SIGNIFICANT SPATIAL HOTSPOT DETECTION AND ANALYSIS OF HIGH-RISK ROAD ACCIDENT ZONES USING OPTICS & HDBSCAN CLUSTERING

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF DEGREE OF

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By
RISHABH JAIN
2K20/CSE/20

under the supervision of

Dr. ARUNA BHAT
(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042
May 2022

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Rishabh Jain**, Roll No. 2K20/CSE/20 student of M.Tech (Computer Science and Engineering), hereby declare that the Project Dissertation titled **"Significant Spatial Hotspot Detection and Analysis of High-Risk Road Accident Zones Using OPTICS & HDBSCAN Clustering"** which is being submitted by me to Delhi Technological University, Delhi, in partial fulfillment of requirements for the degree of Master of Technology in Computer Science and Engineering is a legitimate record of my work and is not copied from any source. The work contained in this report has not been submitted at any other University/Institution for the award of any degree.

Rishabh

Place: Delhi
Date: 25 May 2022

Rishabh Jain
(2K20/CSE/20)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I, hereby certify that the Project titled **"Significant Spatial Hotspot Detection and Analysis of High-Risk Road Accident Zones Using OPTICS & HDBSCAN Clustering",** submitted by Rishabh Jain, Roll No. 2K20/CSE/20, Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of degree of M.Tech in Computer Science and Engineering is a genuine record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree to this University or elsewhere.

Place: Delhi **Dr. Aruna Bhat**
25 May 2022 **Associate Professor**

ABSTRACT

Spatial Clustering is a subclass of Unsupervised Machine Learning algorithms that utilizes a specified criterion to group a set of geographically dispersed data points together. A Spatial hotspot is a confined region/space inside a geographical area which has a higher concentration of activity points than outside the hotspot region within the study area. With technological breakthroughs in spatial data collection and increased computation capabilities, several research publications presented novel methods for Spatial Hotspot detection. The relationship between Spatial Hotspots and time (temporal effect) has also been investigated by various researchers, expanding the horizon of the spatial hotspot to spatiotemporal hotspot. Spatiotemporal hotspots are a valuable tool for determining the highly vulnerable regions (e.g., high-risk crime territories, high accident-prone areas, severe disease-outbreak areas, areas more exposed to natural calamities and many more). The idea of Spatiotemporal hotspot detection gained considerable interest among researchers because of its practical utility in public health, public safety, traffic volume analysis, crime zone analysis, and other essential applications. Our research aims to organize the existing comprehensive literature into a well-organized hierarchical framework to comprehend better the various methodologies widely adopted around the world.

We propose a framework that suitably categorizes the research effort of various scholars into several appropriate categories based on their differences in primary algorithmic approaches. Furthermore, we also present an extensive analysis of the widely utilized evaluation measures adopted in this research domain. Effective clustering algorithms are required in societal applications such as road safety to uncover useful patterns linked with the data. Many applications employ Spatial Scan Statistics to identify spatial clusters, however it needs users to first determine the shape of the cluster, which is ambiguous in the context of road safety and could result in unfavorable outcomes. We outlined a method for discovering statistically significant shape-invariant spatial clusters in order to identify high-risk road accident zones.

The proposed approach incorporates the OPTICS and HDBSCAN clustering algorithms, as well as Cluster density and the Log Likelihood Ratio methods for evaluating the significance of spatial clusters. OPTICS and HDBSCAN are optimal for finding clusters of arbitrary shape and only require a single input parameter, making hyper parameter tuning relatively easier. We incorporated the statistical significance of clusters to eliminate spurious patterns. To illustrate the obtained results, the proposed approach is applied to the UK Road-Accidents data. We also presented a comprehensive analysis of the seasonal and temporal variations of the significant spatial hotspots discovered. Besides, we highlighted the methodology of prioritizing the identified hotspot zones based on the severity of the accidents for situations in which authorities are restricted with limited budget. We further suggest potential solutions for the probable causes of accidents in the hotspot area to guide relevant authorities to take timely and effective action.

ACKNOWLEDGEMENT

I am extremely grateful to my project guide, **Dr. Aruna Bhat,** Associate Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to her for the extensive support and encouragement she provided.

I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout by giving new ideas, providing necessary information and pushing me forward to complete the work.

Rishabh

Rishabh Jain

(2K20/CSE/20)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

STDM	Spatiotemporal Data Mining
ST	Spatiotemporal
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive
AUC	Area under Curve
ROC	Receiver Operating Characteristic Curve
OPTICS	Ordering Points to Identify the Clustering Structure
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
MAPE	Mean Absolute Percentage Error
LLR	Log Likelihood Ratio method
CD	Cluster-Density method
SC	Spatial Cluster
STC	Spatiotemporal Cluster

CHAPTER 1: INTRODUCTION

1.1 Overview

Spatial Clustering and Hotspot Detection is a special class of technique which can group a set of spatially distributed data points together based on pre-defined criteria in order to create a spatial cluster. Over the years, many Spatial Clustering techniques have been rendered to find insights about the behavior of data-points in various applications to provide relevant results. But, the problem arises when these clustering methods are applied on those application domains which have *low-tolerance to false positives*.

For instance, if one identified clusters of crime locations over a region and supposedly one of the identified clusters falsely claimed some area to be a crime hotspot, then it can seriously affect the status of the residents living in that area or tourists visiting that area. Thus, the identified clusters do not prove to be significant for the application, which is undesirable. This example clearly demonstrates that significance detection of spatial clusters is crucial for real-world applications in wide domains like crime, environment, accidents, political applications, natural calamities etc.

We introduce the concept of **"Significant Spatial Hotspot Detection and Analysis of High-Risk Road Accident Zones using OPTICS and HDBSCAN Clustering"** as a model capable to eliminate the mentioned problem with widely adopted traditional approaches. Significant Spatial hotspot detection takes statistical significance of clusters or hotspots into consideration to make the clusters resistant to spurious or false patterns or false positives. Many real-life applications in the domain of public health, transportation, forestry, environmental science, public safety etc. require significant clustering to deal with real datasets and provide authenticated results.

We have taken a serious actual world application of reducing road accidents in the world by analyzing the significant hotspots of road accidents so that areas with frequent road accidents are identified and infrastructure development, road planning and traffic management could be properly channeled and prioritized to such areas to prevent precious lives lost of both humans and animals as well as damage to vehicles and government property caused due to accidents.

1.2 Problem Statement I

Spatial clustering has been widely employed for use in massive datasets where the clusters are formed based on their closeness of the spatial data points or the similarity between these spatial data points. Many applications employ Spatial Scan Statistics as the most powerful technique to identify spatial hotspots of various geometric shapes (circular, elliptical, ring, linear, grid) but it requires the researcher to know the hotspot shape beforehand as an input which is ambiguous for many real-world problem statements and data-sets. It also takes a lot of computation time to be practically used for societal applications. Societal applications are observed to have arbitrary shaped spatial cluster. Thus, the traditionally used "Spatial Scan Statistics" is favorable to limited applications. Also, Significance detection of spatial clusters is crucial for real-world applications in order to remove spurious or chance patterns or false positives. This makes the spatial cluster analysis much more effective and yields out interesting and true insights about the behavior of data points as a whole.
Based on the problem statement 1 following questions is identified:

1. What are recently employed techniques of spatial hotspot detection used for real world applications?

2. Which technique is more suited to be employed in certain particular application domain?

3. What evaluation metrics are to be utilized for the particular application domain?

4. How to detect arbitrary shaped hotspots which is the core of hotspot detection in real world applications?

5. Is it possible not to give predefined shape as an input to the model which generates spatial hotspots?

6. How to test significance of the generated hotspots to ensure true insights only and not adulterated by chance or spurious patterns?

1.3 Problem Statement II

Everyday thousands of road accidents happen globally which always is not fault of driver alone but finds primary cause to be poor road connections, poor road infrastructure, bad management of road traffic, unsatisfactory road planning or any constant other external factor contributing to increased chances of accidents being caused at certain places. Such accidents lead to lives being destroyed with lives lost of both humans and animals. Also, it contributes to damage to vehicles and government infrastructure which adds to financial loss for both public and government. It is peculiar to observe that certain places experience more accidents than others which are gravely alarming for public safety and road safety.

So, the problem statement is to develop a model to eliminate road Accidents caused due to poor infrastructure, ineffective implementation of road safety guidelines and inefficient road traffic management for building safer smart cities.

Based on the problem statement 2, the following questions arise:

1. Can there be a model which points the hotspot locations of these prone to road accidents areas where accidents are not just because of human error but due to poor infrastructure, ineffective implementation of road safety guidelines and inefficient road traffic management?

2. Can such a model be developed which constructs safer roads harnessed with technology that brings out Smart outcomes for Smart City Building and Planning?

The Application domain of Spatial Clustering is sensitive in nature and demonstrates the real-world scenarios wherein the probability of obtaining misleading results is high. An erroneously identified spatial hotspot can have adverse effects. For example, suppose one of the clusters mistakenly labeled a specific area as a hotspot for accidents. As a result, critical resources such as needed traffic force may be wasted instead of being invested in the right region. Furthermore, if the actual hot spot areas are identified, the relevant authorities can more effectively plan and utilize their resources to benefit society.
This demonstrates that merely identifying spatial clusters is insufficient; it is also

necessary to determine the significance of the spatial cluster. Several Research studies suggested the use of a preset geometric shape for hotspot identification based upon past observations. For example, circularly shaped clusters are quite popular in case of Disease surveillance and Ring-shaped hotspots are prevalent in identification of high-risk zones of criminal activities. But, it is important to note that all spatial clusters cannot have a predefined geometric shape. Many real-world applications exist wherein the spatial clusters have an arbitrary shape. Spatial Scan Statistics by M.Kulldorff [16] is one of the most effective methods to determine spatial clusters of a predefined geometric shape (circular, linear, ring etc.) along with their statistical significance to yield accurate results. Major drawback of Spatial Scan Statistics is that it requires a predefined shape for cluster identification. Thus, Spatial Scan statistics becomes unfit for finding significant spatial clusters of arbitrary shape. Furthermore, selection of appropriate clustering algorithms can help in determining clusters of arbitrary shape and provide authentic results. Spatial Hotpot identification has largely relied on clustering based techniques.

1.4 Project Objectives

The emergence of social media, medical imaging, remote sensing, telecommunications, crowd sourcing, and other technologies has substantially facilitated collecting data at several locations (spatial domain) over a given time period (temporal domain). Spatiotemporal data mining is an emerging field of research. Spatiotemporal objects deal with both space and time attributes which are updated continually as time and location changes. A spatiotemporal database comprises a collection of such dynamic spatiotemporal objects. Several researchers studied spatiotemporal databases and designed applications in public health and disease outbreaks, crime and public safety, environmental science, natural disaster management, road safety and transportation, traffic management, and other crucial areas.

G. Atluri et al. [1] presented a comprehensive survey on the details of spatiotemporal data and outlined the problems and methods concerned with Spatiotemporal Data Mining (STDM). Hamdi et al. [2] conducted a detailed research survey on spatiotemporal data mining and discussed the challenges and tasks associated with the field. They classified clustering and hotspot detection as task-related challenges in STDM. The authors of [2] mentioned that spatiotemporal clustering is different from classification and involves the grouping of spatiotemporal objects to form clusters of similar spatiotemporal characteristics. S.Shekhar et al. [3] examined several crucial aspects of spatiotemporal data mining primarily concerned with the field of criminal activity analysis. The authors characterized the three essential attributes associated with spatiotemporal data as follows: non-Spatial attributes (e.g., population, crimes reported in the city, the unemployment rate in the city, city name, etc.), spatial attributes (e.g., latitude, longitude, etc.) and temporal attributes (e.g., timestamp of an event, duration of an event, etc.). Spatial data takes different forms: point, line, polygon, regular/irregular map fields, or graph network. The authors also stated the subtle difference between the concepts of spatiotemporal clustering and spatiotemporal hotspot detection.

Spatiotemporal clustering is a general term used for grouping a large number of spatiotemporal data items on the basis of similarity in their space-time characteristics. On the other hand, the process of spatiotemporal hotspot identification is also described as the phenomenon of identifying regions or clusters in a spatiotemporal domain where the number of activity points are unexpectedly but significantly higher within the time intervals. The major challenge associated with the obtained hotspots is eliminating the false positives so that the results are of practical importance and aligned with the underlying ground truth. Recent research works suggested various evaluation measures for better hotspot prediction in accordance with space-time characteristics.

Clustering techniques have been widely used in different scientific and social sectors, such as market research, pattern identification, fraudulent and criminal activity detection, network traffic analysis, and so on, to uncover patterns or trends of certain actions or events. Spatial Clustering is a subset of clustering algorithms that uses a predefined criterion to group a set of spatially scattered data points into a cluster. The essential concept of spatial clustering is that the spatial data points within a cluster are more similar to one another than the data points outside the cluster bounds. The equality or dependency between data points can all be used as clustering criteria, however it primarily depends on the application domain. The data points in the spatial domain may represent a particular event such as crime, disease, accident, taxi pickup areas etc. Basically, spatial clusters describe the behavior of data points in relation to one another which help in quantifying the variation in geographic patterns. Many spatial clustering approaches have been proposed and used in the past to extract useful information from unstructured data in a variety of applications. It is largely utilized in Disease surveillance, spatial epidemiology, Population genetics, Crime analysis and various other fields. A Spatial Hotspot can also be perceived as a Significant Cluster. A Significant Cluster is a unique cluster in which the frequency of events within the cluster exceeds that outside its bounds. In order to determine Significant Clusters, candidate clusters are formed and significance testing is performed on these clusters. We proposed a hybrid approach that combines Spatial Clustering with Significance Testing. The paper employed two types of clustering techniques, namely density-based clustering and hierarchical clustering.

The nature of geographical data points differs significantly from that of ordinary data points; hence it is essential to use appropriate algorithms to generate spatial clusters. We proposed an approach combining the capabilities of two efficient methods to obtain Significant Spatial Hotspots. The Clustering algorithms are employed to obtain candidate spatial clusters. Clusters that are resistant to false patterns are generated by spatial clustering with statistical significance of detected arbitrarily shaped clusters. We proposed a way for combining the benefits of both methods in order to overcome their respective constraints. The proposed algorithm used a comparison approach to choose two effective Spatial Clustering algorithms, OPTICS and HDBSCAN, from the Density based clustering algorithms and Hierarchical Clustering techniques categories, respectively. By detecting significant arbitrary shaped spatial hotspots of UK road-accidents, we demonstrated the results achieved by the proposed approach. We further proposed a comprehensive analysis of the seasonal and temporal variations of the identified significant hotspot zones.

A severity-based analysis of the significant spatial hotspots is also conduced to assist the relevant authorities in identifying the places that demand considerable attention and prioritizing all other hotspot zones as well. This efficiently assists the relevant authorities in developing a practical strategy that should be implemented in a timely manner in order to provide a better and safer road environment for the general public.

1.5 Role of Machine Learning

Spatiotemporal data mining is the process of extracting meaningful patterns, trends, or other relevant information from the data collected in both spatial and temporal dimensions. Spatiotemporal data mining applications are widely used in a variety of sectors, including criminology, medicine, transportation, public safety, and many more. Spatial Clustering is an important subcategory of Spatial Data Mining. Spatial Clustering is the aggregation of similar event occurrences occurring in space over a period of time. Spatial Hotspot detection is a special subset of Spatial Clustering.

Road safety has been a significant matter of concern worldwide for many years. With increasing number of vehicles on the road, the probability of accident occurrence also increases. Hence, it becomes essential to devise ways to promote road safety by improving upon external contributing factors such as road surface conditions, road signs, traffic management etc., in order to reduce the likelihood of road traffic accidents as much feasible. However, it is critical to identify accident-prone areas not only based on the number of accidents that have occurred in the past, but also by considering other aspects of an accident event such as road surface conditions, lightning conditions, weather conditions, time of the event, traffic police availability, accident severity, and a variety of other important factors. As a result, the spatial characteristics of the region as well as the temporal fluctuations of the event occurrence are the most important factors in determining the magnitude of a road accident at a specific site. Moreover, it becomes vital to identify the appropriate geographical locations having unexpectedly higher occurrence of road accident events over defined period of time such that the valuable resources are directed towards the right direction.

Many researchers have demonstrated interest towards spatiotemporal hotspot detection of traffic accidents from different parts of the world. Several researchers conducted various types of analyses and presented many revolutionary and interesting insights in various domains of technology, with the primary objective of road safety being acknowledged by everyone. Traditionally used statistics-based methods for spatial hotspot detection were merged and further employed with newer forms of Machine Learning or Deep Learning based models to develop an infused strategy that leveraged the benefits of both the techniques. Machine Learning-based models provide an added benefit in a variety of traffic accident analysis tasks, such as classifying the severity of accidents based on feature engineering or extracting patterns from data to anticipate or forecast future accident zones. Furthermore, Deep Learning based models, such as Convolution Neural Networks, give increased accuracy and the ability to handle greater size inputs, as well as perform better in terms of the road network environment. The paper provides a comprehensive survey of recently introduced methods for traffic accident analysis, such as Spatiotemporal road traffic accident hotspot detection, Principal Feature selection from traffic accident datasets that include spatial, temporal, and other influential factors, and also predicting the risk or severity of accidents after training the model using past data to predict future high-risk or low-risk road accident zones.

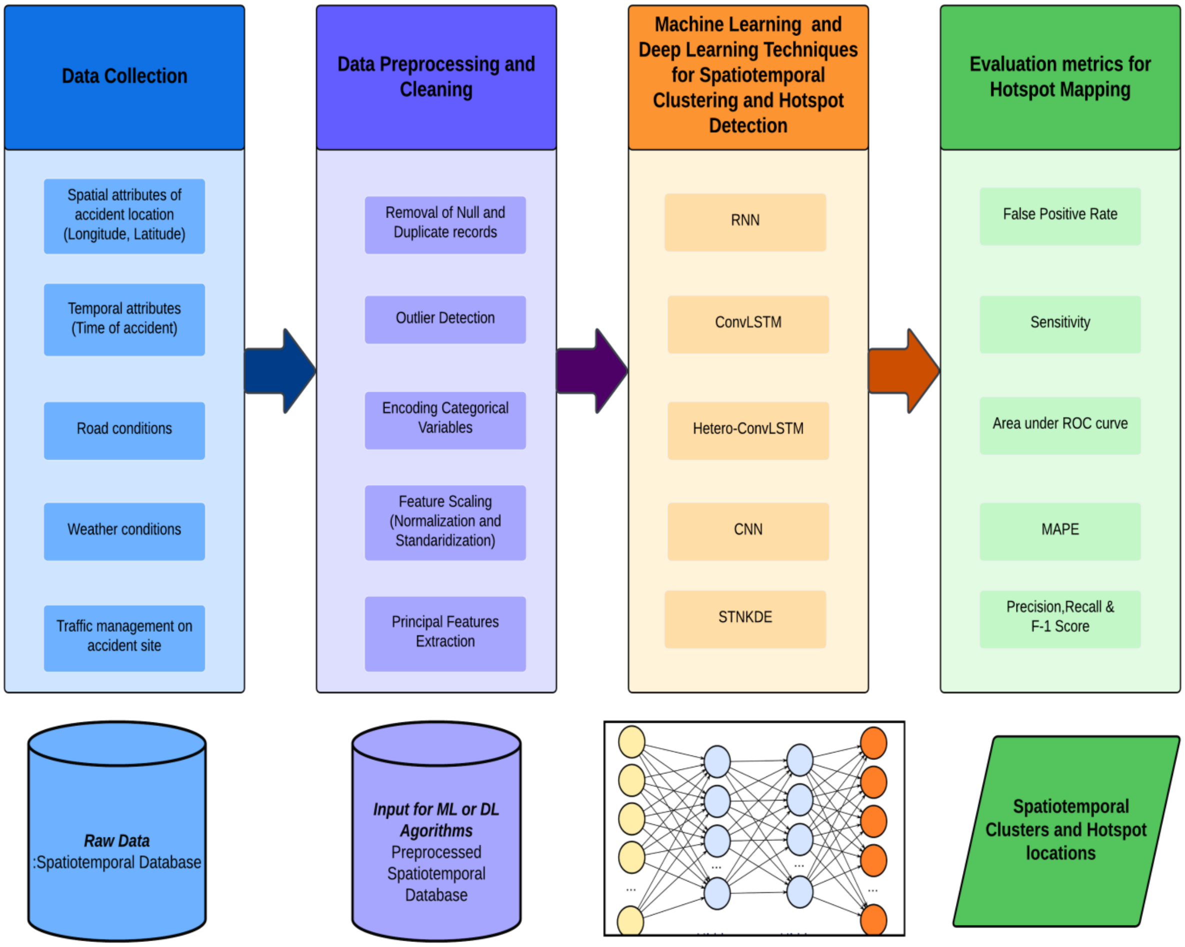

Data Collection
Spatial attributes of accident location (Longitude, Latitude)
Temporal attributes (Time of accident)
Road conditions
Weather conditions
Traffic management on accident site
Data Preprocessing and Cleaning
Removal of Null and Duplicate records
Outlier Detection
Encoding Categorical Variables
Feature Scaling (Normalization and Standaridization)
Principal Features Extraction
Machine Learning and Deep Learning Techniques for Spatiotemporal Clustering and Hotspot Detection
RNN
ConvLSTM
Hetero-ConvLSTM
CNN
STNKDE
Evaluation metrics for Hotspot Mapping
False Positive Rate
Sensitivity
Area under ROC curve
MAPE
Precision,Recall & F-1 Score
Raw Data :Spatiotemporal Database
Input for ML or DL Agorithms Preprocessed Spatiotemporal Database
Spatiotemporal Clusters and Hotspot locations

Fig.1 Diagrammatic Representation of Spatiotemporal Traffic Accident Analysis

CHAPTER 2: RELATED WORK

2.1 Overview of Literature Survey

This section examines a variety of methods for Spatiotemporal Hotspot detection that have been proposed all across the world. By categorizing the existing widely adopted approaches and recently proposed hotspot detection methods, we established a hierarchical framework to explore these methodologies. We classified the hotspot detection methods upon analyzing the two significant features of Spatiotemporal data analysis, i.e., Strategic computation techniques (or algorithms) and Application Areas. Based on the existing literature, we divided the methods for detecting Spatiotemporal Hotspots into the following eight categories demonstrated in Fig.2 below.

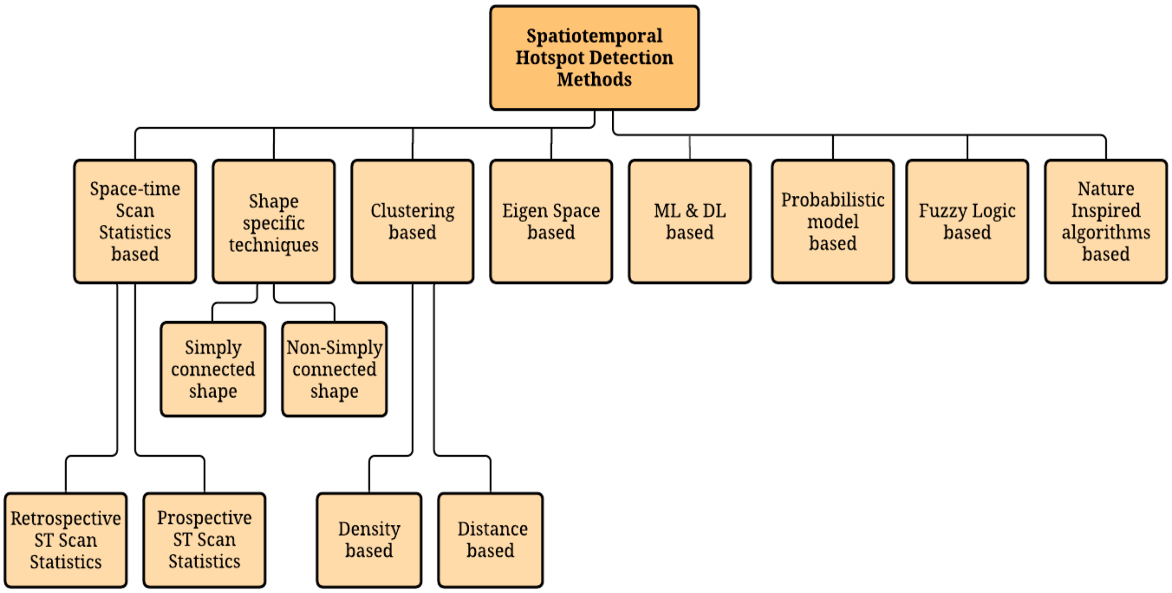

Spatiotemporal Hotspot Detection Methods
Space-time Scan Statistics based
Shape specific techniques
Clustering based
Eigen Space based
ML & DL based
Probabilistic model based
Fuzzy Logic based
Nature Inspired algorithms based
Simply connected shape
Non-Simply connected shape
Retrospective ST Scan Statistics
Prospective ST Scan Statistics
Density based
Distance based

Fig.2 Hierarchical framework for classification of ST hotspot detection methods

The prime objective of this systematic study is to organize the relevant research literature for Spatiotemporal hotspot detection and effective evaluation measures in order to assist the research community in gaining a better understanding of the state-of-the-art methods for determining Spatiotemporal hotspots. We attempted to summarize the Spatiotemporal hotspot detection algorithms. We define the application domain of research studies and suitably categorize them into the developed hierarchical framework after carefully selecting research works in the subject of Spatiotemporal hotspot identification.

