Received: 9 April 2023 | Revised: 27 June 2023 | Accepted: 19 July 2023

DOI: 10.1111/exsy.13415

ORIGINAL ARTICLE

Expert Systems WILEY

Comprehensive analysis of UK AADF traffic dataset set within four geographical regions of England

Victor Chang[1] | Qianwen Ariel Xu[1] | Karl Hall[2] | Olojede Theophilus Oluwaseyi[2] | Jiabin Luo[1]

[1]Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK

[2]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK

Correspondence
Victor Chang, Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK.
Email: victorchang.research@gmail.com and v.chang1@aston.ac.uk

Funding information
VC Resarch, Grant/Award Number: VCR 0000186

Abstract

Traffic flow detection plays a significant part in freeway traffic surveillance systems. Currently, effective autonomous traffic analysis is a challenging task due to the complexity of traffic delays, despite the significant investment spent by authorities in monitoring and analysing traffic congestion. This study builds an intelligent analytic method based on machine-learning algorithms to investigate and predict road traffic flows in four locations in the United Kingdom (London, Yorkshire and the Humber, North East, and North West) with a range of relevant factors. While aiming to conduct the study, the dataset 'estimated annual average daily flows (AADFs) Data—major and minor roads' from the UK government was used. Machine-learning algorithms are used for this research and classification applied consists of Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, and Gradient Boosting. Each of these algorithms achieves an accuracy of over 93% and the F1 score of over 95%, with Random Forest outperforming the other algorithms. This analytical approach helps to focus attention on critical areas to reduce traffic flows on major and minor roads in the area. In summary, the findings on traffic analysis have been discussed in detail to demonstrate the practical insights of this study.

KEYWORDS
algorithms for traffic analysis, machine-learning algorithms, Random Forest, traffic analysis, traffic flow

1 | INTRODUCTION

Traffic flow detection is a component of a freeway traffic surveillance system that detects data such as traffic volume, vehicle type, speed, and possession ratio to control and ensure traffic flow to be unobstructed and safe and provide supervisors with an accurate forecast basis. Traffic congestion is a major concern in today's metropolis (Lingras et al., 2000). They result in severe economic losses, higher pollution, and longer travel times. Authorities spend heavy investments in monitoring and analysing traffic congestion, but it is difficult due to the intricacy of traffic delays. One of the disadvantages is its pervasiveness. There are traffic delays occasionally, but sometimes there are not. This intermittent pattern of congestion makes it difficult to devise effective and proactive congestion management strategies. Another issue is that traffic congestion is ever-changing and interrelated. Impediments in traffic could, for example, stretch from one road to the next (Lubbe et al., 2018). Autonomous traffic analysis is challenging due to these complexities and necessitates a high level of skills, experience and understandable awareness.

Therefore, there is a need for a comprehensive understanding of traffic congestion patterns and identifying specific inefficiencies in congestion management in the crucial regions of the United Kingdom. This research gap has prompted the need to explore the dynamics of traffic congestion, and the factors contributing to congestion in urban areas (Ji & Hong, 2019). In addition, the effectiveness and applicability of machine-learning techniques for traffic analysis has not been well-documented compared with other domains (Quek et al., 2006). By addressing these research gaps, the study can contribute to the development of more efficient and targeted strategies for managing traffic congestion in urban areas. This study creates a visual analytic technique for investigating traffic sequences and circulation in four locations of the United Kingdom in this research (London, Yorkshire and The Humber, Northeast, and Northwest).

London is by far the largest city, with a population of approximately 9 million, compared to the second largest city in the United Kingdom, Birmingham, with 1.1 million. The population and infrastructure in London can be used as a forecast for future development in other regions of the United Kingdom. Therefore, by observing possible causes of congestion in London, this study can predict and avoid certain inefficiencies when planning the construction of major and minor roads, local congestion taxes and investments in public transport. The local authorities of London can be separated into five sections, Central, North, East, South, and West.

Figure 1 shows the Principal London Road network. The network is made up of two subsidiaries, Transport for London (TfL) and the Road Network represented by the Red Route and Borough Principal Road Network. The combination of the two networks makes up 11% of London roads, but during weekdays, between 7 am and 7 pm, these roads account for 54% of the London traffic. As shown in Figure 1, Central London acts as a focal point, representing the location where all roads converge. Around 790,000 people commute from England and Wales to London; among these numbers, nearly 400,000 workers commute to Westminster, and 230,000 workers commute to the City of London (Chow et al., 2014).

Moreover, an analysis of London would be conducted individually by separating the data from the AADF dataset, into five sections, with the 'local_name' of 'London (Central, North, East, South and West)'. Individual analysis using visual techniques will be conducted for each region, highlighting regional outliers, and a general comparison of traffic congestion among the five regions will be conducted. Additionally, the data for the two most congested roads in London, 'A406' and 'A23', will be shown. Drivers lose an average of 117 h per year due to traffic congestion between the two roads; thus, highlighting any potential correlation of congestion in 'A406' and 'A23' would provide more insight into possible reasoning for congestion present in local authorities.

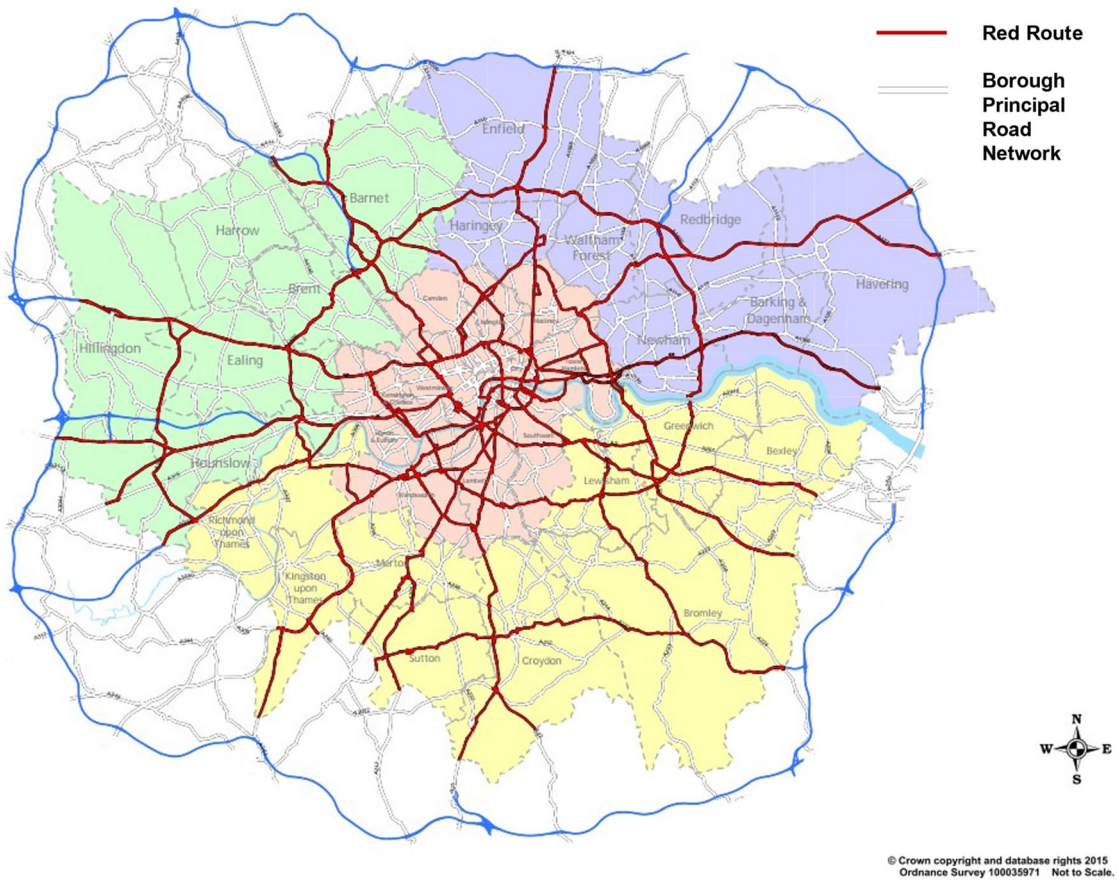

FIGURE 1 Principal London road network.

