ABSTRACT

NIWUNHELLA, DONA HIRUNI HANSINIE. Incident Hotspots Prediction in North Carolina for Effective Incident Management using Deep Learning Techniques. (Under the direction of Dr. Leila Hajibabai).

Traffic incidents have posed a threat to the safety of human life. Losses due to crashes have continued to increase globally. Predicting traffic accidents is a vital requirement which addresses improving safety of travelers, enhancing transportation, and effective routing in transportation systems. Identification of hot spots is the first step to proactively develop traffic safety improvement strategies. Hot spots are mostly considered as points or locations where traffic crashes are concentrated. Prediction of incident hotspots paves the way to be vigilant about potential crashes, reduce the probability of occurrence, use alternative routes, and effectively manage the demands/ requirements of incident management. Thus, this study presents a novel approach to incident hotspots prediction for any given day using deep learning techniques along with a normalized density score. It uses Long Short-Term Memory (LSTM) network method to accommodate the spatiotemporal correlation of hotspots. The model is validated using incident data in North Carolina, from 2017 to 2019. It is further enhanced by using other variables which affect the occurrence of a crash i.e., weather information, holidays, unusual events, and Annual Average Daily Traffic (AADT) Index. The error estimates i.e., Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of the model are then compared with the error estimate values of some common machine learning techniques used in literature proving that proposed model gives a better prediction with a low error value. This model can be used to identify the risky spots for incidents, and proactively reduce the probability of accidents, while optimizing the incident response process.

Incident Hotspots Prediction in North Carolina for Effective Incident Management using Deep Learning Techniques

by
Dona Hiruni Hansinie Niwunhella

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Industrial Engineering

Raleigh, North Carolina
2023

APPROVED BY:

_____	_____
Dr. Leila Hajibabai Committee Chair	Dr. Russell King
_____	_____
Dr. Michael Kay	Dr. Ali Hajbabaie

DEDICATION

I dedicate this thesis to my late father, my mother, my husband, and my brother who continuously supported me to make this journey a success. I am grateful for all the sacrifices they made, and the immense support provided being miles away. I also dedicate this to my advisor Dr. Leila Hajibabai without whom this journey would not have been possible.

BIOGRAPHY

Dona Hiruni Hansinie Niwunhella graduated from University of Kelaniya, Sri Lanka with a B.Sc. in Management and Information Technology, specializing in Business Systems Engineering. She received the highest GPA from the batch and was awarded as the Best Graduating Student in 2018. After graduating she worked as a Data Analyst in one of the leading apparel manufacturing companies in Sri Lanka i.e., MAS Holdings for two years. Later, she joined the University of Kelaniya as a Lecturer since she was passionate in academia and research. After a couple of years, she received the Fulbright Scholarship to pursue the MS in Industrial Engineering at North Carolina State University. After graduating in August 2023, she intends to continue her higher studies to obtain her PhD in Industrial Engineering. Her ultimate goal is to go back to Sri Lanka and serve the less fortunate with the knowledge gained.

ACKNOWLEDGMENTS

I express my heartfelt gratitude to my advisor Dr. Leila Hajibabai for enormously guiding and mentoring me throughout, in conducting this study. Moreover, I sincerely thank Dr. Russell King, Dr. Michael Kay and Dr. Ali Hajbabaie for being in my advisory committee and for continuously supporting me. I am also extremely thankful to my colleague Asya Atik, who collaborated with me and constantly guided me to make this study a success. At last but not least, I am grateful to my family and friends who became the pillar of strength to me through thick and thin.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This chapter provides the background and the motivation for this study, the scope of the research, objectives of the research and justification of the research gap. Finally, it depicts the structure of the thesis, for clear understanding.

1.1. Background and justification of the research problem

Road traffic incidents are non-recurrent, uncertain, and random; thus, they can take place anywhere anytime. They are one of the major concerns of the entire world since they cause enormous losses daily, in terms of lives, property, time and money. Traffic incidents are one of the major causes for non-recurrent traffic congestion, which leads to problematic situations including secondary incidents as well.

According to Global Status Report on Road Safety published by World Health Organization in 2018, deaths from road traffic crashes have increased to 1.35 million a year i.e., nearly 3700 people dying on the world's roads every day. It further states that road traffic incidents are the 8th leading cause of death of people of all ages, and the first cause of death for children and young adults of 5-29 years of age [1].

INRIX 2022 Global Traffic Scorecard states that year over year collisions increased in the US (+4%), the UK (+11%), Germany (+5%), and Canada (+4%) based on incident data [2]. With more drivers on the road, traffic safety remains a top concern. It also states that the US fatality rate (fatalities per 100 million vehicle miles traveled) jumped from 1.07 in 2019 to 1.30 in 2021, a 21% jump, only to drop a marginal 0.03 in the first half of 2022. The rise of incidents around the world justifies that traffic incidents still remain a problem, despite numerous mitigation strategies developed over the years.

Furthermore, according to the US Department of Transportation, National Highway and Traffic Safety Administration 5,250,837 police reported motor vehicle crashes have occurred and at total of 38,824 people have been killed, while 2,282,015 people have been injured in the year 2020. Figure 1.1 depicts the breakdown [3]. Figure 1.2 depicts the national crashes by crash severity from 1975-2020.

POLICE-REPORTED MOTOR VEHICLE CRASHES	
Fatal	35,766
Injury	1,593,390
Property-Damage-Only	3,621,681
Total	**5,250,837**

TRAFFIC CRASH VICTIMS	Killed	Injured
Occupants	**25,536**	**2,093,246**
Drivers	19,519	1,545,689
Passengers	5,966	546,822
Unknown	51	735
Motorcyclists	**5,579**	**82,528**
Nonoccupants	**7,709**	**106,241**
Pedestrians	6,516	54,769
Pedalcyclists	938	38,886
Other/Unknown	255	12,586
Total	**38,824**	**2,282,015**

Figure 1.1 *National statistics of vehicle crashes and victims 2020*

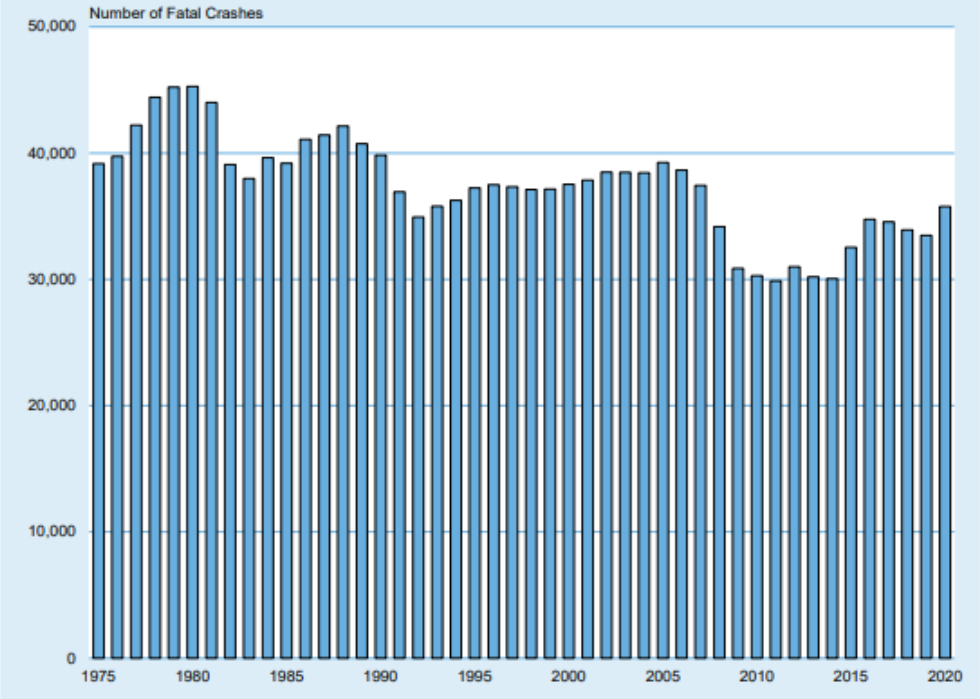


Figure 1.2 *National fatal crashes from 1975-2020*

The above statistics depict that there is no significant decrease in the number of fatal crashes. The reasons could be rapid urbanization, poor safety standards, lack of enforcement, people driving distracted or fatigued, others under the influence of drugs or alcohol, speeding and a failure to wear seatbelts or helmets, etc. According to [1], there is an urgent need to scale up evidence-based interventions and investment, in order to mitigate the risk and reduce the occurrence of road incidents.

This study focuses on identifying a novel solution to address this issue. It is important to manage incidents after they have occurred via incident response management and emergency response optimization. However, the proactive approach is to predict the incident hotspots before they occur, so that incidents can be reduced and better managed when they occur. Thus, this study presents an evaluation of existing predictive models related to incident management and identifying a research gap.

The main focus is incident hotspots prediction which will enable predicting areas where the number of accidents is significantly higher and concentrated than the other areas. This will facilitate incident management in the hotspots effectively by allocating resources. Furthermore, warning mechanisms and incident mitigation techniques can be implemented by predicting the hotspots.

1.2.Scope of the research

This research focuses on incident management in its proactive approach where prediction models are studied in terms of hotspot identification. In incident management there 4 sub arenas namely, 1) incident prediction 2) incident detection 3) resource allocation and 4) emergency response. This research mainly focuses on incident hotspots prediction models.

The results of this study will enable: 1) reduction of the frequency of incidents; 2) better management of emergency response for incidents; 3) implementation of effective warning systems; 4) improve safety; and 5) effective routing.

1.3. Structure of the thesis

The remainder of this thesis is as follows. Chapter 2 provides a summary of the literature reviewed as an interception of the three areas namely, incident management, prediction models and hotspot identification. Subsequently, Chapter 3 elaborates on the research methodology used, where the research approach, research design, and data collection and validation techniques for the study are depicted. Chapter 4 presents the implementation and results of the study. Finally, Chapter 5 concludes the thesis explaining the limitations and future work.

1.4. Summary

This episode presents an introduction to the aforementioned research study, with a clear explanation of its background, scope and rationale. It also discusses the objectives, expected outcomes of the study, and the organization of the thesis.

CHAPTER 2: LITERATURE REVIEW

This chapter presents a systematic review of literature which has been conducted in order to identify the gap in knowledge, in terms of traffic incident hotspots prediction. Studies on incident management, prediction techniques and hotspot identification are studied. This chapter presents the findings of the literature review, overview and the research gap identified.

In identifying the research gap, the following questions are used to approach the review of the literature.

Q1: What are the variables affecting the likelihood/ severity of an incident?

Q2: What are the techniques used for incident prediction?

Q3: What are the categories of incident prediction models?

Q4: What are the existing methods of incident hotspots prediction?

Q5: What are the gaps in current approaches to incident hotspots prediction?

The studies are categorized based on the prediction model categories and each category is explained in terms of the methodologies used. (Figure 2.1)

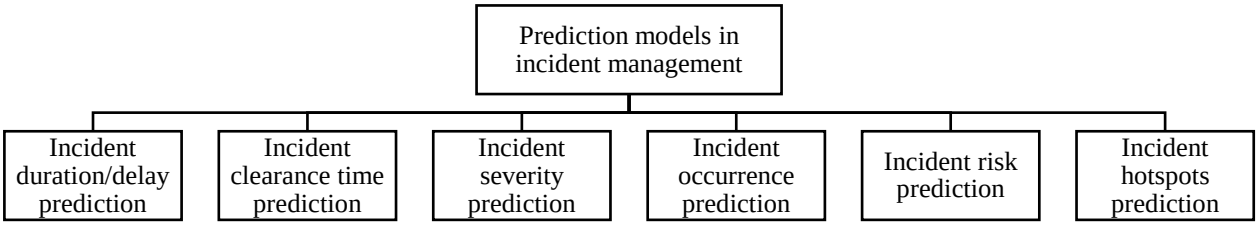


Figure 2.1 *Prediction models in incident management*

2.1.Incident duration/delay prediction models

There are several studies conducted regarding incident duration prediction and the delay caused. Incident duration can be defined as the combination of reporting time, response time and clearance time. Incident delay includes the incident duration and recovery time [4]. Figure 2.2 depicts the process of a traffic incident from incident occurrence to normal operation.

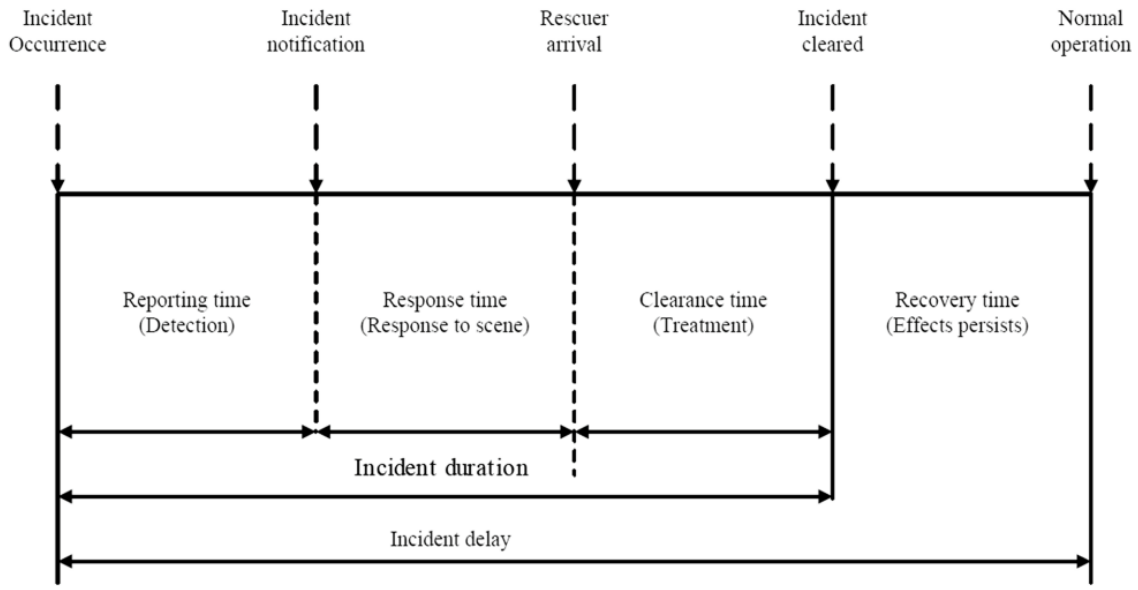


Figure 2.2 *Process of a traffic incident*

A study by Li et al. [5] presents a comprehensive review and discusses the research evolution in terms of incident duration prediction, mainly including the different stages of incident duration, data sources, and methods that are applied in the traffic incident duration influence factor analysis and duration time prediction. It also compares the methodologies used and the accuracy level of each model. It shows that studies are conducted using statistical methods such as regression models to Artificial Neural Networks (ANNs) to predict incident duration. It highlights the importance of capturing all the variables that affect the duration of an incident.

One of the early studies presented by Khattak et al [6] presents a simple time sequential procedure to predict traffic incident duration in freeways, where a series of truncated regression models are for estimation. Another study by Garib et al. [7] provides a detailed study on estimating the magnitude and duration of incident delays. Two multiple regression statistical models are presented to estimate the incident delay and to predict the incident duration. With time, more sophisticated models are developed based on early studies.

Wang et al. [8] propose an incident duration prediction model using Partial Least Squares Regression. The results look promising as a good prediction model. Furthermore, Chung [9] present a log-logistic Accelerated Failure Time (AFT) metric model. This prediction model demonstrates temporal stability suggesting that the model has potential to be used as a basis for making rational diversion and dispatching decisions in the event of an accident. Similarly, an AFT hazard-based model is used by Li [10] to analyze traffic incident duration analysis and predict based on the survival analysis approach. This study considers the unobserved heterogeneity, time-varying covariate and relationship between consecutive traffic incident duration stages. Li et al. [11] propose a mixed model using the multinomial logistic model and parametric hazard-based model to predict incident durations. It also uses a text analysis technique, to process the textual features of the traffic incident to extract time-dependent topics.