Table I. Hotspot detection techniques with different application domains

Application Domain	Shape	Category	Approach	Reference
Disease outbreak	Circular	Space-Time scan Statistics based	SaTScan	[19] [7] [20]
	Circular	Space-Time scan Statistics based	Retrospective Space-time analysis	[9]
	Circular	Space-Time scan Statistics based	Prospective Space-time analysis	[10]
	Circular	Clustering based	Extended Fuzzy C-Means Clustering	[21]
	Elliptical	Space-time scan Statistics based	Spatial Scan Statistic with elliptical scanning window	[16] [22] [23]
	Ring	Clustering based	Grid-Based clustering with pruning	[14]
	Cylindrical	Space-time scan Statistics based	Prospective Space-time analysis	[11] [12]
	Irregular	Space-time scan Statistics based	Flexible Space-time Scan statistic	[24] [5]
	Irregular	Space-time scan Statistics based	GridScan	[17]
	Irregular	Clustering based	Polygon Propagation	[18]
	Irregular	Clustering based	AMOEBA clustering algorithm using modified Getis-Ord statistic	[19]
	Irregular	Eigen Space based	EigenSpot Algorithm	[25]

	Irregular	Eigen Space based	SST-Hotspot algorithm involving eigen vector matching	[26]
	Irregular	Nature-Inspired Algorithms based	PSO optimized scanning window in Space-time scan statistics	[27]
	Point Density based Cluster	Space-time scan Statistics based	Retrospective spatial analysis using General G statistics	[8]
Crime Hotspots	Elliptical	Space-time scan Statistics based	Spatial & temporal analysis of Crime (STAC)	[28]
	Elliptical	Clustering based	RNHH Clustering	[29]
	Cylindrical	Space-time scan Statistics based	SaTScan	[30]
	Irregular	Space-time scan statistics based	SaTScan	[6]
	Irregular	Clustering based	Spatiotemporal Kernel Density Estimation (STKDE)	[31]
	Irregular	Clustering based	Net KDE &Getis-Ord statistic	[32]
	Irregular	Clustering based	Modified Moving Window methods for network detection (EdgeScan&NDScan)	[33]
	Irregular	Machine Learning & Deep Learning based	Gated localized diffusion network (GLDNet)	[34]
	Irregular (Dynamically changing)	Machine Learning & Deep Learning based	Spatiotemporal Deep RNN	[35]
	Irregular	Fuzzy Logic based	Novel Genetic Fuzzy system	[36]
	Irregular	Nature-Inspired Algorithms based	Multiobjective Evolutionary Algorithm	[37]

	Irregular	Machine Learning & Deep Learning based	Deep Inception Residual Networks	[38]
Political Violence & Terrorism	Cylindrical	Space-time scan Statistics based	Space-time Scan statistic with cylindrical window	[39]
Road-Traffic Management & Accident Prevention	Irregular	Probabilistic Model based	Kernel Density estimation	[40]
	Irregular	Probabilistic Model based	Full-Bayesian approach with B-ST-I model	[41]
	Irregular	Clustering based	Weighted Fuzzy-C Means Clustering	[42]
	Irregular	Clustering based	Nearest neighborhood-related Quality clustering	[43]
	Irregular	Clustering based	ST-HDBSCAN	[44]
	Irregular	Clustering based	Network-based Spatiotemporal field clustering	[45]
	Point Density based cluster	Fuzzy Logic based	Weighted Overlay method and Fuzzy Overlay method	[46]
Public Sentiment Detection	Irregular	Machine Learning & Deep Learning based	Emerging Hotspot Analysis tool	[47]
Environmental Science & Disaster Management	Elliptical	Clustering based	Extended Gustafson–Kessel (EGK) clustering algorithm	[48]
Drug activity	Circular	Space-time scan Statistics based	Retrospective Space-time analysis	[13]

2.2 Hierarchical framework for classification of Spatio-Temporal Hotspot detection methods

A. Space-Time Scan Statistics based Techniques:

The Spatial Scan Statistic, introduced by M.Kulldorff et al. [4], is one of the most popular methods for detecting statistically significant hotspots of a predefined shape. It essentially conducts an exhaustive search for spatial activity points throughout the study area and discovers regions where the intensity of activity points/locations is higher than a specific threshold. However, they assumed that the spatial activity points are independent of one another, but it is not always the case in real-world scenarios. Moreover, this method turned out to be highly computationally expensive. Over the years, M.Kulldorff et al. [5] further investigated Spatial Scan Statistics and considered it viable to include the importance of temporal aspect as well, which help in determining highly vulnerable regions corresponding to a specific application.

In real-world situations, clusters can be irregularly shaped too. Hence, the authors of [5] presented a novel approach for detecting major areas of disease outbreaks through an advanced Space-time Scan statistics approach which incorporates flexibly shaped cluster detection. They modified the standard scanning method used in Space-time Scan Statistics into an advanced scan statistic measure that expands cylindrical scanning window by incorporating particular connected regions whose central point's lie within the candidate concentric circle of largest size. The authors mentioned that a limitation of this method is that it is incompatible with emerging hotspots (e.g., hotspots detected when disease grows or shrinks in time). The Space-time Scan statistics approach can be broadly classified into two groups namely, Retrospective Space-time Scan statistics and Prospective Space-time Scan statistics. Y. Kim and M. O'Kelly [6] outlined the differences in these two categories as follows.

1) *Retrospective Space-time Scan statistics:* This type of approach emphasizes on historical records of spatiotemporal data points to predict the spatiotemporal hotspots prevalent in the past. Consequently, the dataset utilized in retrospective analysis remains unchanged throughout time. This analysis approach proved out to be useful in multiple disciplines. In 2017, H. Rao, X. Shi, and X. Zhang, [7] performed a retrospective analysis of M.Kulldorff's Space-time scan statistics approach to find locations of spatiotemporal hotspots having higher incidents of Tuberculosis compared to adjoining study area. In 2020, D. Adham et al. [8] presented a study to demonstrate the retrospective spatial

analysis using the General G statistics to put forth most vulnerable regions suffering from pediculosis. In 2021, A. Guemes et al. [9] employed a retrospective methodology to identify spatial hotspots which are significant historical clusters of COVID-like symptoms by incorporating Space-time scan statistics using a scanning window which is circular.

2) *Prospective Space-time Scan statistics:* This strategy focuses on leveraging dynamic spatiotemporal data to detect the candidate hotspot zones in order to assist concerned authorities in effectively making important management decisions. This approach can be highly useful for planning out activities and businesses in various application areas. This method also aids in the detection of active clusters/hotspots that change over time. In 2014, J. Mosha et al. [10] presented a comparative study of different prospective approaches to detect spatial clusters of malaria disease and results showed that prospective Space-time Scan statistics outperformed other statistical techniques. Many researchers invested into prospective space-time hotspots to discover emerging spatial clusters to serve as a guidance tool to develop efficient management strategies. C. Chen et al. [11] applied prospective Space-time Scan statistics for identification of active hotspots of dengue fever.

With the unforeseen pandemic of COVID-19 infecting the entire globe towards the end of 2019, various research investigations to determine the growing spatiotemporal hotspot zones of COVID-19 were proposed. The applicability of Space-time Scan statistics for finding COVID-19 hotspot areas is obvious, as it has been widely used for disease outbreak detection in the past. A. Hohl et al. [12] utilized the SatScan tool and used daily COVID-19 data to predict emerging and active hotspot for COVID-19 Surveillance. Space-time Scan statistics mainly deal with two types of spatiotemporal hotspots, i.e., Persistent and Emerging. Persistent hotspots can be defined as the regions where the concentration of activity points grows rapidly with a constant rate. Emerging hotspots can be defined as the regions where the event occurrence abruptly or suddenly increases, resulting in a significant outbreak. This technique considers time as a third dimension besides the two primary spatial attributes (latitude and longitude) to identify spatiotemporal hotspots.

Space-time Scan statistics perform well when the data is present in the form of a time series or fixed points/events. For instance, [13] used Space-time scan statistics for drug activity monitoring in order to identify the regions of intense drug usage over a defined time interval. SaTScan is a free tool for analyzing spatiotemporal data using Space-time Scan statistics. This method is accepted by various research studies mainly concerning public health and disease outbreak pattern detection. Table 1 describes various applications which have adopted Space-time scan statistics.

B. Shape specific hotspot detection techniques:

Spatial scan statistics makes restrictive assumptions about predefined shapes of hotspots. Several researchers investigated into Spatial scan statistics and expanded it to construct space-time clusters of various shapes. Determining appropriate shape of hotspot prior to implementation is crucial. The initial circular shaped hotspots were studied in detail and their limitations in practical scenarios led to the development of hotspots of different shapes. Hotspots of diverse shapes correspond to different application requirements and deliver applicable results. It was observed that statistically significant hotspots with predefined shapes can be classified into two categories as follows.

1) *Simply connected shapes*: A Simply connected shape is defined as a two-dimensional region without any holes in it. Spatial Scan statistics or SaTScan is mainly utilized for the detection of statistically significant hotspots having shapes which are simply connected such as rectangles, circles, etc. [14].

2) *Non-Simply connected shapes*: A Non-Simply connected shape is defined as a two-dimensional region with holes in it. S.Shekhar et al. [14] demonstrated a novel approach for ring shaped hotspot detection. The methods for hotspot detection with restrictive shapes cannot detect emerging spatiotemporal hotspots which are capable of adjusting their region with change in activity points over time. IBM Research Division [15] proposed a novel approach that generates space-time clusters of square-pyramid shape which accommodate such changes. This approach works better than the original cylindered structure in detecting the growth or shrinkage of disease outbreak with time.

With an increase in computational power over time, research studies also shifted focus to detect irregular shaped hotspots as well that depicts real-world situations. Spatial Scan Statistics generally do not perform well to detect irregularly shaped space-time clusters. However, M.Kulldorff et al. [16] extended their work on spatial scan statistic in 2006 to identify hotspots of no fix geometric shape with changing scanning window through an elliptical scan statistic. In 2012, W.Dong et al. [17] presented a completely different technique, termed GridScan, based upon a local greedy search algorithm for identification of irregularly shaped spatial clusters from spatial point data. In 2019, S. Katragadda et al. [18] presented another approach for detecting irregularly shaped spatial hotspots on a spatial-point dataset based on polygon propagation. Table 1 demonstrates certain shape-specific hotspots along with their application domain.

C. Clustering based techniques:

Clustering is applied in a large variety of applications in Data Mining. Spatiotemporal clustering mainly groups activity points or regions or events based on spatiotemporal characteristics. Spatiotemporal hotspots are also perceived as a special type of Spatiotemporal clusters with an unexpectedly higher intensity of points within the hotspot region. Spatiotemporal clustering provides numerous methods to detect regions with similar characteristics.

Many researchers developed feasible methods to detect spatiotemporal clusters with improved computational complexity and modified their approach to find spatiotemporal hotspots. Clustering based approaches mainly focus on generating spatial clusters which serve as the candidate areas for a hotspot. Among the many candidate areas detected, spatial hotspots are identified through evaluation of statistically significant spatial clusters. Spatiotemporal clusters can be determined in a similar manner using the timestamp as the third dimension of data. Clustering can be performed on the grounds of several different measures as described below.

1) *Density-based clustering*: Density-based clustering methods determine hotspot regions from candidate cluster through a specified a density threshold which is used to separate relevant data from noise. The well-known DBSCAN algorithm clusters geographical and spatiotemporal data by evaluating density of data points. A well-known clustering technique namely, Risk Adjusted Nearest Neighbor Hierarchical Clustering (RNHH) consolidates the advantages of both Hierarchical clustering approach and Kernel Density estimation (KDE) techniques [29]. This method has been proven to be effective in the analysis of criminal incidents. However, DBSCAN, OPTICS, ST-DBSCAN, and other variations of many such algorithms are applicable to spatial point data.

With location services growing tremendously, large amount of data is trajectory dependent. It was observed that these methods perform well with trajectory data or moving data. Q. Yu et al. [43] recently proposed a density-clustering based method for identification of urban hotspot based upon taxi-trajectory data. Spatiotemporal Kernel density estimation techniques are also one of the popular methods for hotspot detection.

In 2018, Y. Hu et al. [31] proposed a multi-featured framework for determining statistically significant crime hotspots in the spatiotemporal domain using the modified kernel density estimation technique and evaluated the obtained results using the Predictive Accuracy Index curve and demonstrated their findings over the prevalent robberies in Baton Rouge, Louisiana. S. Khalid et al. [32] presented a hybrid model by combining the Spatial Analysis along Networks Tools.

2) *Distance-based clustering*: Clustering can also be performed using distance as threshold to establish boundaries for hotspot regions. J. Baker et al. [33] proposed novel scanning window methods, EdgeScan and Network Density Scan (NDScan), which utilize the eucledian distance measure to locate hotspots in spatial domain.

D. Eigen Space based techniques:

Several applications have been tested using Spatial Scan statistics and clustering algorithms, and the results have been promising. It is important to highlight, however, that these solutions rely on limiting assumptions that appear to be impracticable in some real-world scenarios. H.Fanaee-T et al. [25] suggested a unique computationally efficient method in the realm of spatiotemporal hotspot detection, which revealed some intriguing insights. The authors showed that a comprehensive search across the entire study region cannot produce reliable findings in a timely manner, but it is critical to monitor changes in correlation patterns across space and time dimensions. As a result, the Eigen space method for spatiotemporal hotspot discovery overcomes many of the shortcomings of Spatial Scan statistics. This method provides a shape-independent way for accurately detecting irregularly formed clusters.

The authors also compared Space-time scan statistics with the EigenSpot algorithm on a real dataset to discover the most brain cancer-affected locations in New York. In 2014, eigen vector [26] based hotspot detection method was proposed to detect hotspots using tensor decomposition and matching of the elements of eigen vector representation of spatiotemporal data. The results were demonstrated on the same brain cancer dataset and compared with ST-Scan results which showed more accurate results in lesser time complexity.

E. Machine Learning & Deep Learning based techniques:

The increasing popularity of machine learning in the field of computer science prompted researchers to explore and discover methods that might be applied to spatiotemporal data. U.M. Butt et al. [49] provided a comprehensive analysis of contemporary spatiotemporal hotspot detection approaches for detecting criminal behavior for which they identified a wide range of machine-learning and deep-learning-based methodologies. Random Forest classifier, Random Forest Regressor, Cluster Confidence Rate Boosting (CCRBoost), LDA-KNN, ARIMA, and Regression approaches (Ridge Regression, Lasso Regression, Support vector Regression) were reported by the authors as reliable machine learning models. It is noteworthy that these methods necessitate the conversion of spatiotemporal data points or events into a certain format according to the algorithm. Generally, Deep learning frameworks utilize a grid-based representation of spatiotemporal data.

These models can effectively cope with crime predictions and hotspot identification, according to the authors. Machine learning Models such as Spatio-temporal NN based on LSTM, Spatio-temporal-ResNet, and Spatiotemporal CNN consider both space and time dimensions for feeding the input data. The usage of geo-tagged photographs in conjunction with a neural network approach [47] to identify sentiment hotspot zones resulted in a completely new and unique application concerning spatial emotion detection. Researchers attempted to discover hotspot zones for mobile data points in addition to point data consisting of fixed locations in a specific time span. The focus of researchers has switched to predictive hotspot mapping of sparse spatiotemporal data over time. The first attempt to construct a graph-based deep learning model to predict spatiotemporal hotspots using network structure was given by Y.Zhang[34].

Y. Zhang [34] suggested an innovative deep learning approach to determine Crime Hotspots in South Chicago. The authors emphasized that most real-world spatiotemporal events reflect network structure rather of a grid-like structure. For example, when shown as a graphical network, urban taxi pickup points, urban crime, and road traffic accidents better convey the changes in space and time dimensions. The authors of the research created a gated localized diffusion network (GLDNet), a graph-based DL framework for generating hotspot mapping of spatio-temporal events in the network space. In GLDNet, a gated network models the timeline of historical events using the time dimension of the dataset, and a localized diffusion network captures the corresponding propagation of spatial events in terms of network distance and topology, in order to overcome spatial heterogeneity.

F. Probabilistic model based techniques:

In recent years, probabilistic estimate methods based on Bayesian models have gained popularity. Hotspot detection can also be classified as determining a probability estimate of occurrences that are more likely to occur inside than outside. For identification of spatiotemporal regions with increased likelihood of events, researchers presented both Empirical Bayesian and Full Bayesian approaches. These methods not only focus on identifying hotspot areas based on a historical record of spatiotemporal event occurrences over a specific period of time, but also on evaluating the core causes of changes in differential patterns and forecasting the likely future hotspot areas. These methods have been used to identify crash hotspots in the context of road safety.

Using the Bayesian Spatiotemporal Interaction model employing the full Bayesian technique, N. Dong et al. [41] proposed a unique strategy for improving road safety by identifying traffic crash hotspot spots in a designated zone. The authors stated with evidence that their proposed approach demonstrated better performance in terms of model fit than traditional full Bayesian techniques. While dealing with spatiotemporal hotspot identification, they essentially addressed two major challenges i.e., the evolution of detection hotspots in time and likelihood of future hotspot areas based on the current tendency to become hotspots.

G. Fuzzy logic based techniques:

The computer-based Boolean logic was upgraded to develop many-valued fuzzy logic models in order to include human-brain-like skills into intelligent systems. However, there isn’t much study literature in the fuzzy logic domain concerned with spatiotemporal hotspot detection, but several studies have experimented out recently to detect road accident hotspots. The traffic department has long been concerned about road safety. Researchers aimed to develop accurate strategies for identifying locations most affected by traffic accidents in order to assist policymakers in taking appropriate changes in specific crash zones. A. Alomari et al. [46] proposed a hotspot prediction model for accident-prone zones employing GIS spatial analysis and evaluation tools, as well as fuzzy modelling, in 2019. The authors explained that they used overlay analysis to keep data distribution uniform and suitable fuzzy models to find high-risk areas in the research area. The Weighted Overlay approach was used in conjunction with the Fuzzy membership’s prediction in order to prioritize the prediction probability values.

According to the authors, utilizing proper Fuzzy membership for each feature, their method was successful in detecting high-risk sites. In 2021, Y. Farjami and K. Abdi [36], presented a hybrid model of genetic algorithms and fuzzy logic to develop a fuzzy knowledge base. They demonstrated their approach on the crime dataset from Tehran, Iran for detecting high-risk hotspot zones over space and time. The authors suggested a threefold system involving the division of problem space using fuzzy parameters, selection of principal features, and developing the fuzzy knowledge base. The system was evaluated and turned out to be an effective tool, thereby, justifying its purpose.

H. Nature Inspired Algorithms:

In recent years, nature has inspired optimization and computationally intelligent algorithms, leading to the development of swarm intelligence and nature-inspired computing algorithms. The numerous hotspot detection approaches discussed in this study can be improved further to produce better findings and offer better predictability of spatiotemporal hotspots, directing the respective fields to better judgments. Moreover, M. Kulldorff's spatial scan statistics, which primarily used a scanning window of a preset form (circular, elliptical, cylindrical, etc.) and its various modifications, have been widely used in criminology and epidemiology.

In 2012, H. Izakian and W. Pedrycz [27] developed a flexible shape to be used as a modified scanning window in the Spatial Scan statistics algorithm for detecting arbitrarily shaped disease spatial and spatiotemporal clusters using the Particle Swarm Optimization approach. The authors adopted this approach to address the NP-complete problem of finding all possible irregular shaped clusters in a brute force manner. They showed the results over the Prostate and Liver datasets and the results proved to be highly stable in nature.

2.3 Analysis of the Framework for Spatiotemporal Hotspot Detection

The general execution stages for detecting spatiotemporal hotspots are shown in Fig. 3. The first step is to acquire spatiotemporal data. The second task is to determine what kind of hotspots are required, i.e., if crime events for a given region are recorded for a particular time period, then spatiotemporal crime hotspots predicted will assist us in determining which specific sub regions of the study area have a high risk of crime occurring and at what point in time. The third phase entails adopting the most appropriate technique for efficiently identifying significant spatiotemporal hotspots. The fourth task entails determining the accuracy of the obtained results using applicable evaluation criteria. Inferences can be drawn from the patterns discovered as a result of the data in the fifth step.

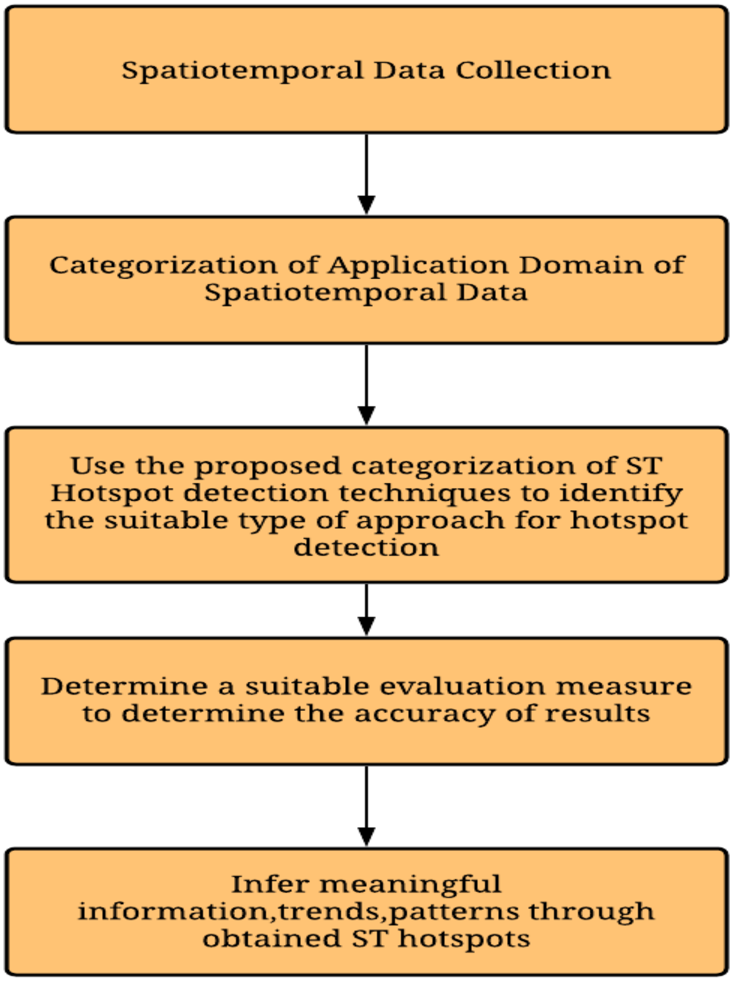

Spatiotemporal Data Collection
Categorization of Application Domain of Spatiotemporal Data
Use the proposed categorization of ST Hotspot detection techniques to identify the suitable type of approach for hotspot detection
Determine a suitable evaluation measure to determine the accuracy of results
Infer meaningful information,trends,patterns through obtained ST hotspots

Fig.3 Flowchart for Analysis of Spatiotemporal Hotspot Detection and Evaluation

Table I shows the numerous strategies for spatiotemporal hotspot detection have been presented for a wide range of applications. While researchers have used several factors to judge the accuracy of spatiotemporal hotspots, there is still a gap between the underlying ground truth and the acquired results. Table II presents a comprehensive survey of the popular evaluation metrics widely utilized by various research studies. Table III emphasizes on the detailed analysis of evaluation measures adopted by various research studies and their significance in distinct categories of ST hotspot detection methods.

Table II. Evaluation measures adopted by related Research Studies

Category	Application	Evaluation metric	Decision criteria	References
Space-time Scan Statistics based	Malaria Infection Hotspot Detection	Area under ROC (AUC)	Higher the AUC in a defined radius, better the outcome	[10](Year-2014)
	Breast cancer Hotspot Detection	Comparison with benchmark synthetic dataset	Higher the similarity with obtained results, better the outcome	[50](Year-2006)
Clustering based	Spatiotemporal Urban Hotspot Detection	Density-based clustering validation index	Higher the value of DBCV, better the outcome	[44](Year-2021)
	Spatiotemporal Urban Hotspot detection using Network based clustering	Precision value upon comparison with NSF and KDE methods	Higher the precision, better the outcome	[45](Year-2019)
	Spatiotemporal Crime Hotspot Detection	Predictive Accuracy Index Curve	Higher the PAI value, better the outcome	[31](Year-2018)
Machine Learning & Deep Learning based	Spatiotemporal Traffic Hotspot Detection	Case study based evaluation with Median Absolute Percentage Error	Lower the MDAPE, better the outcome	[51](Year-2020)
	Spatiotemporal Crime Hotspot Detection using Spatiotemporal Neural Networks	Accuracy, Precision, Recall, F-1 Score	Higher the test metric result, better the outcome	[35](Year-2017)
Probabilistic model based	High-risk Accident Hotspots	Mean Absolute Percentage error	Lower the test statistic, better the outcome	[41](Year-2016)
Eigenspace based	Disease outbreak Hotspot Detection	Comparison with ST-SCAN results	Higher the similarity with ST-SCAN results, better the outcome	[25](Year-2014)
Fuzzy Logic based	Traffic accident Hotspot Detection	Weighted Overlay method and Fuzzy Overlay method	Better the parameter tuning, better the outcome	[46](Year-2019)

2.4 Analysis of Evaluation measures for Hotspot Identification

Using the approaches described above, a valid assessment metric is required to analyze the performance of the detected hotspot. In the retrospective approach to Spatiotemporal hotspot detection, data from the past is collected and analyzed, allowing the results to be compared to existing underlying ground truth or validated using case studies. In the case of a prospective approach to spatiotemporal hotspot detection, however, it is critical to analyze the performance of the found hotspots, which can then aid policymakers from the relevant authorities in applying preventive measures or executing a better future management plan. Our research focuses on identifying evaluation measures that can be used to locate Spatiotemporal hotspots.

Table II lists the evaluation measures and decision criteria used in different application fields. Effective evaluation of Spatiotemporal hotspots is a difficult task that necessitates proper tuning of the model parameters. Diverse evaluation measures appear to function well for multiple areas of applications and methodologies, according to research investigations. Here, we specified the particular domain of concerned application along with the evaluation measures adopted by researchers. We briefly summarized the most common application kinds in each area, as well as the most often used assessment measure for each.

Table III. Significance of Evaluation Metrics adopted by related Research Studies

Evaluation metric	Formula	Significance
Area under ROC (AUC)	$\int Area\ under\ ROC$	AUC is a reliable evaluation metric especially for Space-time scan statistics-based hotspot detection methods and clustering-based hotspot detection methods. Eigenspace based hotspot detection methods cannot be evaluated appropriately using the AUC metric.
Comparison with benchmark synthetic dataset	Use Similarity or Dissimilarity measures for comparison	For evaluation purpose, a synthetic dataset is constructed to create randomly distributed events using Poisson model and relative risk is calculated. All categories of ST hotspot detection methods can use this method for evaluation.

Predictive Accuracy Index Curve	$\frac{HitRate}{Percentage\ Area\ of\ Hotspots}$	Reliable evaluation metric to compare accuracy values among varying hotspot areas. Useful for trend analysis across different regions of the study area
Case study based evaluation	Comparison of predicted results with actual results obtained from case studies	From past records containing actual results and detect mean or median absolute percentage error to determine the correctness of results obtained using any category of ST hotspot detection methods
Accuracy	$\frac{Total\ no.\ of\ correctly\ predicted\ activity\ points}{Totalno.\ of\ predicted\ activity\ points}$	Using either case study based evaluation or any valid record, this metric can evaluate accuracy of predicted hotspots
Precision	$\frac{True\ positives}{True\ positives\ +\ False\ positives}$	Using either case study based evaluation or any record of actual results this metric can be used to evaluate the precision of predicted hotspots
Recall	$\frac{True\ positives}{True\ positives\ +\ False\ negatives}$	Using either case study based evaluation or any record of actual results this metric can be used to detect the most significant hotspots identified among all candidate hotspots
F-1 Score	$2.\frac{Precision.Recall}{Precision + Recall}$	According to the type of ST hotspot detection method, F-1 score is a normalized metric that takes both precision and recall effects into account
Sensitivity	$1 - (False\ Negative\ Rate)$	For determining the power of obtained results after hotspot detection, sensitivity serves as a popular metric in all methods of hotspot detection especially in case of Space-time scan statistics based methods
Mean Absolute Percentage Error	$\frac{\sum \frac{Actual\ activity\ point - Predicted\ activity\ point}{Actual\ activity\ point}, \forall\ activity\ points}{Total\ No.\ of\ activity\ points}$	Used in comparative evaluation of actual and predicted results in any category of ST hotspot detection methods

We also evaluated and analyzed an average value of different evaluation measures. Fig. 4 shows visual analysis of ST hotspot detection methods with Area under ROC. Using the AUC metric as the evaluation criteria of Space-time scan statistics based methods, the score of 0.62 ,0.90 and 0.65 was obtained by [10], [52] and [53] respectively. Based on the analysis from several other research works falling under the category of Space-time scan statistics based hotspot detection methods, we can conclude that the AUC value lies approximately in the range of 0.6 to 0.9. For the Machine learning based methods [54] which are used to identify the spatiotemporal hotspots from crime dataset, it has an average AUC score of 0.80. Similarly, clustering based methods [55] have an AUC score of 0.75. It is important to note that the authors of [25] mentioned that AUC may seem appropriate for the evaluation of eigen space based methods but p-value is a better choice for such evaluation.