The method combines intelligent analysis and insights. This study begins by extracting the interesting areas from the 11 regions included in the United Kingdom Road AADF Count dataset and then creating exploratory data analysis (EDA) propagation for the average annual daily flow. Utilizing the major and minor road categories, this study conducts a systematic examination to determine the annual average daily traffic flows in the chosen locations. This method of analysis assists in focusing attention on crucial regions that will lessen traffic flow on major and minor roads around the region. In this study, machine-learning classification techniques utilized in AADF traffic prediction, including Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbor, and Gradient Boosting, are used to establish a correlation model of road traffic flow with a set of related factors such as pedal cycles, two-wheeled motor vehicles, cars and taxis, buses and coaches, light goods vehicles (LGVs), and heavy goods vehicles (HGVs) which in turn are used to predict the AADF traffic flow within the Four regions of UK. The F1 score and accuracy of the trained models are then used to assess the effectiveness of the developed model as well as the detrimental impact of each prediction model. The ROC curve technique is utilized to evaluate the performance of certain minor elements and the prediction model. Finally, the improved model may be used to forecast AADF traffic flow within the regions and the local authorities associated with them.

This research contributes to the practical implementation of data-driven decision making in traffic analysis. By using traffic flow data from the AADF and employing machine-learning algorithms, this study provides an evidence-based approach to understanding traffic patterns. By demonstrating their effectiveness in analysing traffic flows and making accurate predictions, this research provides city planners, traffic authorities and drivers with a valuable tool for improving traffic conditions and optimizing route choice in real-world scenarios, as well as helps them to make more informed and effective decision-making processes. In addition, this research bridges the gap between academic research and practical applications in the field of traffic analysis. By training and testing the advanced machine-learning algorithms, this study provides practical results that can directly influence traffic planning and decision-making processes, which can ultimately lead to more effective solutions for traffic management.

The breakdown of this paper is as follows. Section 2 describes the related works and literature, and Section 3 presents the methodology for traffic analysis research. Section 4 details the implementation and results of traffic analysis, showing different results for four selected regions. Section 5 provides details on performance evaluation and Section 6 presents the discussion related to the findings. Finally, Section 7 sums up this paper and research contributions.

2 | RELATED WORKS

The literature on traffic flow analysis forecasts is extensive. This work looks at previous technical assessments in the realm of traffic flow explanatory analysis and prediction models.

Andersson and Chapman (2011) employed the Kendall correlation coefficient to evaluate road traffic flow and investigate the link between road networks for road traffic injuries and road obstacle features such as traffic light systems, poor roads, and total road distance population. In a similar study, Fang and Shen (2012) suggested a regression model for estimating road traffic fatalities in London province based on data including local vehicle parks, route distance, and population.

Cai et al. (2015) used the signal strength of cellular phones to calculate the average traffic flow speed by comparing the signal strength traced on the cellular phones to the known trace of roads and computing the average traffic flow. They proceeded to estimate speed using a handover method. They looked for a base station that could handle a high number of road users and calculated the difference in access times between two successive base stations to anticipate traffic flow on the road in real-time. Even though this approach might cover most minor roads, it fails to accurately monitor changes in trace speed based on the road category.

Ramesh et al. (2019) implemented time-series for average traffic flow analysis, predicting the future based on historical parameters associated with numerous predictive techniques. Time-series models find patterns in historical data and extrapolate those patterns into the future. Ramesh et al. constructed and evaluated traffic flow based on road latitude, longitude, and direction using optimization algorithms such as historical average count data, time series, neural networks, and nonparametric regression models. Kalair and Connaughton (2021) proposed an up-to-date nonparametric segmentation method to classify and detect traffic anomalies by identifying the atypical fluctuations in the association between flow and density. They applied their method to the data relating to London's M25 motorway from the UK National Traffic Information Service (NTIS) and evaluated the method by several metrics, including time-to-detect, detection rate, and false alarm rate. The results indicate that their approach outperformed other statistical approaches, particularly on the multi-modal dataset.

Ulbricht (1994) used multi-recurrent neural network models to predict the number of traffic flows going through a highway checkpoint between the early hours of 5 am and 9 am and compared the results to other regression models. Shafiei et al. (2022) introduced a graphical model comprising information from major roads within the local authority, which was the first time Bayesian networks (BN) were used for traffic flow prediction. According to Jomnonkwao et al. (2020), the findings revealed that the BN model outperforms other condensed approaches such as the Random Walk (which considers current traffic flow conditions), the AR model, and a fuzzy-neural model. Likewise, Shepelev et al. (2020) and Monfared et al. (2013) created a graphical model that linked the average daily flow to the congestion status of each road network, as well as a theoretical traffic model that reproduced the delay frequency within a highway.

Based on the edge nodes, Chen et al. (2020) developed a traffic flow detection scheme by utilizing deep learning algorithm. In their study, a vehicle detection algorithm and a multi-object vehicle tracking system were first constructed based on the You Only Look Once (YOLO) model and Deep Simple Online and Real-time Tracking (DeepSORT), respectively. After that, they realized traffic flow detection by developing a real-time vehicle tracking counter by combining these two algorithms. Mehrannia et al. (2023) also employed a deep learning algorithm in their research on traffic accident detection. Based on the real traffic flow and accident data from the Twin Cities Metro freeways of Minnesota, they utilized long-short term memory (LSTM) network to extract the features of traffic flow data and train their model to label the data samples as crash or non-crash. The LSTM is a modified version of RNN, and it can easily remember past data in memory. According to their results, the LSTM model was able to detect accidents within 18 min and achieved better performance than other machine-learning models, such as CNN, RNN, AdaBoost, and so forth.

In the study of Azimjonov and Özmen (2021), they developed an innovative vehicle-tracking algorithm based on the bounding box (Bbox) to improve the performance of Yolo, a general-purpose object detector, in classifying vehicles. They prepared a dataset containing over 7000 images from highway videos and used them to train 10 machine-learning classifiers, with one of them using the CNN algorithm. The classifier with the best performance was then combined with Yolo to develop a new vehicle detector. Their results show that the accuracy of the detector was nearly 40% higher than that of Yolo. In addition, the integration of the detector with the Bbox-based tracking performed great in vehicle counting tasks, contributing to real-time traffic flow detection.

3 | METHODOLOGY

The methodology followed in this paper for developing an effective traffic analysis model is illustrated in Figure 2. The procedure relies on machine-learning techniques and is divided into five crucial steps: data collection, EDA, feature engineering, model training, and model evaluation. The dataset used in this study is traffic flow data collected from the UK government. Following data collection, EDA is performed to understand the data's characteristics. This includes data cleaning, data transformation, evaluating data distribution, investigating relationships between variables, and identifying potential anomalies. Feature engineering is the third step, which aims to refine and optimize the input features for the models to enhance predictive performance. After preparing the dataset and dividing it into a training set and a test set, several traffic analytics models, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, and Gradient Boosting, are trained. The final step in the methodology involves assessing the performance of the trained models. Evaluation metrics such as accuracy, F1-score, and ROC-AUC curve/score are used to quantify how effectively the trained traffic analytics models have learned from the data and predict the target variable.

3.1 | Problem statement and dataset

The aim of this research is to predict which road type will encounter a particular traffic flow rate (major or minor road) using classification models (Logistics Regression, Decision Tree, Random Forest, K-Nearest Neighbor, and Gradient Boosting).

To realize the research purpose, the dataset employed in this study is a primary dataset collected from the UK government. It includes traffic flow data by vehicle type on different roadways in four separate geographic regions of the United Kingdom. The dataset was subjected to EDA to

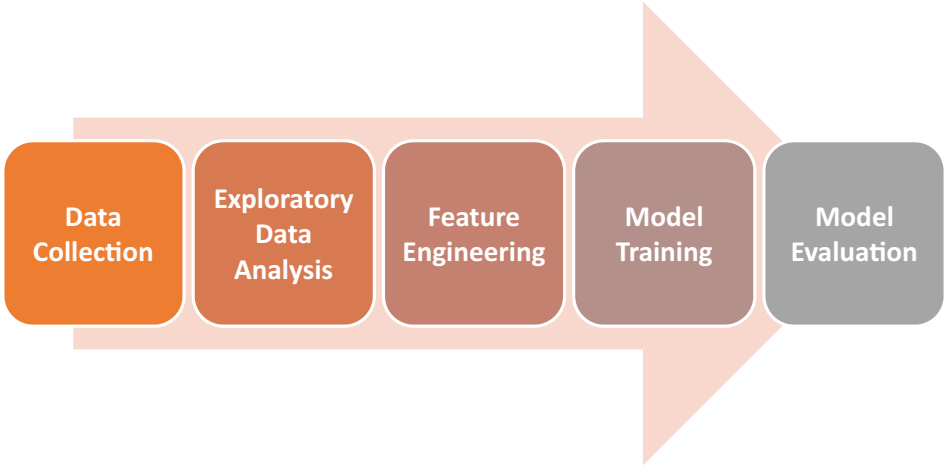

FIGURE 2 Methodology framework.

learn more about it, the key reasons for traffic, and how traffic flow has changed or not changed over time. The challenge is to estimate which road type has the largest traffic flow. Major and small roads are the two categories of roads discussed in this article.