Ghosh et al [12] propose a prediction model based on Bayesian Support Vector Regression (BSVR) using traffic incident data from Singapore, which provides error bars as the measurement of uncertainty including the predicted duration of incidents. Some other studies [13], [14] also use Bayesian methods such as Naive Bayesian Classifier, BSVR and Gaussian Process to predict incident duration effectively. A different approach is taken by Kim and Chang [15] where a hybrid model has been used to develop the primary estimation system for freeways, consisting of a Rule-

Based Tree Model (RBTM), Multinomial Logit Model (MNL), and Naïve Bayesian Classifier (NBC). This study identifies some critical relationships between the set of key factors and the resulting incident duration.

Boyles and Waller [16] present a novel approach where analytical incident delay formulae have been extended to consider uncertain incident duration, and simulation with Monte Carlo sampling has been carried out to study scenarios which are too complicated for exact analysis. This study proves that different demand profiles have an impact in determining the effect of an incident and should also be considered in any delay prediction model.

More studies adapting machine learning algorithms are available in the literature. Lin et al [17] develop a model to predict accident duration using the M5P tree algorithm through the construction of a M5P-Hazard Based Duration Model (HBDM) in which the leaves of the M5P tree model are HBDMs instead of linear regression models. Another recent study by Grigorev et al. [18] presents an incident duration prediction model that uses classification algorithm approach in machine learning. This uses a bi-level machine learning framework with outlier removal and intra-extra joint optimization. In addition, Saracoglu and Ozen [4] provide a comparative study of Decision Tree models (CHAID, CART, C4.5 and LMT) to predict incident duration.

Moreover, Zhao and Deng [19] present an approach of prediction using heterogenous ensemble learning with XGBoost, LightGBM, CatBoost, stacking and elastic network. Studies done by Rahmat-Ullah et al. [20] and Hamad et al. [21] use prediction models based on Random Forests which have demonstrated good prediction capability. [20] also present an approach using Artificial Neural Network (ANN) to predict incident duration.

The approach by Valenti et al. [22] predicts and compares five predictive models, ranging from parametric models to non-parametric and neural network models, evaluating their ability of

predicting incident duration. The five models are 1) Multiple linear regression (MLR) 2) Prediction/Decision tree (DT) 3) ANN 4) Support/Relevance Vector Machine (SVM) and 5) K-Nearest-Neighbor (KNN). The results of the study demonstrate good performance in terms of prediction accuracy, especially for incidents with duration less than 90 min.

Several more approaches using neural networks and deep learning algorithms are used to predict incident duration. An interpretation of Bayesian neural networks for predicting the duration of detected incidents is presented by Park et al. [23]. In this study, network parameters are updated using a hybrid Monte Carlo algorithm and a pedagogical rule extraction algorithm (TREPAN) is applied to extract comprehensible representations from the neural networks. Zhu [24] introduces a dynamic prediction model of traffic incident duration for urban expressways based on Long-Short Term Memory (LSTM) and Multi-Layer Perceptron (MLP) with better prediction results. Furthermore, a study by Li et al. [25] embraces a deep fusion model based on Restricted Boltzmann Machines (RBM) for traffic accident duration prediction. This study states that the spatial–temporal correlations of traffic flow had been ignored in previous studies, and thus the authors incorporate it, in addition to the characteristics of traffic accidents. In this model, a stacked RBM is used to handle the categorical variables, a stacked Gaussian-Bernoulli RBM is used to handle the continuous variables, and a joint layer has been used to fuse the extracted features.

It is noted that there are many studies done to predict incident duration ranging from statistical methods to advanced deep learning techniques.

2.2. Incident clearance time prediction

There are some studies in literature that focus on incident clearance time prediction which is a component of incident duration. As shown in Figure 2.1. clearance time is the third component of incident duration.

Tang et al. [26] provide a detailed study on literature related to statistical and machine-learning methods for clearance time prediction of road incidents. It compares many techniques that are used to predict incident clearance time i.e., statistical models such as Accelerated Failure Time (AFT) model, Quantile Regression (QR) model, Finite Mixture (FM) model, and Random Parameters Hazard-Based Duration (RPHD) model, and machine learning models: K-Nearest Neighbor (KNN) model, Support Vector Machine (SVM) model, Back Propagation Neural Network (BPNN) model, and Random Forest (RF). This study tests all these methods using a dataset and evaluates the performance. The authors state that machine learning methods perform stably in model prediction relative to the statistical methods.

Zhan et al. [27] also present a prediction model to estimate freeway incident lane clearance times using M5P Tree Algorithm. Comparison results of this study demonstrates better prediction results than the traditional regression and decision tree models. Moreover, Ozbay and Noyan [28] present a successful study on prediction of clearance times using Bayesian Networks.

It is noted that there are significant studies done to predict incident duration ranging from statistical methods to advanced deep learning techniques, similar to incident duration prediction models.

2.3. Incident severity prediction

Incident severity is a measurement of the impact an incident has on people, resources, environment, and any other component that is affected by an incident. Several studies are conducted by authors to review about past literature related to incident severity prediction including [29], [30], [31], and [32]. These studies discuss the different methodologies and accuracy of severity prediction models that are in literature ranging from statistical methods to deep learning techniques.

Wang et al. [33] use two-stage mixed multivariate model which combines both accident frequency and severity models. Thus, this study is connected to section 2.4. of this thesis too. A Bayesian spatial model and a mixed logit model are employed at each stage for accident frequency and severity analysis respectively, and the results are combined to produce estimation of the number of accidents at different severity levels. Based on the results from the two-stage model, the accident hotspots are identified which is related to section 2.6. of this thesis. The model is compared with other statistical models to depict that it is able to predict low frequency accidents (such as fatal accidents) better than others.

Studies are conducted using ANN to predict severity of accidents. Umar and Gokcekus [34] use ANN to model fatality and injury index in Nigeria. Furthermore, another study by Alkheder et al. [35] use WEKA (Waikato Environment for Knowledge Analysis) data-mining software to build the ANN classifier, and thereby predict the severity of accidents. Moreover, Shaik et al. [32] present a review on neural network techniques for prediction of accident severity. It states that deep learning methods such as the Recurrent Neural Network (RNN) and the Convolutional Neural Network (CNN) have recently been successfully used for the prediction of road accidents and demonstrate their high accuracy and efficiency.

2.4. Incident occurrence/ frequency prediction

There are several past studies conducted in order to predict the probability or the frequency of accident occurrence. Some studies that have conducted detailed review on existing literature related to frequency/ occurrence of traffic accidents include [29], [30], and [31]. The advanced techniques of predictions ranging from machine learning to deep learning techniques are discussed in [31] . A representation of the aforementioned techniques for prediction is depicted in Figure 2.3.

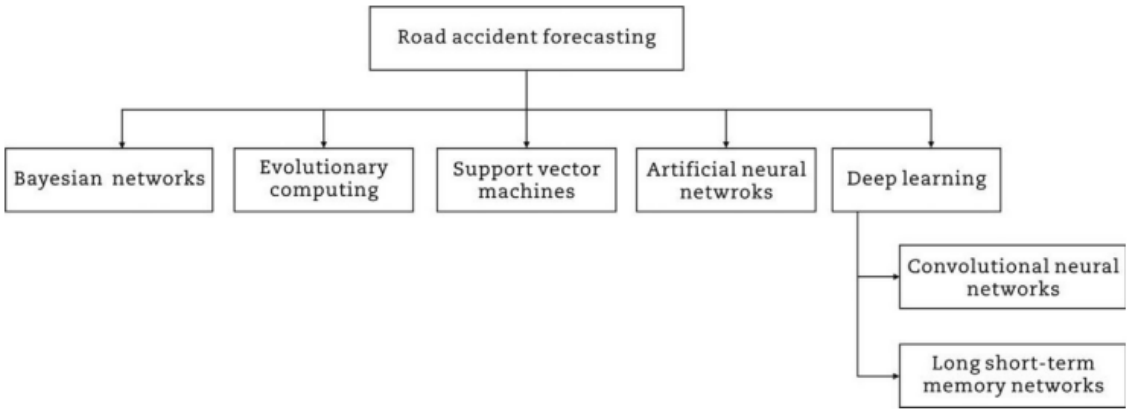


Figure 2.3 *Representative algorithms and methods for incident occurrence prediction*

Accident frequency is predicted in [33] as discussed in section 2.2 above, using a Bayesian special model. Furthermore, Park and Haghani [36] introduce a real time prediction model to predict secondary incidents' frequencies using vehicle probe data. This study achieves the need of quantifying non-recurring congestion and detecting a secondary incident under the adverse influence of a primary incident, using Bayesian structure equation model.

Gutierrez-Osorio et al. [37] use an ensemble model to predict the probability of traffic accidents using social media data. This study uses information gathered from social media and open data, applying an ensemble Deep Learning Model, composed of Gated Recurrent Units and CNN. Zhao et al. [38] use CNN to predict traffic accident possibility. This deep learning algorithm extracts autonomous features from data collected in Vehicular Ad-hoc Network (VANET).

Another study by Ogwueleka et al. [39] use ANN model for the analysis and prediction of accident rates in a developing country using the number of vehicles, accidents, and population as model parameters. Another approach is used by Feng et al. [40] using LSTM to predict the number of traffic accidents in the future as a part of UK Traffic accidents analysis. This study is based on clustering the existing accident incident data (with no prediction) in a map to depict hotspots and visualizing accident attributes to find related causes. Moreover, Roland et al. [41] present a model to predict the vehicle accident occurrence a using a Multilayer Perceptron (MLP) model to predict where accident hotspots are for any given day in the city of Chattanooga, TN. This study is related to section 2.6 too. Similarly, [42] uses Hetero-ConvLSTM to predict incidents based on the count of accidents and presents a hotspot analysis based on the counts of accidents in the grid of the map. (Section 2.6)

2.5. Incident risk prediction

Traffic accident incident risk involves both severity and frequency of occurrence. These prediction models determine the risk of individual incident based on the variables of interest. There are undoubtedly many studies in literature conducted to predict the risk of an accident, based on **traffic accident data, real-time traffic data, social media data and other variables** (such as weather, road network attributes, population data) that affect the risk of an accident. A study [31] presents a detailed review on studies about incident risk prediction. It also presents different data sources and techniques for analysis of road accidents. This is one of the best studies to receive information on techniques used for prediction models.

Huang et al. [43] introduce a highway crash detection and risk estimation model using deep learning. This study uses CNN, with real-world traffic data used to design feature set for the model.

It also considers volume, speed and sensor and occupancy data collected from roadside radar sensors.

Lin et al. [44] present a derived novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. This is done using k-nearest neighbor model and a Bayesian network where a new Frequent Pattern tree (FP tree) based variable selection method has been developed and validated. Another study done by Wang et al. [45] introduce a new technique named GSNet developed based on spatial-temporal correlations from geographical and semantic aspects. Soto et al. [46] present a model using ANN to predict accident risk prediction, with successful results.

LSTM is also a widely used technique to predict incident risk. Ren et al [47] use LSTM along with an invented spatiotemporal correlation to predict the risk of traffic incidents using incident data. However, this study does not consider other factors affecting the incidents but only the incident data. In contrast, a study conducted by Bao et al. [48] use multi-source data including crash data, large-scale taxi GPS data, road network attributes, land use features, population data and weather data for the prediction. It also uses a spatiotemporal deep learning approach for citywide short-term crash risk prediction using a merged approach of Convolutional Long Short-Term Memory (ConvLSTM) network, CNN and LSTM together, which has depicted high accuracy. Studies such as [47] and [48] also approach incident hotspot mapping which will be further discussed in section 2.6. of this thesis. Moosavi et al. [49] also use LSTM to produce a deep neural-network model which has used a variety of data attributes such as traffic events, weather data, points-of-interest, and time. It incorporates multiple components including a recurrent, a fully connected, and a trainable embedding component. Furthermore Chen et al. [50]

use traffic big data to develop a Stack Denoising Convolutional Autoencoder (SDCAE) model for accident risk prediction. It inherits the methodologies of CNN to predict the risk.

It is noted that deep learning techniques have been used effectively to predict traffic incident risk.

2.6. Incident hotspots prediction

Crash hotspots, black spots, or accident-prone locations are defined as hazardous road sites at which crashes often occur and to which must to be given priority [51]. They are road sites at which crashes are concentrated. However, according to literature, unfortunately, there is no universally agreeable method and definition of hot spots [52]. According to [52] the methods used to identify hot spots are also different from country to country. For instance, if a road accident rate exceeds an empirical critical crash rate in India, then the location is classified as a hot spot. In Austria, Belgium, Germany, Hungary, Norway, Switzerland, and Vietnam, accident numbers are used to define the hot spot areas. Denmark, Portugal, and USA use model-based approaches for hot spot identification. Turkey applies crash frequency, rate, and severity methods concurrently.

There are several studies conducted to identify hotspots. Wang et al. [51] discuss about three techniques for Hot Spot Identification (HSID). These techniques are;

1. Crash Frequency Method (CFM) – (widely used) Hazardous sites are ranked in ascending order on the basis of the crash counts.
2. Societal risk-based crash method – Combination of the crash severity and the societal monetary loss, this method is developed on the basis of an estimate of the amounts that individuals are prepared to pay to reduce risk to their lives.
3. Empirical Bayesian method – It integrates a predictive crash model with the recorded crash history to achieve high consistency of results.

Moreover, [51] use the aforementioned Empirical Bayesian method to a case study with crash data, predicting the hotspots. Qu and Meng [53] use societal risk-based simple ranking and empirical Bayesian methods to identify the hotspots. Vadlamani et al. [54] conduct a study focused on large truck accidents. It identifies large truck hot spots in Arizona using negative binomial regression with long crash histories and a newly proposed method using property damage only equivalents (PDOE). Fawcett et al. [55] propose an extension to the classical Bayesian hierarchical model for predicting accident counts in future years at sites within a group of potential road safety hotspots. A recent study by Zarei et al. [56] propose a new non-parametric empirical Bayes approach called CGAN-EB to predict incident hotspots. It also uses negative binomial model to model the crash data.

Krueger et al. [57] develop a new spatial count data model to identify hotspots. It uses a new spatial negative binomial model with Bayesian additive regression trees to endogenously select the specification of the link function. The proposed model is tested using a crash count data set from a metropolitan highway network. Furthermore, Azari et al. [58] present a Geographic Information System (GIS) based approach for accident hotspots mapping in mountain roads using seasonal and geometric indicators. Another study by He et al. [59] introduces a method to generate high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories. This methodology incorporates only the spatial correlation but not the temporal correlation.

As discussed in section 2.3, [33] also presents a method to predict hotspots using the two-stage model explained earlier. Another study by Yu et al. [60] provides an analysis comparing different spatial analysis methods for hotspot identification. These methods include CFM, crash rate method, Empirical Bayes method, local spatial autocorrelation method and Kernel Density Estimation (KDE) method. The KDE method assesses the risk of crashes at a spatial unit given the

crash counts at neighboring spatial units. Moreover, Szénási and Csiba [61] use a modified method using DBSCAN clustering to identify incident black spots.

Some studies use the predicted accident risk using deep learning techniques (as stated in section 2.5 above) to demonstrate a heat map such as [41], [42], [47] and [48].

Machine learning approaches are also used by some studies to predict accident hotspots. Santos et al [62] present a study that has used machine learning techniques to predict hotspots. It uses a clustering approach first (using DBSCAN and hierarchical clustering) and then moves to predict the hotspots using machine learning techniques such as Decision Trees (DT), Random Forests (RF), Logistic Regression (LR), and Naive Bayes (NB). Another study by Lu et al. [63] incorporates Logistic Regression to predict the hotspots using a traffic accident hotspots prediction odds ratio model equation. In this study, location of car in road transects, road safety grade, road surface condition, visual condition, vehicle condition, and driver state are considered as the independent variables and traffic accident hotspot as the dependent variable. Atumo et al. [52] present a model based on spatial statistics (Getis-Ord statistics) and Random Forest to predict incident hotspots. Moreover, Xu and Tao [64] use a clustering ensemble model to identify hotspots. This shows capability to analyze and quantify safety levels of different roads, to obtain principal components and to carry out clustering classification for comprehensive evaluation function of principle components through Canopy-K-means ensemble clustering algorithm to identify accident hotspots.