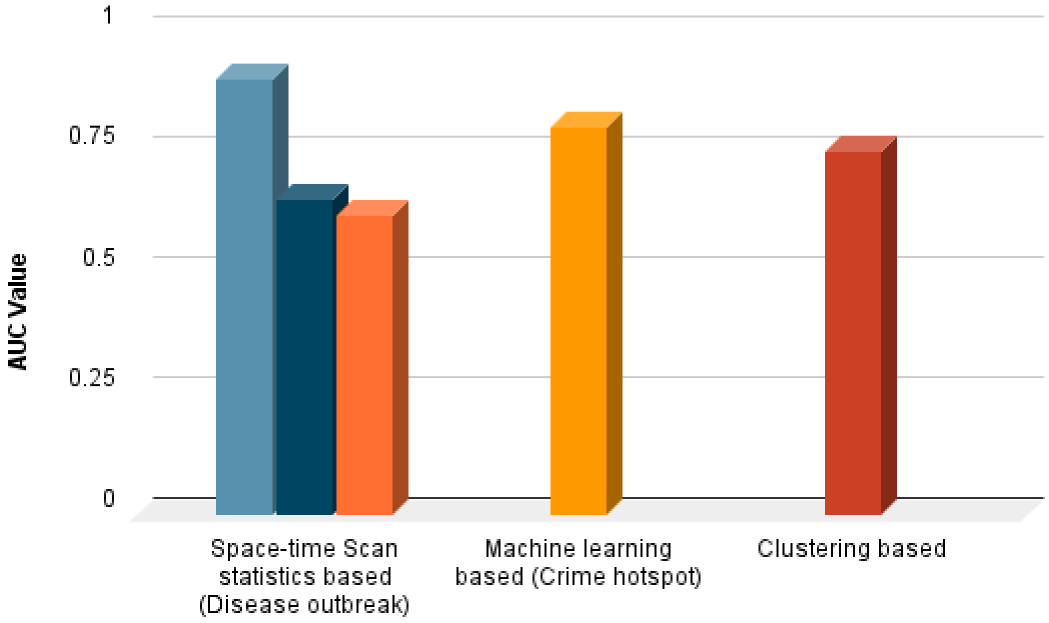


Fig. 4 Analysis of Spatiotemporal Hotspot detection methods with Area under ROC

In case of Case study based evaluation, either mean absolute percentage error or median absolute percentage error is calculated for calculating the relative risk between the actual and predicted hotspots. For finding the similarity between actual and predicted hotspots, several similarity or dissimilarity measures such as cosine similarity can also be used after converting the data into required format. In 2008, a new evaluation metric namely Predictive Accuracy Index Curve (PAI) was introduced which became popular in Machine learning based [56] and Clustering based approaches [31], [57], [58] for

Spatiotemporal hotspot detection. Fig. 8 demonstrates the results for PAI in detecting spatial hotspots. Fig. 5 demonstrates the average precision value obtained from different methods of ST hotspot detection. Machine learning based methods [35] obtained a precision value of 0.80. Space-time scan statistics based methods used for detecting traffic crash hotspots [55] nearly have a precision score of 0.75. As mentioned by the authors of [45], the precision value of clustering based methods varies for different clustering techniques.

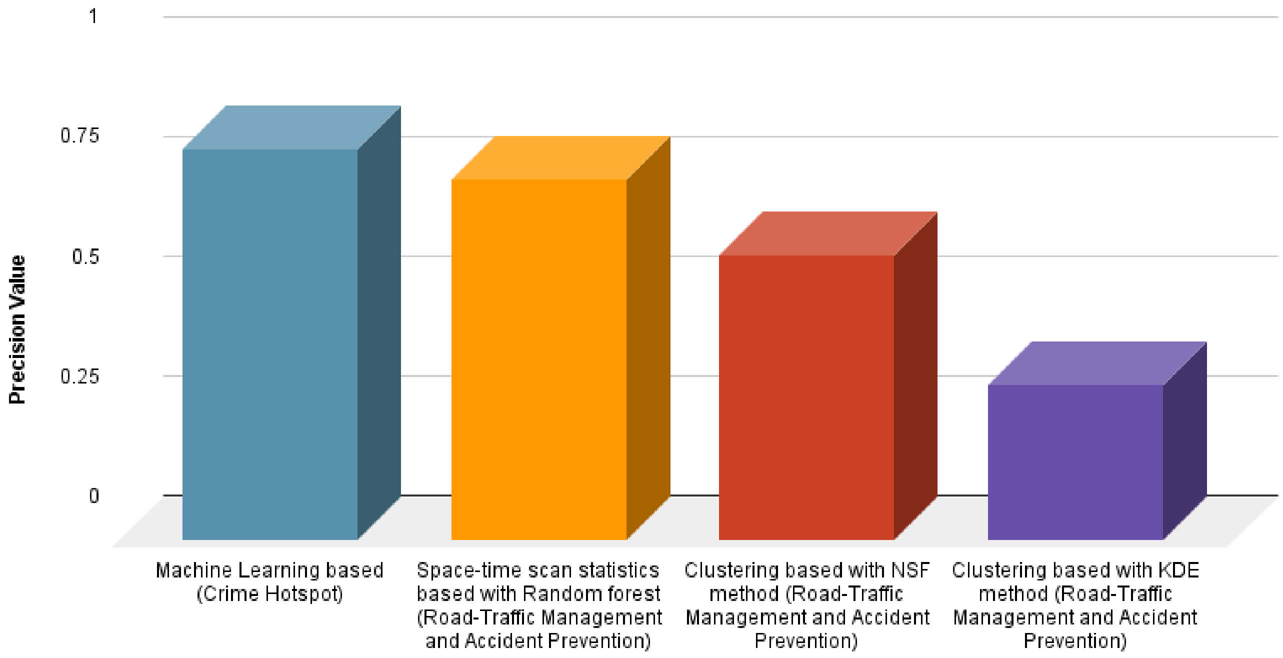


Fig. 5 Analysis of Spatiotemporal Hotspot detection methods with Precision

The authors mentioned that using the Network based Spatiotemporal Field Clustering approach (NSF), the precision value is higher when compared to the precision value obtained by the Kernel Density estimation method (KDE). Fig. 6 demonstrates the usage of Recall [35], [55] as an evaluation metric and Fig.7 describes the combined effect of both precision and recall metrics using the F1-Score evaluation criterion [59], [35], [55]. This evaluation measure has helped in accurately detecting several spatiotemporal hotspots in different application areas within a score range of 0.65 to 0.85. Hotspots are high-risk zones in a study area in varying application fields. In case of disease outbreak or similar applications, sensitivity analysis [5], [11] is crucial to better analyze the utility of predicted spatiotemporal hotspots.

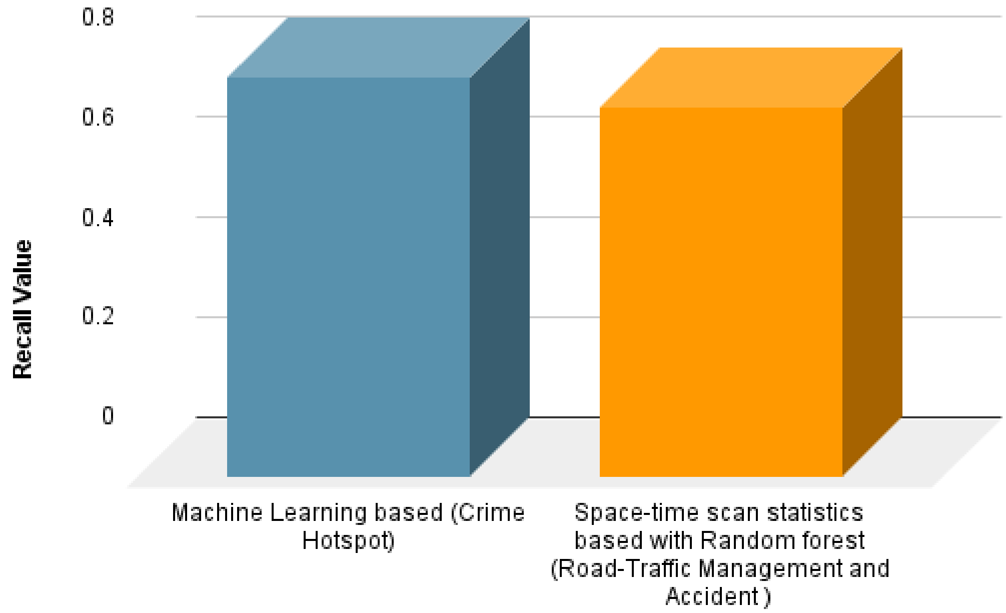


Fig. 6 Analysis of Spatiotemporal Hotspot detection methods with Recall

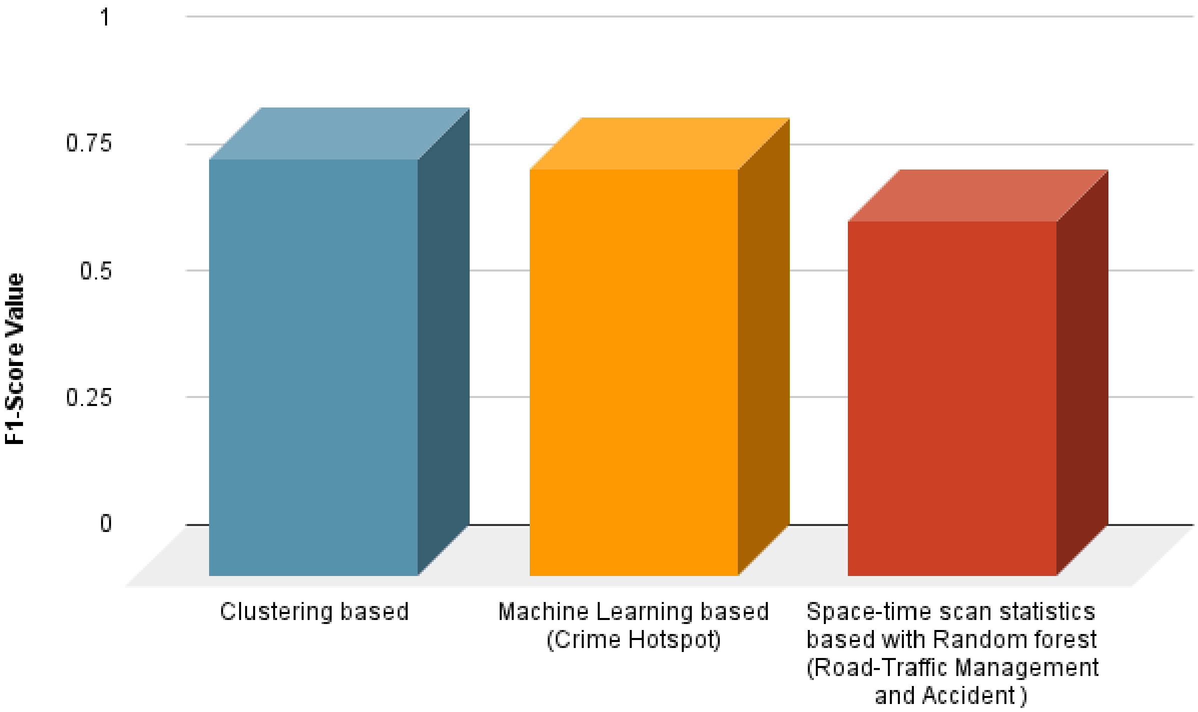


Fig. 7 Analysis of Spatiotemporal Hotspot detection methods with F1-Score

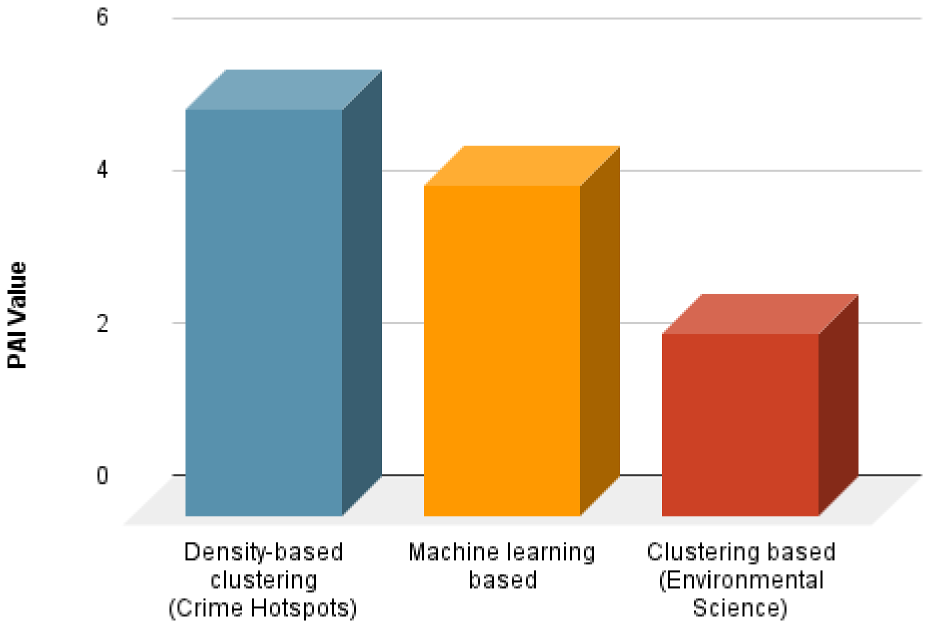


Fig. 8 Analysis of Spatiotemporal Hotspot detection methods with PAI

2.5 Analysis of Machine Learning Approaches in Spatiotemporal Hotspot detection

Our research provides a complete overview of recently suggested Machine Learning algorithms for detecting Spatiotemporal hotspots in road traffic accidents. The survey examines the various algorithms for Spatiotemporal hotspot detection that are based on Machine Learning, Deep Learning and Neural Networks technology. However, understanding the significance of such Artificial Intelligence sub domains for the function of Spatiotemporal Hotspot detection becomes vital. Traditionally, statistics-based methods were used to locate hotspots of any event or activity (crime, disease transmission, epidemic, road accident, etc.). Spatial Scan Statistics has been widely used in the fields of criminology and epidemiology over the years. It has shown to be effective and has developed a standard mechanism for Spatial Hotspot Detection. However, statistics-based hotspot detection algorithms have limitations due to assumptions made, such as having a predetermined input shape parameter and ignoring all other significant dimensions of spatial data apart from the intensity of event recurrence. Consequently, the focus of researchers shifted towards the developing technology such as Machine Learning. D. Santos et al. [60] proposed a review and investigation of numerous Machine Learning models used in the prediction of hotspot areas for road traffic accidents. Data

Table IV. Analysis of Machine Learning models adopted in traffic accident prediction

Approach used	Significance in Traffic accident analysis	Reference
Supervised Machine Learning techniques (Decision Trees, Random Forests, Logistic Regression)	Used for principal feature selection for traffic accident analysis	[60]
Unsupervised Machine Learning techniques (DBSCAN and Hierarchical Clustering)	Used for forecasting future road accident sites based on selected features	[60]
Convolutional Neural Network (CNN)	Traffic Abnormality or Outlier Detection that occurred due to spatial and temporal variations & Prediction of Traffic accidents based on Traffic flow	[62] [69]
Deep Belief Networks (DBN) and Long Short-term Memory model (LSTM)	Pattern Identification and Recognition of tweet data on road traffic accidents	[62]
Convolutional Long Short-term Memory Model (ConvLSTM)	Spatiotemporal Traffic accident location forecasting considering other important influential factors of the event as well	[62]
Hetero-ConvLSTM	Modified ConvLSTM algorithm with added advantage of handling different spatial and temporal variability	[63]
Recurrent Neural Networks (RNN)	Prediction of accident severity and other factors important for forecasting hotspot locations	[65]
Spatiotemporal Convolutional Long Short-Term Memory model (SCLSTM)	Spatiotemporal traffic accident prediction and hotspot mapping	[67]
Kernel Density Estimation (KDE)	Detection of Spatial Hotspots using a combined approach involving Statistics based as well as Machine Learning based techniques	[71]
Spatiotemporal Network Kernel Density estimation (STNKDE)	Spatiotemporal identification of hotspots considering the time dynamics of the identified hotspot locations and the Spatial representation of hotspots in network space rather than linear space	[72]

pre-processing, clustering, application of Machine Learning models and prediction of hotspots are the four stages involved in hotspot identification, according to the researchers. The authors of [60] followed a two-way approach, with the first method attempting to analyze the severity of accidents given a set of other influential factors such as weather and road conditions at the time of incident, and the second method attempting to forecast future accident hotspot locations using the Random Forests Model. The research revealed that crucial external elements such as road signs and adequate road construction have an impact on a substantial portion of the population.

According to the study, the majority of the accident hotspots discovered after clustering are located on rural or urban roads with no dividers This information extracted using the effective Machine Learning models can assist the responsible authorities in taking immediate action at the identified hotspot location. Furthermore, we can comprehend the significance of Machine Learning models for Spatial Hotspot detection over Statistics-based methods because they can take advantage of external contributing factors and do not require the determination of a predefined input shape parameter, as well as produce better results. The researchers explored with Deep Learning Methods as well, due to the significant results gained using the Machine Learning approach. S. Wang et al. [62] presented the most recent survey of Deep Learning techniques in Spatiotemporal Data Mining. The authors covered all different kinds of Spatiotemporal data, their applications, the various Deep Learning task domains, and the popular Deep Learning models employed in the Traffic accident data, according to the survey's authors, fits into the category of event spatiotemporal data.

Event data is simply spatial information of a given event type that occurred at a specific moment of time in the past. They investigated a number of deep learning models that excel at handling event based data. The authors proposed the ConvLSTM Deep Learning model as the most extensively used method for predicting road traffic accidents. It formulates the traffic data's spatial and temporal features before converting it to a 3D tensor data representation format.

The ConvLSTM model uses this 3D tensor data to predict future probable spatiotemporal areas with a greater likelihood of traffic accident count. Z. Yuan et al. [63] proposed the Hetero-ConvLSTM model when the Spatiotemporal data on traffic accidents is heterogeneous in nature, i.e., the data also provides information of the spatial variations of the surrounding region where the accident occurred, i.e., rural or urban neighborhood, and this information is also used in the prediction of future traffic accident areas in given time. To learn the feature representations utilized for estimating the risk level of a traffic collision, a stack denoise Autoencoder model was also suggested by the authors of [63] in their survey.

Moving forward, we examined and focused on the Deep Learning based Traffic Accident Analysis to better comprehend the various factors involved in the road accident event and their impact over time, after going through the Machine Learning based techniques for Traffic accident prediction and the Survey on Deep Learning Models used in Spatiotemporal data mining for different application domains. The essential phases involved in road traffic accident analysis were addressed by A. Naseer et al. [65]. For the prediction of hotspot sites, a recurrent neural network (RNN) approach is used. Data Preprocessing and Data Transformation are critical for the Deep Learning model's performance, according to the authors. They illustrated the major steps involved in identifying the key hotspot areas in the research area, starting with the input data format and ending with the deep learning model output.

J. Bao et al. [67] suggested an innovative spatiotemporal technique based on Deep Learning technology to estimate accident risk zones within a city in 2019. The proposed method takes into account spatial and temporal variables, as well as information about the surrounding region, such as weather, road network data, and land information. The study's authors presented a new Spatiotemporal Convolutional Long Short-Term Memory model (SCLSTM) for forecasting traffic accident hotspot sites in New York City. The authors compared the suggested approach to current state-of-the-art Machine Learning approaches and discovered that the new approach outscore the others in terms of accuracy prediction rate. Deep Learning methods like SCLSTM have the extra benefit of monitoring larger input sizes with better accuracy as data size grows and input sizes fluctuate in different time windows (hourly, daily, weekly, monthly).

A Convolutional Neural Networks-based technique for traffic accident prediction was proposed by H. Zhao et al. [69]. The authors emphasized that these algorithms do not automatically extract features from traffic data, which is a significant improvement over typical Machine learning-based methods. As a result, they proposed using a CNN-based Deep learning approach to extract features from traffic accident data. This aids in the identification of crucial dimensions in the real-time prediction of the likelihood of a traffic accident. They also stressed the applicability of the proposed approach in terms of both real-time accuracy and the potential to deliver a safer driving experience by raising alarms when the probability of accident risk increases.

C. Zhang et al. [74] provided a traffic accident risk analysis approach in a novel structural style for only urban traffic zones in China. They analyzed the traffic network as a graph representation, in which each road crossing is considered a node of the graph and the sections of roads are assumed as the edges, and the edges are weighted in such a way that each section of road accounts for the total number of accidents that have occurred in that section of the road in the past. They thoroughly examined fuzzy-based strategies for accident prediction that have been discussed in the literature. B. Romano et al., [72], addressed a critical issue with the Spatial Hotspot detection approach. They investigated the significance of identifying the primary hotspot areas of traffic accidents in both the spatial and temporal domains. The authors of the research study [72] stated that typical hotspot detection techniques such as Spatial Scan Statistics do not take into account the constraint of road networks while detecting hotspots. However, it is obvious that road network limits are practical and cannot be neglected to the point where the scope of the hotspot detection system becomes irrelevant in real-world circumstances. As a result, the authors proposed a novel Spatiotemporal Network Kernel Density estimation approach that combines two important aspects of road accident hotspots, namely the time dynamics of the identified hotspot locations and the Spatial detection of hotspots in network space rather than linear space. In 2017, the proposed approach was tested on New York City data and outperformed other similar approaches.

The vast majority of research on road traffic accident hotspots has proved the results produced from proposed methodologies in either a simulation-based or lab-controlled context. However, B. Ryder et al. [75] in 2017 revealed the outcomes of traffic accident analysis over actual traffic location analytics to identify traffic accident hotspots and create in-vehicle warnings for potential road risks in the way. They proposed establishing an in-vehicle warning system for risky scenarios in a real-world setting using cloud-based infrastructure and computer vision-based techniques such as object identification and recognition. D. Al-Dogom et al. [71] in 2019 presented a three-step strategy for detecting hotspot sites of road traffic accidents in the United Kingdom region . The method included spatial data collection and representation in GIS format, cluster identification using Getis Ord GI* statistics, and cluster density estimate using the Machine Learning-based Kernel Density Estimation function. In addition, the XGBoost and Extra Trees Classifier techniques are utilized to generate hotspot maps by taking into account selected attributes which prove beneficial in classification of a cluster as a hotspot location. M. Zahid et al. [73] analyzed the importance of spatial and temporal attributes in determining the behavior of taxi drivers on the road. This research study presented with a unique way of analyzing hotspot maps in a manner where a hotspot is classified as the zonal region with increasing traffic violations due to the unperfected behavior of the drivers.

CHAPTER 3: METHODOLOGY

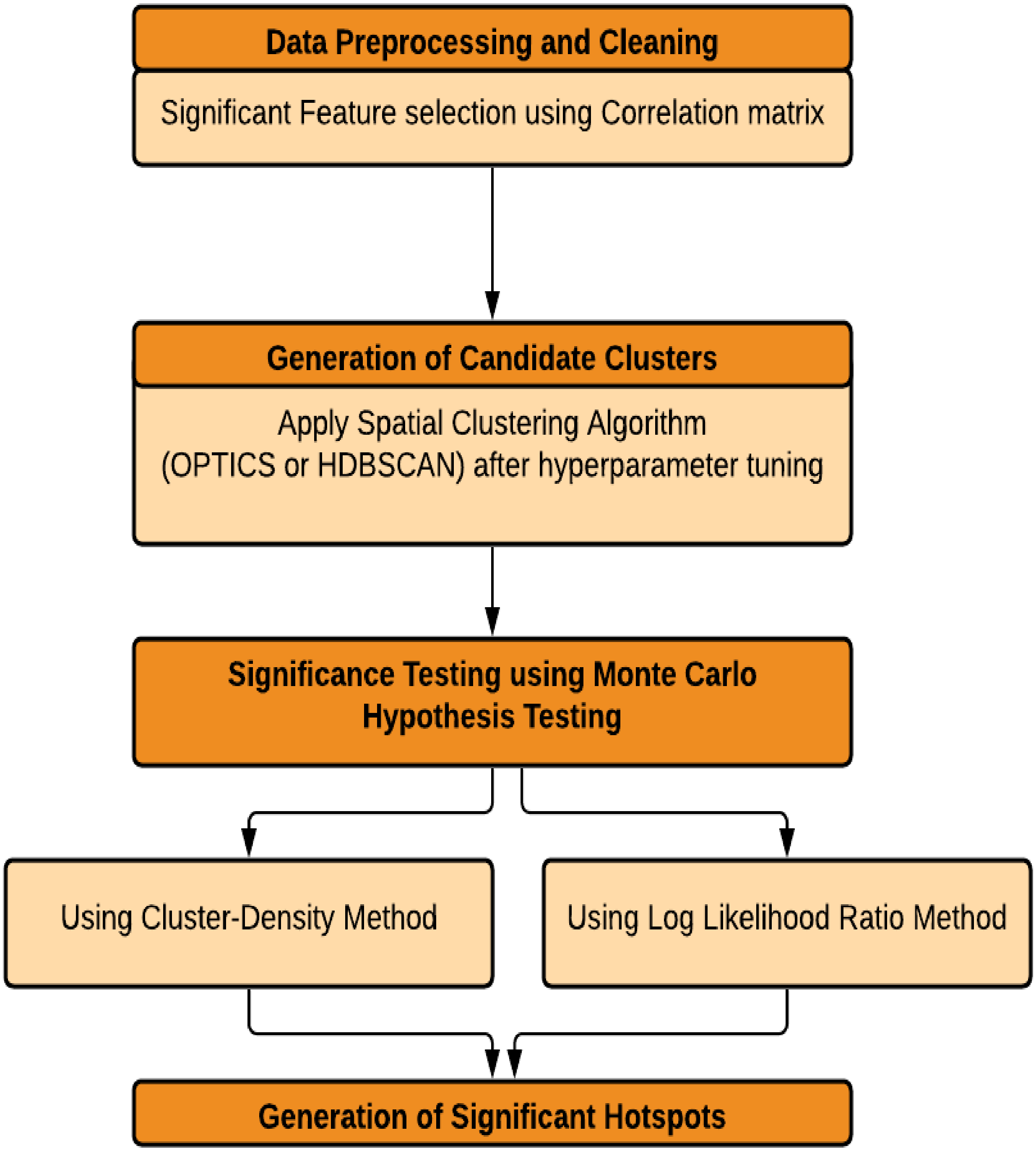

Data Preprocessing and Cleaning
Significant Feature selection using Correlation matrix
Generation of Candidate Clusters
Apply Spatial Clustering Algorithm
(OPTICS or HDBSCAN) after hyperparameter tuning
Significance Testing using Monte Carlo Hypothesis Testing
Using Cluster-Density Method
Using Log Likelihood Ratio Method
Generation of Significant Hotspots

Fig. 9 Overview of Proposed Algorithm

Fig. 9 depicts a high-level overview of the proposed approach. After preprocessing of data and selection of relevant features, the formation of candidate clusters is an important aspect of the proposed approach. Candidate clusters are collections of spatial points with the potential to become significant spatial hotspots. The research employs two effective Spatial Clustering algorithms, namely, OPTICS and HDBSCAN. Our research not only described how to locate candidate clusters, but also how to use a significance testing process to screen the candidate clusters and demonstrate significant hotspots. The proposed methodology is to find statistically significant spatial hotspots so that concerned authorities can make informed decisions.

