IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 May 2023, accepted 7 June 2023, date of publication 9 June 2023, date of current version 20 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284993

RESEARCH ARTICLE

Optimal Ambulance Positioning for Road Accidents With Deep Embedded Clustering

DHYANI DHAVAL DESAI[1], JOYEETA DEY[1], SANDEEP KUMAR SATAPATHY[1], SHRUTI MISHRA[1], SACHI NANDAN MOHANTY[2], (Senior Member, IEEE), PALLAVI MISHRA[3], AND SANDEEP KUMAR PANDA[4]

[1]School of Computer Science and Engineering, Vellore Institute of Technology (VIT Chennai), Chennai, Tamil Nadu 600127, India
[2]Department of Computer Science and Engineering, Vardhaman College Engineering (Autonomous), Hyderabad 501218, India
[3]Department of Computer Science and Engineering, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education, Hyderabad 501203, India
[4]Department of Artificial Intelligence and Data Science, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education, Hyderabad 501203, India

Corresponding author: Sandeep Kumar Panda (Sandeeppanda@ifheindia.org)

ABSTRACT The number of casualties and fatalities brought on by road accidents is one of the most significant concerns in the modern world. Instead of dispatching ambulances only at the time of demand, pre-positioning them can reduce the response time and provide prompt medical attention. Deep learning techniques hold great potential and have proven to be essential for problem-solving and decision-making in the field of healthcare services. This study introduces a deep-embedded clustering-based approach to predict optimal locations for ambulance positing. Various factors and patterns in a geographical region greatly influence the occurrence of road crashes, hence understanding such relationships while model building is crucial. The present study also emphasizes the need of preserving such patterns during model building to ensure real-time results and implements them with the help of another deep-learning-based model, Cat2Vec. The proposed framework is also compared with traditional clustering algorithms like K-means, GMM, and Agglomerative clustering. Moreover, to calculate response time and distance in real time, a novel scoring function has also been introduced for the performance evaluation of various algorithms. The proposed ambulance-positing system exhibits remarkable performance, achieving an accuracy of 95% with k-fold cross-validation and a novel distance score of 7.581 proving the use of the proposed approach is better than all the other traditional algorithms used.

INDEX TERMS Deep embedded clustering (DEC), Cat2Vec, K-means, ambulance positioning, accident hotspots.

I. INTRODUCTION

Today one of the leading causes of death worldwide among children and adults is road accidents. The injuries caused by these fatal accidents cause considerable economic and personal losses to individuals, their families, and the country. An estimated 1.3 million individuals each year die as a result of road accidents. Between 20 and 50 million individuals experience non-fatal injuries, with many of them becoming disabled as a result [1]. The ever-increasing growth in the number of automobiles is certain to have some negative consequences, the most likely of which is an increase in the frequency of fatal road accidents in densely populated places, resulting in a huge burden on the urban infrastructure. It is dreaded that if we fail to take definitive precautionary measures to overcome these statistics, then road accidents will take over as the fifth major cause of death by 2030. Despite these fatal consequences, this problem receives scanty attention and there is a lack of developing systematic methods to improve road safety.

Studies show that over 90 percent of the global traffic accidents occur in medium to lower income countries such as Kenya [2], [3] which is one such example as more than a thousand fatalities occur due to road crashes consisting of a

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

mean of 7 out of 35 casualties each day [4]. Majority of these deaths and severe injuries occur to the population of 15-59 years who are also the economically active citizens of the country, reducing the economic activity of the country.. Kenya, as a country ranking in the range of lower-middle-income, has seen an increase in regional trade deals over the past decade. Based on the reports from National Transport Safety Authority (NTSA), which is the agency responsible for transportation in Kenya, a record of 5186 minor injuries, 6938 major injuries and 3572 deaths was concluded in 2019.

The number of injuries, and fatalities brought on by these deadly accidents can be decreased if preventive measures are taken, the most crucial of which are prompt medical attention for accident victims, information about the precise situation to the aid personnel, and accurate data analysis considering every single factor to diagnose and predict the accident-prone zones in a city. The delay in the arrival of an ambulance has a significant impact on human life especially in the case of emergency response pertaining to road accidents [5]. If the ambulance fails to reach the crash site in the critical hour, casualties may increase, therefore making each second very significant to human life. In every big metropolis, choosing the best places to place emergency responders throughout the day as they wait to be summoned is essential due to the dense traffic patterns and the city's distinctive layout. Monitoring and controlling these killer accidents is even more difficult due to the lack of expertise in stationing emergency response systems. Therefore, the prompt, automated, and timely positioning of ambulances can aid the first responders and doctors by reducing the effort required on their end and enabling earlier treatment decisions.

In this modern era technologies like machine learning and deep learning have always proven to be an emphatic and prevalent approach to decision-making, especially in the field of medical services. The advent of these technologies has been helpful in various road safety problems and their usefulness can also be found in our problem statement. Healing all patients and eliminating casualties in road accidents is needed of the hour as the end goal of improvement in Health Care Output (HCO). This paper considers the optimal positioning of the ambulance (paramedic help) as a clustering problem, since clustering algorithms ensure optimal locations based on distance metrics and, the coordinates of each centroid are the means of the coordinates of the objects in the cluster [6]. Traditional machine learning approaches such as k-means clustering, PAM clustering, Agglomerative clustering, etc. are not adequately versed and practical in all kinds of clustering problems [7]. This causes a novel deep learning-based technique to develop, which is an exercisable way to enhance this process' performance.

In this study, we propose a novel clustering-based approach utilizing Deep Embedded Clustering with Autoencoder (DEC-AE) to address the problem of optimal ambulance positioning in a city. Unlike traditional clustering methods, the DEC-AE method offers a comprehensive framework that combines deep learning, clustering, and autoencoder techniques [23] to optimize ambulance positioning strategies. By reconstructing the input data from the learned latent representations, DEC can effectively capture the essential features and dimensions that contribute to the clustering process. Furthermore, DEC employs a joint optimization objective [24] that integrates clustering assignments and feature learning. This joint optimization facilitates the enhancement of cluster separability and the generation of compact and well- separated clusters in the latent space. DEC-AC combines deep learning and adaptive clustering [25] to provide an effective solution for clustering problems. It leverages deep neural networks to learn meaningful feature representations and adaptively determines the number of clusters based on the data distribution.

Additionally, DEC is scalable and can handle large-scale datasets, making it suitable for real-world applications with high-dimensional and complex data. This enables a more accurate and nuanced understanding of the factors influencing optimal ambulance positioning. The DEC-AE approach also incorporates clustering algorithms, facilitating the identification of clusters or groups of similar patterns within the data [25]. This allows for the identification of hotspot areas with higher accident probabilities or specific risk profiles, aiding in the strategic placement of ambulances to minimize response times and maximize coverage. Additionally, this method has the potential to accommodate diverse data sources, including traffic accident data, road segment characteristics, weather conditions, and other relevant factors. By considering multiple data dimensions, the approach can provide a holistic view of the problem, enhancing the precision and effectiveness of ambulance positioning strategies.

The dataset includes information on traffic accidents that occurred, road segment information, and weather details of Nairobi, Kenya. Performing Exploratory Data Analysis on the dataset of the road surveys, and weather dataset, the paper identifies possible features and attributes affecting the accidents and patterns of risk across the city. To preserve such relationships and patterns of the data we apply a deep learning-based embedding approach called Cat2Vec while converting categorical attributes in the data pre-processing stage. To validate the predicted locations using DEC, the distance from that crash site to the nearest ambulance locations predicted is calculated using a novel Distance Scoring Algorithm. For further evaluation of the algorithm, different clustering metrics have been used and compared with other traditional clustering algorithms.

The proposed methodology in this paper addresses the following aspects:

- Exploratory Data Analysis (EDA) is performed on the real-time accident dataset through which the potential features and attributes which contribute towards accidents and patterns across the city are identified.
- A clustering-based approach using Deep Embedded Clustering (DEC) is developed to identify optimal ambulance positioning locations across Nairobi while

preserving the feature relationships and patterns using Cat2Vec deep learning-based embedding technique which facilitates more accurate clustering.
- A novel Distance Scoring method is developed to validate the DEC model which calculates the distance between crash-site and the nearest predicted ambulance location, thus providing a quantitative measure of effectiveness.
- The performance of the proposed framework is then evaluated with and without the feature selection techniques and compared with existing clustering methods using various clustering metrics, which further validates the effectiveness of the DEC model.

The rest of this paper is organized as follows. Sect. II is a detailed review about related works for the problem statement. In Section III the methodologies and materials used are described, Section IV discusses the experiments performed and results obtained from the proposed framework. The paper delivers the discussion and future scope in Section V.

II. RELATED WORK

Researchers around the world have done a lot of work related to the prediction of crash sites, factors affecting accidents, and choosing the ideal locations for the placement of the paramedic's team. The following literature review section presents methods from various research studies related to clustering techniques used for the optimal positioning of ambulance locations by implementing exploratory data analysis, machine learning and deep learning techniques. This section presents a summary of the background work performed by other scholars regarding the problem and the different clustering techniques and performances achieved with regards to accuracy and other evaluation metrics.

Assi et al. [8] and Xiong et al. [27] proposed Machine learning models for predicting accident vs non-accident patterns in crash sites using Gaussian Mixture models and SVM. They further predicted the severity of the crash injuries by clustering the crashes using fuzzy c-means, Feed Forward Neural Networks and SVM. Data analysis was performed to identify the features from the crash sites to provide inputs to the ML models. The models were evaluated using accuracy, sensitivity and precision. On comparing the models, fuzzy c-means algorithm provided greater accuracy than traditional k-means clustering and SVM models.

Ghandour et al. [9] and Tiwari et al. [10] developed an approach that uses a machine learning hybrid ensemble classifier derived from decision trees and MSO algorithm to identify risk factors that contribute to fatal road accidents. They utilized the Lebanese Road accident platform (LARP) dataset consisting of 8482 accidents and the fatalities occurred in the accidents. To evaluate the impact of the factors causing casualties in a road accident, they performed sensitivity analysis of the attributes. From the selected variables, seven of nine showed significant association to casualties. To evaluate the model performances the metrics used were F1 score, precision, AUC-PR curve and Cohen's Kappa.