It is noted that there are only a few studies dedicated to identify and predict incident hotspots using deep learning techniques.

2.7. Overview of literature and research gap identification

Table 2.1. provides an overview of the studies based on the methodology used for prediction. A detailed study was done about prediction models for incident hotspots.

It is stated in [52] that limited efforts are observed in using machine learning approaches to identify traffic crash hot spots. It is noted that studies incorporating deep learning techniques to predict hotspots are minimal. Some studies use deep learning techniques to predict the risk/severity and then use HSID techniques to demonstrate the hotspots. However, very few studies are dedicated solely with the objective of predicting incident hotspots using deep learning. Thus, this study will contribute to literature by

1. Providing a systematic review on literature related to methodologies of prediction models in incident management i.e., incident duration/delay prediction, incident clearance time prediction, incident severity prediction, incident occurrence/ frequency prediction, incident risk prediction, and incident hotspots prediction.
2. Introducing a deep learning-based model dedicated to predict incident hotspots (based on a normalized density score)

The study provides a solution facilitating 1) effective incident management and emergency response 2) better allocation of resources 3) accident warning mechanisms 4) avoiding traffic accidents by choosing safer regions 5) better routing 6) improving safety.

Table 2.1 *Overview of relevant literature*

Methodology	Prediction Model Description	Incident duration/ delay prediction	Incident clearance time prediction	Incident severity prediction	Incident occurrence prediction	Incident risk prediction	Incident hotspots prediction
Review studies	Comparison of models	[5]	[26]	[29], [30], [31], [32]	[29], [30], [31]	[31]	
Statistical methods	Time sequential procedure	[6]					
	Regression models	[7], [8], [22]					[54]
	Accelerated Failure Time (AFT) metrics	[9], [10]	[26]				
	Mixture model (Multinomial logistic model/ Parametric hazard-based model	[11]	[26]				
	Analytical stochastic formulae	[16]					
	Bayesian prediction (BSVR/ NBC, Gaussian Process)	[12], [13], [14], [15]			[33], [36]		[33]
	Multinomial/ Mixed Logit Model (MNL)	[15]		[33]			[33]
	Rule based Tree Model (RBTM)	[15]					
	Finite Mixture Model		[26]				
	Spatial-Temporal Geographical and Semantic Model					[45]	
	Negative binomial model						[56], [57]
	Bayesian additive regression trees						[57]
	Spatial Statistics (Getis-Ord GI)						[52]
Simulation	Monte Carlo Sampling	[16]					
Machine learning	M5P tree algorithm	[17]	[27]				
	Decision tree/ Classification	[4], [18], [22]					[62]
	Support Vector Machine (SVM)	[22]	[26]		[31]		
	Ensemble learning	[19]			[37]		[64]
	Random Forest (RF)	[20], [21]	[26]				[62], [52]
	Logistic Regression (LR)						[62], [63]

Table 2.1 (continued).

Machine Learning (cont.)	Naïve Bayes (NB)						[62]
	Gradient boosting (XGBoost)	[19]					
	K-Nearest Neighbor (KNN)	[22]	[26]			[44]	
	K-means clustering						
Neural Networks and Deep learning	Artificial Neural Network (ANN)	[20], [22]		[34], [35]	[31], [39]	[46]	
	BPNN		[26]				
	Bayesian Neural Network	[23]	[28]		[31]	[44]	
	Convolutional Neural Network (CNN)			[32]	[31], [37], [38]	[43], [50], [48]	
	Multi-Layer Perceptron (MLP)	[24]		[32]	[41]		
	Recurrent Neural Network (RNN)			[32]			
	Long-Short Term Memory (LSTM)	[24]			[31], [40]	[47], [48], [49]	This work
	Convoluted LSTM (ConvLSTM)				[42]	[48]	
	Deep fusion model based on RBM	[25]					
Hotspot Identification (HSID) and prediction/ Spatial analysis methods	Heat maps based on the predicted risk/ severity						[41], [42], [47], [48]
	Crash Frequency Method (CFM)/ Crash Rate Method						[51], [60]
	Societal risk-based crash method						[51]
	Empirical Bayesian method						[51], [53], [55], [56], [60]
	Local spatial autocorrelation method						[60]
	Kernel Density Estimation Method (KDE)						[60]
	Property Damage Only Equivalents (PDOE)						[54]
	GIS based approach/ Coarse resolution						[58], [59]
	DBSCAN						[62], [61]
	Hierarchical clustering						[62]

2.8. Summary

Chapter 2 provides a systematic review of literature related to prediction models of incident management. Through this the research gap is justified, providing an overview of the studied literature.

CHAPTER 3: METHODOLOGY

Chapter 3 presents the methodology followed in the research including the design of the model, data collection, data pre-processing and the deep learning model details used for the incident hotspots prediction.

3.1. Model design

This research uses traffic incident data to predict the incident hotspots. Figure 3.1 depicts the model design of this study, where heterogenous data is used to predict the hotspots.

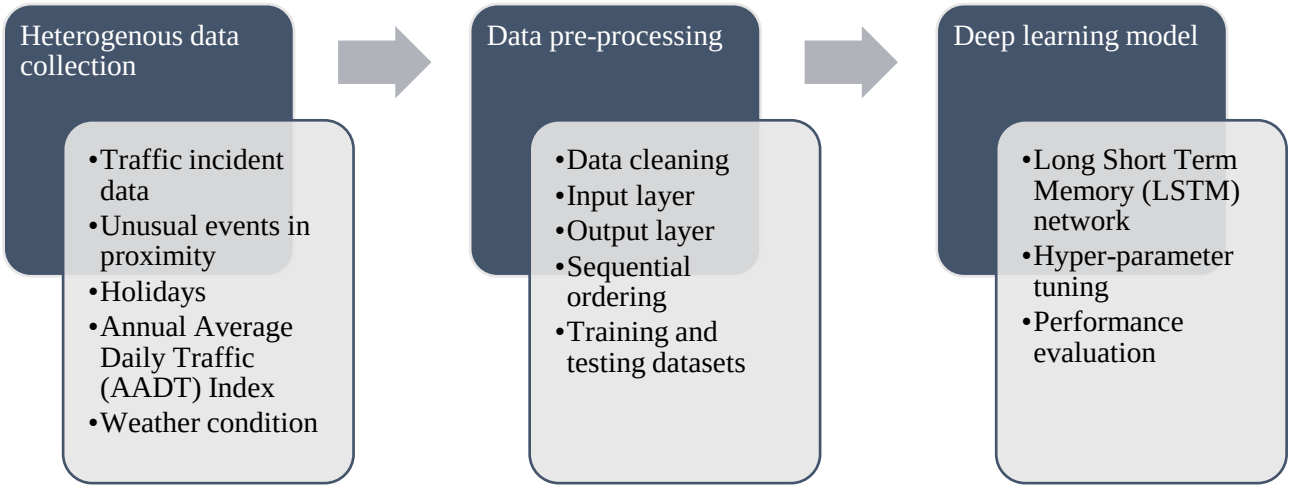


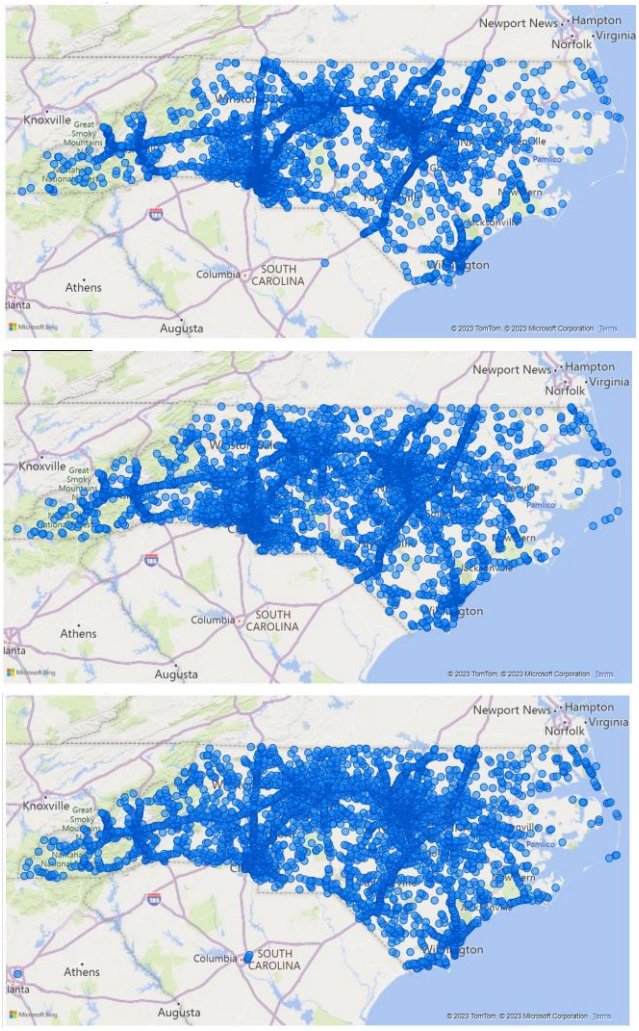
Figure 3.1 *Model design*

3.2. Heterogenous data collection

The model uses heterogeneous data related to traffic accidents. Relevant factors are decided based on past literature. Details of the data collected are as follows.

1. Incident data of North Carolina from 2017 to 2019 acquired from the North Carolina Department of Transportation (NCDOT), which includes the date of the incident, start time of the incident, location of incident in terms of latitude/longitude and the

occurrence of an unusual event in proximity (43,361 incidents). Figure 3.2 presents the locations of incidents in North Carolina from 2017-2019 based on the dataset.



Incident spots of 2017 in NC

Incident spots of 2018 in NC

Incident spots of 2019 in NC

Figure 3.2 *Incident spots in North Carolina (2017-2019)*

2. Holiday information acquired from the Holiday API by Abstract - The API Company including national, and state holidays (based on location - North Carolina) [65].
3. Annual Average Daily Traffic (AADT) Index from 2017 to 2019 (Total volume of vehicle traffic on a highway or a road for a year, divided by 365 days) acquired from NCDOT [66].

4. Weather condition of the location of incident acquired from Weatherbit API which provides the data related to precipitation levels, snow, clouds, temperature, and speed of the wind [67]. Refer to Appendix A for snapshots of datasets.

3.3. Data pre-processing

The proposed model pre-processes the heterogenous data (in section 3.2) before the model development.

3.3.1. Data cleaning

The study involves cleaning the data and removing outliers. The dataset contains incident data with irrelevant dates (before 2017 – 35 records) and invalid latitude and longitude values (13 records). The unwanted data fields are removed and cleaned. Furthermore, the holiday information provides data of several categories of holidays. The study focuses only on national and state holidays, which are filtered based on the location (North Carolina) and obtained from the data. Moreover, the AADT index is matched with each route of the incident, which is incompatible in the original format of the data. This issue is rectified by manual mapping of each route of the index dataset to the incident data. The weather data is also cleaned in terms of irrelevant data fields and proper mapping of variables as required.

3.3.2. Input layer

Based on the data sources, the model utilizes the following input variables derived from the original data after data cleaning. This input shape is utilized by the training dataset to define the output.

1. Date of the incident **(D)** – categorized to season, month, date.
2. Start time of the incident **(S)** – categorized to the 24 hours of the day.
3. Presence/absence of unusual events in proximity (True/False) **(E)**

4. Holiday (True/False) **(H)**
5. AADT Index **(I)**
6. Weather condition **(W)** (Precipitation, snow, clouds, temperature, wind speed)
7. Block **(B)** – This input variable depicts the area of each incident and is defined using the following steps.
    a) Determine the latitude and longitude boundaries of the state of North Carolina. (Longitude range = -84.5 to -75.5, Latitude range = 36.7 to 33.7)
    b) Develop a grid with square blocks with equal area (as an array of blocks (B) which is numbered 0 to n), tangent to the boundaries of the entire state.
    c) Assign each incident to a block (B) based on the latitude and the longitude of the exact location of the incident.
    d) The area of each square block is changed to improve accuracy of the model later, via sensitivity analysis. (Section 4.4).

A set of 108 blocks defined in a 18x6 grid for the North Carolina state map based on the steps above is depicted in Figure 3.3 which is used for the study to gain predictions initially.

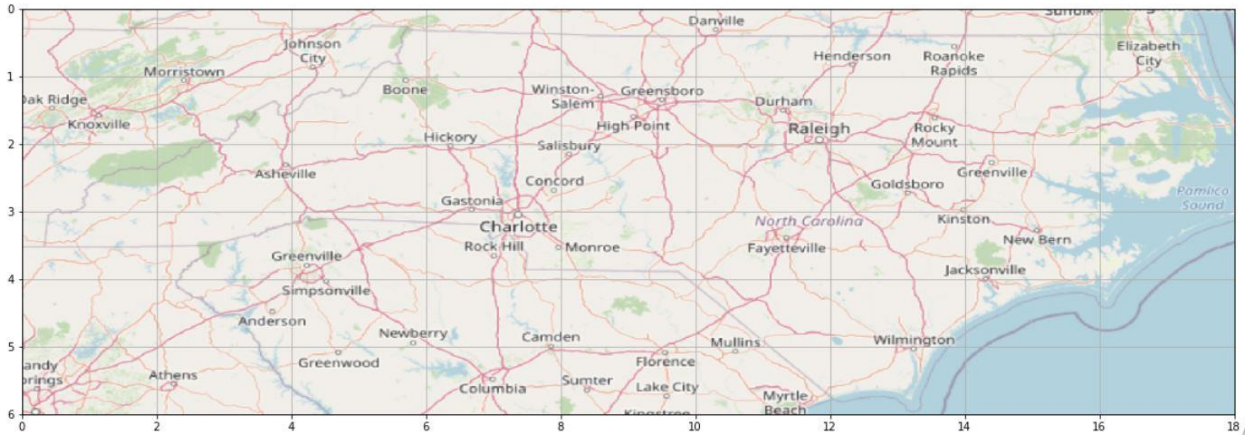


Figure 3.3 *18x6 grid with 108 blocks*

Once the model extracts the above input variables, it converts the data to a processable format for deep learning. All the above categorical variables (D, S, E, H and B) are subjected to **one-hot encoding** which is the numerical conversion of categorical variables into a format that can be fed into deep learning algorithms to improve prediction accuracy. This approach creates a new column for each unique value in the original category column. The zeros and ones are subsequently put in these dummy variables (1 - True, 0 - False).

All the above numerical variables (I and W) go through **min-max normalization** (min-max feature scaling) which is a common way of normalizing numerical data to bring the features to comparable ranges. It includes feature scaling and linear transformation of data ranging from 0 to 1. For every feature, the minimum value of that feature gets transformed into 0, the maximum value gets transformed into a 1, and all the other values are transformed to a value between 0 and 1 linearly as depicted in the equation (1) below forming the scaled variable (x_{scaled}).

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

where $x_{min}$is the minimum value of the feature, x_{max} is the maximum value of the feature, and x is the value of interest.

3.3.3. Output layer

The output layer in this model is defined as a normalized density score based on the incident data available including the frequency, date, and the location. Since the dataset does not provide instances which are not accidents, the model does not have the capability of calculating the probability of incident occurrence. Thus, a score is defined based on the data available which depicts the concentration of incidents based on min-max normalization of count of incidents per

block per date. Min-max normalization of the score allows comparison of the concentration on a linear scale from 0 to 1. This assists in identifying the hotspots effectively.

The model assigns each record of the incident dataset to the blocks (B) based on the location (latitude and longitude) and its date (D) and receives the total number of incidents per date per block. In this way, it obtains a normalized density score of an incident occurrence via min-max normalization (explained in section 3.3.2 above). An example data row with this score is shown in Table 3.1 below.