3.2 | Exploratory data analysis

EDA is a step in the Data Analysis Process that employs a variety of techniques to comprehend the dataset being used properly. EDA allows the analyst to get a better look at the data, and how each attribute affects the other and to derive better insights for the prediction.

To better analyse the dataset, this study first extracts the data samples related to the four regions that need to be studied and then checks for the missing and duplicated values. After data cleaning, the values in text format are transformed into numerical numbers, followed by statistical analysis and visualization. This study analyses the data balance and attribute distribution through the EDA to better understand the datasets.

3.3 | Feature engineering

After EDA, this study performs feature engineering to understand the relationship between the attributes and the dependent variable. Feature engineering is employed to reduce the number of features when constructing classification and prediction models. It helps to identify the variables that contain the most relevant information for a given task (Thakkar & Lohiya, 2021). Moreover, it can also be used to remove redundant and irrelevant variables from a dataset. Through feature selection, classification and prediction models can be more cost-effective, faster, and more accurate simultaneously. This study computes a correlation matrix to learn the relationships and decides whether to drop any variables.

After the data pre-processing, we split the dataset into a training set (60%) and a test set (40%), which is a commonly adopted data split ratio, so that there is a higher amount of training data to ensure diversified inputs to make the model learn the hidden features in the data.

3.4 | Machine-learning classifiers

This study employs five different machine-learning algorithms to build the classifiers using the training set, including Logistics Regression, Decision Tree, K-Nearest Neighbors, Random Forest, and Gradient Boosting. These five algorithms were chosen based on their practical application for traffic analysis. In comparison to neural networks such as RNN and ANN, although they may exhibit slightly lower predictive performance, these algorithms are generally regarded as more easily interpretable and understandable (Díaz-Rodríguez et al., 2022; Garre et al., 2020). The aspect of interpretability is particularly valuable in transportation planning and decision making, as it helps stakeholders comprehend the reasoning behind the algorithmic recommendations. Moreover, they offer cost-effective advantages. These algorithms are computationally efficient and can handle large datasets with reasonable resource requirements compared to deep learning algorithms (Dargan et al., 2020). In addition, they have well-established implementations in popular machine-learning libraries, and their parameter tuning and model selection can be relatively easier. The algorithms are explained as follows in Table 1.

By comparing the output to the data set, an analysis with a confusion matrix would be used to validate the correctness of an ML model. When making correct forecasts, True Negatives, False Positives, and False Negatives are all possible outcomes. These measures are used to evaluate the efficiency of the models and are primarily used to assist in selecting the best-suited model. The number of correct predictions in both classes divided by the number of characteristics in the training set determines the accuracy of a classification model. In this case, the positive class is 1 (main road), while the negative class is 0 (minor road). Based on the confusion matrix, this study evaluates the classifiers' performance by comparing their Accuracy, F1-score, and Roc-AUC curve/score.

4 | IMPLEMENTATION AND RESULTS

4.1 | Data cleaning

The first step is to get a list of the dataset's attributes or features' names. This process is used to provide a broad overview of the data to determine how normalized it is. There are 489,159 samples in the dataset, with 33 different features (7 categorical features and 26 numerical features). The road type features attribute is the focus variable in this study, and it has a categorical variable with values of 'Major' or 'Minor'.

In the next step, this study extracted the four geographic locations needed for this research work from the complete AADF dataset. After this, the newly extracted dataset was inspected for errors and null values. It was observed that there are five columns with missing values. Before this

TABLE 1 Machine-learning models.

ML classifier	Summary
LR	Logistic Regression is a supervised ML algorithm that is used for classification tasks. This algorithm is mostly employed in binary classification. The sigmoid function is applied in the process of Logistic Regression to return the likelihood of a label (Hussein et al., 2021).
DT	The structure of a Decision Tree is similar to a tree. Each internal node represents an attribute test, and each leaf node represents a label. This algorithm builds a tree through the segmentation of the source set into subsets based on the judgements of attributes (Pappalardo et al., 2021). This process is repeated recursively on each derived subset and stops when instances on each node all have the same value or when the partitioning does not add value to the prediction.
KNN	The K-Nearest Neighbour (KNN) is a nonparametric supervised learning algorithm that assumes that similar points can be found in the vicinity of each other. It has been utilized extensively in classification tasks due to its straightforward implementation and few hyperparameters. The KNN algorithm determines the nearest neighbours of a given point by calculating the distance between that point and other data points, thus assigning a category label to that point (Wani & Roy, 2022).
RF	Random Forest is a supervised machine-learning algorithm based on the Decision Tree. It improves the performance of many weak learners by voting for the majority (Gupta et al., 2021). The samples in the original dataset are used as input for each tree in the classifications. After that, features are chosen randomly to be employed in developing the tree at each node. The algorithm does not prune any trees in the forest until the activity has been completed and a decisive prediction is made. By doing so, the Random Forest can construct a powerful classifier based on any classifiers with weak correlations.
GBDT	Gradient Boosting is one technique that excels at capturing high-dimensional relationships. It can automatically identify sophisticated data structures, including non-linearities and higher-order interactions. It can recursively fit a weak learner to the residuals to enhance a model's performance as the number of iterations increases (Zhang et al., 2019). The similarity between GBDT and RF is that they are both ensemble models with Decision Trees as weak learners. The biggest difference is that GBDT uses the boosting technique while RF uses bagging (Li et al., 2021)

was treated, the unnecessary columns were removed to maintain the data integrity and usefulness and remove redundant data. Then, the dataset was inspected for duplicates. There was only one duplicate value in the dataset, and it was dropped. In this stage, it was observed that the dataset has reduced to 166 k samples while retaining 26 features.

4.2 | Data transformation

Before proceeding to the next stage, the values of dependent variables need to be verified in numeric form. The values of the dependent variables are encoded so that machine-learning algorithms can understand and process the input (Brownlee, 2022).

The first step is converting the road type into numeric numbers 0 and 1. For 0 is a major road and 1 indicates a minor road. Apart from the type of road, all the other categorical attributes are encoded. To be specific, all the categorical values in each column are transformed to numerical values (replace object type with integer type) for training and testing. Efficient integration is also achieved before splitting the data for training and testing.

4.3 | Statistical analysis and visualization

4.3.1 | Data balance analysis

This study first checked the balance of the data samples relating to the road type. According to Figure 3, there is an unequal traffic flow; 80% of the average annual daily flow of traffic occurs on the major road in the four regions within this analysis.

As shown in Figure 4, among the four regions, the Northwest has the highest count of road usage and traffic flow, with over 56,000 records. In comparison, the Northeast has the least traffic flow counts, with over 24,000 records.

As the capital of the United Kingdom, London has the highest local authority count, comprising of 32 boroughs and the City of London. Conversely, the North East region has the lowest number of local authorities with 12, see Table 2.

AADF is an estimated measure of the full-year average of the number of vehicles passing a particular point in the road network per day. Investigations were conducted to analyse the AADF of vehicles across different road types.

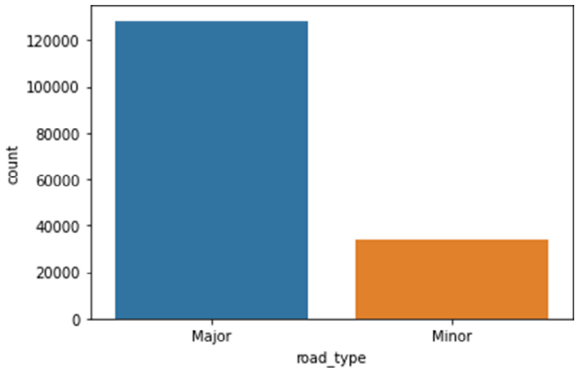

FIGURE 3 Analysis of road type.

FIGURE 4 Analysis of traffic regions.

TABLE 2 Local authority counts.

Region name	Local authority count
London	33
North East	12
North West	23
Yorkshire and The Humber	15

4.3.2 | AADF of vehicles in each road type

The analysis from Figures 5–8 shows that London has the highest annual average daily flow of traffic count on major roads due to its high congestion and commercial activities compared with other regions in the study. It was also observed that buses and coaches have high traffic flow compared with other regions because most citizens prefer to use public transportation over private transportation due to policies and regulations.