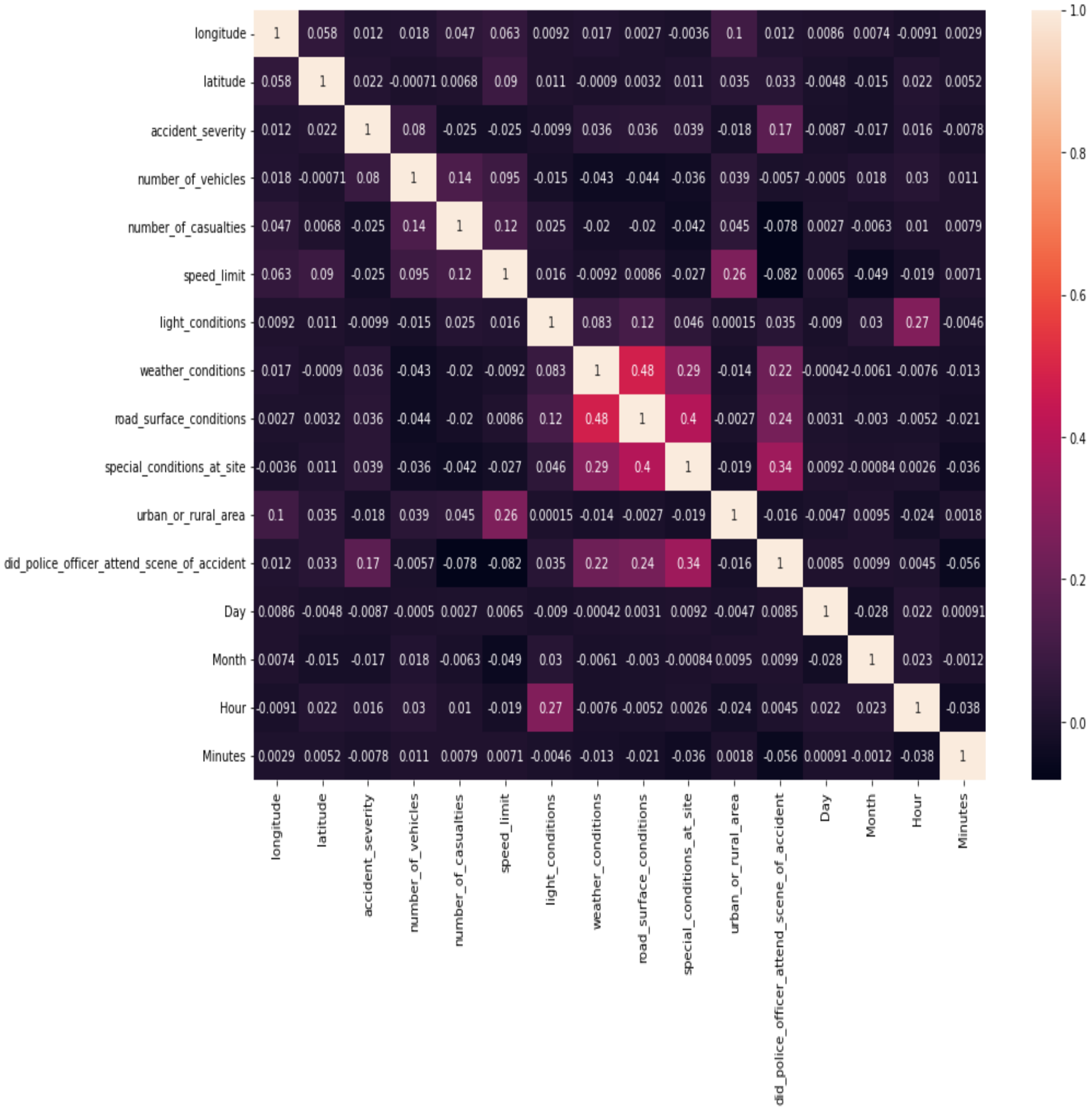


Fig. 10 Correlation matrix for Feature Selection

3.1 Description of Dataset and Data Preprocessing

We demonstrated an implementation of the proposed algorithm using the "Road Safety Data from the Department of Transport, UK" dataset [61] for the year 2020. The UK government has offered public data on road accidents from past several years. The data set extensively describes road accidents by highlighting key aspects of the incident. Latitude and Longitude coordinates, which determine the actual accident location in the region, are the primary spatial parameters utilized to identify hotspots. The dataset's temporal properties include the date and time of the accident. The dataset also includes significant non-spatiotemporal attributes such as accident severity, the concerned region's Districts or Counties, the weather and road surface conditions, and the number of casualties. These features help in conducting a comprehensive analysis incorporating various aspects of the data. The analysis is being carried out for the Metropolitan Area of the United Kingdom.

The Metropolitan Area includes the majority parts of the United Kingdom. The dataset comprises of a total of 20906 records of road accidents in the year 2020. Some of the important regions of the Metropolitan area include:

- Greater London
- North West
- North East
- South West
- South East
- West Midlands
- East Midlands
- East of England

The results are analyzed in two different domains i.e., Seasonal and Temporal. After splitting off the Month field, Seasonal Analysis is performed using the Date component. The Time parameter from the dataset is used for Temporal Analysis. Spatial Clustering is a subclass of the unsupervised machine learning category. Hence, the paper selected correlation matrix for identification of relevant features for clustering. Aside from the dataset's spatial and temporal properties, the data cleaning process includes removing extraneous features.

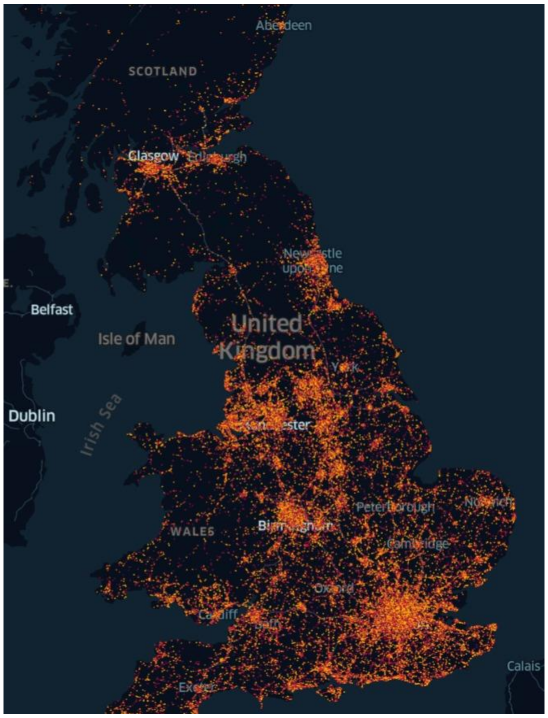


Fig. 11 Visual Representation of Preprocessed Dataset

The correlation matrix shown in Fig.10 was utilized to select important dataset attributes. From the correlation matrix, it can be deduced that the selected attributes have low correlation values with each other. Hence, these features aid in producing better clustering results. A threefold approach has been employed for data preprocessing and cleaning. The first step entails eliminating all records with null value from the entire dataset. The second step entails selecting certain attributes (e.g., Date and Time) and formatting them to relevant data type. The final step is to identify significant features using the Correlation matrix. Fig. 11 demonstrates the data points of the preprocessed dataset where each activity point denotes the spatial location of a road accident in the United Kingdom for the year 2020.

3.2 Generation of Candidate Clusters

Clustering techniques are grouped into various categories such as Partition-based Clustering, Hierarchical Clustering, and Density-based Clustering. The Partition-based Clustering approaches such as K-Means and K-Medoids implement an iterative approach to generate predefined number of clusters based on a particular distance metric until the algorithm reaches a convergence state. In general, the linearly defined distance functions are frequently used in partition-based clustering. Spatial data points, on the other hand, are quite different from the usual data points. Due to the irregularly shaped ellipsoidical surface of the earth, the Latitude and Longitude points possess non-linear characteristics. As a result, linear distance functions are no longer a viable option. Consequently, non-linear geodetic distance functions are required to overcome the abovementioned shortcomings. The Haversine formula determines the non-linear distance between two points on a spherical surface given their longitudes and latitudes. Furthermore, in the case of Spatial Hotspot detection, it is difficult to determine the number of clusters (K) in advance. Although experimentation with the K-Means and K-Medoids algorithms may yield adequate results, the fundamental implementation of these algorithms must be modified to account for spatial features. However, the convergence of these algorithms after modification cannot be guaranteed. Hence, we shifted our focus towards Density-based clustering and Hierarchical clustering techniques.

Density-based Spatial Clustering of Applications with Noise (DBSCAN) is among the popular algorithms used for Spatial Clustering. However, it requires proper estimation of two important parameters, Eps (neighborhood radius around data point) and minpts (min. number of points within Eps). In case of large spatial dataset, it becomes challenging to determine appropriate values of these parameters to yield accurate results.

Ordering Points to Identify the Clustering Structure (OPTICS)[66] is an extension of DBSCAN that overcame the shortcomings of DBSCAN. Unlike DBSCAN, OPTICS only requires minimum number of samples for clustering and works with a wider range of parameters automatically. Hence, the proposed algorithm adopted OPTICS for demonstrating the capability of density-based clustering algorithm. Besides, Hierarchical Density-based Spatial Clustering for Applications with Noise (HDBSCAN) [76] is a hierarchical clustering algorithm which is another extension of DBSCAN algorithm which also requires minimum cluster size as the only parameter.

Both OPTICS and HDBSCAN algorithms are compatible with haversine distance metric. Hence, these algorithms are adopted for the expansion of candidate spatial clusters which is the first step of detection of significant spatial clusters.

ALGORITHM 1: HOTSPOT DETECTION WITH OPTIMAL SPATIAL CLUSTERING ALGORITHMS AND SIGNIFICANCE TESTING USING CLUSTER-DENSITY METHOD

```
    Input: OPTICS Algorithm, HDBSCAN Algorithm, Latitude & Longitude of activity points in radians
    Output: List of Significant Hotspots
    Initialization:
 1  Specify the CLUSTERING_ALGO = OPTICS or HDBSCAN
 2  Algorithm convergence criteria: (No. of candidate clusters = No. of significant clusters) or
    (Count and cluster_label of significant clusters is not changing with change in min_samples) or
    (Input the threshold value of min_samples based on domain knowledge of dataset)
 3  min_samples = 6                          // Specifies min. activity points required to form a cluster
    Procedure for Generating candidate clusters:
 4  while Algorithm convergence criteria is not met do
 5      min_samples_list.append(min_samples)
 6      for min_sample_val in min_samples_list do
 7          clusterer = CLUSTERING_ALGO (min_sample_val, metric = ‘haversine’).fit(Latitude,Longitude)
 8          Cluster_Labels = clusterer.labels_
 9          GENERATE_CANDIDATE_CLUSTERS(Cluster_Labels):
10          for label in Cluster_Labels do
11              if (label ≠ -1) then
12                  Get (Latitude, Longitude) for label
13                  Candidate_Cluster_List.append(Latitude, Longitude)
14              end if
15          end for
16          Calculate cluster size of each candidate cluster
17      end for
18      min_samples = min_samples + 1
        Procedure for Significance testing of Candidate clusters:
19      Using uniform random distribution, generate a set of random (Latitude, Longitude) points having same
        count of activity points as that of original dataset in the same study area
20      Repeat steps 6 to 17 for randomized dataset to obtain list of cluster size for randomized data points
21      Max_Cluster_Size = max(cluster size obtained from Randomization)
22      foriter in Monte_Carlo_Simulations do                    // Monte_Carlo_Simulations = 100
23          Repeat steps 19 to 21 to obtain max cluster size in each iteration and store it in
            max_random_cluster_size_list
24      end for
25      Sort(max_random_cluster_size_list) in descending order of values
26      MONTE_CARLO_SIMULATION_TESTING:   // Significance testing with max Cluster Size
27      Initialize p_value = 0 and alpha_value = 0.01
28      for iter in range(len(Candidate_Cluster_List) do
29          pos = position of Size(candidate_cluster[iter]) in sorted max_random_cluster_size_list
30          p_value = pos / (Monte_Carlo_Simulations + 1)
31          if (p_value <= alpha_value) then
32              Significant_Cluster_List.append(Candidate_Cluster[iter])
33          end if
34      end for
35  end while
36  return Significant_Cluster_List
```

ALGORITHM 2: HOTSPOT DETECTION WITH OPTIMAL SPATIAL CLUSTERING ALGORITHMS AND SIGNIFICANCE TESTING USING LLR METHOD

```
    Input: OPTICS Algorithm, HDBSCAN Algorithm, Latitude & Longitude of activity points in radians
    Output: List of Significant Hotspots
    Initialization:
 1  Specify the CLUSTERING_ALGO = OPTICS or HDBSCAN
 2  A = Number of activity points in study area
 3  B = (||A|| . area(candidate_cluster)) / area(study_region) which is Expected frequency of activities in candidate
    cluster
 4  c = Observed frequency of activities in candidate cluster
 5  I() = 1 , if c < B and 0, otherwise
 6  Algorithm convergence criteria: (No. of candidate clusters = No. of significant clusters) or
    (Count and cluster_label of significant clusters is not changing with change in min_samples) or
    (Input the threshold value of min_samples based on domain knowledge of dataset)
 7  min_samples = 6                              // Specifies min. activity points required to form a cluster
    Procedure for Generating candidate clusters:
 8  while Algorithm convergence criteria is not met do
 9      min_samples_list.append(min_samples)
10      formin_sample_val in min_samples_list do
11          clusterer = CLUSTERING_ALGO (min_sample_val, metric = ‘haversine’).fit(Latitude,Longitude)
12          Cluster_Labels = clusterer.labels_     // A cluster label is assigned to each activity point (AP) to
            specify the generated candidate cluster to which AP belongs
13          GENERATE_CANDIDATE_CLUSTERS(Cluster_Labels):
14          for label in Cluster_Labels do
15              if (label ≠ -1) then
16                  Get (Latitude, Longitude) for label
17                  Candidate_Cluster_List.append(Latitude, Longitude)
18              end if
19          end for
20          Convert all (Latitude, Longitude) activity points into UTM coordinate system.
21          Apply Convex Hull algorithm on the (Latitude, Longitude) activity points of a candidate cluster to
            obtain an arbitrary shaped polygon
22          Calculate area of obtained polygon for all generated candidate clusters
23          Area_list.append(area of obtained polygon)
24          Log LR of candidate cluster = $\text{Log}\left(\left(\frac{c}{B}\right)^{c}\left(\frac{\|A\|-c}{\|A\|-B}\right)^{(\|A\|-c)}.I()\right)$
25          Calculate value of log likelihood ratio test statistic for all candidate clusters with LLR formula
            above
26          LLR_list.append(Log LR of candidate clusters)
27      end for
28      min_samples = min_samples + 1
        Procedure for Significance testing of Candidate clusters:
29      Using uniform random distribution, generate a set of random (Latitude, Longitude) points having same
        count of activity points as that of original dataset in the same study area
30      Repeat steps 10 to 27 to obtain list of Log LR of candidate cluster  for randomized data points
31      Max_Log_LR = max(Log_LR values obtained from Randomization)
32      for iter in Monte_Carlo_Simulations do                    //Monte_Carlo_Simulations = 100
```

```
33          Repeat steps 29 to 31 to obtain max Log LR in each iteration and store it in
            max_random_cluster_size_list
34      end for
35      Sort(max_random_Log_LR_list) in descending order of values
36      MONTE_CARLO_SIMULATION_TESTING:   // Significance testing with max Log LR
37      Initialize p_value = 0 and alpha_value = 0.01
38      for iter in range(len(Candidate_Cluster_List) do
39          pos = position of Log_LR(candidate_cluster[iter]) in sorted max_random_Log_LR_list
40          p_value = pos / (Monte_Carlo_Simulations + 1)
41          if (p_value <= alpha_value) then
42              Significant_Cluster_List.append(Candidate_Cluster[iter])
43          end if
44      end for
45  end while
46  return Significant_Cluster_List
```

3.3 Significance Testing using Monte Carlo Hypothesis Testing

Monte Carlo Hypothesis Testing is a statistical technique used to compute the risk factor involved in the predictions. According to Spatial Scan Statistics, Monte Carlo Hypothesis Testing is standard method for significance testing of hotspots. The proposed algorithm implemented modified Monte Carlo Hypothesis Testing with the following decision criteria:

1) *Cluster-Density Method*: The amount of data points inside a candidate cluster is used in this method of significance testing. As stated in Algorithm 1 of the paper, the density of data points within a cluster is calculated. Following that, the density of the possible clusters in the immediate vicinity is determined. The more significant a candidate cluster region is, the denser it is. The density, on the other hand, is determined not only by the number of data points, but also by the cluster with the most data points among all the candidate clusters in each iteration. This test statistic is suitable for clusters of any shape.

2) *Log Likelihood Ratio Method*: This test statistic can be used in a number of different spatial domains. This type of significance testing is used in the Spatial Scan Statistic to assess the significance of clusters with a predefined shape. However, significance testing was carried out in this study utilising an improvised method. The process of calculating this test statistic was detailed in Algorithm 2 of the study. The greater the cluster's loglikelihood ratio, the more significant it is. As a result, in each iteration, the maximum likelihood of the candidate clusters is computed.

The above mentioned methods for significance testing of candidate clusters are chosen since they have been proven efficient for our application domain. The authors of [79] mentioned in detail about these test statistics. The Likelihood ratio test statistic is less biased towards smaller sized clusters but required area of the candidate cluster for significance testing. However, cluster size test statistic is independent of the area computations, demonstrate lesser time complexity as well as it doesn't incorporate any favor towards smaller sized clusters.

Hence, the efficiency of these two test statistics towards spatial clusters obtained from clustering algorithms such as OPTICS and HDBSCAN is noteworthy in regards to the combined strengths of the significance testing methods and the spatial clustering algorithms for candidate clusters detection. The proposed algorithm of this paper is depicted in Algorithm 1 and Algorithm 2 above. The output obtained from the proposed algorithm is the coordinate points in the form of latitudes and longitudes of the final significant hotspots. We also present a comprehensive analysis of the significant hotspots in the next section.

CHAPTER 4: RESULTS & ANALYSIS

The suggested methodology attempts to solve the problem of discovering Significant Spatial Hotspots of Road Accidents in the United Kingdom. Our research focuses on not only identifying spatial hotspots, but also analyzing spatial hotspots of road accidents in relation to a variety of other external parameters. The primary objective of this study is to guide the government and concerned authorities to take appropriate actions in the required area in order to make policy changes and execute necessary adjustments. Authorities can use the collected data to provide resources to relevant regions for infrastructure, traffic-light systems, street lights, and road improvements, among other things. In order to retrieve deeper facts from the data, we conducted appropriate experiments and subjectively evaluated the results. For significance testing, the Proposed Algorithms 1 and 2 use the Cluster-Density approach and the Log Likelihood Ratio method, respectively. The results of using the OPTICS Spatial Clustering algorithm with both test statistic measures are shown in Table V. Table VI illustrates the results of using the HDBSCAN Spatial Clustering algorithm with both test statistic measures in a similar way. Variations in the minimum cluster size can also cause changes in the Significant Hotspots.

Table V. Results obtained from Proposed Algorithm using OPTICS

Min. Cluster Size method	Candidate Cluster Count	Significant hotspot Count using Cluster-Density	Significant hotspot Count using LLR method
6	1218	387	444
7	999	322	345
8	844	291	297
9	725	248	272
10	639	233	244
11	566	192	236
12	501	168	207
13	440	157	196
14	405	147	180
15	367	128	147
16	325	118	134

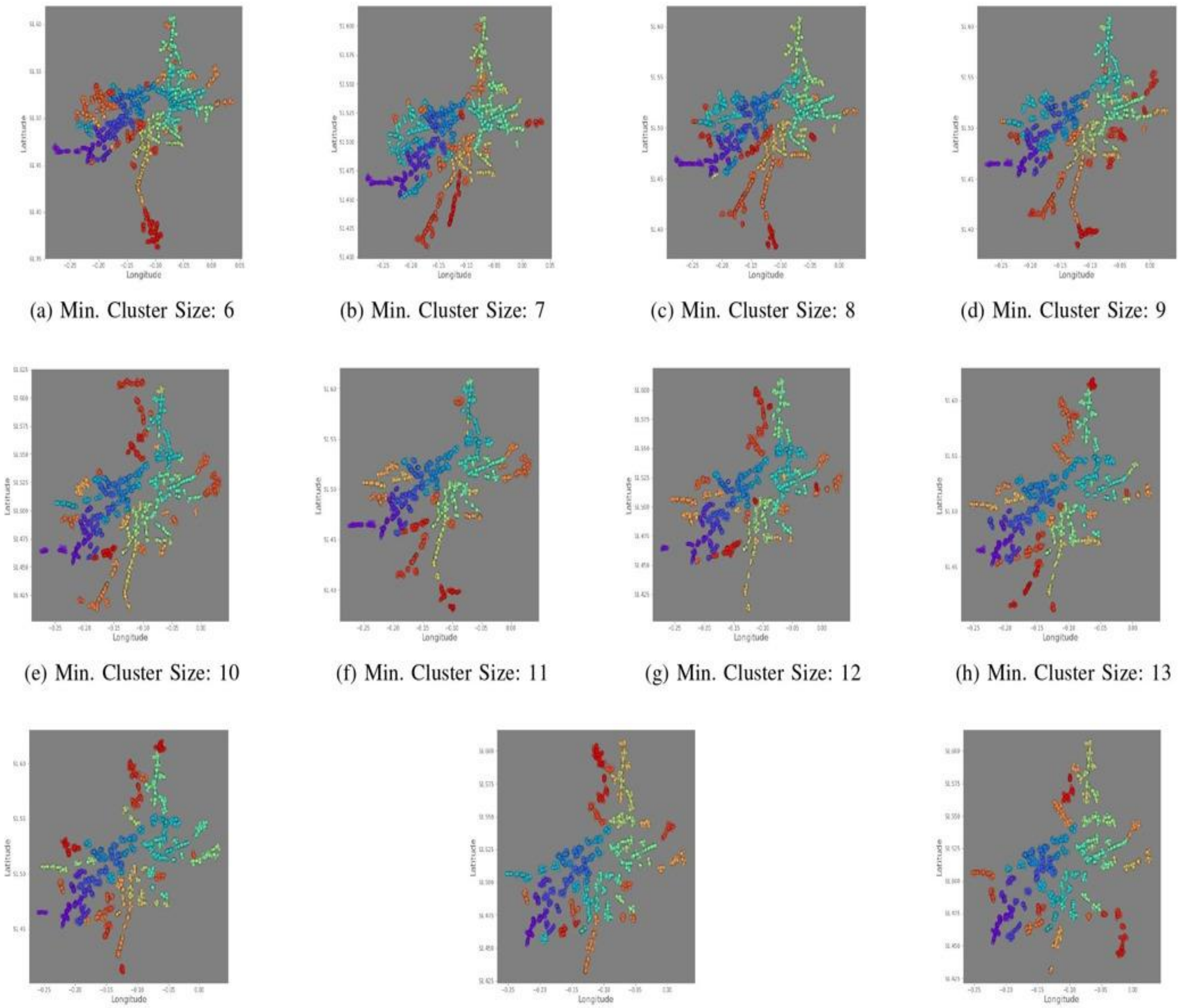
Table VI. Results obtained from Proposed Algorithm using HDBSCAN

Min. Cluster Size method	Candidate Cluster Count	Significant hotspot Count using Cluster-Density	Significant hotspot Count using LLR method
6	1028	323	357
7	824	278	284
8	657	208	223
9	546	172	195
10	473	145	161
11	415	133	156
12	365	115	125
13	325	102	118
14	285	97	103
15	256	76	89
16	237	71	84

In the Metropolitan areas, the minimum cluster size parameter is considered to be 6, which is twice the size of dimensions, because the dataset can be seen as three dimensional with primary features of Longitude, Latitude, and Time. The algorithm converges on a minimum cluster size of 16. Through any of the test statistics, it can be observed that as the cluster size grows, the number of hotspots decreases.

Fig. 12 depicts a demonstration of the results from Table V using the Cluster density method, as well as a scatter plot of the Latitude and Longitude points of the Significant Hotspots. As the algorithm approaches the convergence criteria, the adjacency of clusters improves, yielding improved results. In a similar manner, Fig. 9 depicts a visual representation of the Table VI results using the Cluster density method along with the Latitude and Longitude points. Both the OPTICS and HDBSCAN algorithms perform similarly in terms of the number of significant hotspots. Fig. 8 and Fig. 10 demonstrate the results using the Log Likelihood ratio method from Table V and Table VI respectively. The clustering patterns of both these algorithms, however, differ, and as a result, the positions of cluster points vary. Although the majority of the activity points found in most significant hotspots are intersections of cluster points generated using both OPTICS and HDBSCAN in accordance with the Proposed Algorithm.

The authorities in charge might prioritize work on particular hotspot zones based on the season and the time of year. We analyzed the Significant Spatial Hotspots in two subcategories: Seasonally and Temporally. We divided the Seasons in a year into four categories: Winter (December, January, and February), spring (March, April, May), summer (June, July, August) and autumn (September, October, November) for conducting seasonal analysis. The experiments are also conducted for different times of the day i.e., Morning (6:00 AM - 11:00 AM), Afternoon (11:00AM - 16:00PM), Evening (16:00PM - 21:00PM) and Night (21:00PM - 6:00AM). The Spatiotemporal effect of the identified hotspots is measured and analyzed across different dimensions.