Granberg et al. [11] developed a simulation based predictive model to gauge the emergency ambulance demand in an area using multivariate regression model. The data used for their genetic regression algorithm was collected from census survey of the year 2005 consisting of 2076 small areas. For each location, a distance matrix was developed and used as inputs to the genetic algorithm in order to identify 35 probable ambulance locations using R-statistics software. The proposed model favored a significant R2 value of 0.71 with multiple coefficients. This particular distance matrix based approach provided better results on comparing with the models using nave forecasting techniques. Clustering in machine learning has been explored pertaining feature selection techniques [12], [13], [14], distance functions and cluster validations. A derivative of the popular clustering methods are K-means and Gaussian mixture models. Even though the approach using distance functions was developed earlier, its application and popularity are limited by high dimensionality and dataset space. Cao et al. [16] and Moriya et al. [17] proposed approaches based on batch clustering, fuzzy c-means clustering, and K-means clustering with an FNMF matrix for clustering illustrating the correlation patterns of the initial data points. In both the approaches, the correlation among the crash locations are calculated followed by clustering of the said crash locations based on the factors responsible for the accidents.

Alkheder et al. [18] proposed an approach using decision tree classifier, MLP and Naïve Bayes to identify the significant attributes that impact the prediction of the severity of a road accident. On comparing different models, it was concluded that the decision tree classifier provided a better classification accuracy of 0.08218. The attributes such as year of accident, age, nationality, gender and the type of accident were more significant in determining the accident severity. Hashmienejad et al. [19] utilized decision trees and genetic algorithms to develop a prediction model for predicting the severity of the road accidents. The set of rules induced from the genetic algorithm approach were further provided as input to the decision tree models namely CART, C4.5 and ID3, also validated using the test dataset. The method employed provided an accuracy of 0.8820, recall of 0.889, f-measure of 0.887 and a precision of 0.885 which was better than the alternative methods used namely ANN, SVM, KNN and Naïve bayes.

Ghosh et al. [20] and Sasaki et al. [21] employed Bayesian networks (BN) approaches to develop models based on the relationship of the attributes which were represented as probability distributions. Bayesian Networks are used for identifying the factors responsible for road accidents and also predict the severity of the accidents. These papers also evaluated the sensitivity and specificity of the models along with MAE and RMSE to assess the performance of BSVR. Taamneh et al. [22] utilized Artificial Neural Networks (ANN) along with K-means to predict the severity of road accidents. Other machine learning models were also used to compare the accuracy of the proposed ANN model, concluding that the proposed model provided a higher accuracy of 0.746.

Dizaji et al. [23] and Tian et al. [24] used Auto-encoders to reduce the dimensionality for obtaining the features having higher impact on clustering. Following the dimensionality reduction, Kmeans is employed for clustering the features in groups. The approach employs auto encoders to obtain a representational structures of the locations, then neglects the decoder part to obtain a smooth model and finally adds the Kmeans layer over the encoder layer to obtain the final model. Although this approach uses a deep neural network for mapping the data points in feature space prior to the feature selection and cluster formation, the goals of these two separate processes are not optimized together. Alqahtani et al. [25] proposed an approach using embedded clustering layer in deep auto-encoders. When compared to traditional clustering methods, this technique learns the feature representations and assigns clusters concurrently through deep auto encoders. During the optimization phase, the cluster centers are reassigned to all the data points which are accident locations in this particular problem. Following this, the cluster centers are updated iteratively in order to obtain the final stable clusters and better optimized performance.

Table 1. provides the concise description of the survey conducted on existing methodologies along with the evaluation metrics used.

The existing literature analysis reveals several technical gaps that have prompted the development of our novel research endeavor. Traditional methods of representing categorical data, such as one-hot encoding or numerical encoding, fail to capture the inherent relationships between categories. This limitation results in a loss of valuable information and diminishes model performance, potentially leading to misinterpretations of results. Furthermore, existing research works primarily focus on cluster dynamics. While these metrics like point distance, inter and intra-cluster similarity, and dispersion provide valuable insights into the structure and quality of clustering algorithms, they alone are not sufficient for evaluating the effectiveness of models in real-time data scenarios. The limitations arise from the fact that these metrics primarily focus on geometric properties and overall cluster characteristics, rather than capturing the specific requirements and dynamics of real-time data and may overlook crucial performance aspects related to real-time data, leading to an incomplete assessment of the model&#39;s effectiveness. In the existing works, there is a gap in considering real-time or dynamic data streams, such as live traffic updates, weather conditions, or accident reports, which could significantly improve the accuracy and responsiveness of the models. The identified gaps highlight the necessity for new approaches and methodologies that investigate the interrelationships among categories and address the limitations observed in the current body of knowledge.

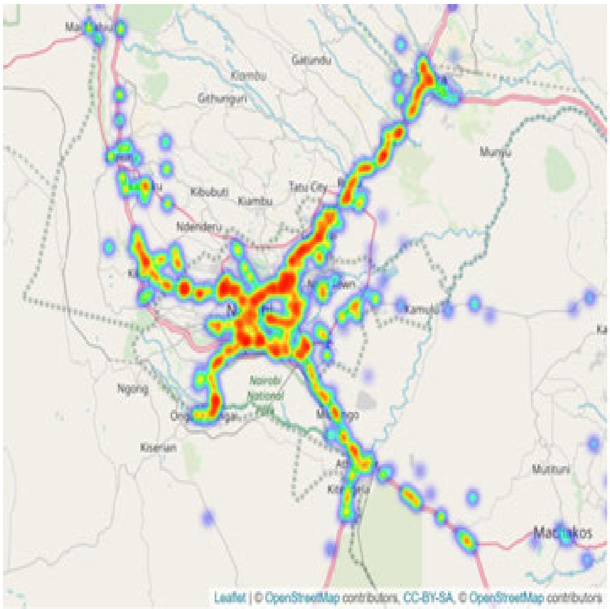


FIGURE 1. Heap Map of road accidents in Nairobi City.

III. STUDY AREA AND MATERIALS

The problem of optimal ambulance positioning is addressed by utilizing real-time crash data from Nairobi City along with additional datasets of real-time road and weather conditions of the crash locations. The identified problem of finding optimal ambulance position is regarded as a clustering problem. By leveraging the combined datasets, the proposed study aims to conduct experiments and analysis to identify city-wide risk patterns and factors that contribute to road accidents in Nairobi. The analysis will involve exploring the relationships between various attributes and features within the datasets to uncover the underlying patterns and correlations. This information will aid in understanding the dynamics of accidents and their associated risks across different areas of the city. The study utilizes advanced machine learning and clustering techniques to analyze the data and identify optimal locations for positioning ambulances. By considering factors such as accident frequency and geographical conditions the study aims to develop a model that can suggest strategic ambulance placement across Nairobi. This approach will help emergency response services improve their efficiency and effectiveness in providing timely medical assistance to accident victims.

A. DESCRIPTION OF THE DATASET

The Nairobi City Accident Dataset includes accidents that have been reported up to 2019.It also consists of supplemental data like road survey data, and weather data. The data collected for the Nairobi City Accident Dataset is acquired in real-time, ensuring its timeliness and relevance. The proposed study utilizes all the mentioned datasets to conduct the experiments and identifies city-wide risk patterns and factors responsible for road accidents. Fig 1. Depicts the intensity of road accidents through a heat map at different coordinates in Nairobi City as a map view. The dataset was

TABLE 1. Representative algorithms and metrics used in other related studies.

Author(s)	Year	Method(s)	Metric(s) used
Jo Olusina [26]	2017	Weighted Severity Index (WSI) Density-Based Clustering Kernel Density Estimation	Spatial autocorrelation z-score MSE
Assi, K., Rahman[8]	2020	FNN, FNN-FCM, SVM, SVM-FCM	Accuracy, F1-score, Sensitivity, Recall
Ghandour[9]	2020	K-means clustering Naïve Bayes Random forest Decision Trees Artificial Neural Network (ANN) Logistic regression	F1-Score, AUC-PR Cohen's Kappa
Granberg, T. A[11].	2018	multivariable regression analysis genetic algorithm	R square minimizing the network distance
Sasaki, S.[21]	2011	Naïve Bayes	mean absolute error (MAE) absolute percentage error (APE) mean absolute percentage error (MAPE) root mean squared error (RMSE)
Mansoor, U.[8]	2020	KNN, Decision Tree AdaBoost feedforward neural network (FNN) SVM	Accuracy, Precision F1 Score and Recall
Taamneh, M. [22]	2017	ANN, Decision trees (J48) Bayesian Network models	Accuracy, F-measure Recall and Precision
Alkheder [18]	2017	Artificial neural network	Accuracy, R-square, MSE and RMSE
Hasheminejad [19]	2017	Pattern Recognition Genetic algorithms using regression ANN	Accuracy, precision recall andF-measure
Ghosh, B.[20]	2016	Bayesian Support Vector Prediction (SVR)	RMSE, MAE, False Positive Rate (FPR)
Xiong et al.[27]	2017	Support vector machines	Accuracy, F-measure, Recall and Precision
Zheng et al. [28]	2019	Deep Learning	Mean absolute error (MAE) Mean Squared error (MSE) Mean Relative Error (MRE) RMSE
Moriya[17]	2018	False Negetive Matrix Factorization (FNMF) + Kmeans clustering Linear Regression	Akaike information criterion (AIC) Bayesian information criterion (BIC)
Cao [16]	2015	Fuzzy C-means clustering Batch clustering	Correlation analysis
Tiwari et al.[10]	2017	lazy classifier multilayer perceptron (MLP) decision tree classifier	Accuracy, F-measure Recall and Precision
Alqahtani[25]	2018	deep convolutional auto-encoder (DCAE)	NMI, ACC

developed to identify the accident locations in 2018 and 2019 to predict potential factors of road accidents and is categorized into accidents, road details, and weather data. The accident data contains 6318 instances with 3 characteristics namely the Datetime, Latitude, and Longitude of the crash sites. The road segment dataset consists of 792 Instances with 4 characteristics including road segment id, road name, sides of the road, and POINT geometry coordinates of the road segments. The weather data contains 728 Instances with 7 characteristics including Date, precipitation, relative humidity, specific humidity, temperature, u component, and v component of wind velocity. Table 2. gives a detailed overview of the attributes in the accident, road information, and weather dataset. Some of these attributes were measured at the traffic accident level, while the others were measured as road properties and environmental data.

TABLE 2. Complete attribute documentation of the nairobi city accident dataset.