Figures 9 and 10 show that London has the highest traffic flow of LGVs. According to the study, this is due to the local movement of goods and services within the region using the road network. In contrast, other regions with high landmarks use train transportation to supplement road transportation.

HGVs use major road networks rather than minor roads to transport industrial products from one location to another, as shown in Figures 11–16. The study also discovered that Yorkshire and the Humber have the highest traffic flow of HGVs, indicating that the region is

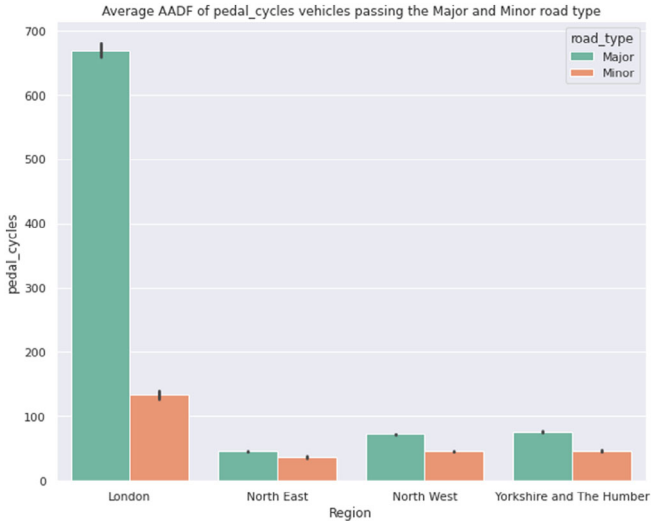

FIGURE 5 AADF pedal cycle wheel.

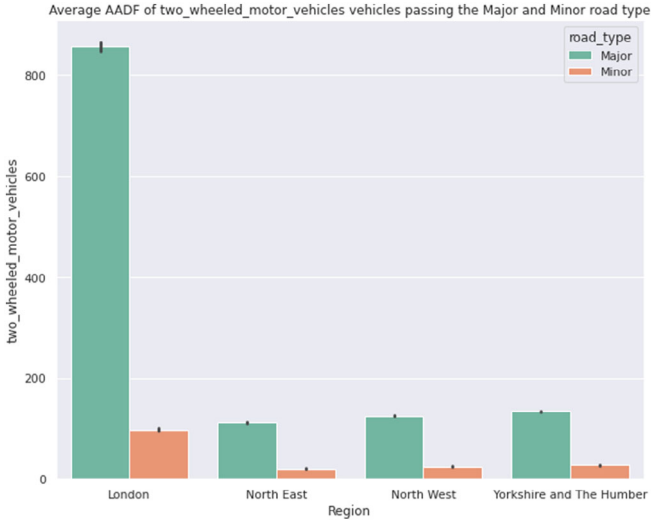

FIGURE 6 AADF two-wheeled motor vehicle.

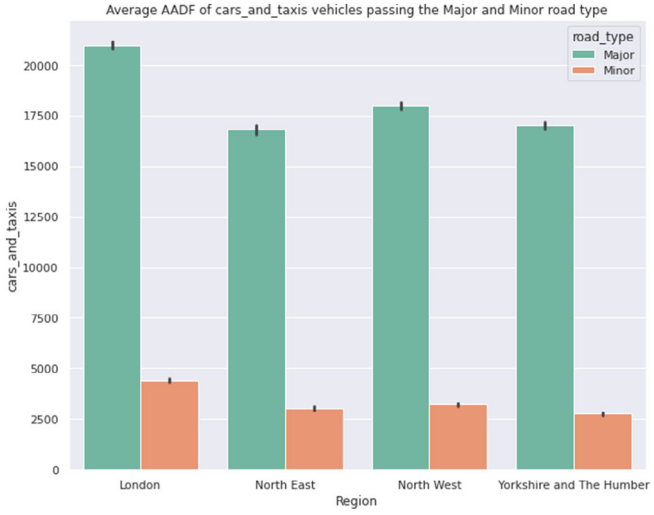

FIGURE 7 AADF cars and taxis.

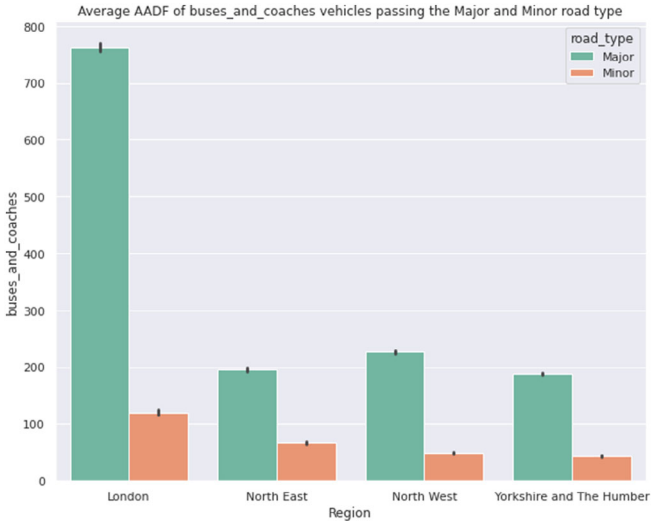

FIGURE 8 AADF buses and coaches.

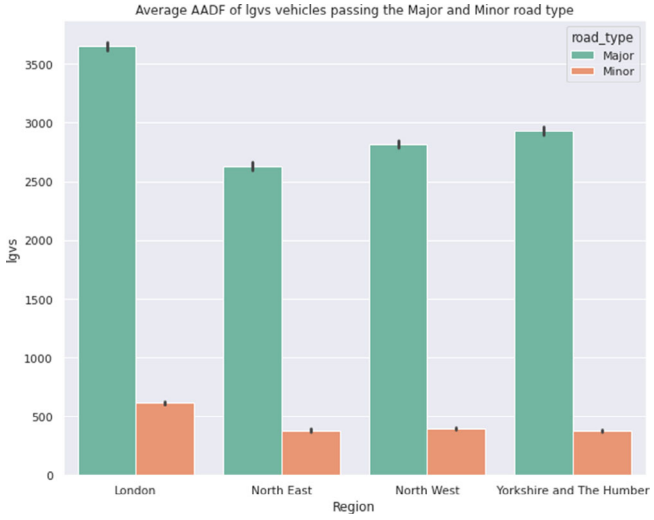

FIGURE 9 AADF LGV vehicles.

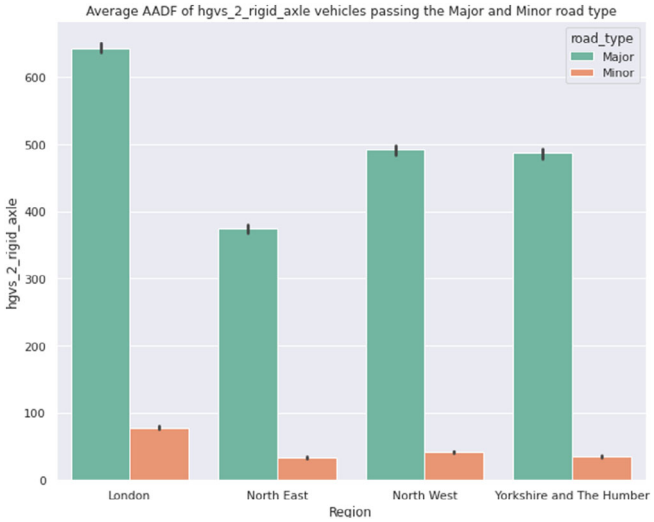

FIGURE 10 AADF 2 rigid axle vehicles.

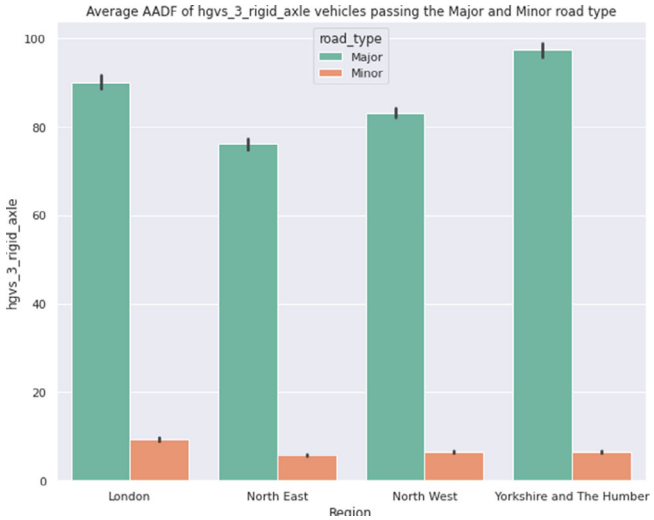

FIGURE 11 AADF HGV 3 rigid axle vehicles.