(a) Min. Cluster Size: 6

(b) Min. Cluster Size: 7

(c) Min. Cluster Size: 8

(d) Min. Cluster Size: 9

(e) Min. Cluster Size: 10

(f) Min. Cluster Size: 11

(g) Min. Cluster Size: 12

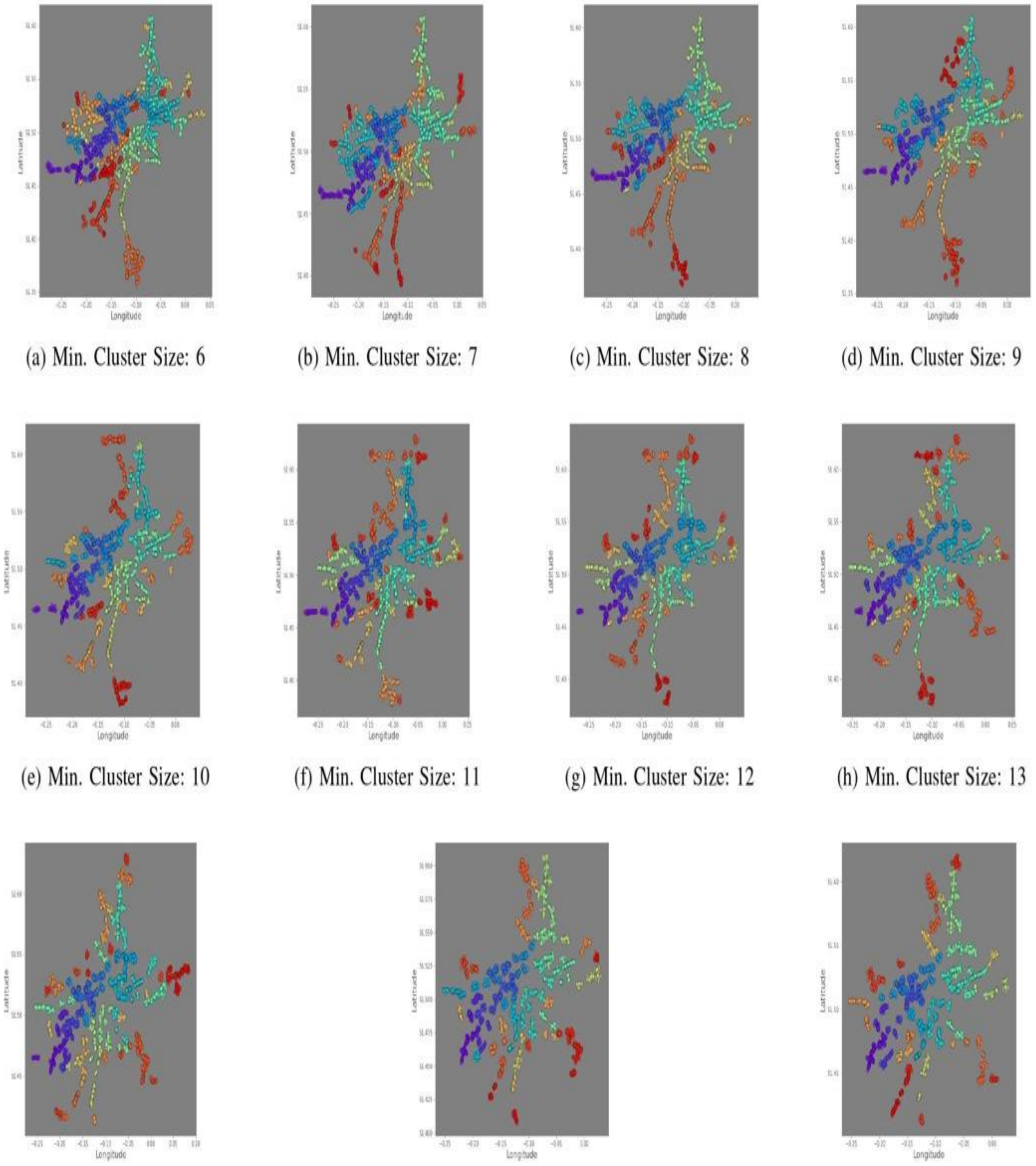
(h) Min. Cluster Size: 13

(i) Min. Cluster Size: 14

(j) Min. Cluster Size: 15

(k) Min. Cluster Size: 16

Fig. 12 Scatter plot of Significant Spatial Hotspots using OPTICS with Cluster Density method for varying values of min cluster size



(a) Min. Cluster Size: 6

(b) Min. Cluster Size: 7

(c) Min. Cluster Size: 8

(d) Min. Cluster Size: 9

(e) Min. Cluster Size: 10

(f) Min. Cluster Size: 11

(g) Min. Cluster Size: 12

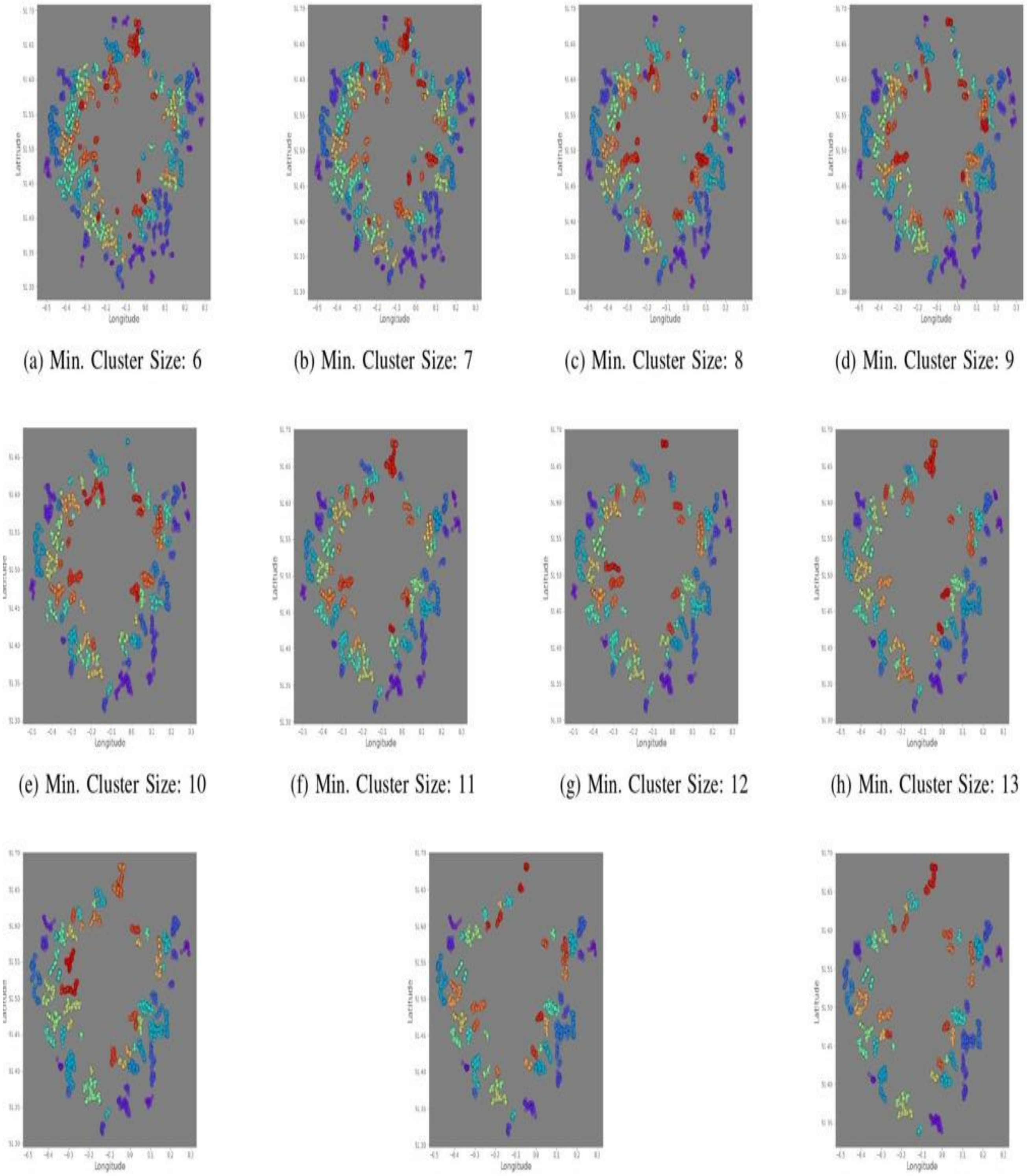
(h) Min. Cluster Size: 13

(i) Min. Cluster Size: 14

(j) Min. Cluster Size: 15

(k) Min. Cluster Size: 16

Fig. 13 Scatter plot of Significant Spatial Hotspots using OPTICS with Log Likelihood Ratio method for varying values of min cluster size



(a) Min. Cluster Size: 6

(b) Min. Cluster Size: 7

(c) Min. Cluster Size: 8

(d) Min. Cluster Size: 9

(e) Min. Cluster Size: 10

(f) Min. Cluster Size: 11

(g) Min. Cluster Size: 12

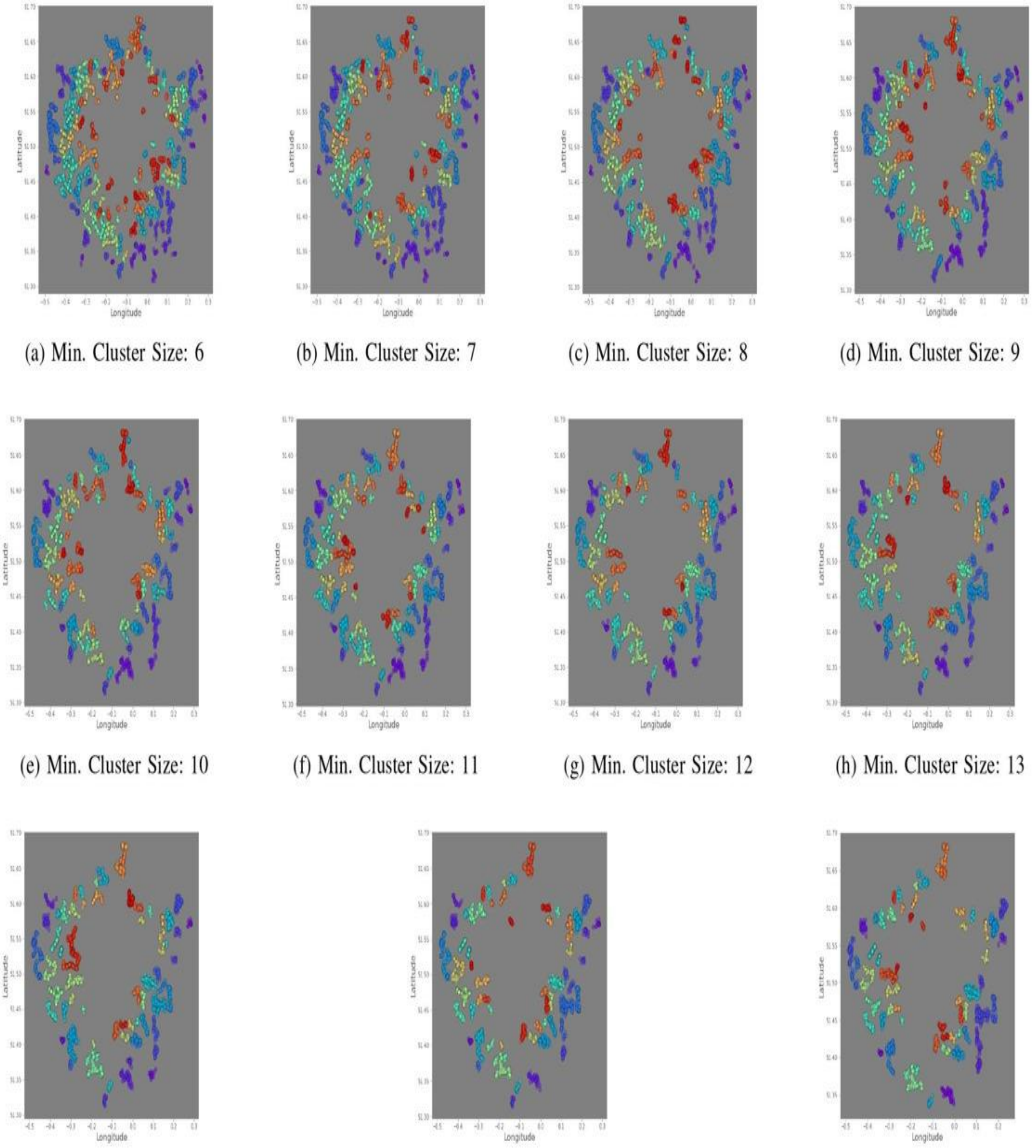
(h) Min. Cluster Size: 13

(i) Min. Cluster Size: 14

(j) Min. Cluster Size: 15

(k) Min. Cluster Size: 16

Fig. 14 Scatter plot of Significant Spatial Hotspots using HDBSCAN with Cluster Density method for varying values of min cluster size



(a) Min. Cluster Size: 6

(b) Min. Cluster Size: 7

(c) Min. Cluster Size: 8

(d) Min. Cluster Size: 9

(e) Min. Cluster Size: 10

(f) Min. Cluster Size: 11

(g) Min. Cluster Size: 12

(h) Min. Cluster Size: 13

(i) Min. Cluster Size: 14

(j) Min. Cluster Size: 15

(k) Min. Cluster Size: 16

Fig.15 Scatter plot of Significant Spatial Hotspots using HDBSCAN with Log Likelihood Ratio method for varying values of min cluster size

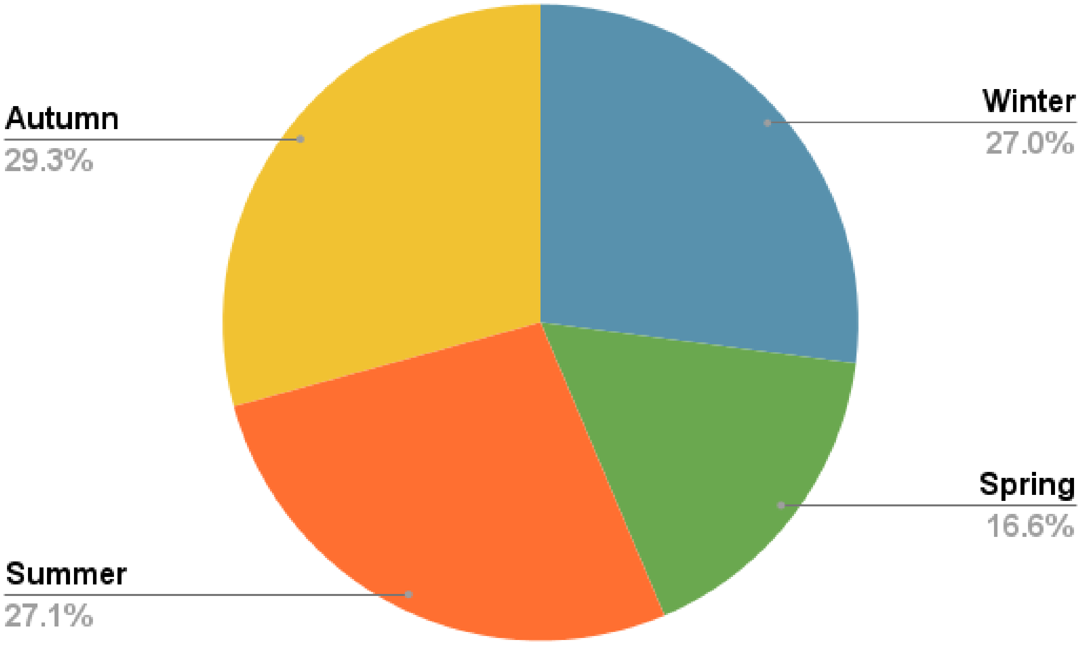


Fig. 16 Percentage of Road Accidents in different Seasons

4.1 Seasonal Analysis of Significant Spatial Hotspots

The road and traffic conditions are influenced by environmental factors. Seasonal changes have a significant impact on meteorological conditions, which has an indirect impact on the road network. The paper presents a comprehensive analysis of the seasonal impact on Road accidents in the UK. Figure 16 depicts the percentage of road accidents in metropolitan areas during various seasons. In the year 2020, Autumn experienced the highest number of accidents, while Summer and Winter experienced a comparable amount of traffic crashes. The interesting observations made are described as follows:

1) *Count of Significant Hotspots per season*: The number of Significant Hotspots observed during each season in a year using each Algorithm (OPTICS using Cluster-Density Significance Testing, OPTICS using Log Likelihood Ratio Significance Testing, HDBSCAN using Cluster-Density Significance Testing, and HDBSCAN using Log Likelihood Ratio Significance Testing) is depicted in Fig. 17. According to the bar plot, Autumn is the season having the highest number of detected significant hotspot zones for road accidents, which might be caused by heavy rain, foggy weather, or autumn foliage. As a result of this observation, we discovered that deer crashes are widespread in the UK during the autumn. However, Spring season witnessed the least number of significant hotspot zones.

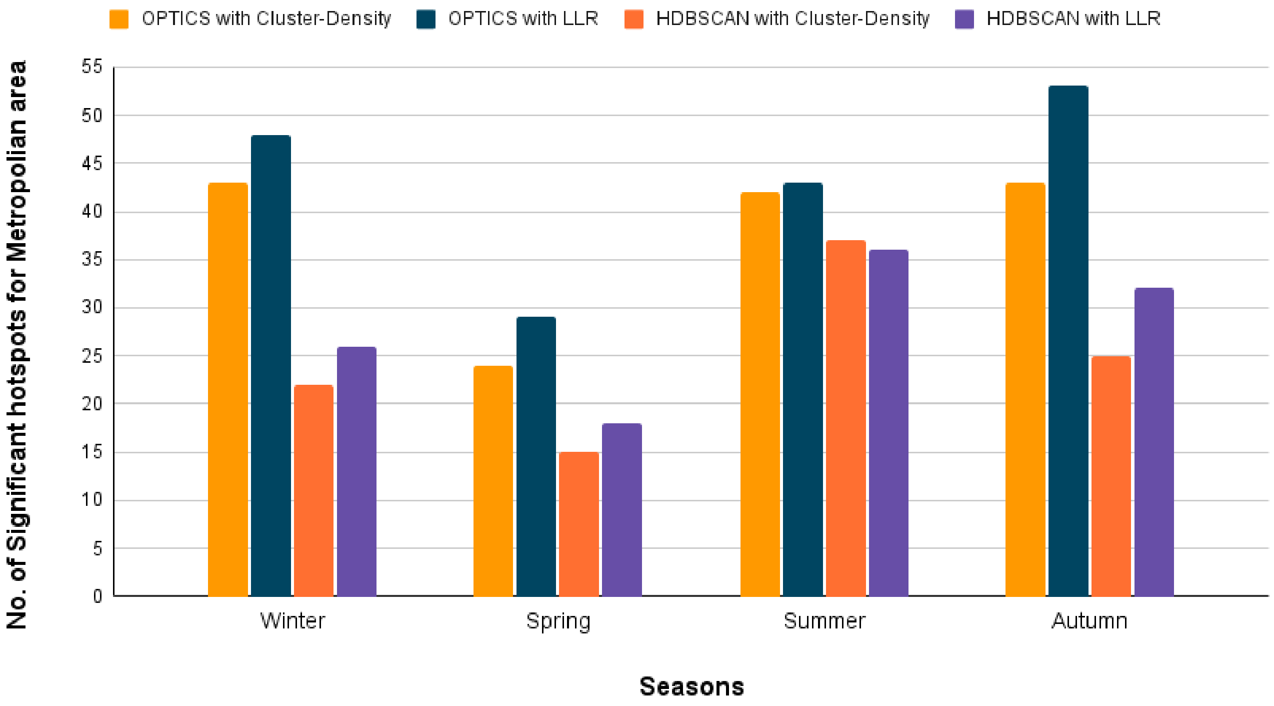


Fig.17 Season wise count of Significant Spatial Hotspots

2) *Identification of the High-Risk Accident Zones by Ranking of Significant Hotspots*:

Using the proposed algorithm, several Significant Hotspot locations in the United Kingdom have been discovered as shown in Fig. 17. However, it is crucial to determine the high-priority zones that require immediate attention by the concerned authorities. The paper identifies such high-risk zones and the results are illustrated in Fig. 20, Fig. 21, Fig. 22, and Fig. 23. Fig. 20(a) shows that the OPTICS with Cluster-Density Significance Testing approach identified Wembley as the most high-risk accident zone in the winter, whereas Fig. 21(b) shows that HDBSCAN with Cluster-Density Significance Testing determined Wembley as the most high-risk zone in the spring.

During the summer and autumn seasons, however, Wembley can be considered as a region with a lower than usual probability of accidents. Similarly, as shown in Fig. 20(c) and Fig. 21(d), Croydon is the region with the highest risk of traffic accidents during the Summer and Autumn seasons. The results of our proposed methodology are consistent with the findings of the UK's National Accident Helpline, which likewise claimed Croydon to be an exceedingly accident-prone location. Fig. 21(a), Fig. 21(c) and Fig. 23(b) depicts the City of London to be a high-risk zone through most part of the year.

Following the identification of these high-risk accident zones, the responsible authorities must take immediate action to recognize the situation and difficulties in road transportation, as well as deliver appropriate resources in accordance with demand of the situation. Better road and highway infrastructure, or prompt cleaning of snow from roads in winters, improved traffic signal management, and proper speed limit management are all potential solutions to challenges faced during seasonal fluctuations.

3) *Identification of Significant Hotspots on the basis of the type of accident severity*: In the previous section, we determined the highest risk accident zones from the large set of hotspot regions generated by the proposed algorithm. These findings assist the relevant authorities in conducting effective road planning in the affected areas. Mostly, the budget of the authorities and the resources of the traffic police department are limited, hence it becomes critical to delve deeper into high risk accident zones and identify those regions that can make the best use of resource deficiency and should be prioritized based on the severity of the accident.

Table VII demonstrates the Seasonal Severity-wise analysis of the Significant Road accident hotspots obtained from all different variations of the proposed algorithm. The severity of an accident is split into two categories: fatal and serious. Accidents having at least one casualty are classified as fatal accidents. Serious accident severity refers to collisions that resulted in major injuries and car damage but no fatalities. Central London was identified as the most significant road accident hotspot for fatal accidents during the Winter Season, while Paddington was identified as the most significant road accident hotspot for serious accidents. Croydon can be designated as the high-risk hotspot for fatal accidents throughout the summer months as well as the Autumn Season. Also, Clapham was the most significant road accident hotspot for serious accidents in Autumn.

Seasonal analysis of road accident hotspots presented in our research work can serve as a guiding tool for the government and concerned agencies in improving road safety conditions that may be impacted by natural seasonal variations. The comprehensive analysis of unpredictable weather conditions is critical for organizing public awareness campaigns for road safety. This contributes to the larger goal of constructing sustainable cities that offer safer roads for the people.

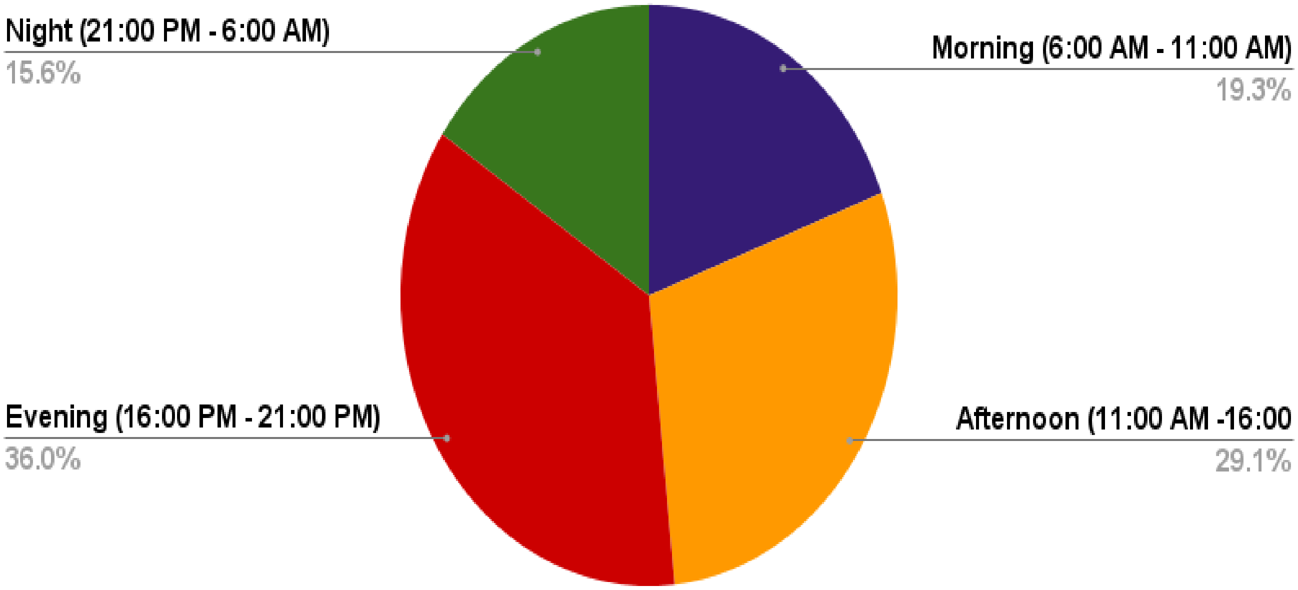


Fig. 18 Percentage of Road Accidents in different time slots of a day

4.2 Temporal Analysis of Significant Spatial Hotspots

The road traffic network is highly influenced by the timestamp of a day. People rush to work in the morning, and as a result, the roads are busy or congested. People also try to compensate for the hours wasted stuck in traffic by breaking the speed limit. Moreover, people rush back home in the evenings, causing the situation to worsen. The night time introduces additional challenges for road drivers because of limited illumination and reduced vision. The paper presents a comprehensive analysis of the temporal effects on Road accidents in the UK. Figure 18 depicts the percentage of road accidents in metropolitan areas during various time slots of a day. In the year 2020, the evening had the largest rate of accidents, while the morning and afternoon had 19.3% and 29.1% of road accidents, respectively. Interestingly, a total of 15.6% road accidents were recorded during night. Hence, it is evident that the typical tendency of the maximum number of crashes occurring at night does not adhere in this situation. The global COVID-19 pandemic prevalent in the year 2020 imposed nighttime curfews in the United Kingdom can help exemplify this pattern. The interesting observations revealed upon analysis are described as follows:

1) *Count of Significant Hotspots in different time slots of a day*: The number of Significant Hotspots observed during four distinct time periods of a day using each Algorithm (OPTICS using Cluster-Density Significance Testing, OPTICS using Log

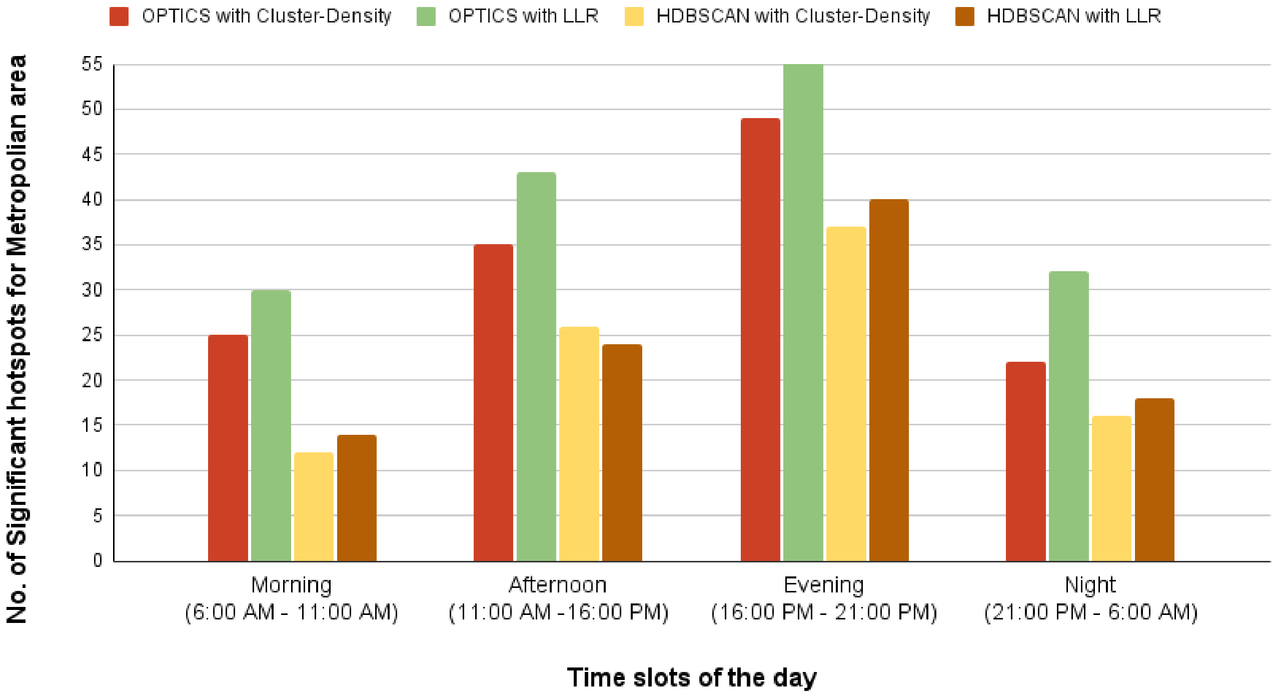


Fig. 19 Time wise count of Significant Spatial Hotspots

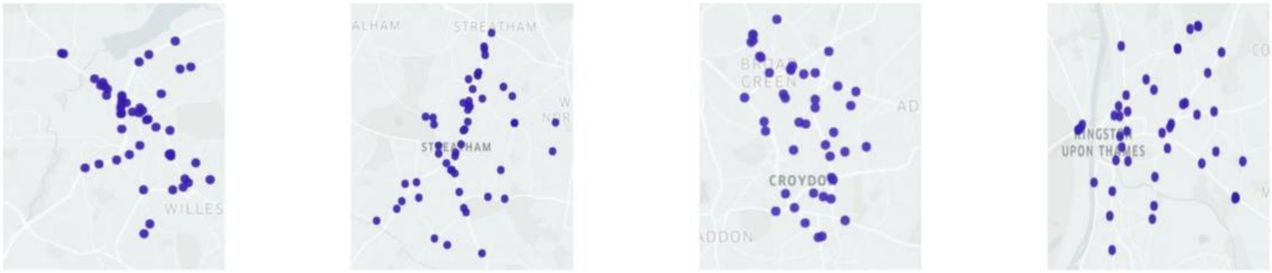
Likelihood Ratio Significance Testing, HDBSCAN using Cluster-Density Significance Testing, and HDBSCAN using Log Likelihood Ratio Significance Testing) is depicted in Fig. 19. Among all four variations of the proposed algorithm, the evening time from 16:00 PM to 21:00 PM accounted for the largest count of Significant road accident hotspots in the Metropolitan region, according to the bar plot. The afternoon hours encountered the next highest significant hotspot areas. During the morning and night, a similar number of Significant hotspot regions were discovered.