Dataset	Variable	Aspect	Values
Accidents Dataset	Datetime	Traffic accident	2018-01-01 00:00:01 to 2019-12-30 23:59:59
	Latitude	Traffic accident	-3.050000 to -0.565402
	Longitude	Traffic accident	36.332202 to 37.879490
Roads Dataset	Segment Id	Road properties	1_1; 1_2 till all the road segments of Nairobi
	Road Name	Road properties	All the road names of Nairobi
	Road sides	Road properties	1 (single lane road); 2(double lane road)
	Road geometry	Road properties	Coordinates of roads in POINT (linestring) format
Weather Dataset	Date	Traffic Accident	2018-01-01 to 2019-12-31
	Precipitation	Environmental	10.80 to 34.00
	Relative Humidity	Environmental	42.200001 to 95.769302 (2m above ground)
	Specific Humidity	Environmental	0.006380 to 0.013284 g.m-3 (2m above ground)
	Temperature	Environmental	11.749994 C to 19.928125 C
	U component of wind	Environmental	-5.880168 to 3.478000 (10m above ground)
	V component of wind	Environmental	-3.796548 to 1.973149 (10m above ground)

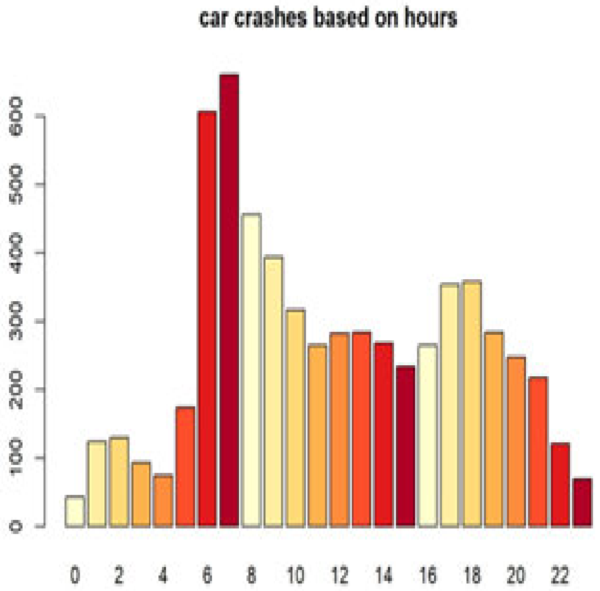


FIGURE 2. Bar chart of accidents (hourly).

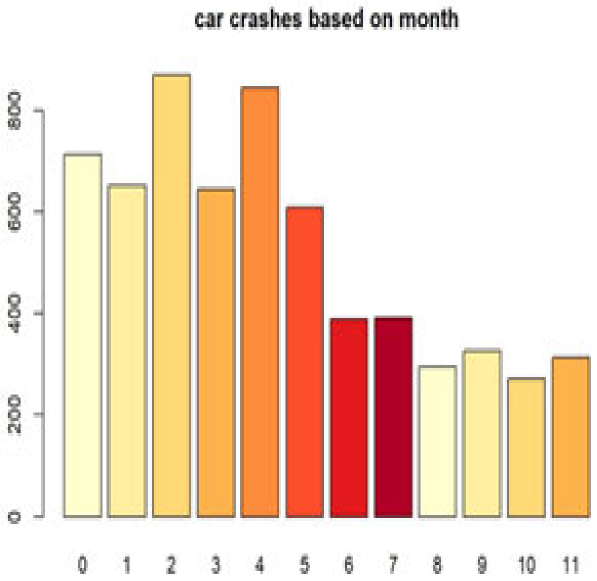


FIGURE 3. Bar chart of accidents (monthly).

B. DATA ANALYSIS

Unlike preliminary information evaluation, exploratory data analysis (EDA) is a method of reading information units to summarize their primary characteristics, frequently with visual methods [29]. Many EDA strategies are followed in big information analytics. In this section, the paper tries to discover the patterns and possible factors influencing road accidents in Nairobi city via exploratory data analysis. The time span of experimental information is from January 2018 to December 2019. Through EDA, it became evident that road accidents follow a distinctive trend where factors like the hour of the day, the month of the year, and the days of the week impact the number of accidents to a certain extent. Table 3 represents the attributes generated from the original dataset after data manipulation and their summary statistics. This obtained dataframe is utilized as the final input of the proposed model in further sections of the paper.

1) ACCIDENTS DATA

On performing exploratory data analysis on the accidents dataset, the road crashes showed trends and patterns which were useful if included in further modeling. Fig 2 and 3 represent the car crashes on an hourly and monthly basis from 2018 and 2019. It can be concluded that the maximum number of accidents occur between 6:00 am to 7:00 am. Although the traffic is less at that time, people Overspeed due to empty

TABLE 3. Summary statistics of input dataset.

attributes	count	mean	std	min	25%	50%	75%	max
latitude	6318	-1.27034	0.125157	-3.05	-1.31659	-1.27171	-1.23375	-0.5654
longitude	6318	36.85546	0.112866	36.3322	36.80205	36.84459	36.89564	37.87949
sec	6318	29.31197	17.01483	0	9	29	39	59
min	6318	29.54147	17.14912	0	15	30	44	59
hour	6318	11.63865	5.65358	0	7	11	17	23
day	6318	15.10399	8.818403	1	7	15	23	31
month	6318	4.317031	3.194254	0	2	4	7	11
year	6318	2018.304	0.460109	2018	2018	2018	2019	2019
wday	6318	3.126148	1.886122	0	2	3	5	6
weekend	6318	0.096391	0.29515	0	0	0	0	1
precipitation	6287	23.72764	4.974721	10.8	19.8	24	27.5038	34
relative_humidity	6287	82.24113	10.01541	42.2	77.2	85.2	89.8	95.7693
specific_humidity	6287	0.010837	0.001371	0.00638	0.00986	0.0111	0.0119	0.013284
temp	6287	15.18853	1.248962	11.74999	14.38061	15.23254	15.99716	19.92813
wind_u 6287	-1.77646	1.456756	-5.88017	-2.78831	-1.84344	-0.89509	3.478	
wind_v	6287	-1.19801	1.143214	-3.79655	-2.12117	-1.08737	-0.31131	1.638513
wind_res	6287	2.492517	1.344323	0.116455	1.378081	2.436066	3.508103	6.088397
distance_from_centre	6318	0.107086	0.136578	0.000227	0.039121	0.070923	0.125485	1.995093
elevation	6318	-45.7121	0.18223	-46.4924	-45.8234	-45.6916	-45.6015	-43.9446

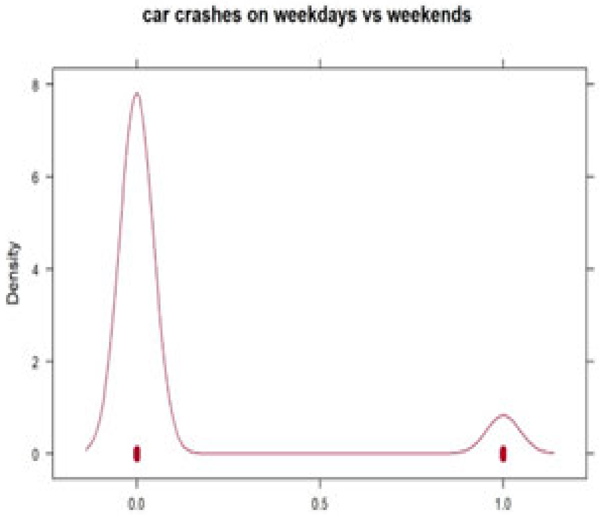


FIGURE 4. Accidents on Weekdays vs Weekends.

roads, causing more accidents. From that monthly data, it can be seen that accidents mostly occur in the months of March and May which is the rainy season in Nairobi. Fig 4. Shows that the number of accidents occurring on weekends is quite less as compared to the accidents occurring on weekdays. Fig 5. Represents the crash sites on a weekly basis. Maximum accidents occur in the region −1.55 to −1.05 latitude and −3.05 to −0.57 longitude region with a count of 6070 accidents. This location is very near the center of the city.

2) WEATHER DATA

On considering the weather factors namely precipitation, humidity, temperature, and wind velocity associated which contributes to road accidents, the precipitation intensity was found to be the most important factor. Table 4. Denotes the precipitation levels and the number of accidents occurring in the years 2018 and 2019. The maximum number of accidents occur during high rainfall as the roads become slippery and visibility decreases.

TABLE 4. Precipitation levels in different years.

Precipitation levels	high	moderate	low
2018	1890	1005	1470
2019	728	630	564
Total accidents	2618	2034	1835

From the weather data analysis, the following inferences can be made which are listed in Table 5.

3) ROAD SEGMENTS DATA

Apart from the accident locations and weather conditions, road accidents also depend on the type and location of the roads. It was found that the number of accidents was significantly high on single-lane roads rather than the two-way roads. The maximum number of accidents occurred on the Waiyaki way-trunk street which is a one-lane road and also closer to the city center.

The real-time nature of the data collection is facilitated by utilizing various instruments and sources. Speedometer data, obtained from vehicles or traffic monitoring systems, provides up-to-date information on vehicle speeds and traffic conditions. Satellite images, which offer a comprehensive view of the road network, are regularly captured and updated to reflect the current state of the city's infrastructure. Weather data, obtained from the weather department, ensures that the environmental conditions associated with the accidents are accurately captured and analyzed.