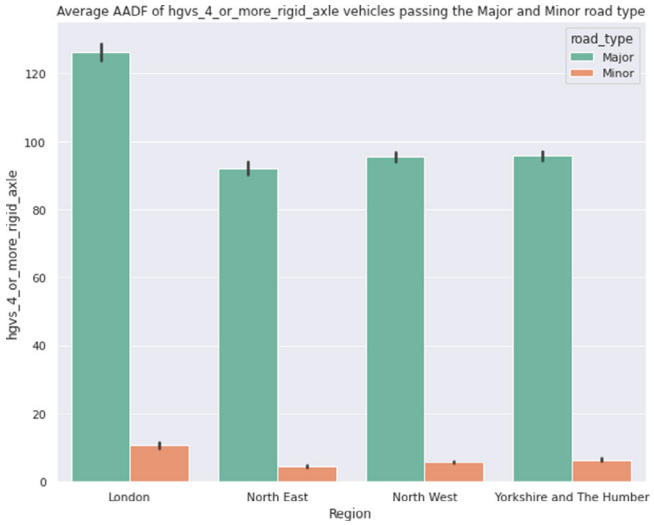

FIGURE 12 AADF HGV 4+ rigid axle vehicles.

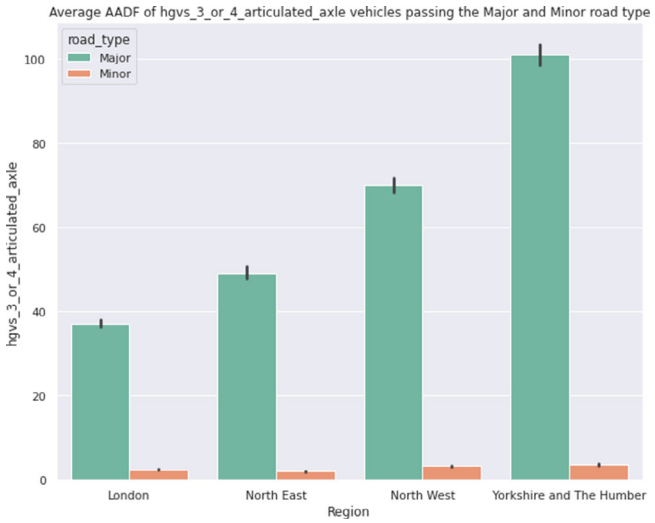

FIGURE 13 AADF HGV 3 or 4 articulated axles vehicles.

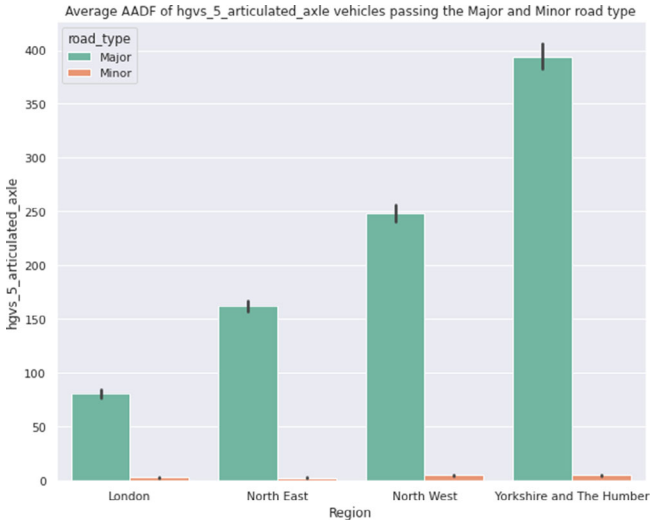

FIGURE 14 AADF HGV 5 articulated axle vehicles.

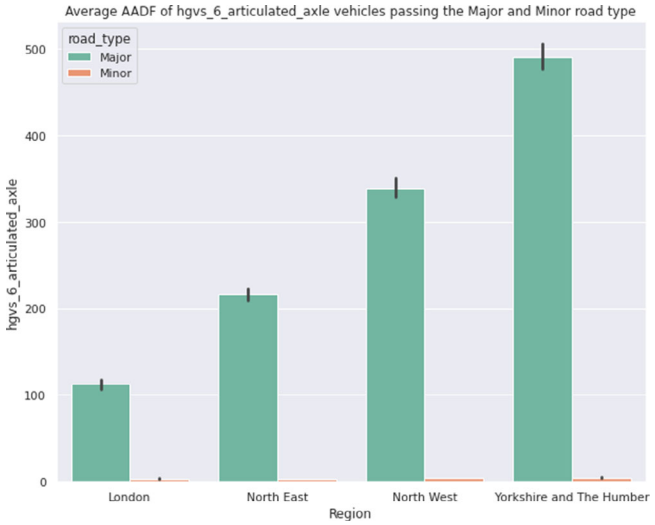

FIGURE 15 AADF HGV 6 articulated axle vehicles.

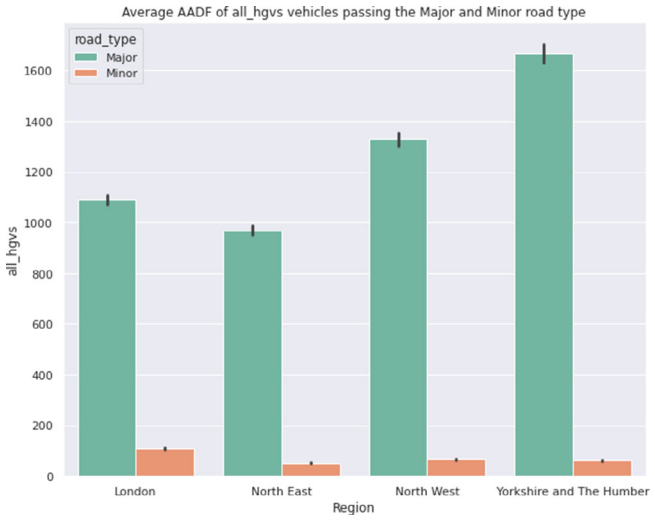

FIGURE 16 AADF all HGV vehicles.

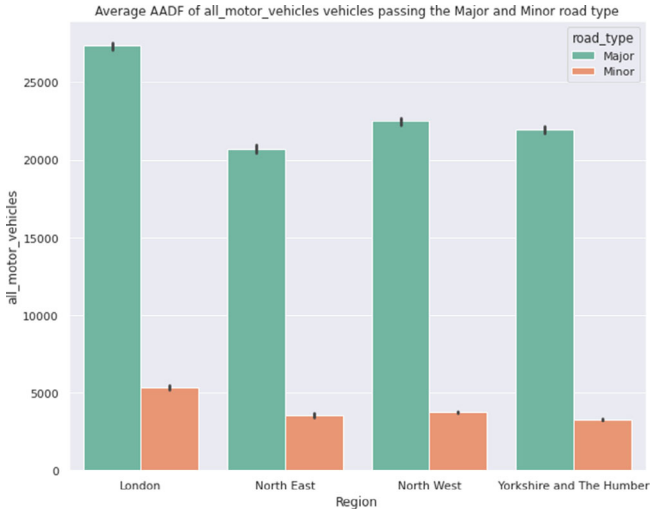

FIGURE 17 AADF all motor vehicles.

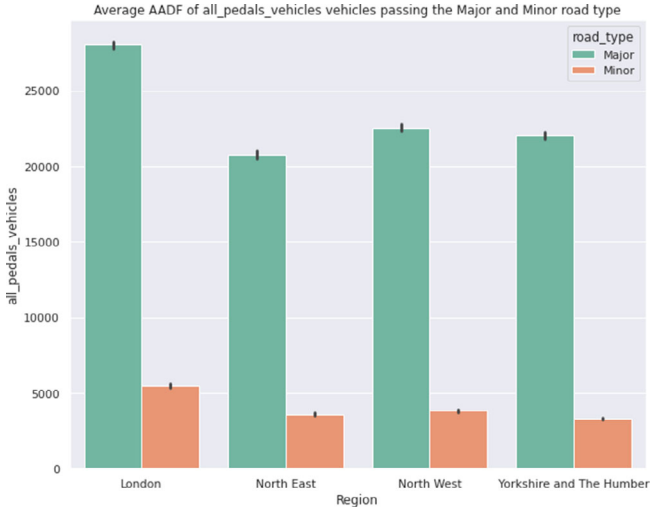

FIGURE 18 AADF all pedals vehicles.

well-known for its industrialization activities compared with other regions. Another reason is that due to its central locations, HGVs that transport goods from north to south and south to north use major roads in Yorkshire and the Humber.