2) *Identification of the High-Risk Accident Zones by Ranking of Significant Hotspots*: Using the proposed algorithm, several Significant Hotspot locations in the United Kingdom have been identified during varying time slots of a day as shown in Fig. 19. Relevant resources, tools, and other required entities must be provided to appropriate high-risk zones at different times of the day. The paper identified such high-risk zones and the results are illustrated in Fig. 24, Fig. 25, Fig. 26, and Fig. 27. Fig. 24(a) shows that the OPTICS with Cluster-Density Significance Testing approach as well as Fig. 27(a) shows that HDBSCAN with Log Likelihood Ratio Significance Testing method identified Croydon as the most high-risk accident zone in the morning time from 6:00 AM to 11:00 AM. According to Fig. 25(c), Croydon has also been identified as a major road accident hotspot region during the rush hours in the evening time from 16:00 PM to

21:00 PM. Croydon has emerged as a prominent accident hotspot zone during rush hours, demanding the execution of important measures to address the dense traffic volume scenario and deadlocks caused due to traffic congestion. Preplanned divergence of congested roads by traffic police, construction of new routes via road and highway networks from one accident cluster to another, and the installation of traffic light monitoring systems and speed limit monitoring systems are all possible countermeasures. In the afternoon hours i.e., 11:00 AM to 16:00 PM, Fig. 24(b) and Fig. 26(b) have identified the Stamford Hill whereas Fig. 25(b) and Fig. 27(b) have identified Lancester West as the relevant road accident hotspot region. Typically, when the roads are busiest, the probability of an accident increases. Clapham road junction remains busy mostly during evenings and hence, the identification of busiest roads in these high-risk zones can aid the public in safer travel.

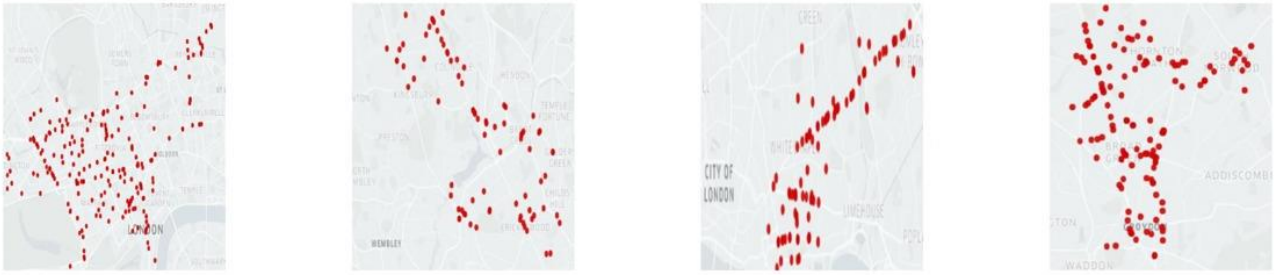
3) *Identification of Significant Hotspots on the basis of the type of accident severity*: The above section discussed the prevalence of high-risk road accident zones in the UK. The important of Severity analysis has already been discussed during the seasonal analysis. Table VIII demonstrates the Temporal Severity-wise analysis of the Significant Road accident hotspots obtained from all different variations of the proposed algorithm. Croydon and Harlington were identified as the most significant road accident hotspot zones for fatal and serious accidents respectively during the morning hours. During the afternoon time, all the four variations of the proposed algorithm discovered Tottenham as the hotspot zone for accidents of serious type. In the evening hours, Clapham and Walthamstow were detected as the accident prone regions for serious and fatal accidents respectively. Based on the data collected for road accidents for the year 2020, four different hotspot zones for the night time are identified to be Enfield Town, Broad Green, St. James's and St. Luke's.

Regardless of the type of application, temporal analysis is an integral aspect of identifying Significant Spatial Hotspots because it reveals hidden patterns and trends from the spatial data. The prominent feature of Temporal analysis is observing the dynamic nature of significant spatial hotspots. Through the introduction of time domain in the spatial data analysis for significant hotspot detection, one can observe how significant hotspots in the study area changing with time and also how a particular hotspot evolves with time. This adds a dynamic component to the proposed algorithm and we are able to get a wider view of the spatial data which helps us to derive relevant solutions for the application of road safety.



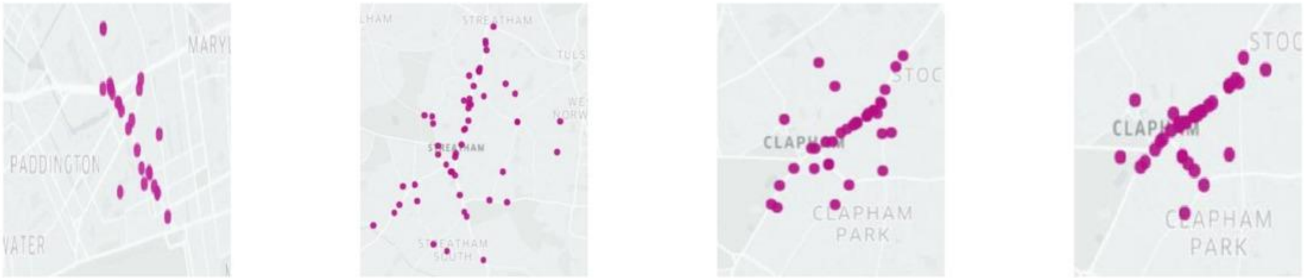
(a) Winter: *Wembley* (b) Spring: *Streatham* (c) Summer: *Croydon* (d) Autumn: *Kingston*

Fig. 20 Most Significant Hotspots using OPTICS with Cluster-Density method seasonally



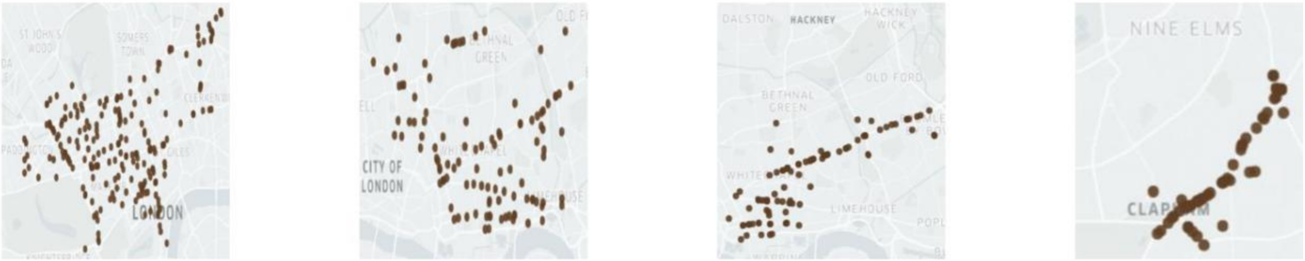
(a) Winter: *City of London* (b) Spring: *Wembley* (c) Summer: *City of London* (d) Autumn: *Croydon*

Fig. 21 Most Significant Hotspots using HDBSCAN with Cluster-Density method seasonally



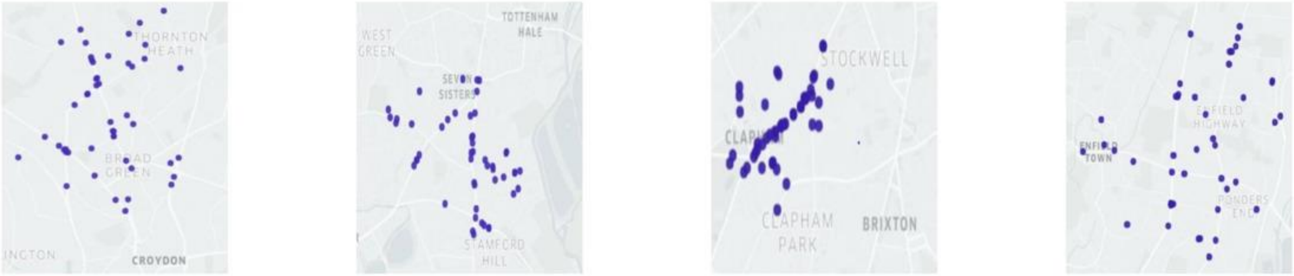
(a) Winter: *Paddington* (b) Spring: *Streatham* (c) Summer: *Clapham* (d) Autumn: *Clapham*

Fig. 22 Most Significant Hotspots using OPTICS with Log Likelihood Ratio method seasonally



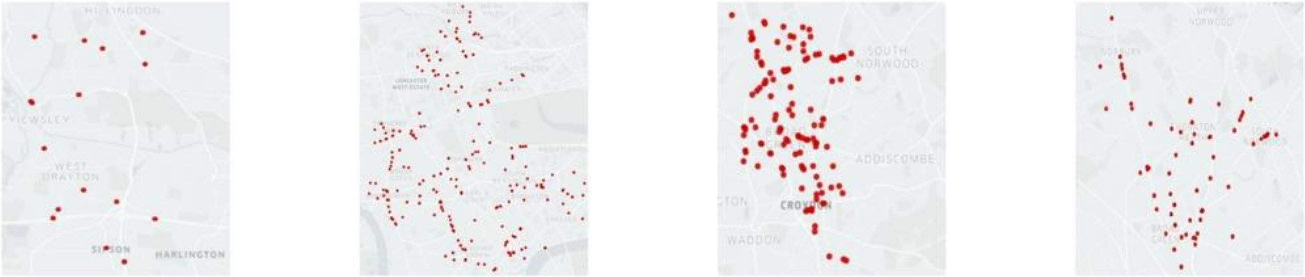
(a) Winter: *City of London* (b) Spring: *City of London* (c) Summer: *Hackney* (d) Autumn: *Clapham*

Fig. 23 Most Significant Hotspots using HDBSCAN with Log Likelihood Ratio method seasonally



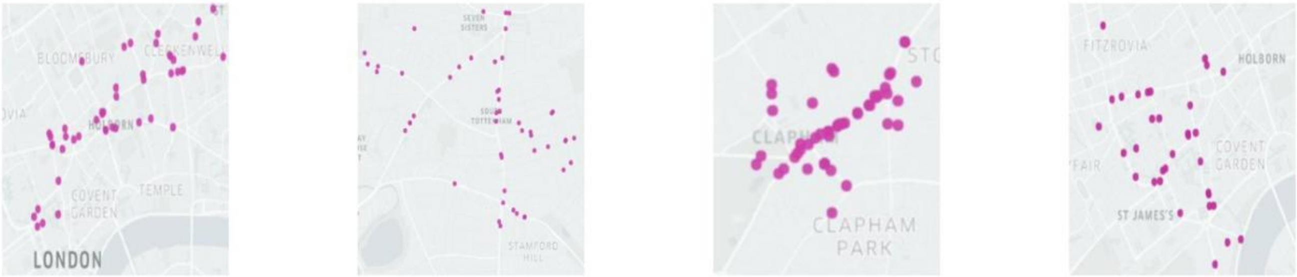
(a) Morning: *Croydon* (b) Afternoon: *Stamford Hill* (c) Evening: *Clapham* (d) Night: *Enfield Town*

Fig. 24 Most Significant Hotspots using OPTICS with Cluster-Density method temporally



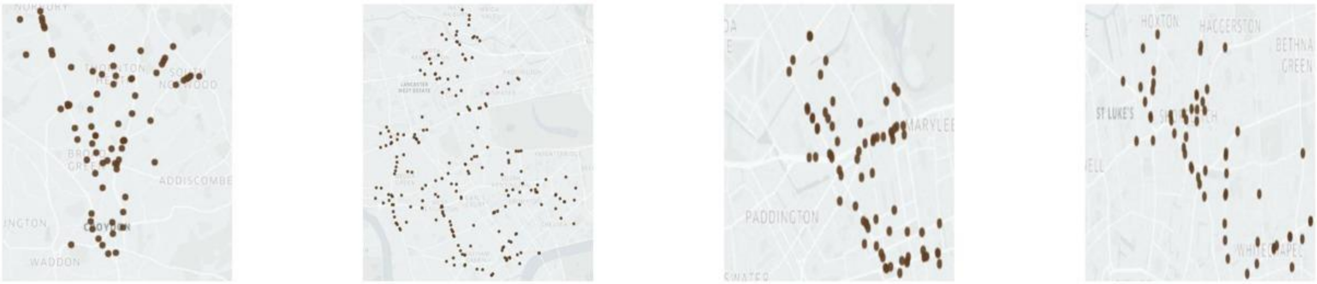
(a) Morning: *Harlington* (b) Afternoon: *Lancester West* (c) Evening: *Croydon* (d) Night: *Broad Green*

Fig. 25 Most Significant Hotspots using HDBSCAN with Cluster-Density method temporally



(a) Morning: *City of London* (b) Afternoon: *Stamford Hill* (c) Evening: *Clapham* (d) Night: *St James's*

Fig. 26 Most Significant Hotspots using OPTICS with Log Likelihood Ratio method temporally



(a) Morning: *Croydon* (b) Afternoon: *Lancester West* (c) Evening: *Paddington* (d) Night: *St Luke's*

Fig. 27 Most Significant Hotspots using HDBSCAN with Log Likelihood Ratio method temporally

TABLE VII. Seasonal Severity-wise analysis

	Winter		Spring		Summer		Autumn	
	Fatal	Serious	Fatal	Serious	Fatal	Serious	Fatal	Serious
OPTICS with Cluster Density method	Central London	Ilford	Woodgreen	Westminster	Croydon	Westminster	Norbiton	Clapham
HDBSCAN with Cluster Density method	ST Luke’s	Paddington	Wembley	Catford	City of London	Clapham	Croydon	Clapham
OPTICS with LLR method	Central London	Paddington	Streatham	East London	Croydon	Clapham	Croydon	Clapham
HDBSCAN with LLR method	Central London	Paddington	City of London	Catford	Hackney	Catford	Croydon	Clapham

TABLE VIII. Temporal Severity-wise analysis

	Morning		Afternoon		Evening		Night	
	Fatal	Serious	Fatal	Serious	Fatal	Serious	Fatal	Serious
OPTICS with Cluster Density method	Croydon	Harlington	Stamford Hill	Tottenham	Walthamstow	Clapham	Enfield Town	St James’s
HDBSCAN with Cluster Density method	Croydon	Harlington	Stamford Hill	Tottenham	Walthamstow	Clapham	Enfield Town	St James’s
OPTICS with LLR method	City of London	Harlington	Lancester West	Tottenham	Croydon	Clapham	Broad Green	ST Luke’s
HDBSCAN with LLR method	Croydon	Harlington	Lancester West	Tottenham	Walthamstow	Paddington	Broad Green	ST Luke’s

CHAPTER 5: COMPARATIVE ANALYSIS OF THE PROPOSED APPROACH

TABLE IX. Comparison of the variants of the proposed method

	Computation	Clustering pattern	Precision
OPTICS with Cluster-Density method	Less input parameters required and O(1) for Significance testing	Varies slightly with increase in min cluster size	0.68-0.72 for min cluster size in [6,16]
HDBSCAN with Cluster-Density method	Faster computation and O(1) for Significance testing	Less variation with increase in min cluster size	0.73-0.75 for min cluster size in [6,16]
OPTICS with LLR method	Less input parameters required and O(n) for Significance testing where n is the total no. of data points	Varies greatly with increase in min cluster size	0.67-0.69 for min cluster size in [6,16]
HDBSCAN with LLR method	Faster computation and O(n) for Significance testing where n is the total no. of data points	Less variation with increase in min cluster size	0.69-0.71 for min cluster size in [6,16]

Several research papers have incorporated road accidents data [61] provided by the government of United Kingdom. Table X shows the comparison of our proposed strategy to other approaches suggested in relevant research studies. Various researchers have identified substantial high-risk zones for road accidents in different parts of the world, with relevant results. The Identification of road accident clusters using Kulldorff's Spatial Scan Statistic was presented by Junxian Song et al. [78] in 2018. This technique used significance testing and did not require a specified cluster count, thus all possible clusters could be found. It also included an investigation of other external factors. However, the discovered clusters were shape-dependent, which may not provide relevance in real-world applications of road safety.

In 2021, Christopher Sinclair et al. [77] employed a partition-based clustering technique to detect high-risk accident zones. For the purpose of identifying spatial clusters, K-Means clustering was applied to the UK road accident dataset. They created shape-invariant clusters but had additional requirements of predetermined cluster count, multiple other input parameters and performed no significance testing. On the other hand, the authors conducted an exhaustive analysis by taking into account a variety of external elements such as the driver's age, the vehicle's age, the vehicle's speed limit, and engine capacity, among others.

To identify possible causes of accidents in India, Reeta Bhardwaj et al. [70] presented a hybrid clustering technique based on K-Mode partitioning and association rule mining. The density-based clustering approach [64] is also adopted for the purpose of finding road accident clusters. The approach yielded shape-invariant clusters and does not necessitate a predetermined cluster count. It does not, however, perform the significance testing of discovered clusters, which is an important criterion for determining statistical significance of clusters and prioritize accident hotspot regions for the concerned authorities.

Determining the statistical importance of clusters is important in eliminating out candidate clusters that are false positives, i.e., clusters that appear to be significant hotspot zones but are actually less relevant than the others. Furthermore, it assigns a strong statistical confidence factor to the clusters discovered by machine learning approaches, solidifying the decision-making process for the appropriate authorities. Our proposed approach features shape-invariant clusters with no requirement of predefined cluster count. It eliminates the difficulties in hyperparameter tuning by only demanding a single input parameter to the proposed algorithm. Moreover, the proposed approach brings out the capabilities of cutting-edge clustering algorithms for utilization in applications of spatial data domain.

We selected OPTICS and HDBSCAN unsupervised machine learning clustering algorithms which work quite effectively for spatial data and are also compatible with haversine distance metric. This produces more accurate clustering results on top of which robust significance testing is employed using two efficient test statistic measures, i.e.,

- Cluster-Density Method
- Log Likelihood Ratio Method.

Table IX shows a comparison of the variants of the proposed four methods in terms of three crucial factors in determining the practical application of the proposed algorithms. The Computation factor is determining the time complexity of the proposed algorithms. The Clustering pattern defines the cluster variations with increase in min. cluster size. The Accuracy is calculated after comparing the predicted results for the year 2020 with the actual results. The range of values for accuracy was calculated for varying values of min. cluster size. Among the four methods, HDBSCAN with Cluster Density methods outperform from the other three methods in terms of all three factors.

After obtaining significant spatial hotspots, we also presented a comprehensive analysis of the seasonal and temporal variations of the identified hotspot zones. In addition, a severity magnitude study of the seasonal and temporal hotspot zones is carried out to identify the depth of accident severity in the hotspot areas. Through the recommended solution strategy outlined in the following section, we also aim to effectively guide the relevant authorities in making critical decisions for road safety and public awareness.

Table X. Comparison of proposed approach with related research studies

Reference	Shape-invariant clusters	No predefined cluster count	Single input parameter	Significance Testing	Analysis
Junxian Song et al. [78]	✗	✓	✗	✓	✓
Christopher Sinclair et al. [77]	✓	✗	✗	✗	✓
Reeta Bhardwaj et al. [70]	✓	✗	✓	✗	✗
Abdullah S. et al.[64]	✓	✓	✗	✗	✓
Our Proposed approach	✓	✓	✓	✓	✓

CHAPTER 6: CONCLUSION & FUTURE SCOPE

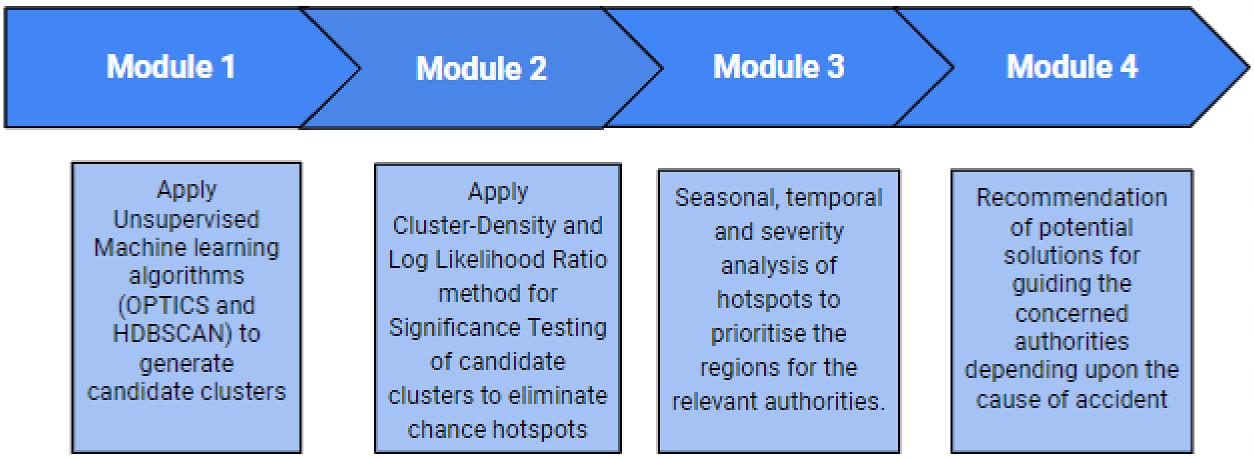

Module 1
Module 2
Module 3
Module 4
Apply Unsupervised Machine learning algorithms (OPTICS and HDBSCAN) to generate candidate clusters
Apply Cluster-Density and Log Likelihood Ratio method for Significance Testing of candidate clusters to eliminate chance hotspots
Seasonal, temporal and severity analysis of hotspots to prioritise the regions for the relevant authorities.
Recommendation of potential solutions for guiding the concerned authorities depending upon the cause of accident

Fig. 28 Proposed Solution approach

We critically examined and analyzed the existing research literature on spatiotemporal hotspot detection techniques in this systematic study of Spatiotemporal hotspot detection and evaluation measures, and developed a hierarchical framework for organizing and summarizing the widely used methods for spatiotemporal hotspot detection. We have summarized the most often used evaluation metrics for measuring the accuracy of Spatiotemporal hotspots. Our findings will aid future researchers in developing and improving more robust strategies for detecting Spatiotemporal Hotspots, as well as maximizing the use of existing techniques for detecting Spatiotemporal Hotspots.

The proposed algorithm focuses on the combined strengths of Spatial Clustering algorithms with efficient significance testing methods. The hybrid combination of these two mechanisms yields better accuracy in terms of significant hotspot identification. The proposed algorithm described as Algorithm 1 and Algorithm 2 of this paper reduced the probability of detecting false positives to a reasonably lower extent which makes it suitable for real world datasets. For demonstration of the results obtained by the proposed algorithm, the paper presented a comprehensive analysis of the UK-Road accident hotspots. The paper also presented a comparative analysis of the OPTICS algorithm and HDBSCAN algorithm for Spatial Clustering as well as the significance testing of the hotspots performed using two efficient test statistic measures i.e., Cluster Density method and Log Likelihood Ratio Method. The results are demonstrated in detail in the sections above. We further identified the likely causes of the problems after investigation and propose the following solutions:

1) *Weather conditions*: It is natural for weather to have a significant impact on traffic accidents. On marshland, a high number of accidents occur, causing road damage. Aside from that, the weather effects in some parts of the United Kingdom are concerning. However, improving road fog illumination, repairing slick roads, and improving pedestrian pathways are all potential solutions. The authors of [68] performed a detailed investigation of the impact of weather on road accidents.

2) *Road Surface conditions*: The road surface conditions of sites where incidents occurred were also reported by the UK road safety data. It was observed that dry roads had seen the majority of road accidents in the identified significant spatial hotspot zones. Aside from that, wet or damp roads, as well as snowy roads, have been known to cause traffic accidents in the detected hotspot areas. As a result, it is critical to execute quick road damage repair as well as fast snow removal from roads during the particular seasons and time of the day as determined from the Seasonal and Temporal analysis.