IV. METHODOLOGY

This paper proposes an approach (optimal ambulance positioning framework) for the automatic placement of paramedic help using Deep Embedded Clustering (DEC). To ameliorate the classification accuracy, this study utilizes Cat2vec(a deep learning-based model) to represent high cardinality categorical variables using low-dimension embedding while

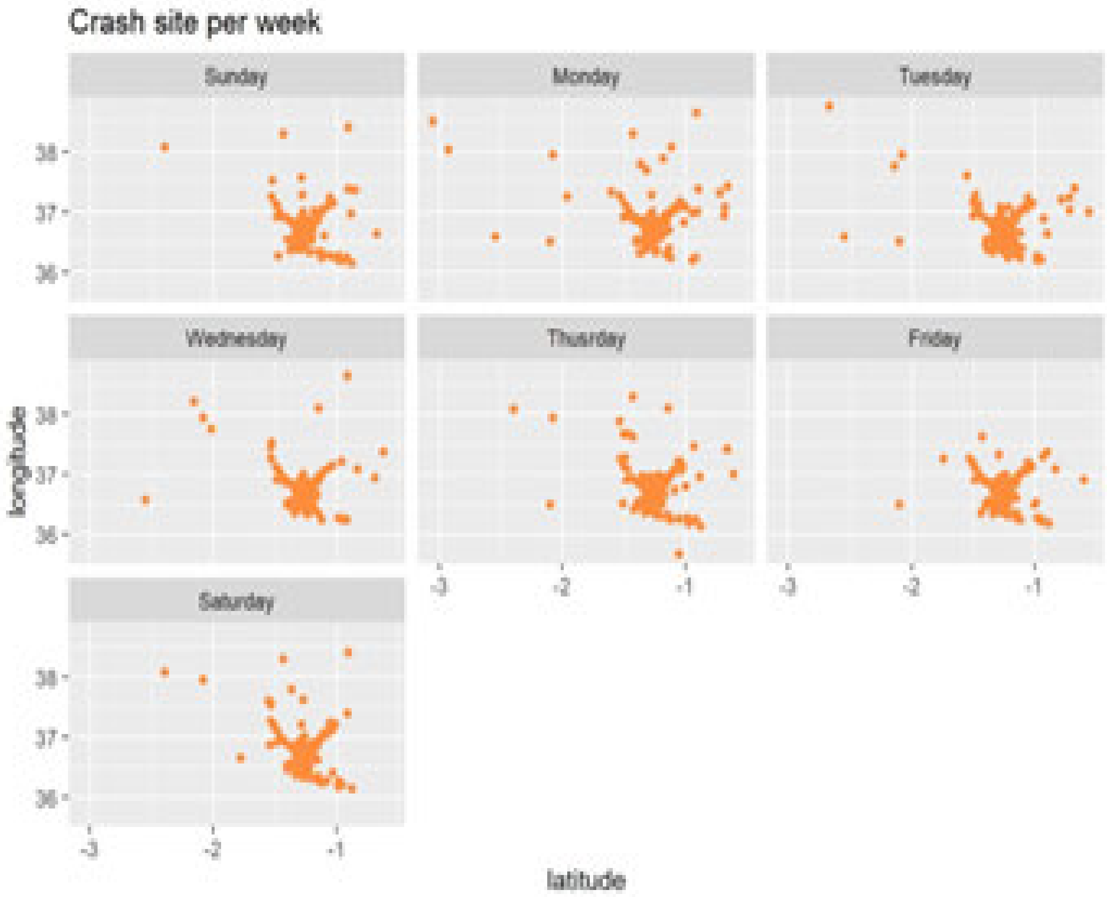


FIGURE 5. Accident coordinates (weekly).

TABLE 5. Inferences from weather data variables.

Road Weather Variables	Roadway Impact	Traffic Flow Impact	Accident Impact
Temperature and Humidity	No Impact	Less traffic	Fewer accidents
Wind Speed	Visibility decreases	Traffic speed	Moderate Accident Risk
Precipitation	Visibility decreases Friction between road vehicle decreases	Traffic speed Travel time delay	High accident risk

preserving the relationship and patterns obtained through exploratory data analysis between each of the categories. This study employs K-fold cross-validation for dividing the dataset into training and test sets.

A. Cat2Vec

Cat2Vec is a deep learning-based method of learning distributed representation for multi-field categorical data. Using Cat2Vec a low-dimensional continuous vector is automatically learned for each category in each field [30]. In simpler terms, Cat2Vec is a use of deep learning for creating embedding for categorical variables of tabular data. The use of embedding for categorical variables allows for capturing the relationship between categories. As shown in section III-B road accidents follow a distinct pattern affected by factors like an hour, month, or day. While working with such categorical variables researchers usually use traditional transformations like one-hot encoding, binary encoding, etc. However such transformations fail to capture relationships between categories.

Through data analysis, it was very evident that weekdays show more accidents than weekends or that the seasonal patterns of each month affect the number of accidents occurring. To ensure the most promising results the preservation of such relationships among categories while applying transformations during the pre-processing stages should be maintained. Embedding after applying Cat2Vec in our proposed framework helps capture these rich relationships and complexities.

This study uses Cat2Vec for all selected categorical features, the model consists of a perceptron network with a dense layer network and the chosen activation function is 'relu' and 'Adam' optimizer with a mean square loss function. The neural network learns the best representations for each category during the training phase to preserve the

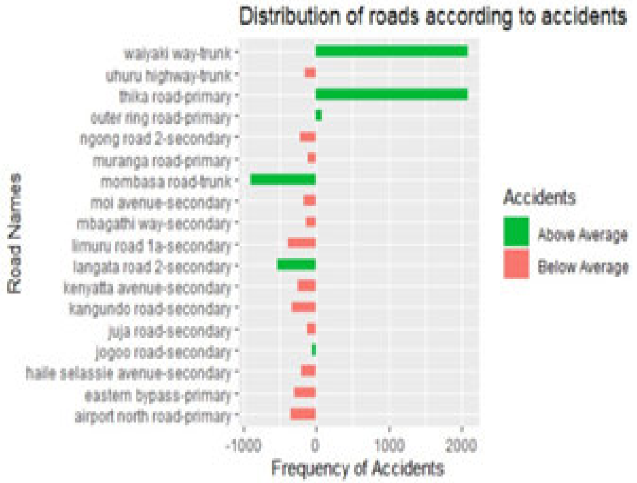


FIGURE 6. Distribution of road accidents on different roads.

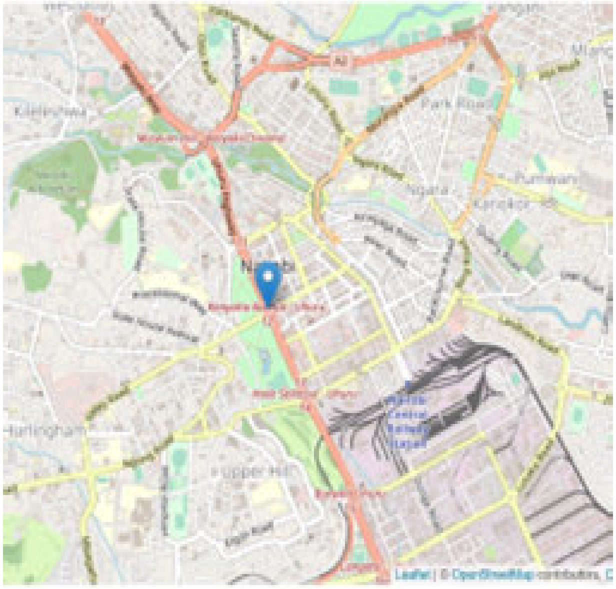


FIGURE 7. Roads in Nairobi and city center.

TABLE 6. Architecture of Cat2vec model.

Layers	# Parameters
Embedding layer	36
Flatten	0
Dense_22	200
Dense_23	765
Dense_24	16

relationships in these distributed dimensions. Fig 8 represents the network parameter and architecture for the used Cat2Vec model. We specify an embedding size of 3 for all categorical variables as they are enough to capture the relationship. Fig 9 and Fig 10 visualize this using a 3D plot, where one can see a clear relationship between crashes occurring during the week. The days with higher number of crashes like 2(Wednesday),5(Saturday) is grouped together and similarly those with lesser number crashes are grouped together like 0 (Monday),1(Tuesday),6(Sunday). Fig 9 shows the number of crashes occurring throughout the week obtained from Data Analysis of the given dataset. Fig 10 shows the embeddings of the weekday attribute obtained through Cat2Vec.One can clearly understand that the Cat2Vec model preserves the relationship between each category and hence extremely helpful for the proposed work, that demands pattern identification and understanding for prediction. The same model is applied to other categorial variables present in the dataset i.e., months and hour of day.

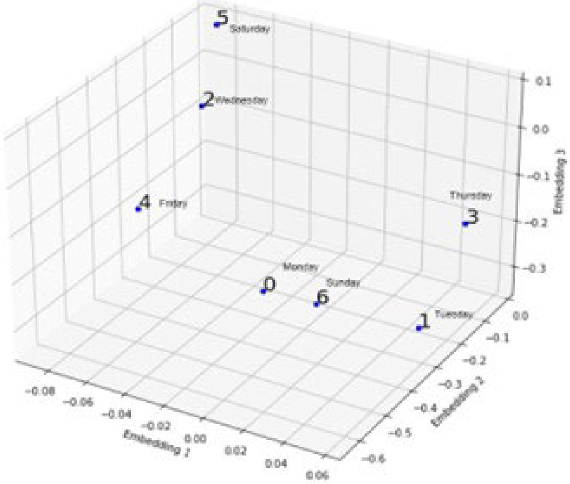


FIGURE 8. DEC architecture.

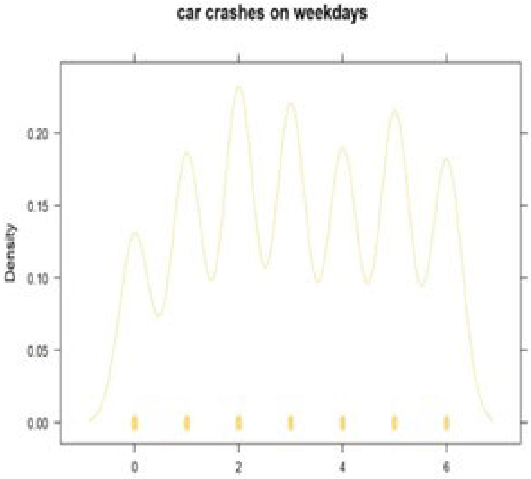


FIGURE 9. DEC architecture.