Table 3.1 *Output layer*

Date of Incident (D)	Block (B)	No of incidents in the block on each date	Normalized density score
4/1/2017	30	13	0.315789

3.3.4. Sequential ordering

Since the model uses the Long-Short Term Memory network (LSTM) which is discussed later in section 3.4.2. the dataset is arranged in a sequential manner. Depicting the periodic pattern of incidents, the dataset is grouped in the order of blocks (B) first and then ordered from the earliest to the latest date within each block. This reflects the spatiotemporal correlation of the incidents since it considers both the spatial arrangement and time of accidents.

3.3.5. Training and testing datasets

Once the model goes through data preprocessing, the dataset is divided into a training dataset (70% of data 1/1/2017 to 2/5/2019)) and a testing dataset (30% of data - 2/6/2019 to 12/31/2019). The training dataset is the subset of the actual dataset that is fed into the deep learning model to discover and learn patterns, training the model. The testing dataset is the subset with

unseen data to test the model, evaluating the performance of the algorithm. While maintaining the chronological sequence of the ordered data covering all blocks, suitable for the LSTM model (as described in section 3.3.4 above, this model separates the training and the testing datasets prior to feeding to the deep learning model.

3.4. Deep learning model

3.4.1. Long Short-Term Memory (LSTM) network

Neural networks are designed to simulate the behavior of human brains, detecting various non-linear relationships in the data. Deep learning with neural networks is a popular machine learning technique that simulates the mechanism of learning in biological organisms [68]. Deep learning is based on a set of algorithms that try to model high level abstractions in data. When the input layer receives an input, it passes on a modified version of the input to the next layer. In a deep network, there are many layers between the input and output allowing the algorithm to use multiple processing layers, composed of multiple linear and non-linear transformations.

The deep learning model in this study utilizes the algorithm of Long Short-Term Memory (LSTM) network. The reason for using LSTM network is because of its ability to capture the periodic nature of traffic incidents along with the spatial proximity of the blocks (B), utilizing the sequence of data in order. It can solve the long-term dependencies, without being subjected to intrinsic difficulties in training unlike RNNs [69]. LSTM also handles the issue of vanishing/exploding gradient problem present in traditional RNNs, since LSTM has an explicit memory cell as the cell state (long term memory) without biases or weights to modify it. It allows backpropagation of the error through time and layers preserving the information over time as required. It has feedback connections allowing it to process entire sequences of data, and not only

single data points. The LSTM model consists of the input layer, hidden layer(s) and the output layer. The deep model of this study is depicted in Figure 3.4.

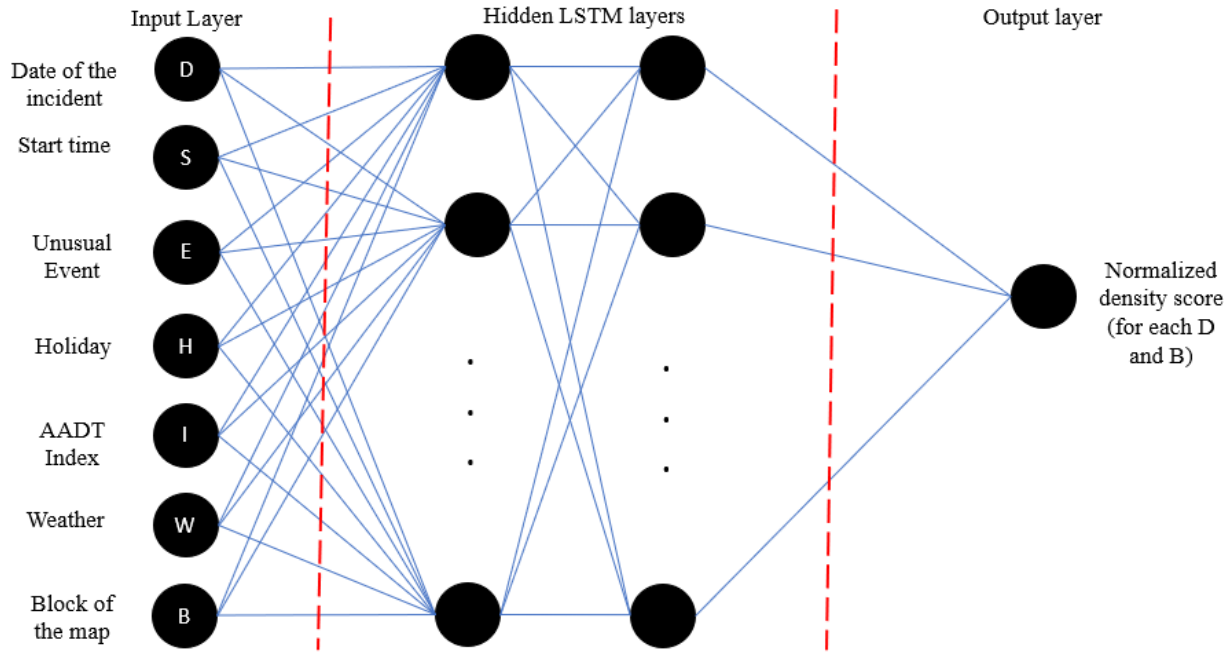


Figure 3.4 *Deep model using the LSTM network.*

The fundamental principle of LSTM lies in the hidden layer(s). The LSTM unit consists of a cell, an input gate, a forget gate and an output gate [69], which give it the ability to remember the past over arbitrary time intervals. The three gates control the flow of information connected with the cell. Figure 3.5 depicts the process within the LSTM cell [69].

The activation function of a node defines the output of that node given an input or set of inputs (Figure 3.6). Gates in LSTM goes through sigmoid activation function which outputs a value between 0 or 1 (Logistic sigmoid $\sigma(x) = 1/(1 + e^{-x})$). It is used since a gate must provide only positive values and should be able to give a decision whether a particular feature is kept or discarded (0 or 1). The hyperbolic tangent $g(x) = tanh(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x})$ is used as the input and output activation functions, which converts any input to a value between -1 and +1.

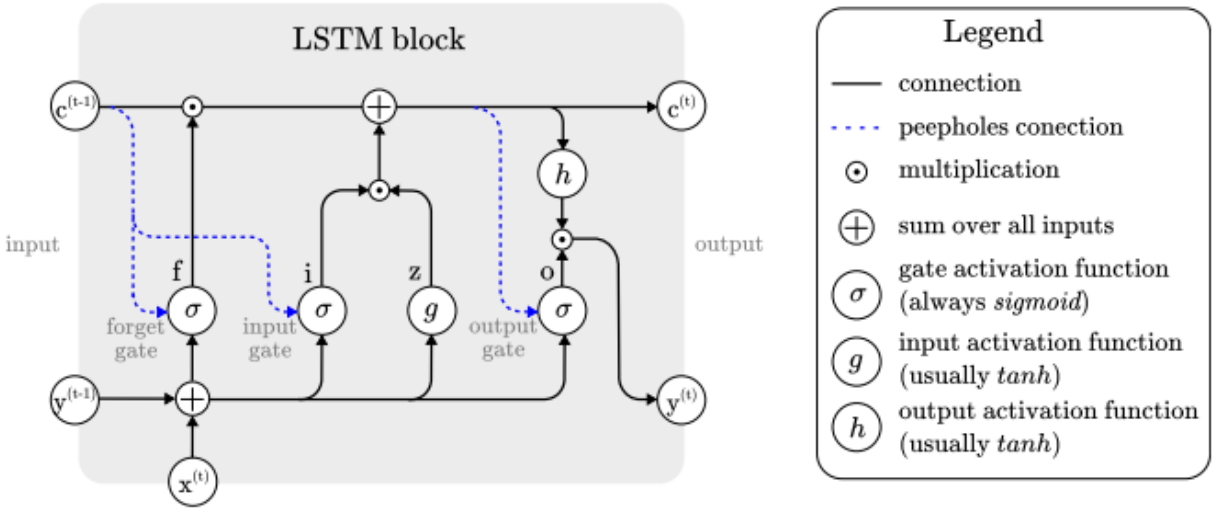


Figure 3.5 *Process within a cell with LSTM algorithm*

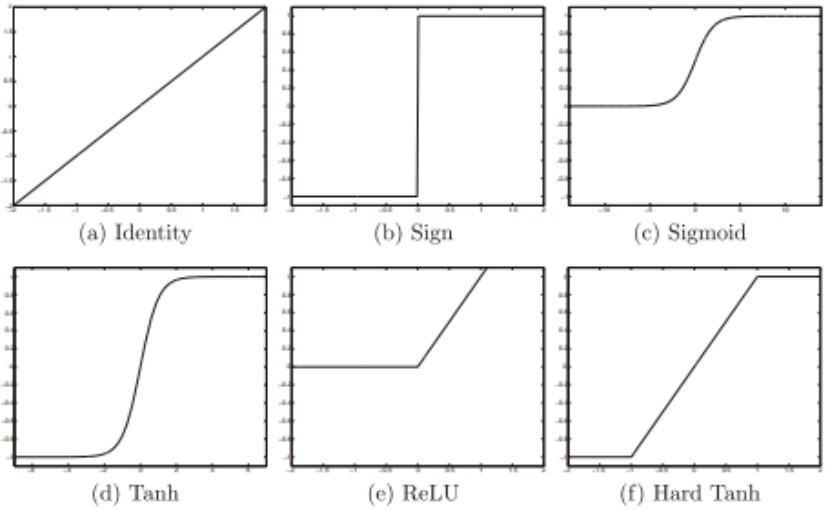


Figure 3.6 *Various activation functions*

The forget gate, input gate and output gate in LSTM follow the equations (2), (3) and (4) respectively. Forget defines information to remove from the cell state (long term memory). Input gate defines what new information is stored in the cell state. Output gate provides the activation to the final output (hidden state/short term memory) of the LSTM cell at timestamp *t*.

$$f_t = \sigma(W_f . x_t + R_f . y_{t-1} + b_f) \quad (2)$$

$$i_t = \sigma(W_i . x_t + R_i . y_{t-1} + b_i) \quad (3)$$

$$o_t = \sigma(W_o . x_t + R_o . y_{t-1} + b_o) \quad (4)$$

Here f_t is the forget gate, i_t is the input gate, $o_t$is the output gate, σ represents the sigmoid function, W_x and R_x represent the weight for the respective gate(x) neurons for the input and output, y_{t-1} is the output (hidden state/ short term memory) of the previous LSTM block (at timestamp *t-1*), x_t is the input at current timestamp and b_x represents the bias vectors the for the respective gates (x).

The equations for the candidate cell state, cell state (long term memory), and the final output (hidden state/short term memory) are stated in the equations (5), (6), (7)

$$z = tanh(W_c . x_t + R_c . y_{t-1} + b_c) \quad (5)$$

$$c_t = f_t * c_{t-1} + i_t * z \quad (6)$$

$$y_t = o_t * tanh(c_t) \quad (7)$$

Here z is the candidate cell state at timestamp (t), c_t is the cell state at timestamp *t*, y_t is the output at timestamp *t*, c_{t-1} is the cell state at timestamp *t-1*, and * represents element-wise multiplication of vectors. At any timestamp, the cell state knows that what it needs to forget from the previous state ($f_t * c_{t-1}$) and what it needs to consider from the current timestamp ($i_t * z$). Short term memory is directly connected to weights and can modify the value. Lastly, the cell state is filtered and then it is passed through the tanh activation function which predicts what portion should appear as the output of current LSTM unit at timestamp *t*.

The above equations are for only one time step, and they have to be recomputed for the next time step. The weight matrices and biases are not time-dependent and do not change from one

time step to another. Pseudocode for the LSTM computation is provided in Figure 3.7. LSTM generates the c_t (long term memory) and y_t (short term memory) for the consumption of the next time step LSTM.

Backpropagation through time [69] is used to calculate the weights that connect the several components in the network. Therefore, during the backward pass, the cell state c_t gets gradients from y_t as well as the next cell state c_{t+1}, which are collected before being backpropagated to the current layer.

```
# Pseudo code for the LSTM computation
sequence_len = 10
for i in range (0, sequence_len):
# If it is the initial step initialize y(t-1) and c(t-1) randomly
if i==0:
     yt_1 = random ()
     ct_1 = random ()
else:
     yt_1 = y_t
     ct_1 = c_t
f_t = sigmoid (matrix_mul(Wf, xt)+ matrix_mul(Rf, yt_1) + bf)
i_t = sigmoid (matrix_mul(Wi, xt)+ matrix_mul(Ri, yt_1) + bi)
o_t = sigmoid (matrix_mul(Wo, xt)+ matrix_mul(Ro, yt_1) + bo)
z    = tanh   (matrix_mul(Wc, xt)+ matrix_mul(Rc, yt_1) + bc)
c_t  = element_wise_mul(f_t, ct_1) + element_wise_mul(i_t, z)
h_t  = element_wise_mul(o_t, tanh(c_t))
```

Figure 3.7 *Pseudocode for the LSTM computation*

This study uses the Rectified Linear Unit (ReLu) Activation Function (Figure 3.4) for the overall final output, which is defined as *f(x) = max (0, x).* ReLU is one of most used activation function in the world right now for deep learning [69]. ReLU is half rectified (from bottom). *f(z)* is zero when *z* is less than zero and *f(z)* is equal to z when *z* is above or equal to zero. Thus, it is the most suitable activation function to predict the normalized density score as the final output.

The proposed model uses the above LSTM algorithm for the testing dataset to predict the output layer, using the input shape defined. The predicted normalized density score provides the

concentration of potential likelihood of incident occurrence in each block (B) per day (D). This score is depicted in a geographical map with the predefined blocks (B), representing the magnitude of the score for each block using an appropriate color palette.

3.4.2. Hyper-parameter tuning

Once the deep learning model is built, the hyperparameters of the LSTM model go through tuning to improve the performance of the model. Hyperparameters are the variables which determine the network structure. The model tunes the following hyper parameters to optimize the results, minimizing the error of the prediction.

1. **Number of nodes** (hidden neurons) **and hidden layers** (layers between the input and output layers) - More nodes within a layer can increase accuracy. Fewer number of nodes can cause underfitting while a high number of nodes may cause overfitting. Thus, the right balance must be identified.
2. **Number of units in a dense layer** - A dense layer has each neuron receiving input from all neurons in the previous layer.
3. **Dropout rate** - Every LSTM layer is accompanied by a dropout layer, which avoids overfitting in training by bypassing randomly selected neurons, thereby reducing the sensitivity to specific weights of the individual neurons. Dropout can address overfitting discussed in (1) above.
4. **Learning rate** - This represents the speed with which the network renews its parameters. A higher learning rate speeds up the learning, but the model might not diverge or converge (a state during training where the loss settles to within an error range around the final value). A lower learning rate will slow down the learning,

making the achievement of minimum of loss function slow but it will allow the model to converge effortlessly [68]. Thus, the learning rate must be optimized.

5. **Number of epochs** - This sets the number of complete iterations of the dataset to be run. This must be increased until the validation accuracy starts to decrease (even though training accuracy increases).

3.4.3. Performance evaluation

Once the deep learning model is built and optimized, it is evaluated using three performance metrics to determine the accuracy and precision of the prediction: Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Equations 8, 9 and 10 below provide the definition of the metrics.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \widehat{Y}_i| \qquad (8)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \qquad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2} \qquad (10)$$

where n is the sample size, Y_i is the actual/ observed value and $\widehat{Y}_i$ is the predicted value.

For benchmark analysis, the developed model is compared to some baseline machine learning models (Decision Tree Regression and Random Forest). Moreover, the results of the study are subjected to sensitivity analysis for further evaluation.

3.5. Summary

This chapter presents the methodology of the study in detail including the model design, data collection, data pre-processing, deep learning algorithm and the metrics used for evaluating the performance of the model.

CHAPTER 4: IMPLEMENTATION AND RESULTS

This chapter provides details of the model implementation and the results obtained from the deep learning approach to predict incident hotspots. Furthermore, it presents benchmark analysis of the study with other techniques and a sensitivity analysis of the model.