3) *Violation of Traffic rules*: Over speeding, crossing red lights, and other traffic violations are common. Object detection and CCTV monitoring systems can be used to track whether or not passengers are wearing seat belts or helmets while driving. It can also help the traffic police to impose penalties on anyone who break traffic signals on the road. Utilizing the identified road accident hotspot zones, the concerned authorities can direct traffic police force or required personnel at these high-risk accident hotspot regions to ensure road safety is well maintained at varying times of the day.

4) *Driver’s Carelessness*: Some reasons of road accidents are unfortunate, such as the driver’s irresponsibility. Public awareness campaigns can help raise knowledge about the issue of road safety by using safety hoardings or relevant posters in hotspot zones. Furthermore, as the seasons change, posters can be digitized and updated, particularly in places where substantial road accident hotspots have been identified.

As previously stated, the dataset contained attributes that may be used to determine the most likely causal factors of road accidents. After identifying major spatial hotspots, we refined the dataset. The latitude and longitude points of significant spatial hotspots assist in defining the characteristics that can aid in determining the likely causes of road accidents. Fig. 29 shows that the most likely cause of an accident is the driver’s carelessness while driving. Policymakers must ensure that road safety initiatives and information campaigns are organized to spread awareness of the crisis and its devastating impact on people's lives. Furthermore, poor road surface conditions are the second significant component in the causes of road accidents that must be addressed. Following seasonal changes, a thorough evaluation of accident-prone highways and adjacent routes is required. Proper road traffic signs and illumination also aid in the creation of a safer road environment. Weather is unpredictable, and while it is the least prevalent cause of road accidents in the United Kingdom, it is always preferable to avoid mistakes that may occur as a result of it. As a result, it's critical to take the required actions to eliminate accident causes and ensure that everyone is safe on the roads.

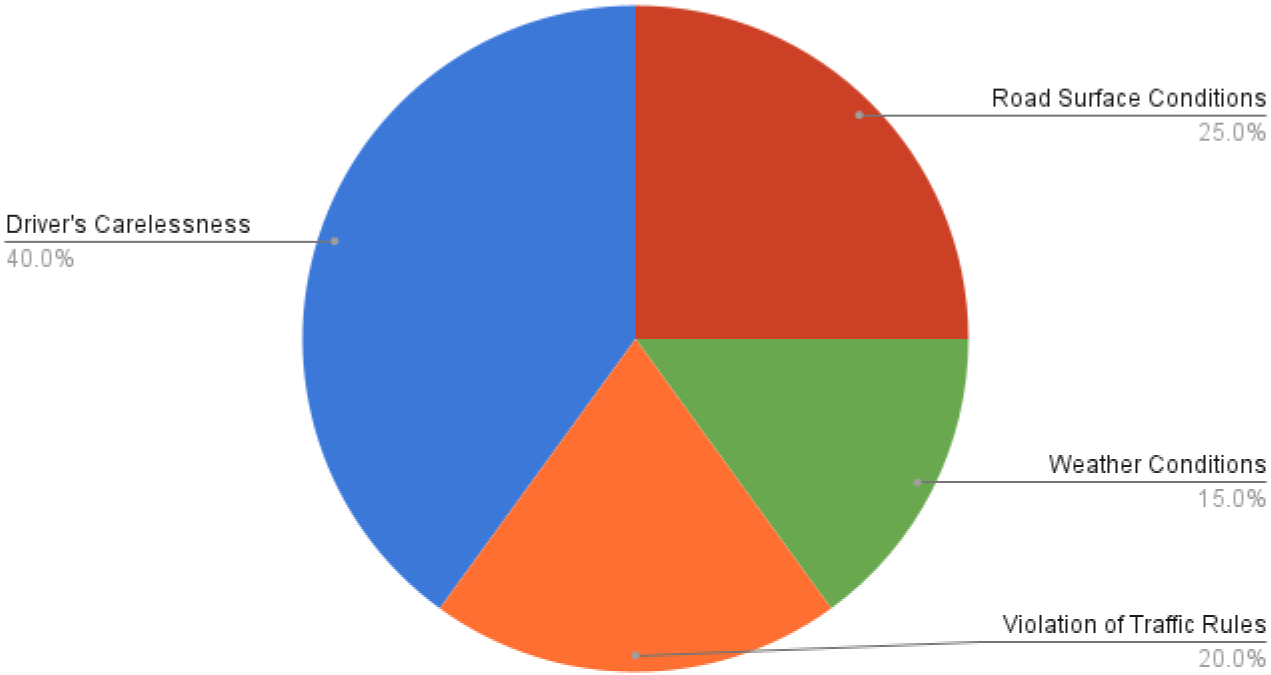


Fig. *29 Piechart describing the likely causes of road accidents*

The proposed solution approach, depicted in Fig.28, summarized the complete modular approach used in this paper to avoid road accidents. Module 1 demonstrated how OPTICS and HDBSCAN Clustering techniques can be used to find candidate clusters. Module 2 then showed how to use the proposed algorithm to discover statistically significant Hotspot zones in the United Kingdom. Module 3 examined the identified major spatial hotspots in terms of season, time, and severity. Module 4 concluded the paper by offering viable options for the concerned authorities to focus their resources and attention on necessary and required hotspot regions in order to respond quickly and effectively so as to create a sustainable and safer road network for all people.

BIBLIOGRAPHY

[1] G. Atluri, A. Karpatne, and V. Kumar, “Spatio-temporal data mining: A survey of problems and methods,” ACM Comput. Surv., vol. 51, no. 4, Aug. 2018.

[2] A. Hamdi, K. Shaban, A. Erradi, A. Mohamed, S. K. Rumi, and F. Salim, “Spatiotemporal data mining: A survey on challenges and open problems,” 2021.

[3] S. Shekhar, Z. Jiang, R. Y. Ali, E. Eftelioglu, X. Tang, V. M. V. Gunturi, and X. Zhou, “Spatiotemporal data mining: A computational perspective,” ISPRS International Journal of GeoInformation, vol. 4, no. 4, pp. 2306–2338, 2015.

[4] M. Kulldorff, “A spatial scan statistic,” Communications in Statistics - Theory and Methods, vol. 26, no. 6, pp. 1481–1496, 1997.

[5] K. Takahashi, M. Kulldorff, T. Tango, and K. Yih, “A flexibly shaped space–time scan statistic for disease outbreak detection and monitoring,” International journal of health geographics, vol. 7, p. 14, 02 2008.

[6] Y. Kim and M. O’Kelly, “A bootstrap based space–time surveillance model with an application to crime occurrences,” Journal of Geographical Systems, vol. 10, pp. 141–165, 02 2008.

[7] H. Rao, X. Shi, and X. Zhang, “Using the kulldorff’s scan statistical analysis to detect spatio-temporal clusters of tuberculosis in qinghai province, china, 2009–2016,” BMC Infectious Diseases, vol. 17, 08 2017.

[8] D. Adham, E. Asl, M. Abazari, A. Saghafipour, and P. Alizadeh, “Forecasting head lice (pediculidae: Pediculus humanus capitis) infestation incidence hotspots based on spatial correlation analysis in northwest iran,” Veterinary World, vol. 13, pp. 40–46, 01 2020.

[9] A. Guemes, S. Ray, K. Aboumerhi, M. Desjardins, A. Kvit, A. Corrigan, ¨ B. Fries, T. Shields, R. Stevens, F. Curriero, and R. Etienne-Cummings, “A syndromic surveillance tool to detect anomalous clusters of covid-19 symptoms in the united states,” Scientific Reports, vol. 11, p. 4660, 02 2021.

[10] J. Mosha, H. Sturrock, B. Greenwood, C. Sutherland, N. Gadalla, S. Atwal, S. Hemelaar, J. Brown, C. Drakeley, G. Kibiki, T. Bousema, D. Chandramohan, and R. Gosling, “Hot spot or not: A comparison of spatial statistical methods to predict prospective malaria infections,” Malaria journal, vol. 13, p. 53, 02 2014.

[11] C.-C. Chen, Y.-C. Teng, B.-C. Lin, I.-C. Fan, and T.-C. Chan, “Online platform for applying space–time scan statistics for prospectively detecting emerging hot spots of dengue fever,” International

Journal of Health Geographics, vol. 15, 11 2016.

[12] A. Hohl, E. M. Delmelle, M. R. Desjardins, and Y. Lan, “Daily surveillance of covid-19 using the prospective space-time scan statistic in the united states,” Spatial and Spatio-temporal Epidemiology, vol. 34, p. 100354, 2020.

[13] S. L. Linton, J. M. Jennings, C. A. Latkin, M. B. Gomez, and S. H. Mehta, “Application of space-time scan statistics to describe geographic and temporal clustering of visible drug activity,” Journal of Urban Health, vol. 91, pp. 940–956, 2014.

[14] E. Eftelioglu, S. Shekhar, J. M. Kang, and C. C. Farah, “Ring-shaped hotspot detection,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 12, pp. 3367–3381, 2016.

[15] V. S. Iyengar, “On detecting space-time clusters,” in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’04. New York, NY, USA: Association for Computing Machinery, 2004, p. 587–592.

[16] M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal, “An elliptic spatial scan statistic,” Statistics in medicine, vol. 25, pp. 3929–43, 11 2006.

[17] W. Dong, X. Zhang, L. Li, C. Sun, L. Shi, and W. Sun, “Detecting irregularly shaped significant spatial and spatio-temporal clusters,” Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012, pp. 732–743, 04 2012.

[18] S. Katragadda, J. Chen, and S. Abbady, “Spatial hotspot detection using polygon propagation,” International Journal of Digital Earth, vol. 12, pp. 1–18, 06 2018.

[19] A. Barro, I. Kracalik, L. Malania, N. Tsertsvadze, J. Manvelyan, P. Imnadze, and J. Blackburn, “Identifying hotspots of human anthrax transmission using three local clustering techniques,” Applied Geography, vol. 60, pp. 29–36, 06 2015.

[20] I. Gwitira, N. Karumazondo, M. Shekede, C. Sandy, N. Siziba, and J. Chirenda, “Spatial patterns of pulmonary tuberculosis (tb) cases in zimbabwe from 2015 to 2018,” PLOS ONE, vol. 16, p. e0249523, 04 2021.

[21] F. Di Martino, R. Mele, U. Barillari, M. R. Barillari, I. Perfiljeva, and S. Senatore, “Spatiotemporal hotspots analysis for exploring the evolution of diseases: An application to oto-laryngopharyngeal diseases,” Advances in Fuzzy Systems, vol. 2013, 01 2013.

[22] L. Roche, X. Niu, A. Stroup, and K. Henry, “Disparities in female breast cancer stage at diagnosis in new jersey: A spatial-temporal analysis,” Journal of Public Health Management and Practice, vol. 23,

p. 1, 12 2016.

[23] C. Bationo, J. Gaudart, S. Dieng, M. Cissoko, P. Taconet, O. Boukary, A. Some, I. Zongo, D. Soma, G. Tougri, R. Dabir ´ e, A. Koffi, C. Pen- ´ netier, and N. Moiroux, "Spatio-temporal analysis and prediction of malaria cases using remote sensing meteorological data in diebougou ´ health district, burkinafaso, 2016–2017," Scientific Reports, vol. 11, p. 20027, 10 2021.

[24] V. S. Iyengar, "On detecting space-time clusters," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 587–592.

[25] H. Fanaee-T and J. Gama, "Eigenspace method for spatiotemporal hotspot detection," CoRR, vol. abs/1406.3506, 2014.

[26] H. Fanaee-T and J. Gama, "An eigenvector-based hotspot detection," 06 2014.

[27] H. Izakian and W. Pedrycz, "A new pso-optimized geometry of spatial and spatio-temporal scan statistics for disease outbreak detection," Swarm and Evolutionary Computation, vol. 4, p. 1–11, 06 2012.

[28] S. Fan, "The spatial-temporal prediction of various crime types in houston, tx based on hot-spot techniques," 2014.

[29] D. Zeng, W. Chang, and H. Chen, "A comparative study of spatiotemporal hotspot analysis techniques in security informatics," in Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No.04TH8749), 2004, pp. 106–111.

[30] N. Malleson and M. Andresen, "Malleson, n., andresen, m.a. (2015). spatio-temporal crime hotspots and the ambient population. crime science, 4, article 10." vol. 4, p. Article 10., 12 2015.

[31] Y. Hu, F. Wang, C. Guin, and H. Zhu, "A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation," Applied Geography, vol. 99, 08 2018.

[32] S. Khalid, F. Shoaib, T. Qian, Y. Rui, A. I. Bari, M. Sajjad, M. Shakeel, and J. Wang, "Network constrained spatio-temporal hotspot mapping of crimes in faisalabad," Applied Spatial Analysis and Policy, vol. 11, pp. 599–622, 2018.

[33] J. Baker, C. Andris, and D. DellaPosta, "Spatial social network (ssn) hot spot detection: Scan methods for non-planar networks," 11 2020.

[34] Y. Zhang and T. Cheng, "Graph deep learning model for network-based predictive hotspot mapping

of sparse spatio-temporal events," Computers Environment and Urban Systems, vol. 79, 11 2019.

[35] Y. Zhuang, M. Almeida, M. Morabito, and W. Ding, "Crime hot spot forecasting: A recurrent model with spatial and temporal information," 08 2017, pp. 143–150.

[36] Y. Farjami and K. Abdi, "A genetic-fuzzy algorithm for spatio-temporal crime prediction," Journal of Ambient Intelligence and Humanized Computing, 01 2021.

[37] X. Wu and T. H. Grubesic, "Identifying irregularly shaped crime hot-spots using a multiobjective evolutionary algorithm," Journal of Geographical Systems, vol. 12, pp. 409–433, 2010.

[38] X. ye, L. Duan, and Q. Peng, "Spatiotemporal prediction of theft risk with deep inception-residual networks," Smart Cities, vol. 4, pp. 204– 216, 01 2021.

[39] P. Gao, D. Guo, K. Liao, J. Haney (Webb), and S. Cutter, "Early detection of terrorism outbreaks using prospective space–time scan statistics*," The Professional Geographer, vol. 65, 11 2013.

[40] K. G. Le, P. Liu, and L.-T. Lin, "Determining the road traffic accident hotspots using gis-based temporal-spatial statistical analytic techniques in hanoi, vietnam," Geo-spatial Information Science, vol. 23, no. 2, pp. 153–164, 2020.

[41] N. Dong, H. Huang, J. Lee, M. Gao, and M. AbdelAty, "Macroscopic hotspots identification: A bayesian spatiotemporal interaction approach," Accident Analysis & Prevention, vol. 92, pp. 256–264, 2016.

[42] Y. Shen, W. Pedrycz, R. M. De Moraes, X. Hu, X. Wang, and A. Gacek, "Clustering of information granules in hotspot identification," in 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019, pp. 1–6.

[43] Q. Yu, C. Chen, L. Sun, and X. Zheng, "Urban hotspot area detection using nearest-neighborhood-related quality clustering on taxi trajectory data," ISPRS International Journal of Geo-Information, vol. 10, no. 7, 2021.

[44] F. Li, W. Shi, and H. Zhang, "A two-phase clustering approach for urban hotspot detection with spatiotemporal and network constraints," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. PP, pp. 1–1, 03 2021.

[45] Xia, Li, G. Chen, and Liao, "Identify and delimitate urban hotspot areas using a network-based spatiotemporal field clustering method," ISPRS International Journal of Geo-Information, vol. 8, p. 344, 07 2019.

[46] A. Alomari, N. Shatnawi, T. Khedaywi, and T. Miqdady, “Prediction of traffic accidents hot spots using fuzzy logic and gis,” Applied Geomatics, vol. 12, 12 2019.

[47] Y. Zhu and S. D. Newsam, “Spatio-temporal sentiment hotspot detection using geotagged photos,” CoRR, vol. abs/1609.06772, 2016.

[48] F. Di Martino and S. Sessa, “Extended gustafson–kessel granular hotspot detection,” Granular Computing, vol. 5, 01 2020.

[49] U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, and H. H. R. Sherazi, “Spatio-temporal crime hotspot detection and prediction: A systematic literature review,” IEEE Access, vol. 8, pp. 166 553–166 574, 2020.

[50] L. Duczmal, M. Kulldorff, and L. Huang, “Evaluation of spatial scan statistics for irregularly shaped clusters,” Journal of Computational and Graphical Statistics, vol. 15, no. 2, pp. 428–442, 2006.

[51] W. Ding, Y. Xia, Z. Wang, Z. Chen, and X. Gao, “An ensemble-learning method for potential traffic hotspots detection on heterogeneous spatiotemporal data in highway domain,” Journal of Cloud Computing, vol. 9, pp. 1–11, 2020.

[52] K. Liu, H. Zhou, R.-X. Sun, H.-W. Yao, Y. Li, L.-P. Wang, D. Mu, X.- L. Li, Y. Yang, G. Gray, N. Cui, W.-W. Yin, L.-Q. Fang, H.-J. Yu, and W.-C. Cao, “A national assessment of the epidemiology of severe fever with thrombocytopenia syndrome, china,” Scientific reports, vol. 5, p. 9679, 04 2015.

[53] T. Bousema, C. Drakeley, S. Gesase, R. Hashim, S. Magesa, F. Mosha, S. Otieno, I. Carneiro, J. Cox, E. Msuya, I. Kleinschmidt, C. Maxwell, B. Greenwood, E. Riley, R. Sauerwein, D. Chandramohan, and R. Gosling, “Identification of hot spots of malaria transmission for targeted malaria control,” The Journal of infectious diseases, vol. 201, pp. 1764–74, 06 2010.

[54] M. J. C. Baculo, C. S. Marzan, R. de Dios Bulos, and C. Ruiz, “Geospatial-temporal analysis and classification of criminal data in manila,” in 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), 2017, pp. 6–11.

[55] E. Atumo, T. Fang, and X. Jiang, “Spatial statistics and random forest approaches for traffic crash hot spot identification and prediction,” International journal of injury control and safety promotion, pp. 1–10, 10 2021.

[56] P. Stalidis, T. Semertzidis, and P. Daras, “Examining deep learning architectures for crime classification and prediction,” Forecasting, vol. 3, pp. 741–762, 10 2021.

[57] H. Yu, Z. Li, G. Zhang, P. Liu, and J. Wang, “Extracting and predicting taxi hotspots in

spatiotemporal dimensions using conditional generative adversarial neural networks," IEEE Transactions on Vehicular Technology, vol. 69, no. 4, pp. 3680–3692, 2020.

[58] P. Rashidi, A. Skidmore, A. Vrieling, R. Darvishzadeh, B. Toxopeus, S. Ngene, and P. Oduor, "Spatial and spatiotemporal clustering methods for detecting elephant poaching hotspots," Ecological Modelling, vol. 297, pp. 180–186, 12 2014.

[59] C.-H. Yu, W. Ding, M. Morabito, and P. Chen, "Hierarchical spatiotemporal pattern discovery and predictive modeling," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 4, pp. 979–993, 2016.

[60] D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, "Machine learning approaches to traffic accident analysis and hotspot prediction," Computers, vol. 10, no. 12, 2021.

[61] Road safety data - data.gov.uk

[62] S. Wang, J. Cao, and P. Yu, "Deep learning for spatio-temporal data mining: A survey," IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2020.

[63] Z. Yuan, X. Zhou, and T. Yang, "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data," 07 2018, pp. 984–992.

[64] Abdullah S. Alotaibi. Density-based clustering for road accident data analysis. International Journal of ADVANCED AND APPLIED SCIENCES, 5(8):113–121, 2018.

[65] A. Naseer, M. K. Nour, and B. Y. Alkazemi, "Towards deep learning based traffic accident analysis," in 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), 2020, pp. 0817– 0820.

[66]MihaelAnkerst, Markus M. Breunig, Hans-Peter Kriegel, and Jorg Sander. ¨ Optics: Ordering points to identify the clustering structure. SIGMOD Rec., 28(2):49–60, jun 1999.

[67] J. Bao, P. Liu, and S. V. Ukkusuri, "A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data," Accident Analysis Prevention, vol. 122, pp. 239–254, 2019.

[68]Ruth Bergel-Hayat, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. Explaining the road accident risk: Weather effects. Accident Analysis Prevention, 60, 04 2013.

[69] H. Zhao, H. Cheng, T. Mao, and C. He, "Research on traffic accident prediction model based on convolutional neural networks in vanet," in 2019 2nd International Conference on Artificial Intelligence

and Big Data (ICAIBD), 2019, pp. 79–84.

[70]Reeta Bhardwaj, Ridhi R, and Rajeev Kumar. Modified Approach of Cluster Algorithm to Analysis Road Accident. International Journal of Computer Applications, 166(2):24–28, 2017.

[71] D. Al-Dogom, N. Aburaed, M. Al-Saad, and S. Almansoori, "Spatiotemporal analysis and machine learning for traffic accidents prediction," in 2019 2nd International Conference on Signal Processing and Information Security (ICSPIS), 2019, pp. 1–4.

[72] B. Romano and Z. Jiang, "Visualizing traffic accident hotspots based on spatial-temporal network kernel density estimation," in Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ser. SIGSPATIAL '17. New York, NY, USA: Association for Computing Machinery, 2017.

[73] M. Zahid, Y. Chen, S. Khan, A. Jamal, M. Ijaz, and T. Ahmed, "Predicting risky and aggressive driving behavior among taxi drivers: Do spatio-temporal attributes matter?" International Journal of Environmental Research and Public Health, vol. 17, no. 11, 2020.

[74] C. Zhang, J. jin, Q. Huang, Z. Du, Z. Yuan, S. Tang, and Y. Liu, "The spatiotemporal traffic accident risk analysis in urban traffic network," in Mobile Wireless Middleware, Operating Systems and Applications, W. Li and D. Tang, Eds. Cham: Springer International Publishing, 2020, pp. 92–97.

[75]B. Ryder and F. Wortmann, "Autonomously detecting and classifying traffic accident hotspots," in Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, ser. UbiComp '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 365–370.

[76] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software, 2, 03 2017.

[77] Christopher Sinclair and Saptarshi Das. Traffic accidents analytics in uk urban areas using k-means clustering for geospatial mapping. In 2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET), pages 1–7, 2021.

[78]Junxian Song, Rong Wen, and Wenjing Yan. Identification of traffic accident clusters using kulldorff's space-time scan statistics. In 2018 IEEE International Conference on Big Data (Big Data), pages 3162–3167, 2018.

[79]YiqunXie and Shashi Shekhar. Significant dbscan towards statistically robust clustering. In Proceedings of the 16th International Symposium on Spatial and Temporal Databases, SSTD '19, page 31–40, New York, NY, USA, 2019. Association for Computing Machinery.

LIST OF PUBLICATIONS

COMMUNICATED (In SCI-Indexed Journal)

(Revisions under Review)

[1] "Significant Spatial Hotspot Detection and Analysis for High-Risk Road Accidents using OPTICS & HDBSCAN Clustering", GeoInformatica, Impact Factor: 2.60.

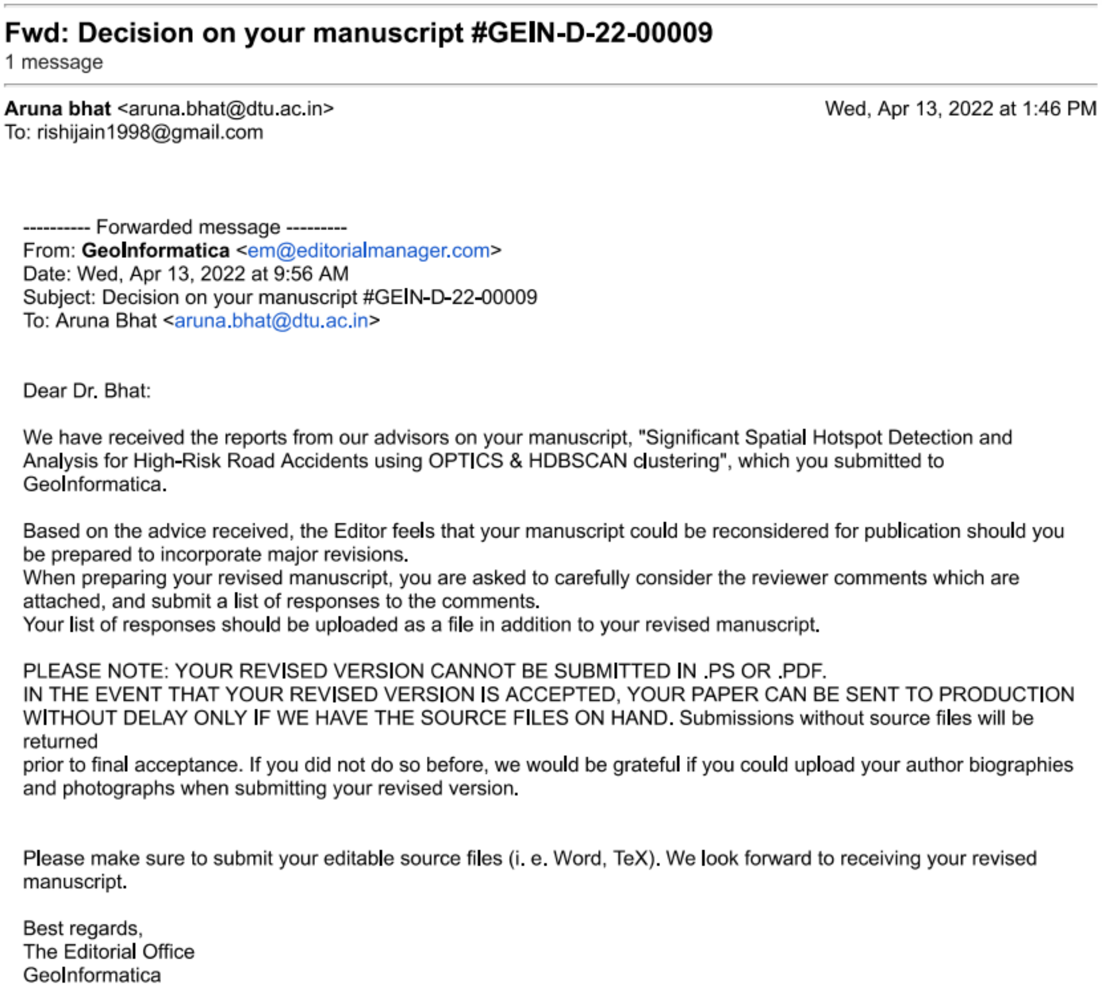
ACCEPTED (In IEEE Scopus Indexed Conference)

[2] R. Jain and A. Bhat, "A Systematic Study on methods of Spatiotemporal Hotspot Detection and Evaluation Metrics", communicated and accepted at 4th IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N-22) to be held on 16th-17th December 2022, Galgotias College of Engineering and Technology at Uttar Pradesh, India.

[3] R. Jain and A. Bhat, "Determining Statistically Significant Road Accident Spatial Hotspots using Machine Learning Approaches", communicated and accepted at 4th IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N-22) to be held on 16th-17th December 2022 at Galgotias College of Engineering and Technology, Uttar Pradesh, India.