B. DEEP EMBEDDED CLUSTERING

Deep embedded clustering (DEC) is an unsupervised clustering algorithm using deep neural networks [31] that utilizes and auto encoder configure the attributes of the initial data and then form clusters. Deep embedded clustering (DEC) has been attracting attention due to its efficient performance in the end-to-end clustering problems. The proposed study employs a deeply embedded clustering (DEC) method to identify

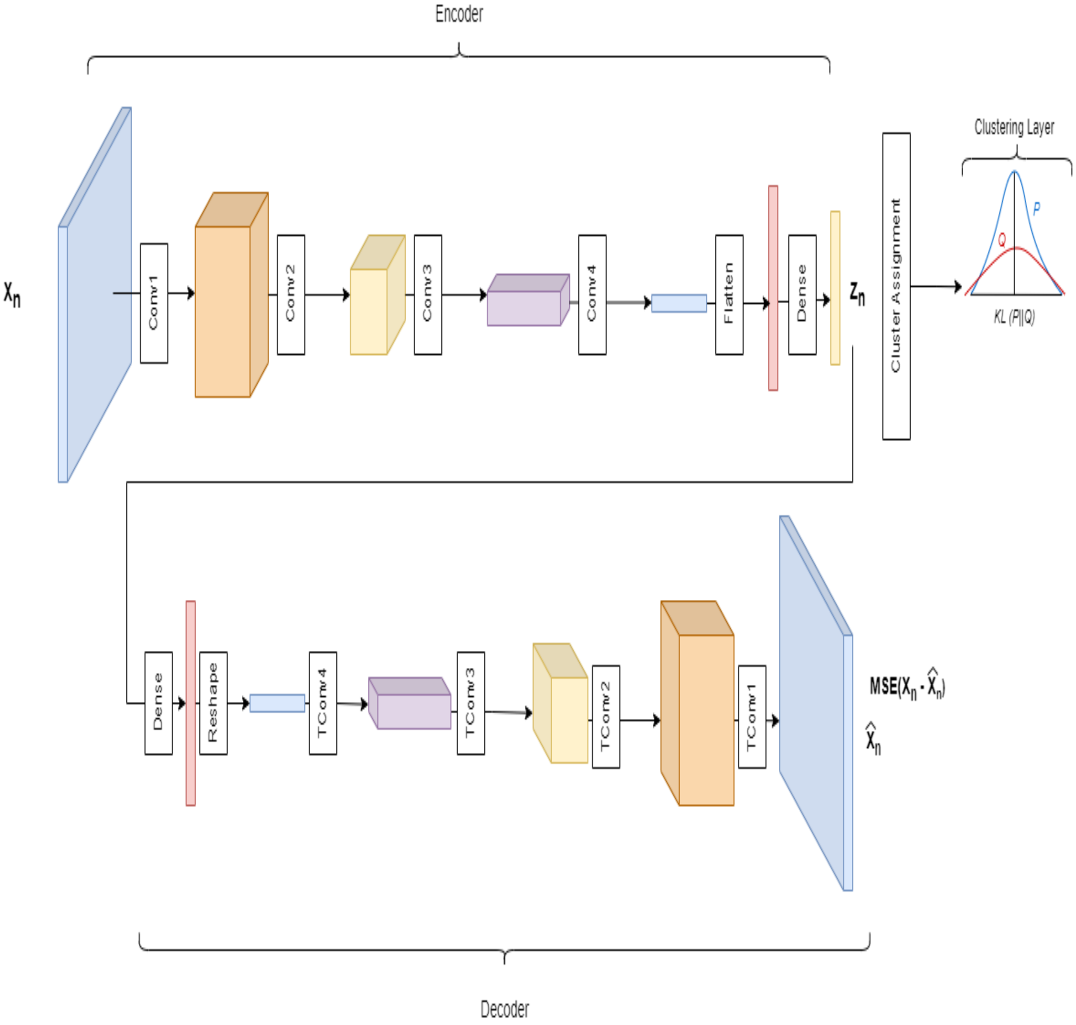


FIGURE 10. DEC architecture.

optimal positions for ambulance positioning. This paper identifies the problem of finding the coordinates to position ambulances as a clustering problem and DEC assists in finding the clusters based on the existing crash points present in the dataset. There are two phases of DEC namely (1) initialization of parameters with a deep auto-encoder (2) optimization of parameters (i.e. clustering) in which it iterates by calculating auxiliary targets and minimizing the Kullback–Leibler (KL) divergence to it. The proposed DEC method combines Autoencoder with a custom clustering layer with a K-means algorithm. Fig12 represents the architecture of the proposed DEC model.

During the initialization of the DEC model, the deep autoencoder is trained to produce the low-dimensional latent feature of the images. The dimensions of the autoencoder network in the proposed study are d-500-500-2000-10, where d is the dimension of the input data. The decoded layers are dropped and the encoded layers are connected to the clustering layer after training the autoencoder. The autoencoder network is trained using the Adam optimizer. The batch size is set at 128 and the deep autoencoder model is finetuned for 200 epochs with a constant learning rate of 0.0001. The embedded data from the autoencoder network is fit into the K-means model to initialize the centroids of the clusters.

C. PROPOSED FRAMEWORK FOR AMBULANCE POSITIONING SYSTEM

1) DATA PREPROCESSING

While dealing with the dataset few missing values were discovered. In order to deal with this problem a mean fill approach is used to impute the missing values. The missing values are imputed with the row average. The next step after data imputation in data preprocessing is attribute conversion. Attribute conversion is applied if the attribute type

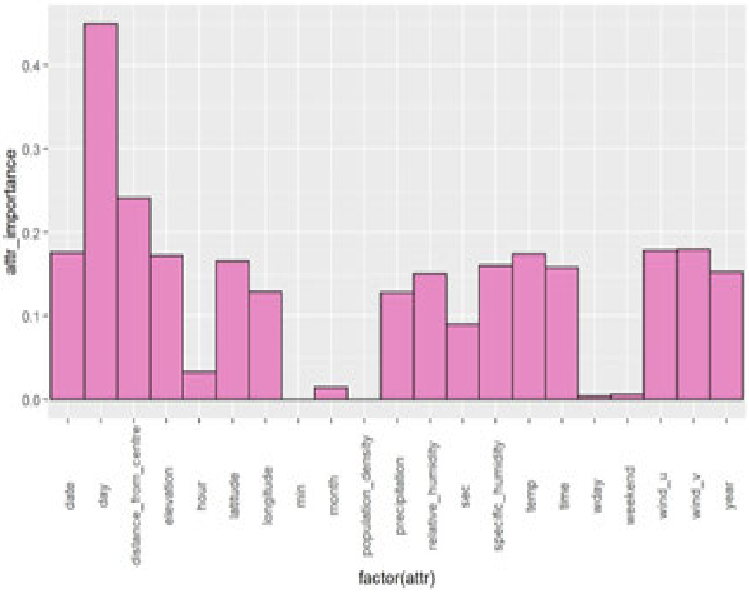


FIGURE 11. Bar graph for information gain of each attribute.

is categorical. To emphasize the impact of transformation techniques used on categorical variables and how the choice of transformation can show noticeable changes in model performance this paper conducts research in 2 phases. Phase 1 is applying traditional transformation like one-hot encoding in attribute conversion for categorical variables and Phase 2 is using Cat2Vec, a deep embedding-based approach for attribute conversion.

2) FEATURE SELECTION

The primary objective of Feature selection is to identify the most significant feature set that enhances the creation of models. For unsupervised ML problems such as clustering, filter methods procures the fundamental properties of the attributes which is evaluated using univariate statistics instead of cross-validation performance. Information gain calculates the reduction in entropy from the transformation of a dataset [32]. The strategy is employed for feature selection by evaluating the data gain of every variable within the context of the target. In the proposed method, information gain is calculated for all the attributes of the input dataset with respect to the initial clusters of k-means clustering. Fig 13 displays the bar graph for the information gain of each attribute in the dataset. Based on the results obtained, the 10 best features are selected for further processing steps.

3) MODELLING

Section 3.4 mentions that the last step while applying the DEC model is using a custom clustering layer with K-means. It is important to rescale the features in the range [0,1] since K-means uses Euclidean distances as a similarity measure which makes it sensitive to the scale of its features. Min-Max Scaler is used in the proposed work for Normalization.

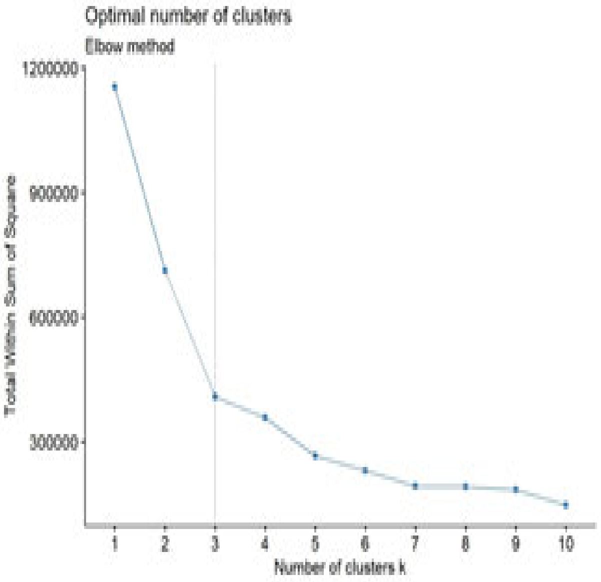


FIGURE 12. Elbow plot for the number of clusters.

The scaled and transformed dataset is then used as input for the deep embedded clustering model (DEC). The DEC model predicts 5 clusters over the city, whose centroids are considered as the predicted ambulance locations. Choosing the number of ambulances required for each city or area involves decision-making on different levels. In this study, the number of clusters chosen which in terms determines the number of ambulance locations has been determined using both statistical and real-time logical facts.

Considering the use of the elbow method and silhouette method to establish the ideal number of clusters, from

Algorithm 1: Proposed framework for optimal ambulance positioning system

```
Input: Nairobi City Dataset D
Output: Coordinates of ambulance positions
1. Data Preprocessing
        1.1 For each attribute in the dataset
                if attribute_value==NULL
                  then fill missing value(s) by mean(attribute)
        1.2 if attribute==categorical
                If Phase 1 then apply One-Hot Encoding
                If Phase 2 then apply Cat2Vec model
        1.3 For each attribute in the dataset calculate InformationGain(attribute)
        1.4 Select top 10 features based on InformationGain for further processing
        1.5 Scale all values of each attributes in the dataset using MinMaxScaler(values)
        Output:New dataset (filtered) is used for subsequent phases.
2.Apply elbow method and average silhouette score to choose the optimal number of clusters
3.Apply K-fold cross-validation with K value at 10 and 30
4.DEC Model Building
        4.1 Estimating the number of clusters
        4.2 Creating and training a K-means model
        4.3 Creating and training an autoencoder
        4.4 Creating a new DEC model
        4.5 Training the New DEC Model
        4.6 Using the Trained DEC Model for Predicting Clusters
        4.7 Calculating centroids of each cluster (ambulance locations)
5.Model Evaluation and comparison with standard algorithms
        Phase 1. Without Cat2Vec for categorical embedding
        Phase 2. With Cat2vec for categorical embeddings
        Output: Optimal Coordinates of ambulance positions
```

Fig 14. it is evident that the elbow occurs at k=3, but Van Essen et al. [33] shows that the number of ambulances in a city is also dependent on the factors like the accident occurring per second, the range of area of the accidents, and the demand of ambulances. Based on the Nairobi dataset, and keeping the above-mentioned factors in consideration the number of clusters determined was five which is also the second elbow in Fig 15. The experiments are conducted using K-Fold cross-validation with two different K values, 10 and 30.The semantic view of the proposed ambulance positioning system is presented in Algorithm 1.