4.1.Model implementation

The model design described in chapter 3 is implemented using the Google Colab environment with Cloud Jupyter notebook (System RAM – 12.7 GB, Disk 107.7 GB). Python deep learning library is used incorporating libraries such as Tensorflow, Keras, matplotlib, pandas, numpy to develop the deep learning model. The integrated development environment enables the process of data pre-processing, development of the deep learning model, hyper-parameter tuning and evaluation of performance.

Hyper-parameter tuning is done using the Keras Tuner which is a hyper parameter optimization framework searching for the optimized values of hyper parameters. It enables defining the search space and it is embedded with Bayesian Optimization, Hyperband, and Random Search algorithms built in. All baseline models for benchmark analysis are developed using scikit-learn.

4.2.Results of the deep learning approach

The results of the study provided incident hotspots predicted for a selected date in North Carolina. The 18x6 grid depicted in section 3.3.2 was used for the model. Hotspots were depicted using the value of the normalized density score predicted using deep learning. Based on the

magnitude of the aforementioned score, the colors are depicted in each block. If the score was zero, the block would not have any color. The blocks which had a score which was not zero (i.e., greater than zero) are colored using hot colors ranging from yellow to red (yellow being the lowest and red being the highest). The color spectrum used for the hotspot depiction is presented in Figure 4.1, where the score range is divided into 256 portions and the respective color is given to the block based on the score value.

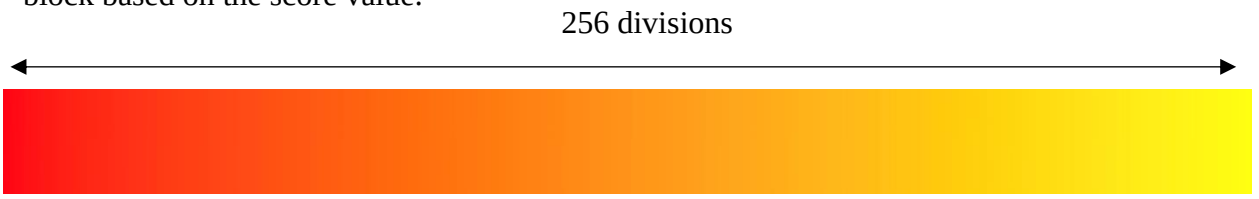


Figure 4.1 *Color spectrum of blocks*

The LSTM model used to predict incident hotspots provided satisfactory performance as depicted in Table 4.1. The North Carolina maps depicting the hotspots with the color coding defined above are presented below for some of the selected dates in the testing dataset. (Figures 4.2, 4.3, 4.4, 4.5 and 4.6)

Table 4.1 *Performance of the LSTM model*

Prediction Model	**MAE**	**MSE**	**RMSE**
LSTM	0.00212	0.00012	0.01091

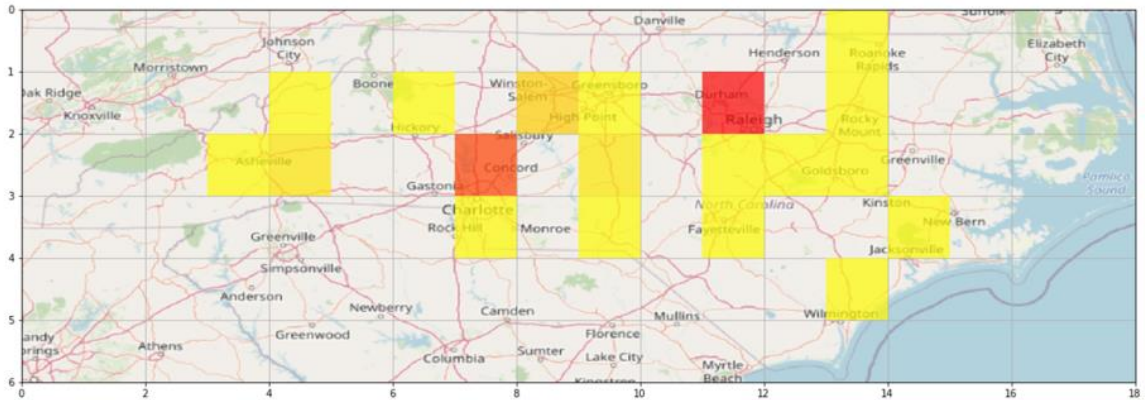


Figure 4.2 *Predicted incident hotspots for February 7, 2019*

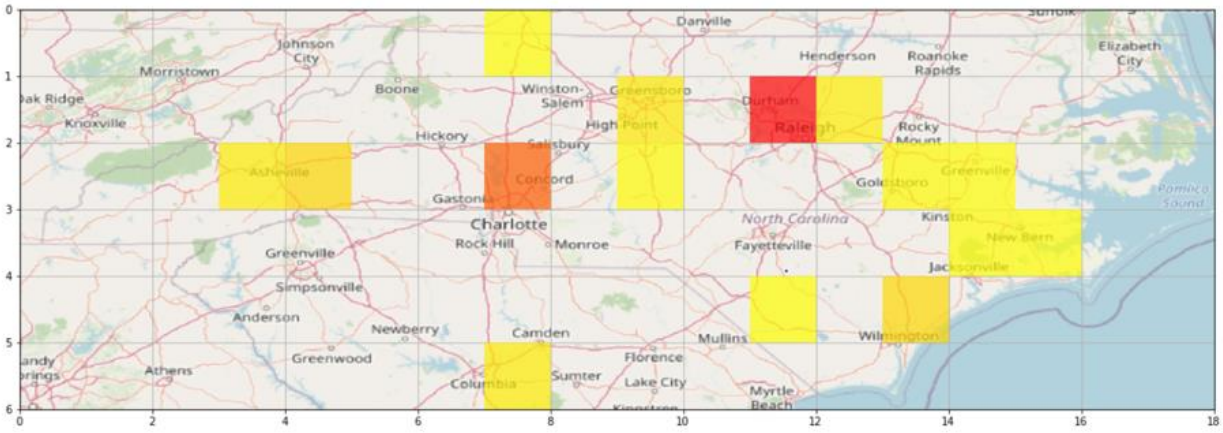


Figure 4.3 *Predicted incident hotspots for February 14, 2019*

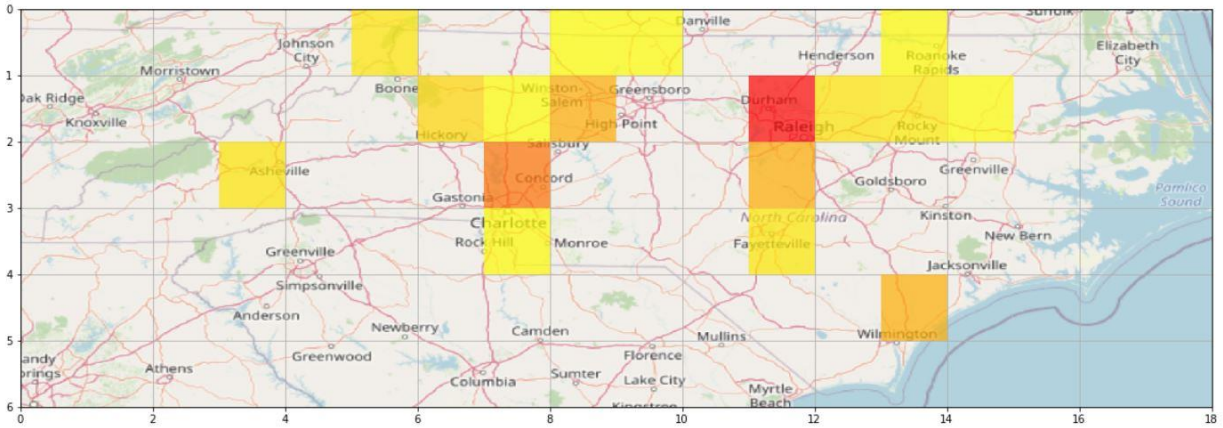


Figure 4.4 *Predicted incident hotspots for May 18, 2019*

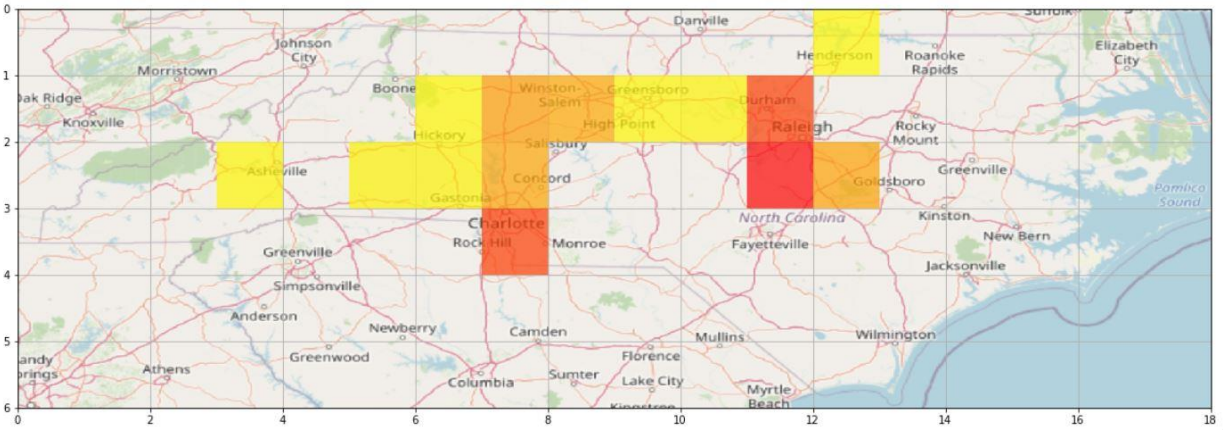


Figure 4.5 *Predicted incident hotspots for August 26, 2019*

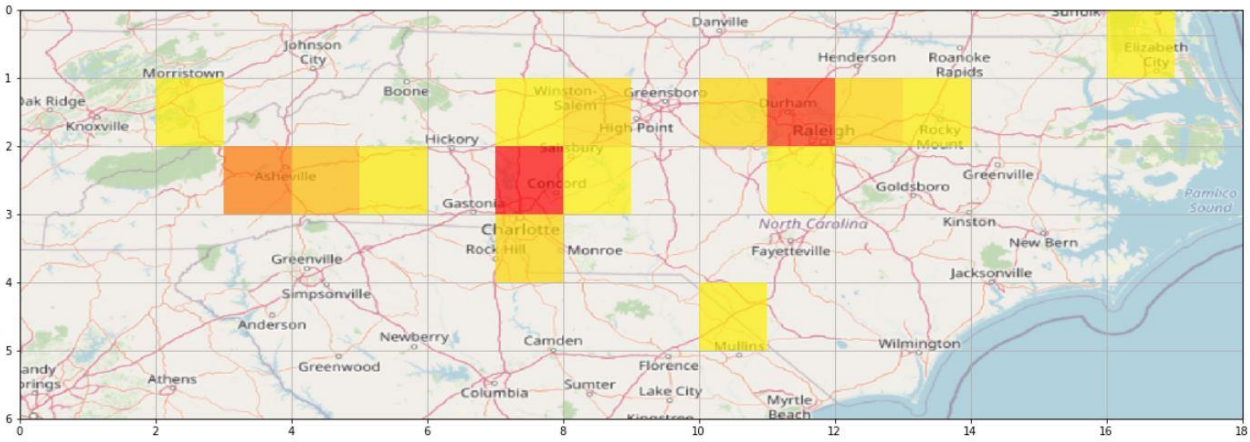


Figure 4.6 *Predicted incident hotspots for October 15, 2019*

The plots below in Figure 4.7 and Figure 4.8 depict the comparison of the actual value with the predicted value of the normalized density score for the month of March 2019 for blocks 29 and 43 respectively. Block 29 covers Raleigh/Durham cities and block 43 covers Concord, Salisbury, Gastonia. The model developed is able to provide the normalized density score values predicted for a given date and block, which enables better management of incidents.

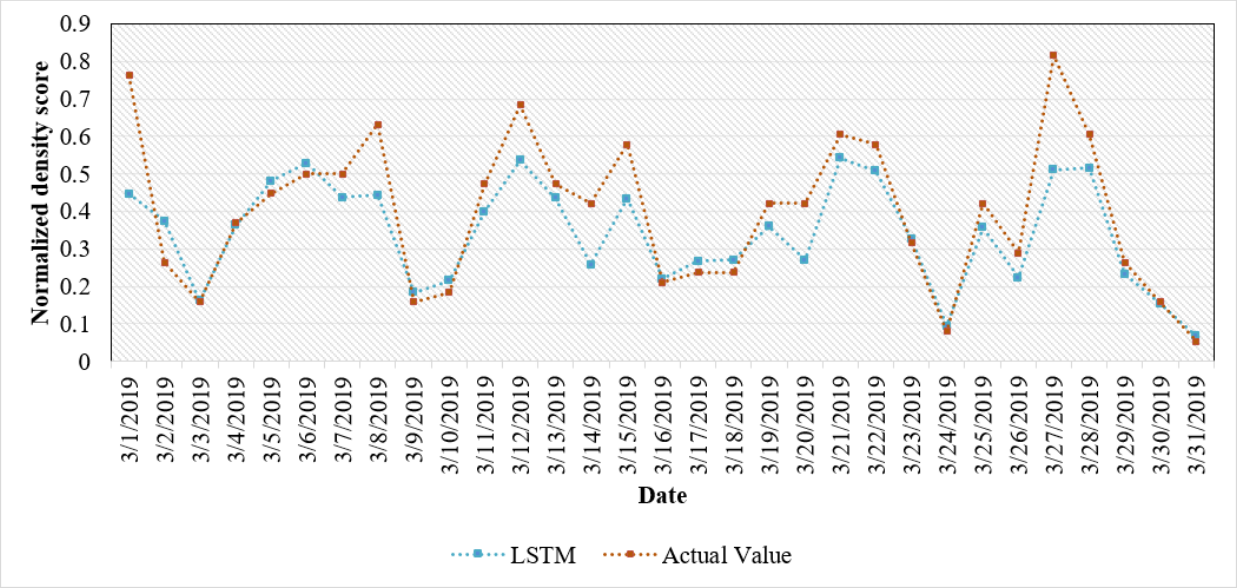


Figure 4.7 *Predicted score vs actual score for March 2019 in Block 29*

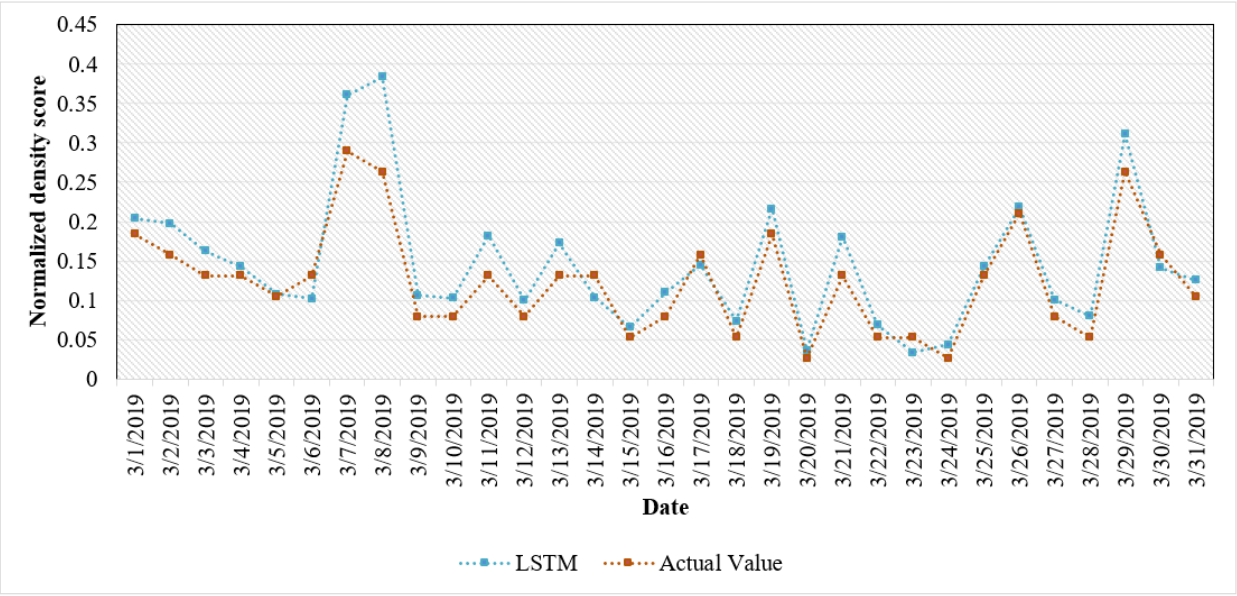


Figure 4.8 *Predicted score vs actual score for March 2019 in Block 43*

4.3.Benchmark analysis

The results of the developed deep learning model are compared with two baseline machine learning models for benchmark analysis. The selected two machine learning techniques are Decision Tree Regression (DTR) and Random Forest (RF). DTR technique goes through recursive binary splitting, while minimizing the loss function in each split. RF technique is an ensemble approach, which reduces the correlation of trees produced by randomly selecting variables that are considered at each split (to avoid splitting of dominating variables only).