As can be observed from Figures 17 and 18, most vehicles in the four locations prefer going through the major roads rather than the minor roads. This means that the traffic flow on minor roads will be less than on major roads.

4.3.3 | Top 5 busiest locations in each region

This study performed an analysis to identify the top five busiest locations in each region. Analyses between Figures 19–22 show the top give busiest average annual daily traffic flow within each region, with Leeds being the busiest in Yorkshire, Durham with the busiest in the Northeast, Lancashire the busiest in the Northwest and Westminster the busiest in London.

4.3.4 | Yearly trend on total traffic volume on road type

Figures 23–25 show that traffic flow on major roadways has significantly increased since 2007, affecting all pedal, motor, and HGVs. The above trends show that traffic congestion on major roads remained steady from 2007 to 2020, while after 2020, it began to decrease due to the

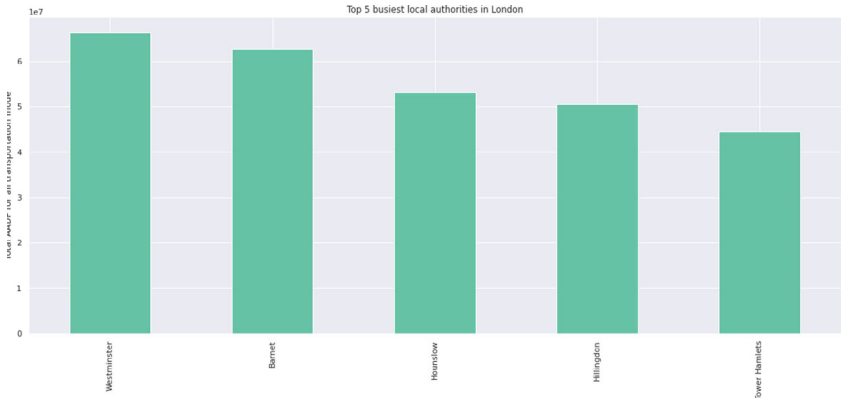

FIGURE 19 Busiest local authorities in London.

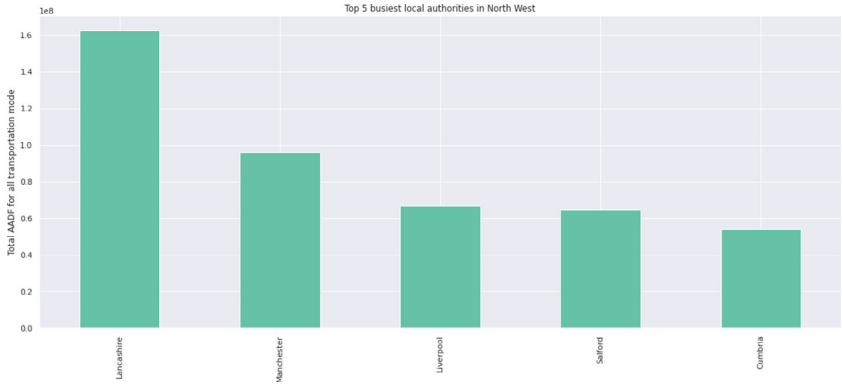

FIGURE 20 Busiest local authorities in the North West.

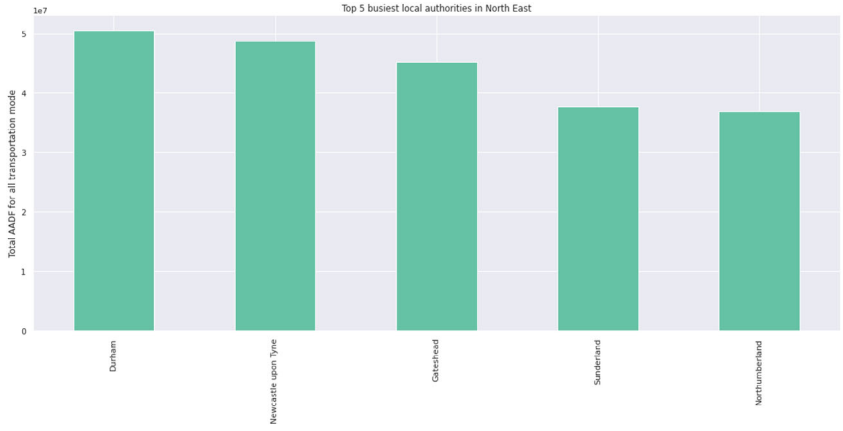

FIGURE 21 Busiest local authorities in the North East.

lockdown in response to the COVID-19 pandemic. Similarly, the data shows a sharp increase in the traffic flow on minor roads from 2006 to 2008, followed by a simultaneous decrease until 2010. Then, the traffic flow remained constant until 2016, followed by a massive increase during the period to 2019 and a large reduction after 2019 due to the COVID-19 lockdown.

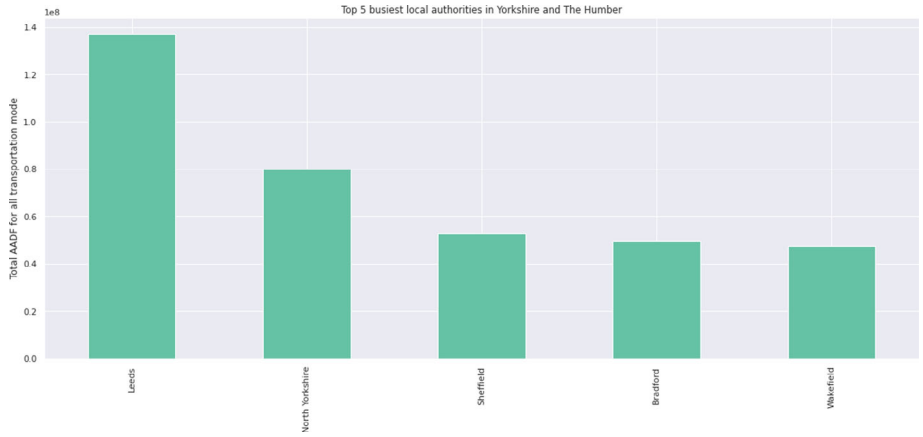

FIGURE 22 Busiest local authorities in Yorkshire and the Humber.

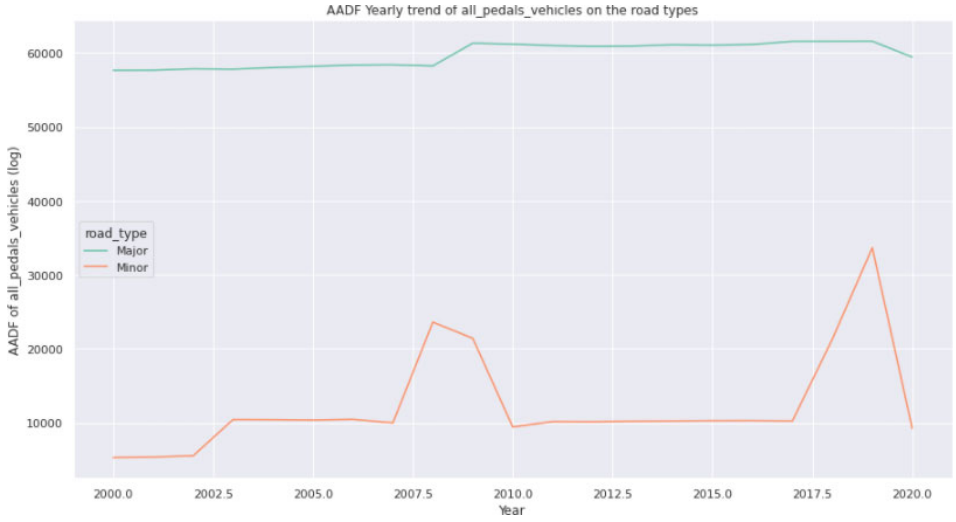

FIGURE 23 AADF trends of pedal vehicles.