V. EXPERIMENTS AND MODEL EVALUATION

Experimental Setup: This section discusses the results from the analysis of the Nairobi City Accident dataset using the proposed ambulance positioning system. The Python 3.8 platform was taken into account with certain fundamental packages like NumPy, scipy, and matplotlib and data-analysis packages like Keras, scikit-learn, and TensorFlow for the implementation of different algorithms, data analysis, and prediction. The study uses an i7 core CPU running at 3.4 GHz and 8 GB of RAM to conduct all experiments To validate the experiment, the proposed work the suggested work is compared with state-of-the-art clustering algorithms like K-Means, Gaussian Mixture Model, and Agglomerative clustering.

A. PERFORMANCE ANALYSIS METRICS

In this research work 4 evaluation and performance measures including Silhouette Score, Davies-Bouldin Index,, Calinski-Harabasz Index and a novel scoring measure have been used to compute the efficacy of the proposed model.

1) SILHOUETTE SCORE

The Silhouette Score is a very helpful tool to prospect the similarities and differences within and across clusters visually by measuring the separation distance between the clusters. It ranges from $[-1,1]$ and shows the proximity of each point in a cluster to the points of the neighboring cluster. Values closer to $+1$ i.e higher Silhoutees Coefficients indicate the cluster's sample is further away from the neighboring clusters' sample. The formula used to calculate the Silhouette Score is:

$$SC = \frac{(b-a)}{max(a,b)}$$

where a = mean intra-cluster distance b = mean nearest-cluster distance

2) CALINSKI-HARABASZ INDEX

Calinski-Harabasz Index is a great tool for evaluating the performance of clustering algorithms as the a-priori

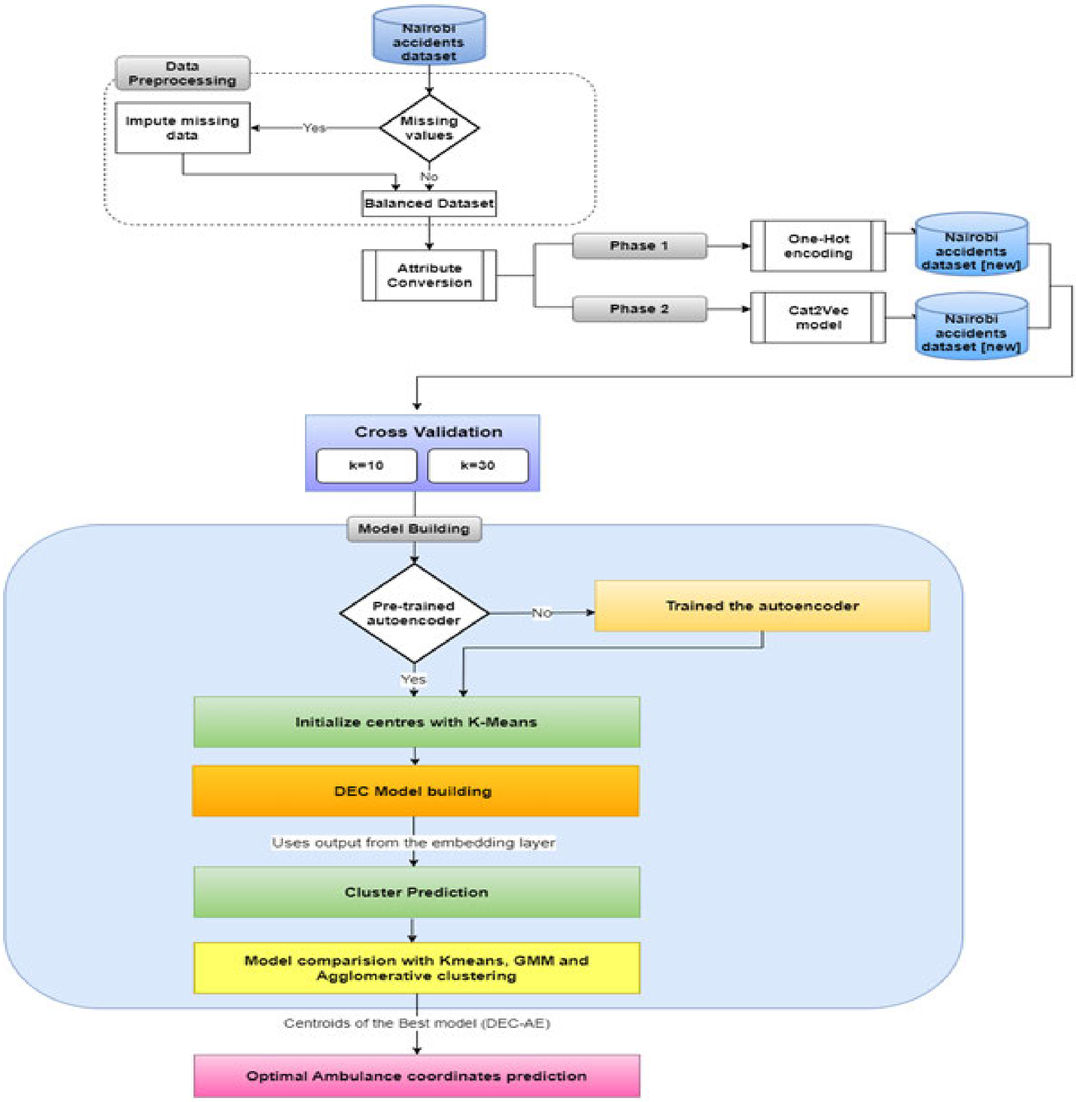

Nairobi accidents dataset
Data Preprocessing
Impute missing data
Yes
Missing values
No
Balanced Dataset
Attribute Conversion
Phase 1
One-Hot encoding
Nairobi accidents dataset [new]
Phase 2
Cat2Vec model
Nairobi accidents dataset [new]
Cross Validation
k=10
k=30
Model Building
Pre-trained autoencoder
No
Trained the autoencoder
Yes
Initialize centres with K-Means
DEC Model building
Uses output from the embedding layer
Cluster Prediction
Model comparision with Kmeans, GMM and Agglomerative clustering
Centroids of the Best model (DEC-AE)
Optimal Ambulance coordinates prediction

FIGURE 13. Proposed ambulance position prediction system.

knowledge of ground truth label information is not required, which suits the proposed study where the ground truth labels are not available. C-H index is a ratio between the within-cluster dispersion and the between-cluster dispersion. A higher C-H index indicates a better performance.

The formula used to calculate the CB index is:

$$SC = \frac{tr(B_a)}{tr(W_a)} \times \frac{n_E - a}{a - 1}.$$

Here $tr(B_k)$ represents the trace of the between-group dispersion matrix and $tr(W_k)$ represents the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{a} \sum_{x \in C_q} (x - C_q) \cdot (x - C_q)^T$$

$$b_k = \sum_{q=1}^{a} \sum_{x \in C_q} (C_q - C_E) \cdot (C_q - C_E)^T$$

3) DAVIES-BOULDIN INDEX

Like C-H index Davies-Bouldin Index does not require a ground truth label. It is an average similarity measure of each cluster with its most similar cluster wherein the similarity defines the ratio of within-cluster distances to between-cluster distances. A lesser score results in better performance as it shows that the clusters are less dispersed and further apart from each other. This way, clusters that are less dispersed and farther apart will provide a better score.

The formula used to calculate the DB index is:

$$DB \equiv \frac{1}{N} \sum_{i=1}^{N} D_i$$

where,

$$D_i \equiv \max_{(j \neq i)} R_{i,j}$$

Algorithm 2: Distance score calculation

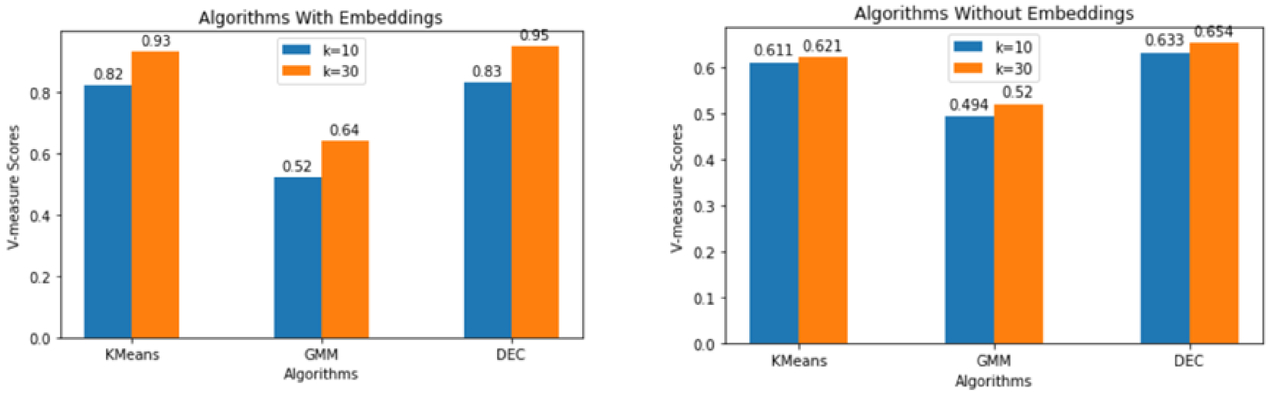
```
Input: Test dataset and result dataset
Output: Average distance score
1.1 total_distance=0
1.2for each latitude, longitude in test dataframe
        minimum_distance = 0
        for each in result_coordinates
                distance = $((latitude - result\_lat)^2 + (longitude - result\_long)^2)^{(0.5)}$
                inv_distance = 1/dist //to maximize the score function
Add inv_distance in list D
Output: average of D
```

TABLE 7. Performance metrics of DEC-AE model without Cat2Vec for categorical embedding.

Algorithm Used	Silhouette Score	Distance Score	Davies Bouldin	Callinski Harbasz
K-Means	0.1845	3.052	1.241	2759.71
GMM	0.1410	1.599	1.219	2714.82
Agglomerative	0.1748	2.528	1.275	2686.49
DEC-AE	0.1960	7.325	1.472	2195.50

TABLE 8. Performance metrics of DEC model with Cat2vec for categorical embeddings.