The performance values received for the benchmark analysis are presented in Table 4.2. The results show that the error value is the least in the proposed deep learning model. This proves that the deep learning model performs better than the baseline machine learning models which were selected for the study.

Table 4.2 *Benchmark analysis*

Prediction Model	**Parameters**	**MAE**	**MSE**	**RMSE**
Decision Tree Regression (DTR)	max_depth=4, min_samples_leaf=0.1, random_state=3	0.00790	0.00125	0.03543
Random Forest (RF)	n_estimators = 50, random_state = 0	0.00227	0.00018	0.01358
LSTM (This work)	dropout rate=0.2, activation='relu', loss='mse', optimizer=Adam, Learning rate=0.0001, epochs = 8, Number of nodes =200, Hidden layers = 3, Number of units in a dense layer=2	**0.00212**	**0.00012**	**0.01091**

Moreover, Figures 4.9 and 4.10 depict plots which compare the actual value of the normalized density score with the predicted values of the three models i.e., LSTM deep learning model, DTR and RF models. The plots depict that the actual value is closer to the LSTM model's predicted value in comparison. The aforementioned two figures depict the normalized density score for the month of March 2019 for blocks 29 and 43 respectively. Block 29 covers Raleigh/Durham cities and block 43 covers Concord, Salisbury, Gastonia.

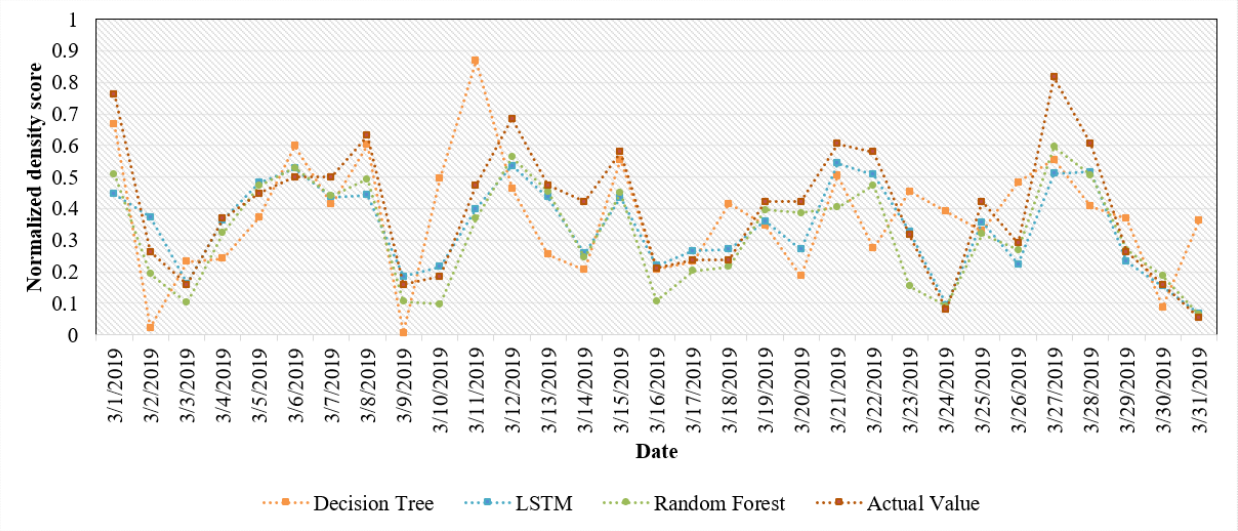


Figure 4.9 *Benchmark analysis for March 2019 in Block 29*

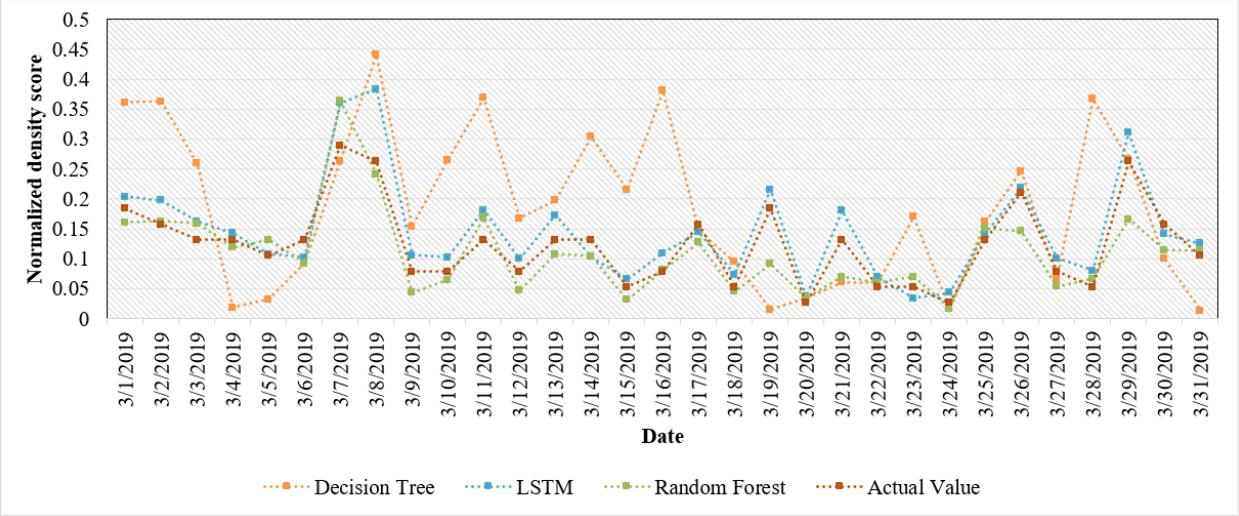


Figure 4.10 *Benchmark analysis for March 2019 in Block 43*

4.4.Sensitivity analysis

In order to study the sensitivity of the model results by changing the size of the blocks, a detailed analysis was done. The initial results depicted in section 4.2 were conducted using 18 x 6 grid which contained 108 blocks. This grid was amended to make the block size smaller and the

number of blocks higher. The hotspots predicted using the changed block sizes for February 14 2019 are depicted in Figures 4.11, 4.12, 4.13 and 4.14.

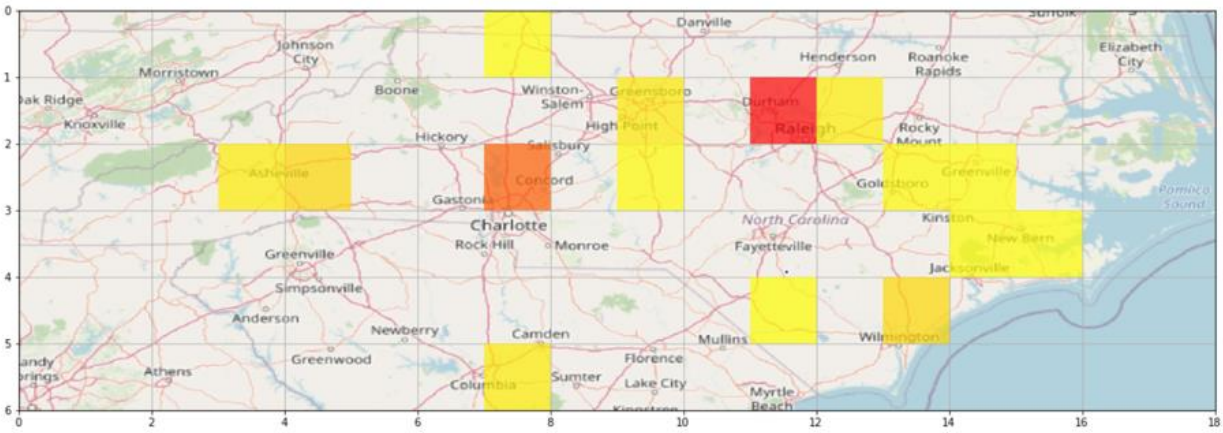


Figure 4.11 *Predicted 18x6 grid for February 14, 2019*

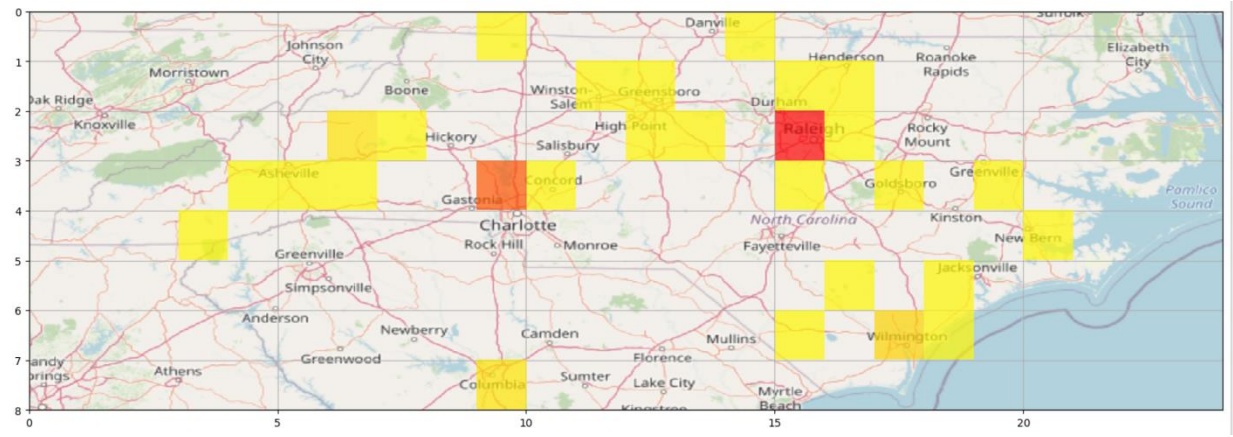


Figure 4.12 *Predicted 24x8 grid for February 14, 2019*

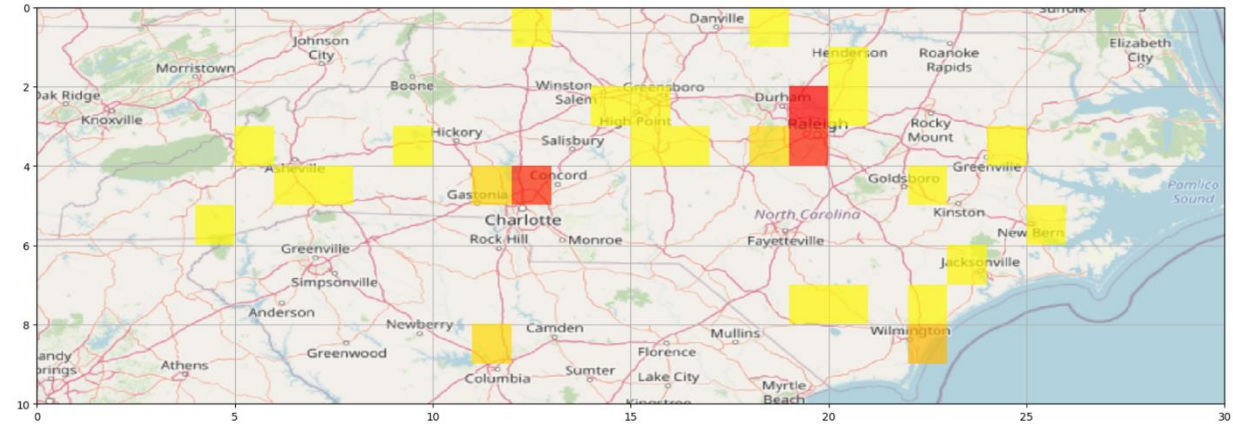


Figure 4.13 *Predicted 30x10 grid for February 14, 2019*

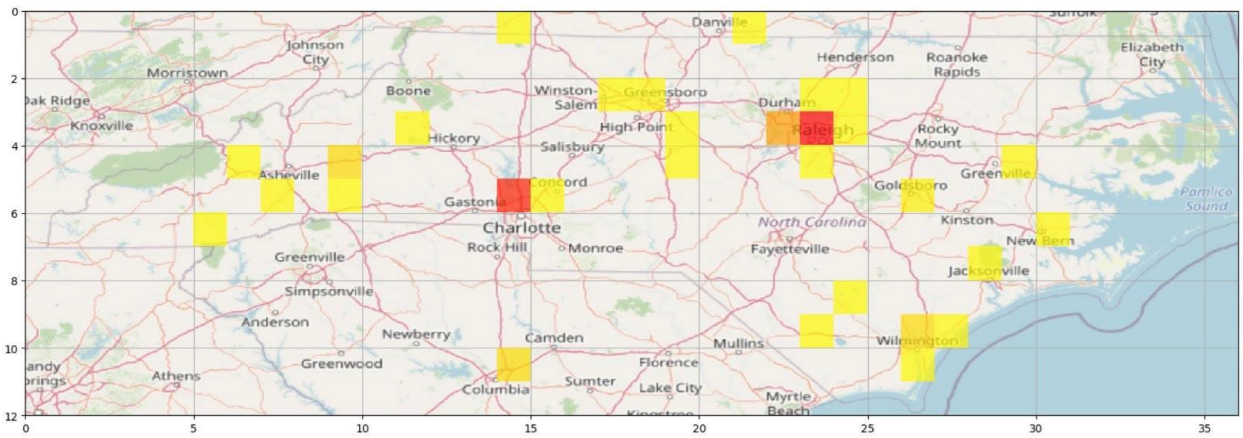


Figure 4.14 *Predicted 36x12 grid for February 14, 2019*

It was noted the error values were reduced when the block sizes are made smaller in the model (Table 4.3). This implied that the model accuracy increased when the blocks are smaller, and it proved to be more beneficial in predicting the hotspots accurately. When the block is smaller, better management of incidents, emergency response handling and resource allocation can be done. Furthermore, the hotspots were predicted focusing the Raleigh city map using the same model. (Longitude range = -78.82 to -78.47; Latitude range = 35.97 to 35.71). It was noted that 8201 incidents were reported from Raleigh. The results obtained for different dates are depicted in Figures 4.15, 4.16 and 4.17. The same prediction was tested for the Raleigh district map, where the 6 districts are depicted. These maps are depicted in Figures 4.18, 4.19 and 4.20.