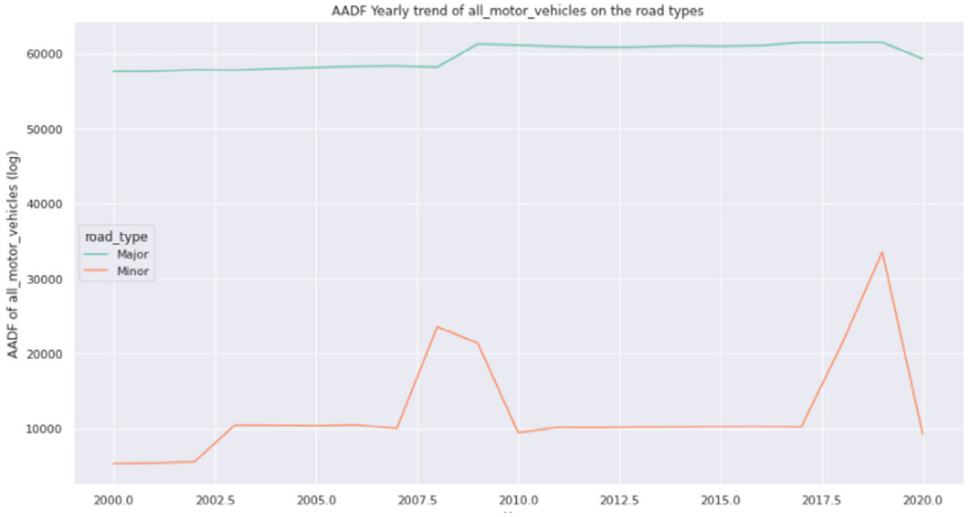

FIGURE 24 AADF trends of all motor vehicles.

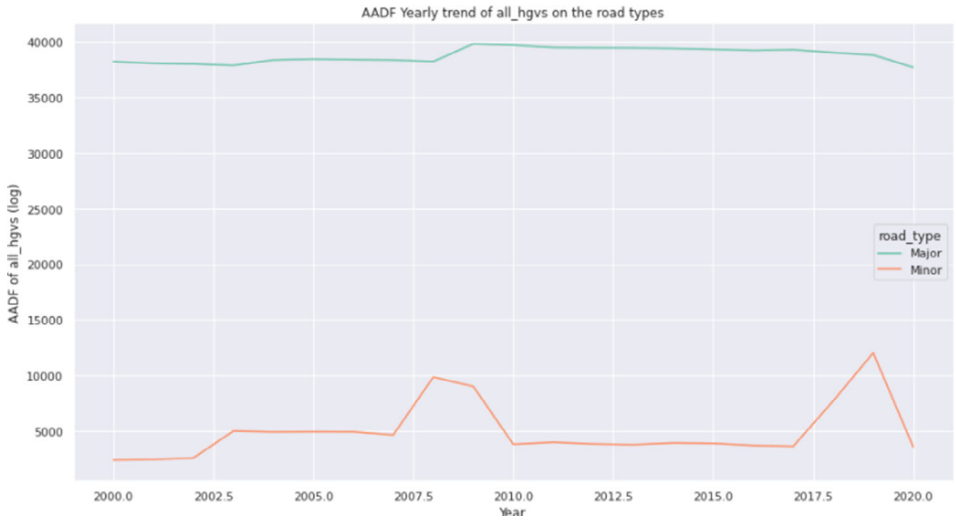

FIGURE 25 AADF trends of all HGVs.

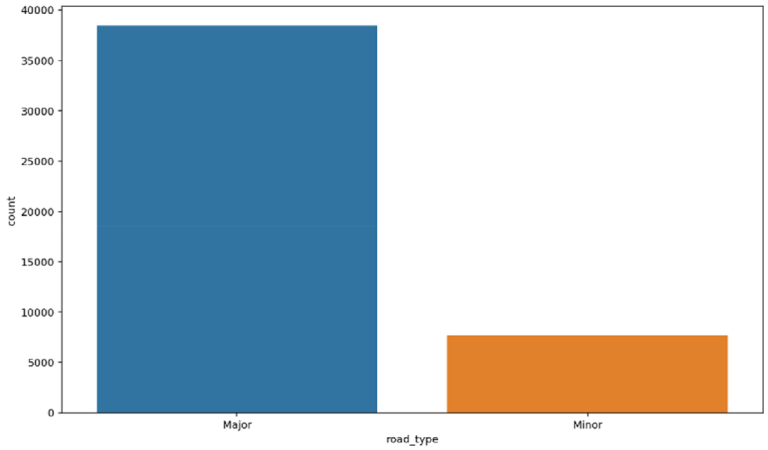

FIGURE 26 Analysis of road type in London.

4.3.5 | London regions

Figure 26 shows a similar distribution of road types between London as in Figure 3. Thus, this study can make relevant inferences using London and apply any possible findings to the rest of the United Kingdom.

Currently, the most congested roads within the United Kingdom are present in London, 'A406' and 'A23'. The most common vehicles traversing these roads are shown in Figures 27 and 28. Highway A406 and A23 circle around central London and connect all of North, East, South, and West, acting as a 'ring'. Initially, there were meant to be several 'ring' highways, but due to construction costs and pushbacks, only one was completed. That 'ring' highway was split into two sections, A406 and A23; thus, these highways now accommodate most vehicles, resulting in heavy congestion.

The following Figures 29–31, show a distinct difference between the distribution of vehicles passing major and minor road types. East and West London had more vehicles in all three categories compared to central London, likely due to the larger population present in outer London (Outer London contains a population of approximately 4.4 million while inner London has a population of 2.7 million; Brownlee, 2022). Surprisingly, central London does not have the highest number of vehicles passing on its roads in the three categories. In other words, there is an effective management of restricted usage, particularly restricted usage of HGVs. High fees are also applied to all vehicles if passing central London in peak hours.

From Figure 29, the busiest local authority is shown to be Westminster, located in central London, followed by Barnet, located in North London. Next is Hounslow and Hillingdon in West London and Tower Hamlets in East London. This study then analyses the AADF of all motor

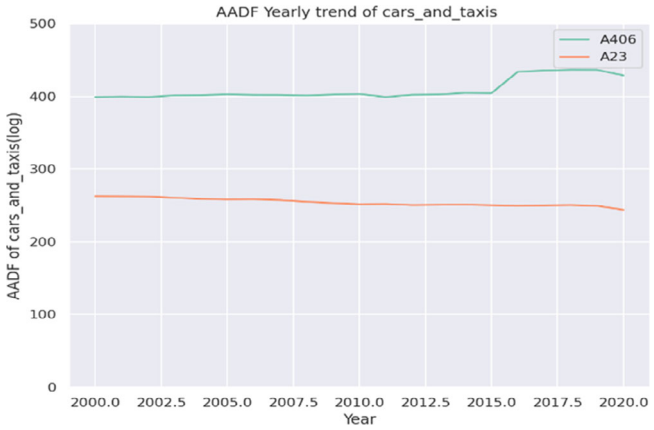

FIGURE 27 Recording of cars and taxis.

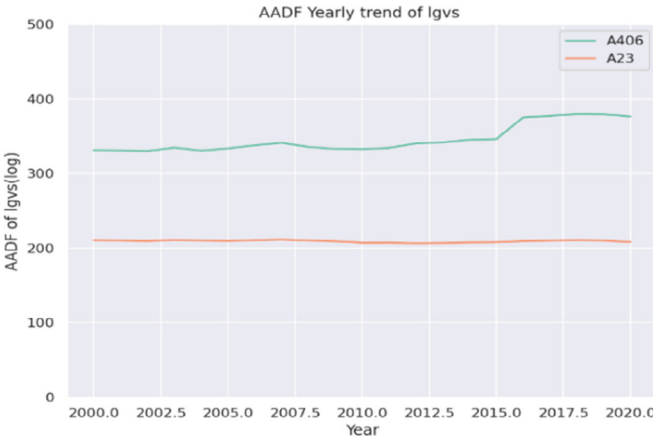

FIGURE 28 Recording of large ground vehicles.

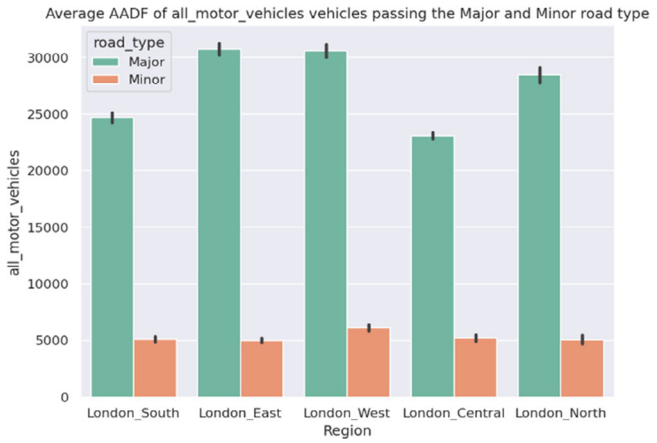

FIGURE 29 AADF of all motor vehicles.

vehicles, pedals, and heavy ground vehicles. Figures 29 and 30 show that East London and West London have the highest and second-highest vehicles on the major road, and West London has the highest number of vehicles on the minor road. Results show that the traffic controls in Central London have worked well on the major road but not on the minor road. More people have moved to East London since before London Olympics in 2012. Figure 31 shows East London and North London have the highest and second-highest heavy ground vehicles. North London has more offloading of goods and services since it is nearer to some industrial areas nearby. Therefore, it has the second-highest heavy ground vehicles.