Algorithm Used	Silhouette Score	Distance Score	Davies Bouldin	Callinski Harbasz
K-Means	0.3416	3.641	2.096	712.17
GMM	0.3489	1.529	2.320	660.42
Agglomerative	0.3353	3.233	2.114	646.88
DEC-AE	0.4853	7.581	2.523	500.29



(a)

(b)

FIGURE 14. Overall performance evaluation of Phase 1 and Phase 2.

N = number of clusters

i,j = numbers of clusters which come from the same partitioning

4) NOVEL SCORING MEASURE

The above classification metrics evaluate cluster dynamics like the distance between cluster points, inter and intra-cluster similarity, and dispersion but do not evaluate how well the proposed model works in a real-time simulation. The problem of finding optimal locations to place the ambulances can only be resolved if the response time and distance between ambulance and crash sites are assessed. Since the above evaluation metrics fail to measure any such parameters, the novel scoring function introduced solves this issue. It calculates the average separation distance between crash sites and predicted centroids using Euclidean distance. With a higher score indicating greater performance, it shows how close each point in a cluster is to the predicted cluster centroid. Shown below is the algorithm used as the novel scoring function.

B. WITHOUT Cat2Vec CATEGORICAL EMBEDDING

In this section, the results after pre-processing the data without deep encoding on the Nairobi city accident dataset are described. The acquired results are based on the average

TABLE 9. Overall accuracy comparison using K-fold cross-validation.

Algorithm	k	V-measure Without Embeddings	V-measure With Embeddinggs
K-Means	k=10	0.611	0.82
	k=30	0.621	0.93
GMM	k=10	0.611	0.82
	k=30	0.621	0.93
DEC-AE	k=10	0.633	0.83
	k=30	0.654	0.95

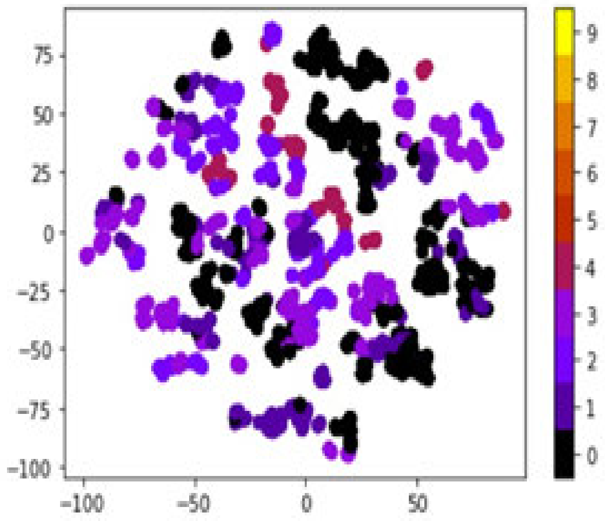


FIGURE 15. Predicted clusters from Phase 1.

TABLE 10. Final coordinates of cluster-centroids (ambulance locations).

S. No.	Latitude	Longitude
1	-1.280920	36.854787
2	-0.679048	37.032580
3	-1.901356	36.671511
4	-1.282354	36.854364
5	-0.870903	36.975906

findings from the test data set for each fold after applying 10 and 30-fold cross-validation to each model. Table 6 reports the comparison of individual classifiers including KMeans, DEC-AE, and Gaussian Mixture Model with the proposed deep-embedded clustering framework. It shows how the proposed model proves to give better performance when compared with standard clustering algorithms. Table 6. Shows the performances of different clustering algorithms used for the said problem. In order to measure the efficiencies, cluster scoring metrics namely Silhouette score, novel distance score, Davies Bouldin score, and Callinski Harbasz score have been applied. The DEC-AE performs consistently well in all the scoring criteria giving a silhouette score of 0.1960, distance score of 7.325, Davies Bouldin score of 1.472, and Calinski Harbasz score of 2195.50.

C. With Cat2vec FOR CATEGORICAL EMBEDDINGS

The aim of this approach is to emphasize that using the Cat2vec model for transforming categorical variables to deep embeddings can enhance the performance of the positioning framework and corroborate its effect on optimizing ambulance locations. The Cat2vec model preserves the relationship between the categories and feeds the dataset with the patterns identified in data analysis to the DEC model. The performance of the model using Cat2Vec for categorical transformations to deep embedding and then feeding the dataset along with the deep embeddings to the DEC model are detailed in this section. To verify its supremacy, 10 and 30 cross-validations is employed to assess performance. The below-given Table 7. Depicts the performances of different clustering algorithms used for the said problem. In order to measure the efficiencies, cluster scoring metrics namely Silhouette score, novel distance score, Davies Bouldin score, and Callinski Harbasz score has been applied. The DEC-AE performs consistently well in all the scoring criteria giving a silhouette score of 0.4853, distance score of 7.581, Davies Bouldin score of 2.523, and Calinski Harbasz score of 500.29. The results demonstrated that Cat2vec helps significantly in increasing the efficiency of the model.

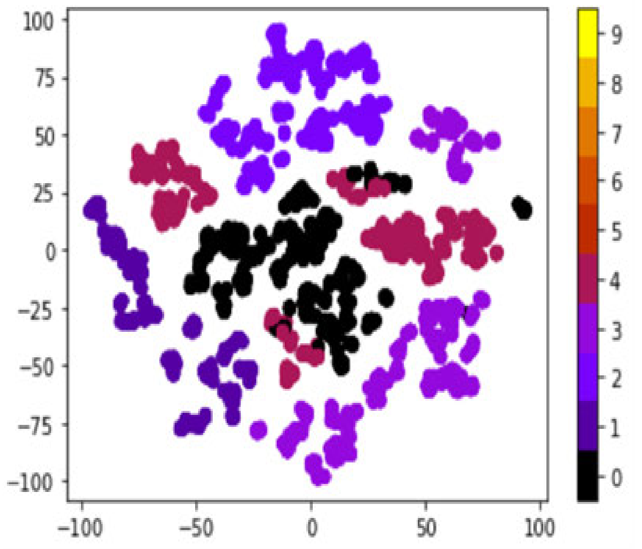


FIGURE 16. Predicted clusters from Phase 2.

D. OVERALL ACCURACY COMPARISON

Fig 16 depict the overall comparison of the proposed ESD prediction system with and without Cat2Vec using a K-fold cross-validation variant with values of 10 and 30. For the evaluation of the homogeneity and completeness of the clusters calculated, V-measure is used as a scoring function k-cross validation for both with and without Cat2Vec embeddings. The main advantage of using V-measure is that this evaluation metric is independent of the number of clusters, the number of data points in each cluster as well as the number of class labels of the dataset. Hence, it makes this metric very suited for the given problem.

Table 8. depicts that the DEC-AE has reached the maximum score of 0.95 with k-fold cross-validation accuracy, showing that the predicted position for ambulance locations will help the paramedics reach the crash site faster resulting in lesser casualties.

Fig 15 and Fig 16 supports the evidence that on applying Cat2Vec embedding, the clusters obtained are much more compact and clear in comparison to those without the embeddings.

Table 9. Accounts for the optimal coordinates obtained to position the ambulances in Nairobi city after applying DEC-AE with Cat2Vec embeddings.

The latitude and longitude represent the optimal positions in the city from where paramedic help can easily reach the crash locations in minimum time and thereby reducing road accident fatalities.

VI. CONCLUSION

Over the past 20 years, methods for identifying accident hotspots and determining optimal paramedic positions have evolved and now plays a significant role in the successful implementation of traffic safety management programs. This study aimed to develop and compare models for predicting optimal locations for positioning ambulances in Nairobi city, based on the Nairobi accidents dataset from 2018 to 2019. The final model utilized the Cat2Vec model for converting categorical data to numerical data in the form of embeddings for respective categorical attributes. Following data preprocessing and feature selection, a clustering-based approach was followed using Deep Embedded Clustering along with standard machine learning algorithms like K-Means clustering, GMM, and Agglomerative clustering to identify five clusters, the centroids of which provided the optimal ambulance positions. In order to evaluate the clustering algorithms, performance metrics including the Silhouette score, Calinski-Harbasz score, Davies Bouldin Score, and V-measure were used. To evaluate the distance of the centroid and the predicted ambulance locations, a novel scoring method namely Distance score was implemented. Among the developed model the DEC-AE model with Cat2Vec embeddings provided the highest accuracy of 95% in k-fold cross-validation. The distance score of 7.581 for the DEC-AE model which is higher than standard machine learning algorithms depicts that the distance between possible crash locations and ambulance positions is minimum. The analysis of various clustering metrics mentioned above reveals that the proposed DEC-AE model consistently outperforms other models in terms of clustering performance. This finding highlights the effectiveness and robustness of the DEC-AE model in accurately clustering the data and capturing underlying patterns. The study will advise decision-makers on where best to invest or implement security measures.

VII. FUTURE SCOPE

The study on predicting optimal ambulance positions in Nairobi using DEC has opened up several avenues for future research. Firstly, the dataset used in this study can be enhanced by including additional variables such as road type, road construction, speed limit, accident severity, driving behaviour, and road conditions that were absent due to data scarcity. The incorporation of these variables would provide a more comprehensive understanding of the factors influencing accident occurrences and improve the accuracy of ambulance positioning models. Additionally, expanding the analysis period beyond the limited years of 2018 to 2019 would allow for the identification of temporal trends and patterns in accidents. A longer time span would capture seasonal variations, changes in traffic patterns, and the impact of evolving road infrastructure or safety measures. Including cities with varying levels of urbanization and traffic conditions would help understand how the models perform in diverse settings. This longitudinal analysis can provide valuable insights into the effectiveness of safety measures, policy interventions, and changing patterns of road accidents.

Furthermore, conducting comparative studies using datasets from other cities, particularly in different socio-economic contexts, would provide insights into the generalizability and robustness of the proposed approach. It would allow for a more comprehensive understanding of the factors that contribute to accidents and their variations in different settings. Moreover, further analysis is recommended to examine trends in accidents over time. In this regard, the use of time series analysis is recommended. Integrating real-time data feeds, such as traffic flow, weather updates, and accident reports, would further enhance the responsiveness and adaptability of ambulance positioning models. By incorporating live data streams, the models could dynamically adjust ambulance positions based on changing traffic conditions, accident hotspots, and real-time demands.

FUNDING AND/OR CONFLICTS OF INTERESTS/COMPETING INTERESTS

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] *Global Status Report on Road Safety*, World Health Organization, Geneva, Switzerland, 2015.