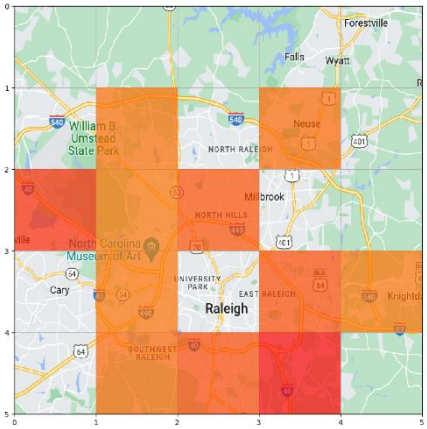


Figure 4.15 *Predicted hotspots of Raleigh for February 14, 2019*

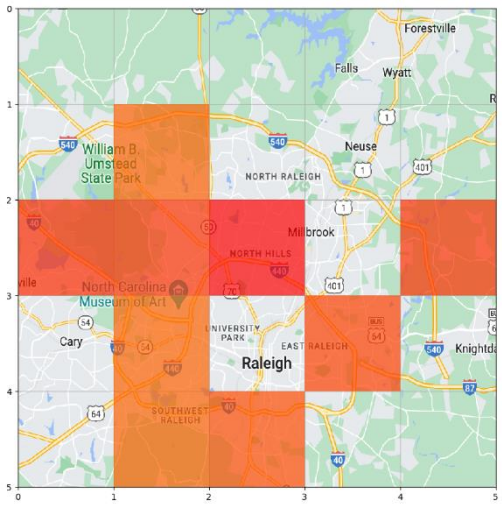


Figure 4.16 *Predicted hotspots of Raleigh for May 18, 2019*

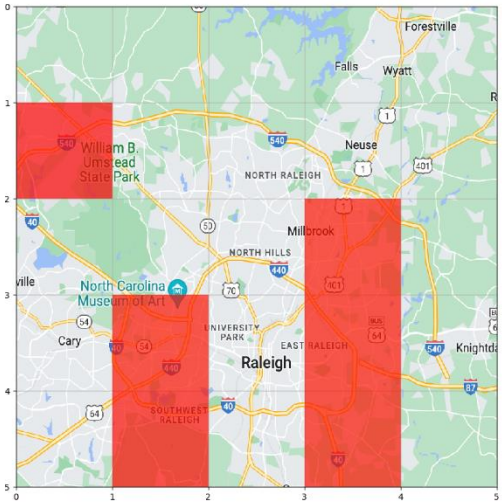


Figure 4.17 *Predicted hotspots of Raleigh for October 15, 2019*

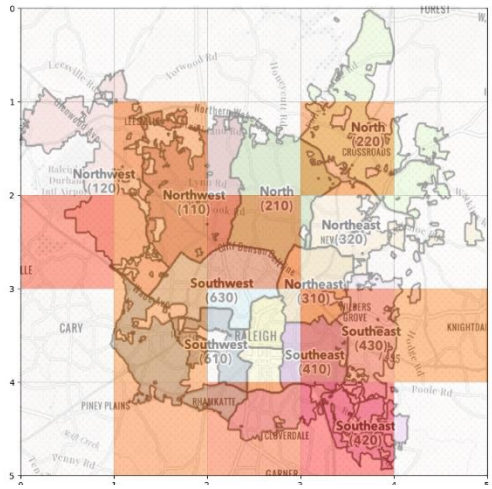


Figure 4.18 *Predicted hotspots of Raleigh for February 14, 2019 (District map)*

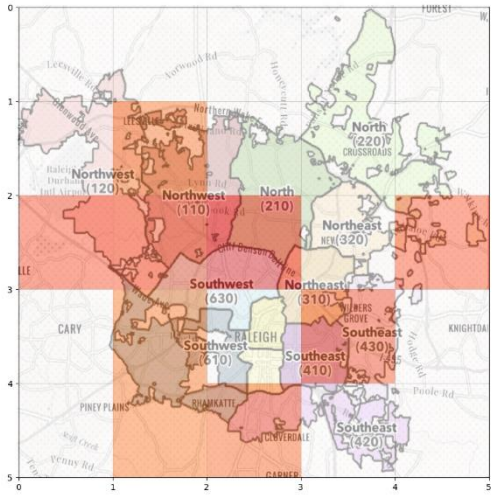


Figure 4.19 *Predicted hotspots of Raleigh for May 18, 2019 (District map)*

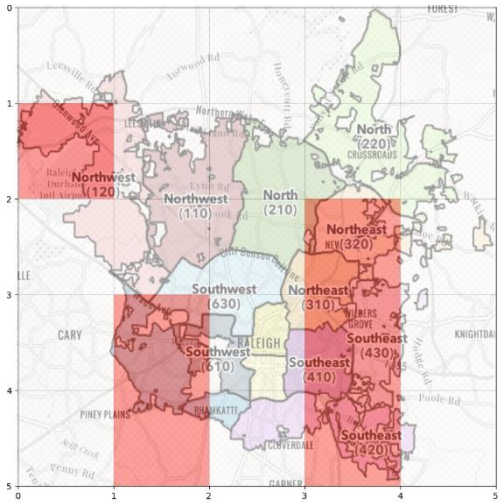


Figure 4.20 *Predicted hotspots of Raleigh for October 15, 2019 (District map)*

Considering the Raleigh map, the smallest area of prediction of this model was determined as 19.1 square miles. Thus, the maximum number of blocks possible for the entire North Carolina state map is 207 x 69 = 14,283 blocks. Table 4.3 depicts the performance of these models with different grid sizes.

Table 4.3 *Sensitivity analysis with different grid sizes*

Grid size	Model runtime (minutes)	MAE	MSE	RMSE
18 X 6	48	0.00212	0.00012	0.01091
24 X 8	61	0.00137	0.00006	0.00797
30 X 10	74	0.00084	0.00004	0.00649
36 X 12	88	0.00047	0.00002	0.00532
Raleigh (5x5)	37	0.00110	0.00001	0.00372

4.5.Summary

This chapter presents the implementation details of the model and the results obtained. Later, a benchmark analysis is presented where the deep learning model is compared with other machine learning models proving that the proposed model performs better. Finally, sensitivity analysis of the model is done by changing the block sizes of the grid, and the impact of it to the performance of the model is discussed.

CHAPTER 5: CONCLUSION AND FUTURE WORK

This chapter concludes the study and provides a summary of the approach taken in the research. It also presents the limitations and future work of the thesis.

5.1. Conclusions

This research focuses on incident management prediction models. A systematic review of literature is done to identify six categories of prediction models. Out of these models, it is identified that there is a gap in the current incident hotspots prediction models available. Limited studies are conducted using deep learning techniques to predict incident hotspots. Thus, this study presents a novel approach using deep learning techniques to predict the hotspots, by introducing a normalized density score to define a hotspot.

The methodology of the study comprises of a threefold approach i.e., heterogenous data collection, data pre-processing and deep learning model. Test cases from literature are used to identify significant variables that determine a hotspot. Incident data from North Carolina for 3 years, presence/absence of unusual event(s) in proximity, holiday information, AADT index and weather information are collected as heterogenous data. The data is pre-processed to be entered to the deep learning model. The input and the output layers are defined by aggregating the incidents to blocks and catering them as potential hotspots. The deep learning algorithm used for the model is the LSTM network which is able to capture the periodic nature of incidents and reflect the spatiotemporal correlation of incidents. The study predicts the normalized density score for each defined block in the map for a given date.

The results of the study provide a hotspot map for a given date. The predicted results are compared with two other baseline machine learning models (DTR and RF), where it proved that

the LSTM model performed better than the machine learning approaches. Moreover, a sensitivity analysis is done for the study which analyzed the effect on the performance of the model when the size of the blocks was changed in the prediction. It is observed that the error values reduced when the block size was made smaller. This implies a better result since incident management would be more effective when the prediction provided a smaller area in terms of the block.

The study provides an approach to predict the incident hotspots effectively which would undoubtedly support incident management, emergency response, better routing, accident safety mechanism implementations and ultimately to improve safety.

5.2. Limitations

The study uses data only for a period of 3 years. This provides a small dataset to train the model with changing patterns. The accuracy will improve if more data could be collected.

Furthermore, the model uses one-hot encoding in data pre-processing. Since one-hot encoding procedure generates several new variables, it is liable to cause many predictors if the original column has many unique values. It also may cause multicollinearity among the various variables, lowering the model's accuracy if a higher number of variables are used.

The results of the study provide the block level hotspots and not the road level hotspots. If road network structure data could be collected along with real time traffic data, it could improve its accuracy.

5.3. Future work

The study could be improved by using more data for increased time periods. This study uses only the data for 3 years. If the time period is expanded, accuracy will be increased. Furthermore, more variables could be incorporated such as lane curvature, road network structure,

drivers' behaviour and speed limit to improve accuracy. Data could be combined with road network structure and its factors to receive road level hotspots. Most significant variables could be filtered for a higher number of variables by using techniques such as Principal Components Analysis.

Moreover, the prediction could be taken on an hourly basis (which is daily in the current model). This would provide a hotspot map every hour for a given date.

5.4. Summary

The final chapter presents the conclusion of the entire work conducted along with its limitations and possible future work. It concludes that the results of the deep learning model presented could be used effectively to predict incident hotspots with higher accuracy.

REFERENCES

[1] World Health Organization, "Global Status Report on Road Safety," ISBN 978-92-4-156568-4, Geneva, 2018.

[2] INRIX, "INRIX Global Traffic Scorecard 2022," 17 2 2023. [Online]. Available: https://inrix.com/blog/2022-traffic-scorecard/.

[3] National Center for Statistics and Analysis, "Traffic safety facts 2020: A compilation of motor vehicle crash data (Report No. DOT HS 813 375)," National Highway Traffic Safety Administration, Washington, DC 20590, 2022.

[4] A. Saracoglu and H. Ozen, "Estimation of Trafc Incident Duration: A Comparative Study of Decision Tree Models," *Arabian Journal for Science and Engineering,* 2020.

[5] R. Li, F. C. Pereira and . M. E. Ben-Akiva, "Overview of traffic incident duration analysis and prediction," *European Transport Research Review,* vol. 10, no. 2, pp. 1-13, 2018.

[6] A. J. Khattak, J. L. Schofer and M.-H. Wang, "A SIMPLE TIME SEQUENTIAL PROCEDURE FOR PREDICTING FREEWAY INCIDENT DURATION," *Journal of Intelligent Transportation Systems,* no. 2, pp. 113-138, 1995.

[7] A. Garib, A. E. Radwan and H. AI.Deek, "ESTIMATING MAGNITUDE AND DURATION OF INCIDENT DELAYS," *Journal of Transportation Engineering,* no. 123, pp. 459-466, 1997.

[8] X. Wang, S. Chen and W. Zheng, "Traffic Incident Duration Prediction Based On Partial Least Squares Regression," *Procedia-Social and Behavioral Sciences ,* no. 96, pp. 425-432, 2013.

[9] Y. Chung, "Development of an accident duration prediction model on the Korean Freeway Systems," *Accident Analysis and Prevention,* no. 42, pp. 282-289, 2010.

[10] R. Li, "Traffic incident duration analysis and prediction models based on the survival analysis approach," *IET Intelligent Transport Systems,* vol. 9, no. 4, pp. 351-358, 2015.

[11] R. Li, F. . C. Pereira and M. E. Ben-Akiva, "Competing risk mixture model and text analysis for sequential incident duration prediction," *Transportation Research Part C,* no. 54, pp. 74-85, 2015.

[12] B. Ghosh, M. T. Asif and J. Dauwels, "Bayesian Prediction of the Duration of Non-recurring Road Incidents," in *IEEE Region 10 Conference (TENCON)*, 2016.

[13] S. Boyles, D. Fajardo and S. T. Waller , "A naive Bayesian classifier for incident duration prediction," 86th Annual Meeting of the Transportation Research Board, Washington, DC, 2007.

[14] B. Ghosh and J. Dauwels, "Comparison of different Bayesian methods for estimating error bars with incident duration prediction," *Journal of Intelligent Transportation Systems,* 2021.

[15] W. Kim and G.-L. Chang, "Development of a hybrid prediction model for freeway incident duration: a case study in Maryland," *International journal of intelligent transportation systems research,* vol. 10, pp. 22-33, 2012.

[16] S. Boyles and S. T. Waller, "A stochastic delay prediction model for real-time incident management," *ITE Journal,* vol. 77, no. 11, pp. 18-24, 2007.

[17] L. Lin, Q. Wang and A. W. Sadek, "A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations," *Accident Analysis and Prevention,* vol. 91, pp. 114-126, 2016.

[18] A. Grigorev, A.-S. Mihaita, S. Lee and F. Chen, "Incident duration prediction using a bi-level machine learning framework with outlier removal and intra-extra joint optimisation," *Transportation research part C: emerging technologies,* vol. 141, 2022.

[19] Y. Zhao and W. Deng, "Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning," *APPLIED ARTIFICIAL INTELLIGENCE,* vol. 36, no. 1, 2022.

[20] Z. Rahmat-Ullah, S. Alsmadi and K. Hamad, "Classifying and Forecasting Traffic Incident Duration Using Various Machine Learning Techniques," in *14th International Conference on Developments in eSystems Engineering (DeSE)*, 2021.

[21] K. Hamad, R. Al-Ruzouq, W. Zeiada, S. A. Dabous and M. A. Khalil, "Predicting Incident Duration Using Random Forests," *Transportmetrica A: Transport Science,* 2020.

[22] G. Valenti, M. Lelli and D. Cucina, "A comparative study of models for the incident duration prediction," *Eur. Transp. Res. Rev.,* vol. 2, pp. 103-111, 2010.

[23] H. Park, A. Haghani and X. Zhang, "Interpretation of bayesian neural networks for predicting the duration of detected incidents," *Journal of Intelligent Transport Systems,* vol. 0, no. 0, pp. 1-16, 2015.

[24] W. Zhu, "Dynamic prediction of traffic incident duration on urban expressways: a deep learning approach based on LSTM and MLP," *Journal of Intelligent and Connected Vehicles,* vol. 4, no. 2, pp. 80-91, 2021.

[25] L. Li, X. Sheng, B. Du, Y. Wang and B. Ran, "A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction," *Engineering Applications of Artificial Intelligence,* vol. 93, 2020.

[26] J. Tang, L. Zheng, . C. Han, W. Yin, Y. Zhang, Y. Zou and H. Huang, "Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review," *Analytic Methods in Accident Research,* vol. 27, 2020.

[27] C. Zhan, A. Gan and M. Hadi, "Prediction of Lane Clearance Time of Freeway Incidents Using the M5P Tree Algorithm," *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS,* vol. 12, no. 4, 2011.

[28] K. Ozbay and N. Noyan, "Estimation of incident clearance times using Bayesian Networks approach," *Accident Analysis and Prevention,* vol. 38, pp. 542-555, 2006.

[29] A. Mukhopadhyay, G. Pettet, S. M. Vazirizade, . D. Lu, A. Jaimes, S. E. Said, H. Baroud, Y. Vorobeychik, M. Kochenderfer and A. Dubey, "A Review of Incident Prediction, Resource Allocation, and Dispatch Models for Emergency Management," *Accident Analysis & Prevention,* vol. 165, 2022.

[30] D. Huang, S. Wang and Z. Liu, "A systematic review of prediction methods for emergency management," *International Journal of Disaster Risk Reduction,* vol. 62, 2021.

[31] C. Gutierrez-Osorio and . C. Pedraza, "Modern data sources and techniques for analysis and forecast of road accidents: A review," *Journal of Traffic and Transportation Engineering,* vol. 7, no. 4, pp. 432-446, 2020.

[32] M. E. Shaik, M. M. Islam and Q. S. Hossain, "A review on neural network techniques for the prediction of road traffic accident severity," *Asian Transport Studies,* vol. 7, 2021.

[33] C. Wang, M. A. Quddus and S. G. Ison, "Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model," *Accident Analysis & Prevention,* vol. 43, no. 6, pp. 1979-1990, 2011.

[34] I. K. Umar and H. Gokcekus, "Modelling Severity of Road Traffic Accident in Nigeria using Aritificial Neural Network," *Jurnal Kejuruteraan,* vol. 31, no. 2, pp. 221-227, 2019.

[35] S. Alkheder, M. Taamneh and . S. Taamneh, "Severity Prediction of Traffic Accident Using an Artificial Neural Network," *Journal of Forecasting,* vol. 36, pp. 100-108, 2017.

[36] H. Park and A. Haghani, "Real-time prediction of secondary incident occurrences using vehicle probe data," *Transportation Research Part C,* vol. 70, pp. 69-85, 2016.

[37] C. Gutierrez-Osorio, F. A. González and C. A. Pedraza, "Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data," *Computers,* vol. 11, no. 126, 2022.

[38] H. Zhao, H. Cheng, T. Mao and . C. He, "Research on Traffic Accident Prediction Model Based on Convolutional Neural Networks in VANET," in *2nd International Conference on Artificial Intelligence and Big Data*, 2019.

