sum of squared distances between data points and their respective cluster centroids. However, it requires the number of clusters to be predefined, is sensitive to noise and outliers, and assumes spherical cluster shapes, which may not align with real-world data distributions.

GMM offers a probabilistic approach to clustering, providing flexibility in the capture of complex and overlapping shapes of clusters. It assigns probabilities to data points, which can be used for more subtle interpretations. However, it is computationally intensive, requires careful initialization to avoid suboptimal solutions, and assumes that the data follows Gaussian distributions, which may not always be the case.

DBSCAN is very efficient on datasets with irregular shapes of clusters and varying densities. It groups closely packed points into clusters and marks outliers as noise, which makes it robust to anomalies. Another advantage of DBSCAN is that it does not need the number of clusters to be predefined. However, its performance depends on careful selection of parameters like epsilon and minimum points and can be problematic for datasets showing wide variations in density.

By this project, the cluster and evolution, an integration will be built across real-time data flows adjusting ambulance placement in anticipation and relation to pattern changes through use of evolving patterns by evaluation of clustered results utilizing various metrics silhouette score Davies- Bouldin, tests on it using simulations involving response time, effectiveness at the stages of applying relevant selection methods.

By addressing the specific limitations and exploiting the strengths of each clustering method, this project demonstrates potential data-driven approaches in optimizing ambulance positioning. Scalable and adaptable to different scenarios, it provides a robust framework for improving public safety and health outcomes, ultimately revolutionizing EMS systems.