[2] T. Sivakumar and R. Krishnaraj, "Road traffic accidents due to drunken driving in India–challenges in prevention," in *Proc. Int. J. Res. Manage. Technol.*, vol. 2, no. 4, p. 1, 2012.

[3] C. Baguley, "The importance of a road accident data system and its utilization," Tech. Rep., Nov. 2001, pp. 1–2.

[4] W. Odero, M. Khayesi, and P. M. Heda, "Road traffic injuries in kenya: Magnitude, causes and status of intervention," *Injury Control Saf. Promotion*, vol. 10, nos. 1–2, pp. 53–61, Apr. 2003.

[5] A. F. G. G. Ferreira, D. M. A. Fernandes, A. P. Catarino, and J. L. Monteiro, "Localization and positioning systems for emergency responders: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2836–2870, 4th Quart., 2017.

[6] M. F. N. Maghfiroh, M. Hossain, and S. Hanaoka, ''Minimising emergency response time of ambulances through pre-positioning in Dhaka city, Bangladesh,'' *Int. J. Logistics Res. Appl.*, vol. 21, no. 1, pp. 53–71, Jan. 2018.

[7] J. Kauffmann, M. Esders, L. Ruff, G. Montavon, W. Samek, and K. Müller, ''From clustering to cluster explanations via neural networks,'' *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 7, 2022, doi: 10.1109/TNNLS.2022.3185901.

[8] K. Assi, S. M. Rahman, U. Mansoor, and N. Ratrout, ''Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol,'' *Int. J. Environ. Res. Public Health*, vol. 17, no. 15, p. 5497, Jul. 2020.

[9] A. J. Ghandour, H. Hammoud, and S. Al-Hajj, ''Analyzing factors associated with fatal road crashes: A machine learning approach,'' *Int. J. Environ. Res. Public Health*, vol. 17, no. 11, p. 4111, Jun. 2020.

[10] P. Tiwari, H. Dao, and N. G. Nguyen, ''Performance evaluation of lazy, decision tree classifier and multilayer perceptron on traffic accident analysis,'' *Informatica*, vol. 41, no. 1, pp. 39–46, 2017.

[11] T. A. Granberg and H. T. N. Nguyen, ''Simulation based prediction of the near-future emergency medical services system state,'' in *Proc. Winter Simul. Conf. (WSC)*, Dec. 2018, pp. 2542–2553.

[12] C. Boutsidis, P. Drineas, and M. W. Mahoney, ''Unsupervised feature selection for the k-means clustering problem,'' in *Proc. NIPS*, 2009, pp. 1–9.

[13] H. Liu and L. Yu, ''Toward integrating feature selection algorithms for classification and clustering,'' *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[14] S. Alelyani, J. Tang, and H. Liu, ''Feature selection for clustering: A review,'' in *Data Clustering: Algorithms and Applications*, 1st ed. Taylor & Francis, 2013, p. 32.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[16] G. Cao, J. Michelini, and K. Grigoriadis, ''Cluster-based correlation of severe braking events with time and location,'' *J. Intell. Transp. Syst.*, vol. 20, no. 6, 2015, Art. no. 187e192.

[17] K. Moriya, S. Matsushima, and K. Yamanishi, ''Traffic risk mining from heterogeneous road statistics,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 11, pp. 3662–3675, Nov. 2018.

[18] S. Alkheder, M. Taamneh, and S. Taamneh, ''Severity prediction of traffic accident using an artificial neural network,'' *J. Forecasting*, vol. 36, no. 1, pp. 100–108, Jan. 2017.

[19] S. H.-A. Hashmienejad and S. M. H. Hasheminejad, ''Traffic accident severity prediction using a novel multi-objective genetic algorithm,'' *Int. J. Crashworthiness*, vol. 22, no. 4, pp. 425–440, Jul. 2017.

[20] B. Ghosh, M. T. Asif, and J. Dauwels, ''Bayesian prediction of the duration of non-recurring road incidents,'' in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2016, pp. 87–90.

[21] S. Sasaki, A. J. Comber, H. Suzuki, and C. Brunsdon, ''Using genetic algorithms to optimise current and future health planning–the example of ambulance locations,'' *Int. J. Health Geographics*, vol. 9, no. 1, pp. 1–10, Dec. 2010.

[22] M. Taamneh, S. Taamneh, and S. Alkheder, ''Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks,'' *Int. J. Injury Control Saf. Promotion*, vol. 24, no. 3, pp. 388–395, Jul. 2017.

[23] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, ''Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5747–5756.

[24] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu, ''Learning deep representations for graph clustering,'' in *Proc. 28th AAAI Conf. Artif. Intell.*, Jun. 2014, pp. 1–6.

[25] A. Alqahtani, X. Xie, J. Deng, and M. W. Jones, ''A deep convolutional auto-encoder with embedded clustering,'' in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4058–4062.

[26] J. Olusina and W. A. Ajanaku, ''Spatial analysis of accident spots using weighted severity index (WSI) and density-based clustering algorithm,'' *J. Appl. Sci. Environ. Manage.*, vol. 21, no. 2, pp. 397–403, Apr. 2017.

[27] X. Xiong, L. Chen, and J. Liang, ''A new framework of vehicle collision prediction by combining SVM and HMM,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 699–710, Mar. 2018.

[28] M. Zheng, T. Li, R. Zhu, J. Chen, Z. Ma, M. Tang, Z. Cui, and Z. Wang, ''Traffic accident's severity prediction: A deep-learning approach-based CNN network,'' *IEEE Access*, vol. 7, pp. 39897–39910, 2019.

[29] C. Yu, ''Research of time series air quality data based on exploratory data analysis and representation,'' in *Proc. 5th Int. Conf. Agro-Geoinformatics (Agro-Geoinformatics)*, Jul. 2016, pp. 1–5, doi: 10.1109/agro-geoinformatics.2016.7577697.

[30] Y. Wen, J. Wang, T. Chen, and W. Zhang, ''Cat2Vec: Learning distributed representation of multi-field categorical data,'' Tech. Rep., 2016.

[31] J. Xie, R. Girshick, and A. Farhadi, ''Unsupervised deep embedding for clustering analysis,'' in *Proc. Int. Conf. Mach. Learn.*, vol. 48, 2015, pp. 478–487.

[32] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, ''Feature selection using information gain for improved structural-based alert correlation,'' *PLoS ONE*, vol. 11, no. 11, Nov. 2016, Art. no. e0166017.

[33] J. T. Van Essen, J. L. Hurink, S. Nickel, and M. Reuter, ''Models for ambulance planning on the strategic and the tactical level,'' Univ. Eindhoven, Beta Res. School Oper. Manage. Logistics, Eindhoven The, Netherlands, Beta Work. Paper WP-434, 2013.

DHYANI DHAVAL DESAI is currently pursuing the B.Tech. degree in computer science engineering from the Vellore Institute of Technology, Chennai, India. Her research interests include data analytics, cloud computing, and machine learning and its applications in semi supervised clustering methods.

JOYEETA DEY is currently pursuing the B.Tech. degree in computer science engineering from the Vellore Institute of Technology, Chennai, India. Her research interests include data analytics and deep learning and its applications in deep embedded clustering and image processing.

SANDEEP KUMAR SATAPATHY received the Ph.D. degree in data mining and machine learning. His Ph.D. thesis include 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 did his doctorate in the field of data mining and machine learning, where his thesis included a detailed classification of brain EEG signals using machine learning techniques. He is currently an Assistant Professor (Senior) with the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai. He is highly engrossed into the areas of deep learning, image processing, and machine learning. He has many research publications to his credit, i.e., 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 currently a member of many professional organizations and society. He has been an active reviewer in various peer-reviewed journals and prestigious conferences. He has also reviewed many research articles and books in Elsevier for possible publication.

SHRUTI MISHRA 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 School of Computer Science and Engineering, Vellore Institute of Technology, Chennai. She has more than 30 papers 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 have also served as a reviewer for many reputed journals and conferences.

SACHI NANDAN MOHANTY (Senior Member, IEEE) received the Ph.D. degree from IIT Kharagpur, India, in 2015, with MHRD Scholarship from the Government of India, and the Ph.D. degree from IIT Kanpur, in 2019. He has guided six Ph.D. Scholar. He has published 60 international journals of international repute. He has edited 24 books in association with Springer and Wiley. His research interests include data mining, big data analysis, cognitive science, fuzzy decision making, brain–computer interface, cognition, and computational intelligence. He was elected as a fellow of the Institute of Engineers and a Senior Member of the IEEE Computer Society Hyderabad Chapter. He has received three best paper awards during his Ph.D. degree at IIT Kharagpur from the International Conference in Beijing, China, and the other from the International Conference on Soft Computing Applications organized by IIT Rookee, in 2013. He was a recipient of the Best Thesis Award First Prize Award from the Computer Society of India, in 2015. He is also a Reviewer of *Robotics and Autonomous Systems* (Elsevier), *Computational and Structural Biotechnology Journal* (Elsevier), *Artificial Intelligence Review* (Springer), and *Spatial Information Research* (Springer).

PALLAVI MISHRA is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education (Deemed to be University), Hyderabad, Telangana, India. Her research interest includes natural language processing. She has published two works with CRC Press (Taylor & Francis Group): her first article, "Big Data Digital Forensics and Cyber Security," was published in 2020, and her second book chapter, "Collaborative Filtering Techniques - Algorithms and Advances," was published in 2021.

SANDEEP KUMAR PANDA is currently an Associate Professor and the Head of the Department of Artificial Intelligence and Data Science, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education (Deemed to be University), Hyderabad, Telangana, India. He has published 50 papers in international journals and international conferences and book chapters in repute. He has 17 Indian patents on his credit. He has four edited books named *Bitcoin and Blockchain: History and Current Applications* (USA: CRC Press), *Blockchain Technology: Applications and Challenges* (Springer ISRL), *AI and ML in Business Management: Concepts, Challenges, and Case Studies* (USA: CRC Press), and *The New Advanced Society: Artificial Intelligence and Industrial Internet of Things Paradigm* (USA: Wiley Press), in his credit. He has ten lakh Seed money projects from IFHE. His research interests include blockchain technology, the Internet of Things, AI, and cloud computing. He is a member of ACM and a Life Member of IAENG. He was a recipient of the Research and Innovation of the Year Award 2020 from MSME, Government of India, and DST, Government of India at New Delhi, in 2020. He was also a recipient of the Research Excellence Award from Brand Honchos, in 2022.