[39] F. N. Ogwueleka, S. Misra, T. C. Ogwueleka and L. Fernandez-Sanz, "An Artificial Neural Network Model for Road Accident Prediction: A Case Study of a Developing Country," *Acta Polytechnica Hungarica,* vol. 11, no. 5, pp. 177-197, 2014.

[40] M. Feng, J. Zheng, J. Ren and Y. Liu, "Towards Big Data Analytics and Mining for UK Traffic Accident Analysis, Visualization & Prediction," in *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, 2020.

[41] J. Roland, P. D. Way, C. Firat, T.-N. Doan and M. Sartipi, "Modeling and predicting vehicle accident occurrence in Chattanooga, Tennessee," *Accident Analysis and Prevention,* vol. 149, 2021.

[42] Z. Yuan, X. Zhou and T. Yang, "Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[43] T. Huang, S. Wang and A. Sharma, "Highway crash detection and risk estimation using deep learning," *Accident Analysis and Prevention,* vol. 135, 2020.

[44] L. Lin, Q. Wang and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction," *Transportation Research Part C,* vol. 55, pp. 444-459, 2015.

[45] B. Wang, Y. Lin, S. Guo and H. Wan, "GSNet: Learning Spatial-Temporal Correlations from Geographical and Semantic Aspects for Traffic Accident Risk Forecasting," *Proceedings of the AAAI conference on artificial intelligence,* vol. 35, no. 5, 2021.

[46] B. G. d. Soto, A. Bumbacher, M. Deublein and B. T. Adey, "Predicting road traffic accidents using artificial neural network models," *Infrastructure Asset Management,* vol. 5, no. 4, pp. 132-144, 2018.

[47] H. Ren, Y. Song, J. Wang, Y. Hu and J. Lei, "A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction," in *A Deep Learning Approach to the Citywide Traffic Accident Risk*, Maui, Hawaii, USA, 2018.

[48] 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 and Prevention,* vol. 122, pp. 239-254, 2019.

[49] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu and R. Ramnath, "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019.

[50] C. Chen, X. Fan, C. Zheng, L. Xiao, M. Cheng and C. Wang, "SDCAE: Stack Denoising Convolutional Autoencoder Model for Accident Risk Prediction via Traffic Big Data," in *Sixth international conference on advanced cloud and big data*, 2018.

[51] X. Wang, X. Qu and S. Jin, "Hotspot identification considering daily variability of traffic flow and crash record: A case study," *Journal of Transportation Safety & Security,* vol. 12, no. 2, pp. 275-291, 2020.

[52] E. A. 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,* vol. 29, no. 2, pp. 207-216, 2022.

[53] X. Qu and Q. Meng, "A note on hotspot identification for urban expressways," *Safety Science,* vol. 66, pp. 87-91, 2014.

[54] S. Vadlamani, E. Chen, S. Ahn and S. Washington, "Identifying Large Truck Hot Spots Using Crash Counts and PDOEs," *Journal of transportation engineering,* vol. 137, no. 1, pp. 11-21, 2011.

[55] L. Fawcett, N. Thorpe, J. Matthews and K. Kremer, "A novel Bayesian hierarchical model for road safety hotspot prediction," *Accident Analysis and Prevention,* vol. 99, pp. 262-271, 2017.

[56] M. Zarei, B. Hellinga and P. Izadpanah, "CGAN-EB: A Non-parametric Empirical Bayes Method for Crash Hotspot Identification Using Conditional Generative Adversarial Networks: A Simulated Crash Data Study," *International Journal of Transportation Science and Technology,* 2022.

[57] R. Krueger, P. Bansal and P. Buddhavarapu, "A new spatial count data model with Bayesian additive regression trees for accident hot spot identification," *Accident Analysis and Prevention,* vol. 144, 2020.

[58] M. Azari, A. Paydar, B. Feizizadeh and V. G. Hasanlou, "A GIS-based approach for accident hotspots mapping in mountain roads using seasonal and geometric indicators," *Applied Geomatics,* pp. 1-13, 2023.

[59] S. He, M. A. Sadeghi, S. Chawla, M. Alizadeh, H. Balakrishnan and S. Madden, "Inferring high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[60] H. Yu, . P. Liu, . J. Chen and . H. Wang, "Comparative analysis of the spatial analysis methods for hotspot identification," *Accident Analysis and Prevention,* vol. 66, pp. 80-88, 2014.

[61] S. Szénási and P. Csiba, "CLUSTERING ALGORITHM IN ORDER TO FIND ACCIDENT BLACK SPOTS IDENTIFIED BY GPS COORDIANTES," in *14TH SGEM GEOCONFERENCE ON INFORMATICS, GEOINFORMATICS AND REMOTE SENSING*, Bulgaria, 2014.

[62] 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.

[63] T. Lu, Y. Lixin, Z. Dunyao and Z. Pan, "The traffic accident hotspot prediction: Based on the logistic regression method," in *3rd International Conference on Transportation Information and Safety*, Wuhan, China, 2015.

[64] Q. Xu and G. Tao , "Traffic Accident Hotspots Identification Based on Clustering Ensemble Model," in *2018 5th IEEE International Conference on Cyber Security and*

Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), 2018.

[65] Abstract - The API Company, "Holidays API," [Online]. Available: https://www.abstractapi.com/api/holidays-api. [Accessed 30 11 2022].

[66] North Carolina Department of Transportation (NCDOT), "Connect NCDOT," [Online]. Available: https://connect.ncdot.gov/resources/State-Mapping/Pages/Traffic-Survey-GIS-Data.aspx. [Accessed 23 1 2023].

[67] "Weather API," [Online]. Available: https://www.weatherbit.io/. [Accessed 2 12 2023].

[68] C. C. Aggarwal, Neural Networks and Deep Learning, Springer, 2018.

[69] G. V. Houdt, C. Mosquera and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, pp. 5929-5955, 2020.

APPENDICES

Appendix A - Snapshots of datasets used for the model

Incident dataset:

IncidentID	StartTime (EST)	EndTime (EST)	RoadName	Suffix	Direction	CrossStreetName	CrossStreetSuffix	CrossStCommonName	MM From	MM To	County	City	Division	Latitude	Longitude
505505	4/30/2017 7:18:00 PM	4/30/2017 8:57:28 PM	I-40		Heading West			NC-96	332.0	-999.0	Johnston	Near Benson	4	35.348941790706 1	-78.4775044304548
505502	4/30/2017 6:02:00 PM	4/30/2017 7:34:28 PM	I-77		Heading North			Clanton Rd	7.0	-999.0	Mecklenburg	In Charlotte	10	35.1945053507159	-80.8841498895093
505501	4/30/2017 5:38:00 PM	4/30/2017 8:51:11 PM	SR-1550		Both Directions	US-264	ALT	US-264 ALT	-999.0	-999.0	Wilson	Saratoga	4	35.6342562224589	-77.7342068789673
505500	4/30/2017 4:56:00 PM	4/30/2017 5:49:33 PM	I-85		Heading South			S Cameron Blvd	58.0	-999.0	Cabarrus	In Concord	10	35.4468415584024	-80.6065341946121
505498	4/30/2017 1:56:00 PM	4/30/2017 2:24:10 PM	I-85		Heading North	US-70	BUS		170.0	-999.0	Orange	Hillsborough	7	36.0385625092622	-79.0133606725421
505497	4/30/2017 1:47:00 PM	4/30/2017 5:09:51 PM	I-85		Heading North			Satterwhite Point Rd	217.0	-999.0	Vance	Henderson	5	36.3660450678441	-78.3675776556577
505496	4/30/2017 1:40:00 PM	4/30/2017 2:05:13 PM	I-77		Heading North	I-485			19.0	-999.0	Mecklenburg	In Charlotte	10	35.3573921152475	-80.8435074174552
505495	4/30/2017 11:36:00 AM	4/30/2017 1:00:00 PM	I-26		Heading East	US-25		US-25	56.0	-999.0	Henderson	Near Flat Rock	14	36.003731	-79.831272
505494	4/30/2017 9:51:00 AM	4/30/2017 10:53:20 AM	I-40		Heading West	SR-1973		Page Road	282.0	-999.0	Durham	Near Durham	5	36.0824794622844	-78.8396976250 24
505493	4/30/2017 9:50:00 AM	4/30/2017 10:07:29 AM	I-95		Heading North	NC-82		NC-82	54.0	-999.0	Cumberland	Eastover	6	35.1816578356394	-78.6923343378558
505489	4/30/2017 3:40:00 AM	4/30/2017 8:07:34 AM	US-1		Heading North	SR-3977		SR-3977	99.0	-999.0	Wake	In Cary	5	35.747746919135	-78.7737453148247
505486	4/30/2017 12:21:00 AM	4/30/2017 2:06:25 AM	I-77		Heading South				6.0	-999.0	Mecklenburg	In Charlotte	10	35.478431	-80.874607
505482	4/29/2017 10:33:00 PM	4/29/2017 11:04:02 PM	I-485		Both Directions	SR-1009		SR-1009	52.0	-999.0	Mecklenburg	In Charlotte	10	35.1068176102473	-80.6983288513 7
505481	4/29/2017 10:09:00 PM	4/30/2017 12:21:22 AM	NC-50		Outer Loop				-999.0	-999.0	Wake	Near Wake Forest	5	36.023102	-78.695207
505480	4/29/2017 10:08:00 PM	4/29/2017 10:43:08 PM	I-485		Outer Loop	SR-1009		SR-1009	52.0	-999.0	Mecklenburg	In Charlotte	10	35.0994226277591	-80.7151680967765
505476	4/29/2017 3:56:00 PM	5/5/2017 12:34:00 PM	SR-2051		Both Directions	SR-2050		Rock Rd	-999.0	-999.0	Wayne		4	35.2708999113138	-77.9306210443077
505475	4/29/2017 3:52:00 PM	5/3/2017 12:00:00 AM	SR-1932		Both Directions	SR-1917		Casey Mill Rd	-999.0	-999.0	Wayne		4	35.3011590573033	-77.9749482960473
505474	4/29/2017 3:21:00 PM	4/29/2017 4:26:32 PM	I-77		Heading North	SR-1138		SR-1138	4.0	-999.0	Mecklenburg	Near Charlotte	10	35.147612	-80.896835
505472	4/29/2017 2:01:00 PM	4/29/2017 3:21:07 PM	US-74		Both Directions				-999.0	-999.0	New Hanover	Near Wilmington	3	34.2339176513825	-77.8410677358962
505470	4/29/2017 12:36:00 PM	4/29/2017 1:13:54 PM	I-85		Heading South	NC-49		NC-49	144.0	-999.0	Alamance	In Burlington	7	36.0650750475252	-79.4497323618833
505469	4/29/2017 12:15:00 PM	4/29/2017 1:27:12 PM	I-40		Heading East				53.0	-999.0	Buncombe	Near Asheville	13	35.5668731621441	-82.4971834102628
505468	4/29/2017 11:32:00 AM	4/29/2017 12:22:42 PM	NC-73		Both Directions				-999.0	-999.0	Lincoln	Near Lincolnton	12	35.4801064261 94	-81.1057861007473
505467	4/29/2017 11:13:00 AM	4/29/2017 8:11:02 PM	US-74		Heading East				172.0	-999.0	Polk	Near Columbus	14	35.2877499324568	-82.0300227944866

Holiday and weather datasets:


```
[
    {
      "name": "New Year's Day",
      "name_local": "",
      "language": "",
      "description": "",
      "country": "US",
      "location": "United States",
      "type": "National",
      "date": "01/01/2017",
      "date_year": "2017",
      "date_month": "01",
      "date_day": "01",
      "week_day": "Sunday"
    },
    {
      "name": "Thanksgiving Day",
      "name_local": "",
      "language": "",
      "description": "",
      "country": "US",
      "location": "United States",
      "type": "National",
      "date": "11/23/2017",
      "date_year": "2017",
      "date_month": "11",
      "date_day": "23",
      "week_day": "Thursday"
    },
    {
      "name": "Christmas Day",
      "name_local": "",
      "language": "",
      "description": "",
      "country": "US",
      "location": "United States",
      "type": "National",
      "date": "12/25/2017",
      "date_year": "2017",
      "date_month": "12",
      "date_day": "25",
      "week_day": "Monday"
    }
  ]
```

```
[
  {
    "city_id": "4464368",
    "city_name": "Durham",
    "country_code": "US",
    "data": [
      {
        "clouds": 85,
        "datetime": "2019-08-03",
        "dewpt": 20,
        "dhi": 53.5,
        "dni": 426.1,
        "ghi": 335.2,
        "max_dhi": 122.7,
        "max_dni": 923.9,
        "max_ghi": 986,
        "max_temp": 29.8,
        "max_temp_ts": 1564866000,
        "max_uv": 3.2,
        "max_wind_dir": 87,
        "max_wind_spd": 1.6,
        "max_wind_spd_ts": 1564866000,
        "min_temp": 19.5,
        "min_temp_ts": 1564822800,
        "precip": 0.5,
        "precip_gpm": 0.5,
        "pres": 1005.4,
        "revision_status": "final",
        "rh": 79.8,
        "slp": 1016,
        "snow": 0,
        "snow_depth": null,
        "solar_rad": 116.2,
        "t_dhi": 1283.9,
        "t_dni": 10227.1,
        "t_ghi": 8044.3,
        "t_solar_rad": 2788.5,
        "temp": 23.9,
        "ts": 1564804800,
        "wind_dir": 87,
        "wind_gust_spd": 1.7,
        "wind_spd": 1.1
      }
    ],
```

AADT Index dataset:

X	Y	FID	LocationID	COUNTY_1	RTE_CLS	ROUTE	LOCATION	AADT_2017	AADT_2018	AADT_2019	AADT_2020	AADT_2021
-79.5185	35.89297	1	10000002	ALAMANCE	4	SR 1103	WEST OF NC 49	180		200		200
-79.4186	36.14746	2	10000003	ALAMANCE	3	NC 62	EAST OF SR 1001	2900	2900	2900		3200
-79.3407	35.94038	3	10000004	ALAMANCE	3	NC 87	WEST OF SR 2172	4400	4500	4800		4200
-79.4264	35.97531	4	10000005	ALAMANCE	4	SR 2321	SOUTH OF SR 2330	1600		1700		1500
-79.4296	36.01241	5	10000006	ALAMANCE	4	SR 2387	EAST OF SR 1136		1700			
-79.4914	36.01626	6	10000007	ALAMANCE	4	SR 1113	SOUTH OF NC 62	2100		2100		2100
-79.521	36.16877	7	10000008	ALAMANCE	4	SR 1554	WEST OF SR 1500	870		900		950
-79.3052	35.9073	8	10000009	ALAMANCE	4	SR 1005	EAST OF NC 87	3000		3400	2900	3000
-79.2723	35.85108	9	10000010	ALAMANCE	3	NC 87	SOUTH OF SR 2102	2500	2700	2700		3000
-79.5152	35.88334	10	10000012	ALAMANCE	4	SR 2375	EAST OF NC 49		550			
-79.4464	36.18881	11	10000013	ALAMANCE	4	SR 1002	WEST OF SR 1605	1100		1100		1200
-79.3415	35.93876	12	10000014	ALAMANCE	4	SR 2172	WEST OF NC 87					
-79.4444	36.02266	13	10000015	ALAMANCE	3	NC 49	NORTH OF SR 1136	7900		7400		7500
-79.5011	36.059	14	10000017	ALAMANCE	4	SR 1213	EAST OF SR 1158	6200		9700		9000
-79.5095	36.18411	15	10000018	ALAMANCE	4	SR 1002	EAST OF NC 87		1700		1500	
-79.2998	36.23958	16	10000019	ALAMANCE	3	NC 119	NORTH OF SR 2009	1800	2000	1800	1600	1800
-79.3789	35.97525	17	10000020	ALAMANCE	4	SR 2116	EAST OF NC 87	620				
-79.4844	35.93308	18	10000022	ALAMANCE	4	SR 1117	NORTH OF NC 49	970		1000		1100