[4] R. Jain and A. Bhat, "Survey on Spatiotemporal Hotspots Detection Methods for Road Accidents: A Machine Learning and Deep Learning Perspective", communicated and accepted at International Conference on Bio-Neuro Informatics Models and Algorithms (ICBNA-22) to be held on 21st-22nd June 2022 at SIDTM, Pune, India.

Journal Communication and Revisions Under Review Mails for paper [1].



Fwd: Decision on your manuscript #GEIN-D-22-00009
1 message

Aruna bhat <aruna.bhat@dtu.ac.in>
To: rishijain1998@gmail.com

Wed, Apr 13, 2022 at 1:46 PM

---------- Forwarded message ---------
From: **GeoInformatica** <em@editorialmanager.com>
Date: Wed, Apr 13, 2022 at 9:56 AM
Subject: Decision on your manuscript #GEIN-D-22-00009
To: Aruna Bhat <aruna.bhat@dtu.ac.in>

Dear Dr. Bhat:

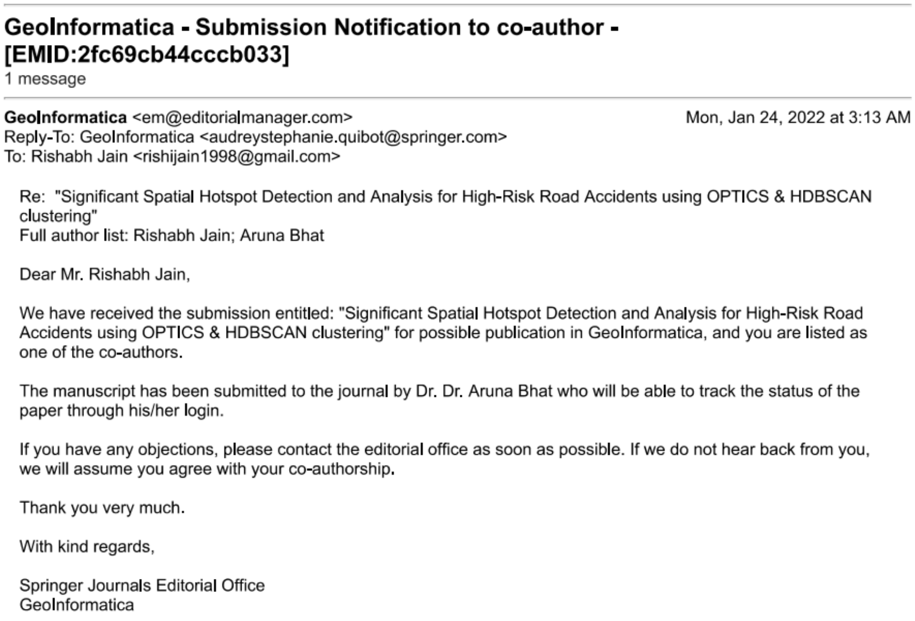
We have received the reports from our advisors on your manuscript, "Significant Spatial Hotspot Detection and Analysis for High-Risk Road Accidents using OPTICS & HDBSCAN clustering", which you submitted to GeoInformatica.

Based on the advice received, the Editor feels that your manuscript could be reconsidered for publication should you be prepared to incorporate major revisions.
When preparing your revised manuscript, you are asked to carefully consider the reviewer comments which are attached, and submit a list of responses to the comments.
Your list of responses should be uploaded as a file in addition to your revised manuscript.

PLEASE NOTE: YOUR REVISED VERSION CANNOT BE SUBMITTED IN .PS OR .PDF.
IN THE EVENT THAT YOUR REVISED VERSION IS ACCEPTED, YOUR PAPER CAN BE SENT TO PRODUCTION WITHOUT DELAY ONLY IF WE HAVE THE SOURCE FILES ON HAND. Submissions without source files will be returned
prior to final acceptance. If you did not do so before, we would be grateful if you could upload your author biographies and photographs when submitting your revised version.

Please make sure to submit your editable source files (i. e. Word, TeX). We look forward to receiving your revised manuscript.

Best regards,
The Editorial Office
GeoInformatica



GeoInformatica - Submission Notification to co-author - [EMID:2fc69cb44cccb033]
1 message

GeoInformatica <em@editorialmanager.com>
Reply-To: GeoInformatica <audreystephanie.quibot@springer.com>
To: Rishabh Jain <rishijain1998@gmail.com>

Mon, Jan 24, 2022 at 3:13 AM

Re: "Significant Spatial Hotspot Detection and Analysis for High-Risk Road Accidents using OPTICS & HDBSCAN clustering"
Full author list: Rishabh Jain; Aruna Bhat

Dear Mr. Rishabh Jain,

We have received the submission entitled: "Significant Spatial Hotspot Detection and Analysis for High-Risk Road Accidents using OPTICS & HDBSCAN clustering" for possible publication in GeoInformatica, and you are listed as one of the co-authors.

The manuscript has been submitted to the journal by Dr. Dr. Aruna Bhat who will be able to track the status of the paper through his/her login.

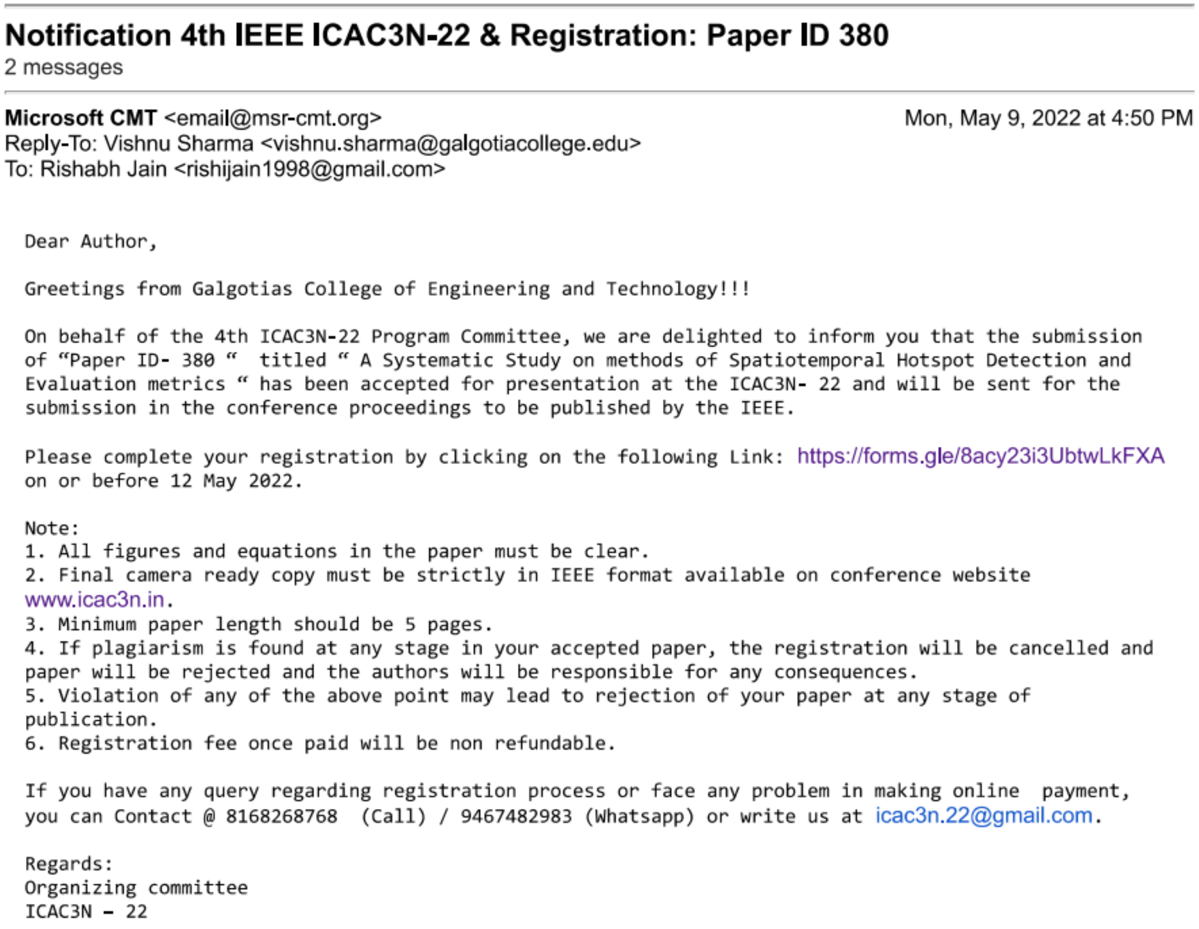
If you have any objections, please contact the editorial office as soon as possible. If we do not hear back from you, we will assume you agree with your co-authorship.

Thank you very much.

With kind regards,

Springer Journals Editorial Office
GeoInformatica

Acceptance Mail and Conference Registration Fee Slip for paper [2].



Notification 4th IEEE ICAC3N-22 & Registration: Paper ID 380
2 messages

Microsoft CMT <email@msr-cmt.org> Mon, May 9, 2022 at 4:50 PM
Reply-To: Vishnu Sharma <vishnu.sharma@galgotiacollege.edu>
To: Rishabh Jain <rishijain1998@gmail.com>

Dear Author,

Greetings from Galgotias College of Engineering and Technology!!!

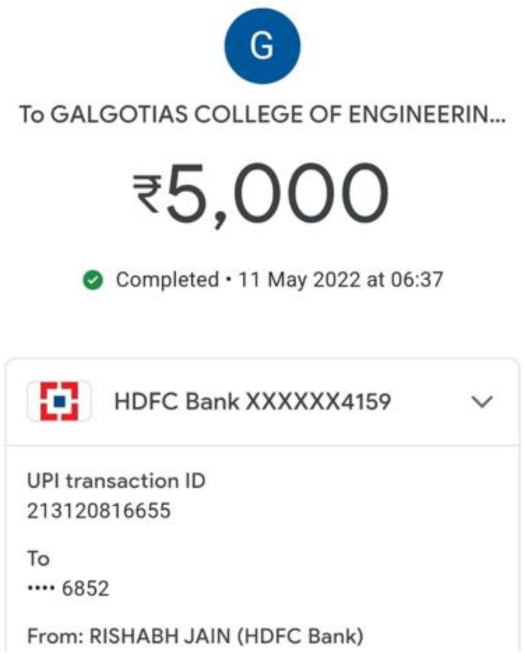
On behalf of the 4th ICAC3N-22 Program Committee, we are delighted to inform you that the submission of “Paper ID- 380 “ titled “ A Systematic Study on methods of Spatiotemporal Hotspot Detection and Evaluation metrics “ has been accepted for presentation at the ICAC3N- 22 and will be sent for the submission in the conference proceedings to be published by the IEEE.

Please complete your registration by clicking on the following Link: https://forms.gle/8acy23i3UbtwLkFXA on or before 12 May 2022.

Note:
1. All figures and equations in the paper must be clear.
2. Final camera ready copy must be strictly in IEEE format available on conference website www.icac3n.in.
3. Minimum paper length should be 5 pages.
4. If plagiarism is found at any stage in your accepted paper, the registration will be cancelled and paper will be rejected and the authors will be responsible for any consequences.
5. Violation of any of the above point may lead to rejection of your paper at any stage of publication.
6. Registration fee once paid will be non refundable.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8168268768 (Call) / 9467482983 (Whatsapp) or write us at icac3n.22@gmail.com.

Regards:
Organizing committee
ICAC3N – 22



To GALGOTIAS COLLEGE OF ENGINEERIN...

₹5,000

✓ Completed • 11 May 2022 at 06:37

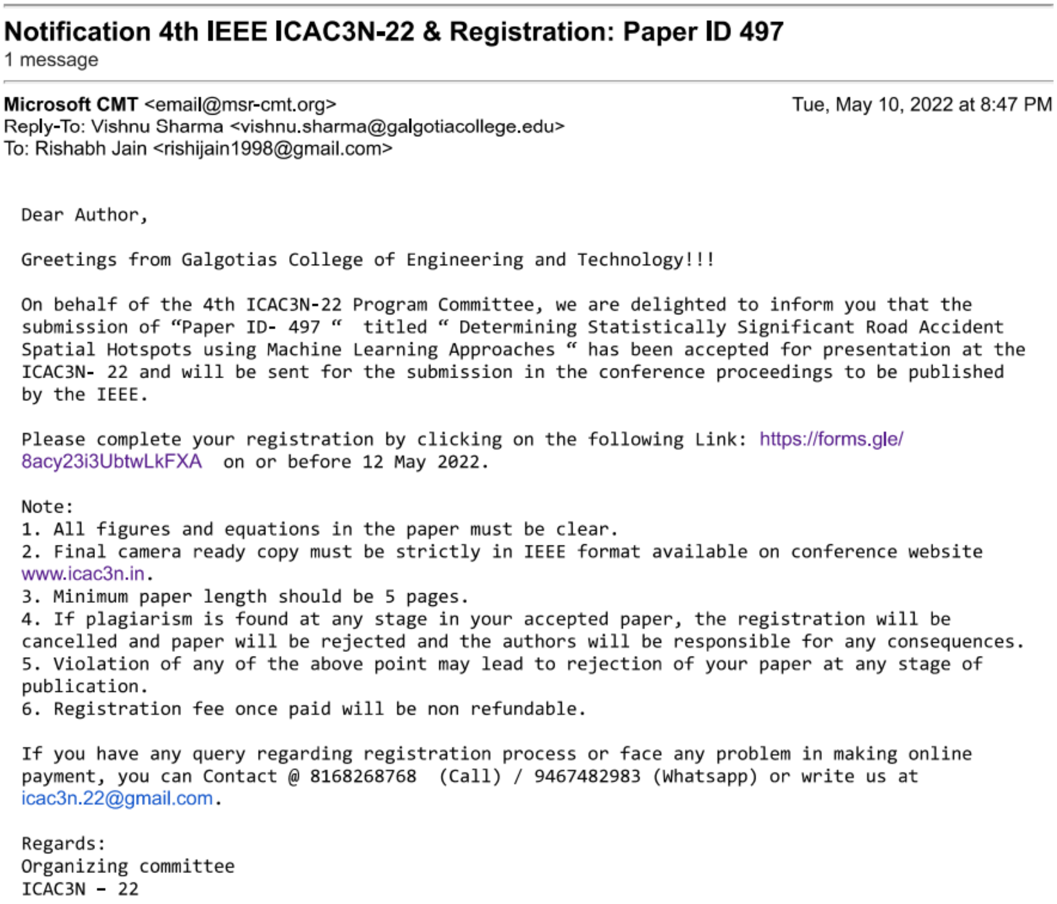
HDFC Bank XXXXXX4159

UPI transaction ID
213120816655

To
•••• 6852

From: RISHABH JAIN (HDFC Bank)

Acceptance Mail and Conference Registration Fee Slip for paper [3].



Notification 4th IEEE ICAC3N-22 & Registration: Paper ID 497
1 message

Microsoft CMT <email@msr-cmt.org>
Reply-To: Vishnu Sharma <vishnu.sharma@galgotiacollege.edu>
To: Rishabh Jain <rishijain1998@gmail.com>

Tue, May 10, 2022 at 8:47 PM

Dear Author,

Greetings from Galgotias College of Engineering and Technology!!!

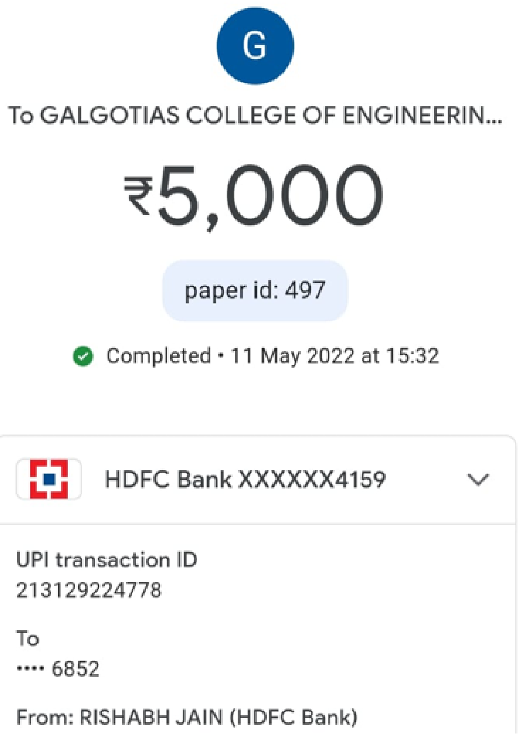
On behalf of the 4th ICAC3N-22 Program Committee, we are delighted to inform you that the submission of “Paper ID- 497 “ titled “ Determining Statistically Significant Road Accident Spatial Hotspots using Machine Learning Approaches “ has been accepted for presentation at the ICAC3N- 22 and will be sent for the submission in the conference proceedings to be published by the IEEE.

Please complete your registration by clicking on the following Link: https://forms.gle/8acy23i3UbtwLkFXA on or before 12 May 2022.

Note:
1. All figures and equations in the paper must be clear.
2. Final camera ready copy must be strictly in IEEE format available on conference website www.icac3n.in.
3. Minimum paper length should be 5 pages.
4. If plagiarism is found at any stage in your accepted paper, the registration will be cancelled and paper will be rejected and the authors will be responsible for any consequences.
5. Violation of any of the above point may lead to rejection of your paper at any stage of publication.
6. Registration fee once paid will be non refundable.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8168268768 (Call) / 9467482983 (Whatsapp) or write us at icac3n.22@gmail.com.

Regards:
Organizing committee
ICAC3N - 22



G

To GALGOTIAS COLLEGE OF ENGINEERIN...

₹5,000

paper id: 497

Completed • 11 May 2022 at 15:32

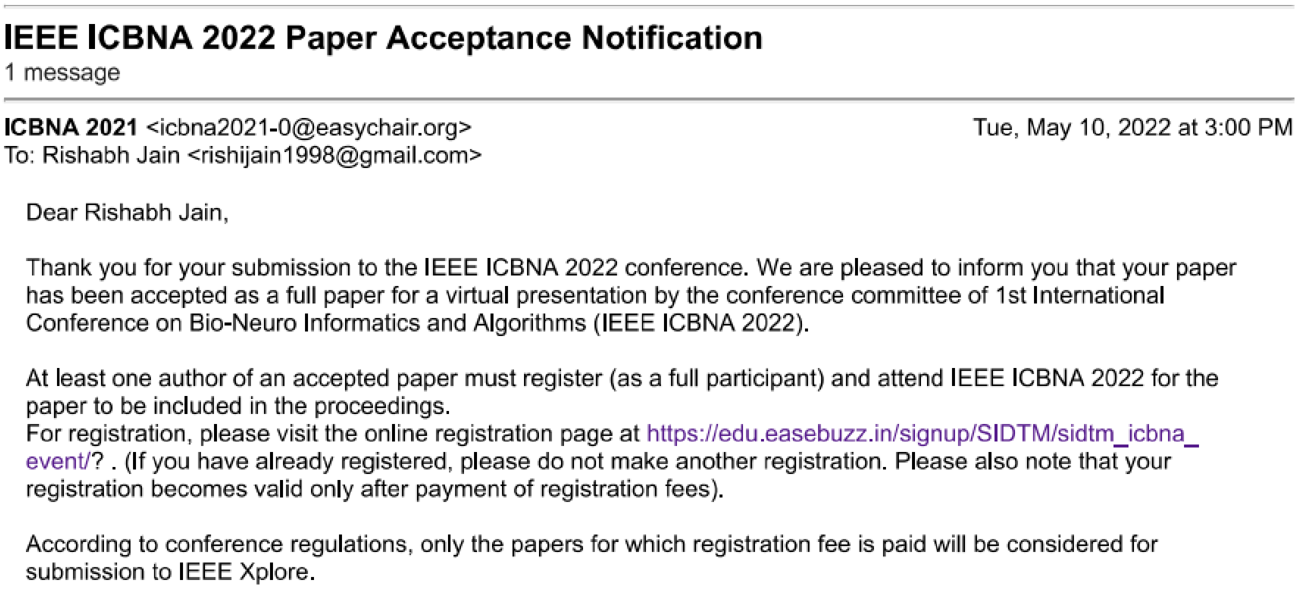
HDFC Bank XXXXXX4159

UPI transaction ID
213129224778

To
•••• 6852

From: RISHABH JAIN (HDFC Bank)

Acceptance Mail and Conference Registration Fee Slip for paper [4].



IEEE ICBNA 2022 Paper Acceptance Notification

1 message

ICBNA 2021 <icbna2021-0@easychair.org>
To: Rishabh Jain <rishijain1998@gmail.com>

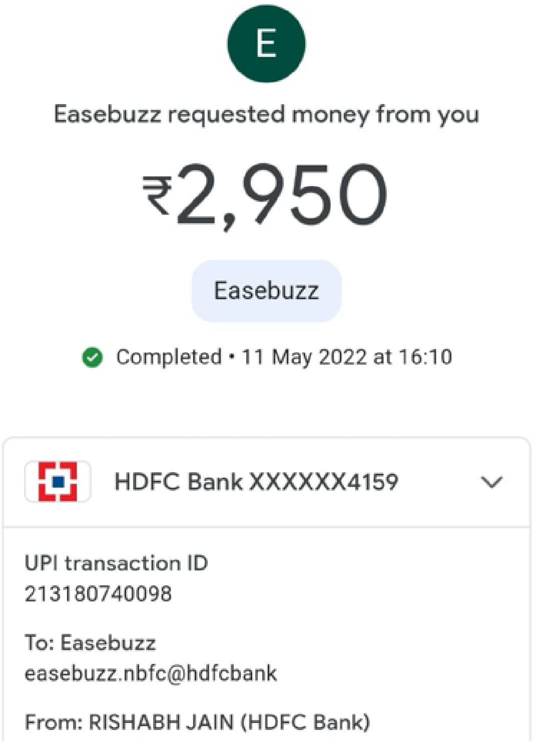
Tue, May 10, 2022 at 3:00 PM

Dear Rishabh Jain,

Thank you for your submission to the IEEE ICBNA 2022 conference. We are pleased to inform you that your paper has been accepted as a full paper for a virtual presentation by the conference committee of 1st International Conference on Bio-Neuro Informatics and Algorithms (IEEE ICBNA 2022).

At least one author of an accepted paper must register (as a full participant) and attend IEEE ICBNA 2022 for the paper to be included in the proceedings.
For registration, please visit the online registration page at https://edu.easebuzz.in/signup/SIDTM/sidtm_icbna_event/? . (If you have already registered, please do not make another registration. Please also note that your registration becomes valid only after payment of registration fees).

According to conference regulations, only the papers for which registration fee is paid will be considered for submission to IEEE Xplore.



E

Easebuzz requested money from you

₹2,950

Easebuzz

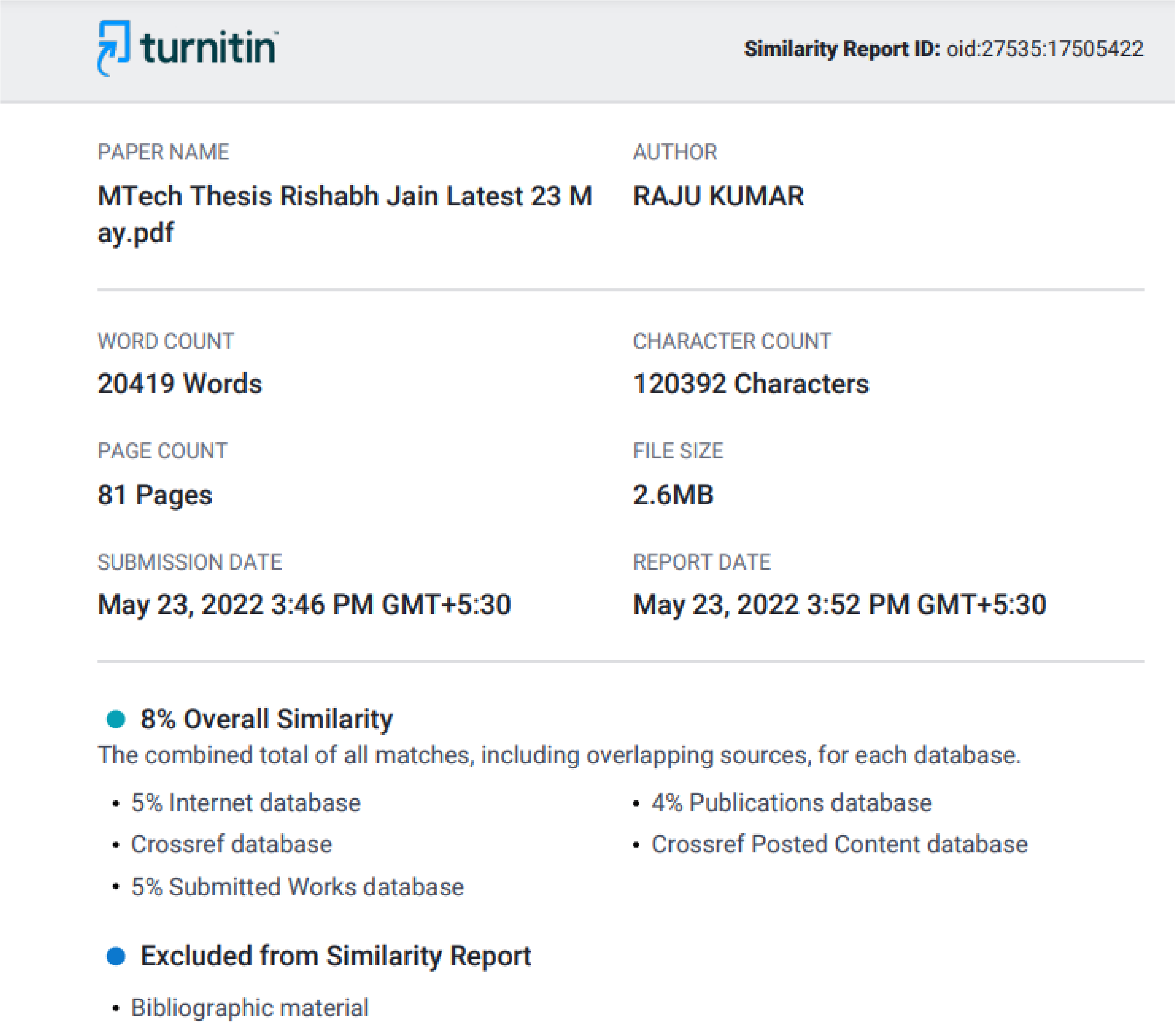
✓ Completed • 11 May 2022 at 16:10

HDFC Bank XXXXXX4159

UPI transaction ID
213180740098

To: Easebuzz
easebuzz.nbfc@hdfcbank

From: RISHABH JAIN (HDFC Bank)



turnitin

Similarity Report ID: oid:27535:17505422

PAPER NAME

MTech Thesis Rishabh Jain Latest 23 May.pdf

AUTHOR

RAJU KUMAR

WORD COUNT

20419 Words

CHARACTER COUNT

120392 Characters

PAGE COUNT

81 Pages

FILE SIZE

2.6MB

SUBMISSION DATE

May 23, 2022 3:46 PM GMT+5:30

REPORT DATE

May 23, 2022 3:52 PM GMT+5:30

● **8% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- Crossref database
- 5% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

● **Excluded from Similarity Report**

- Bibliographic material