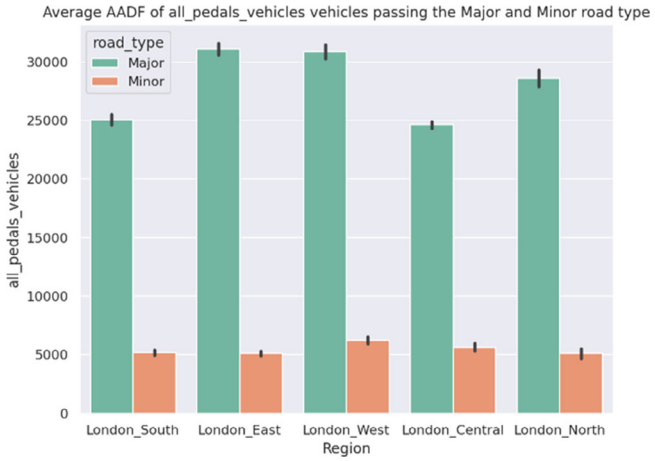

FIGURE 30 AADF of all nonmotor transport.

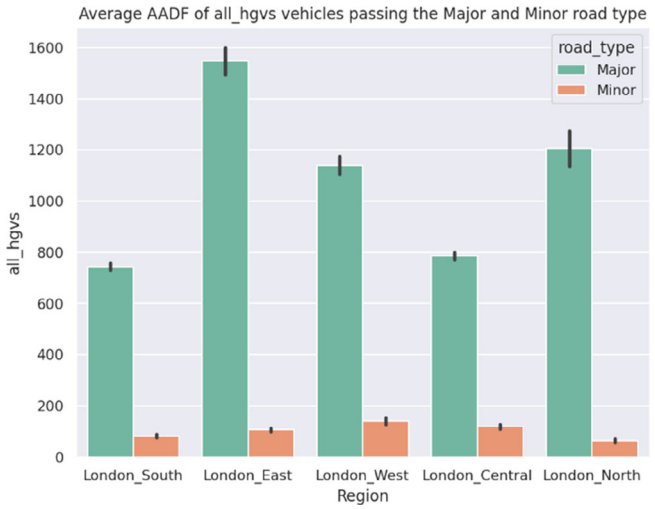

FIGURE 31 Heavy ground vehicles include 2–6-axle ground vehicles.

4.4 | Machine learning models

4.4.1 | Logistic Regression

The LR algorithm is shown in Figure 32 below.

4.4.2 | Decision Tree

The DT algorithm is shown in Figure 33 below.

4.4.3 | Random Forest

The RF algorithm is shown in Figure 34 below.

Input: Training data

1. **For i** ← 1 to **k**
2. **For** each training data instance $\mathbf{d}_i$:
3. Set the target value for the regression to
   $z_i \leftarrow \frac{y_j - P(1|d_j)}{[P(1|d_j) \times (1 - P(1|d_j))]}$
4. initialize the weight of instance $\mathbf{d}_j$ to P(1|$\mathbf{d}_j$) × (1 - P(1|$\mathbf{d}_j$))
5. finalize a f(j) to the data with class value ($\mathbf{z}_j$) & weights ($\mathbf{w}_j$)
6. Assign (class label:1) if P(1|$\mathbf{d}_j$) > 0.5, otherwise (class label: 2)

FIGURE 32 Logistic Regression algorithm.

GenDecTree(Sample **S**, Features **F**)

Steps :

1. **If** stopping_condition(**S**, **F**) = true **then**
   a. Leaf = createNode()
   b. leafLabel = classify(s)
   c. **return** leaf
2. root = createNode()
3. root.test_condition = findBestSpilt(**S**, **F**)
4. **V** = {**v**|**v** a possible outcome of root.test_condition}
5. **For each** value **v** ϵ **V** :
   a. $\mathbf{S}_v$ = {**s**|root.test_condition(s) = **v** and **s** ϵ **S**};
   b. Child = TreeGrowth($\mathbf{S}_v$, F):
   c. Add child as descent of root and label the edge {root → child} as **v**
6. **return** root

FIGURE 33 Decision Tree algorithm.

Precondition: A training set $\mathbf{S} := (x_1, y_1), \ldots, (x_n, y_n)$, features **F**, and number of trees in forest **B**.

```
1  function RandomForest(S , F)
2      H ← ∅
3      for i ∈ 1, . . . , B do
4          S^(i) ← A bootstrap sample from S
5          h_i ← RandomizedTreeLearn(S^(i), F)
6          H ← H ∪ {h_i}
7      end for
8      return H
9  end function
10 function RandomizedTreeLearn(S , F)
11     At each node:
12         f ← very small subset of F
13         Split on best feature in f
14     return The learned tree
15 end function
```

FIGURE 34 Random Forests algorithm.

4.4.4 | K-Nearest Neighbor

The KNN algorithm is shown in Figure 35 below.

Let (Xi, Ci) where i = 1, 2....., n be data points. X_i denotes feature values & C_i denotes labels for X_i for each i.
Assuming the number of classes as 'c'
$C_i \in$ {1, 2, 3, ..., c} for all values of i

Let x be a point for which label is not known, and we would like to find the label class using k-nearest neighbor algorithms.

1. Calculate "d(x, x_i)" i =1, 2, ..., **n**; where **d** denotes the Euclidean distance between the points.
2. Arrange the calculated **n** Euclidean distances in non-decreasing order.
3. Let **k** be a +ve integer, take the first **k** distances from this sorted list.
4. Find those **k**-points corresponding to these **k**-distances.
5. Let $\mathbf{k}_i$ denotes the number of points belonging to the i^{th} class among k points i.e. **k** ≥ 0
6. If $k_i > k_j \; \forall$ i ≠ j then put x in class i.

FIGURE 35 K-Nearest Neighbor algorithm.

Boost Algorithm under Loss Function L
Inputs : Joint distribution Q on $Z = x \times \{1, -1\}$ and initial predictor $H^{(0)} \in S[H]$.
In practice, Q is the empirical probability of training samples.

Loop For m = 1,...,M
1. Find a hypothesis $h^{(m)} \in H$ such that
$$h^{(m)} = \arg\min_{h \in H} \int_Z Q(dz) L'(-yH^{m-1}(x)) I(y \neq h(x))$$
The minimization does not need to be exact.

2. Find a coefficient $\alpha^{(m)} \in$ \mathbb{R} such that
$$\alpha^{(m)} = \arg\min_{\alpha \in \mathbb{R}} R_L(Q, H^{m-1} + \alpha h^{(m)})$$
Line search methods or Newton methods can be applied to slove the one-dimensional optimization problem.

3. Update the predictor: H^{(m)} = H^{m-1} + \alpha ^{(m)} h^{(m)}

Output: H^{(M)} as an estimated predictor.

FIGURE 36 Gradient Boosting Decision-Tree algorithm.

There are several distance metrics to conduct the calculation, including Euclidean distance, Manhattan distance, Minkowski distance, and Hamming distance. In addition to the distance metric, the other aspect is the *K* value, which defines the number of neighbours to be checked. It is essential to determine a suitable *K* as it can cause overfitting or underfitting and then influence the performance of the classifier.

4.4.5 | Gradient Boosting Decision Tree

The GBDT algorithm is shown in Figure 36 below.

5 | PERFORMANCE EVALUATION

5.1 | Feature selection and engineering

The label encoder turns all the category values in each column into numerical values for training and testing. The column road type was then standardized before the data was trained and tested. A heatmap of the correlation matrix was constructed to check if there was a positive or negative correlation between the attributes and the goal (i.e., the type of road), as well as if there was an association between attributes.

As shown in Figure 37, all the features positively correlate with the target label. Moreover, the correlation coefficient between any two features does not reach 1 or −1, indicating that these features are not perfectly correlated with each other. Therefore, this study keeps all the features because they contain different information related to the dependent variable in different degrees. These features will be used for creating the machine-learning model.

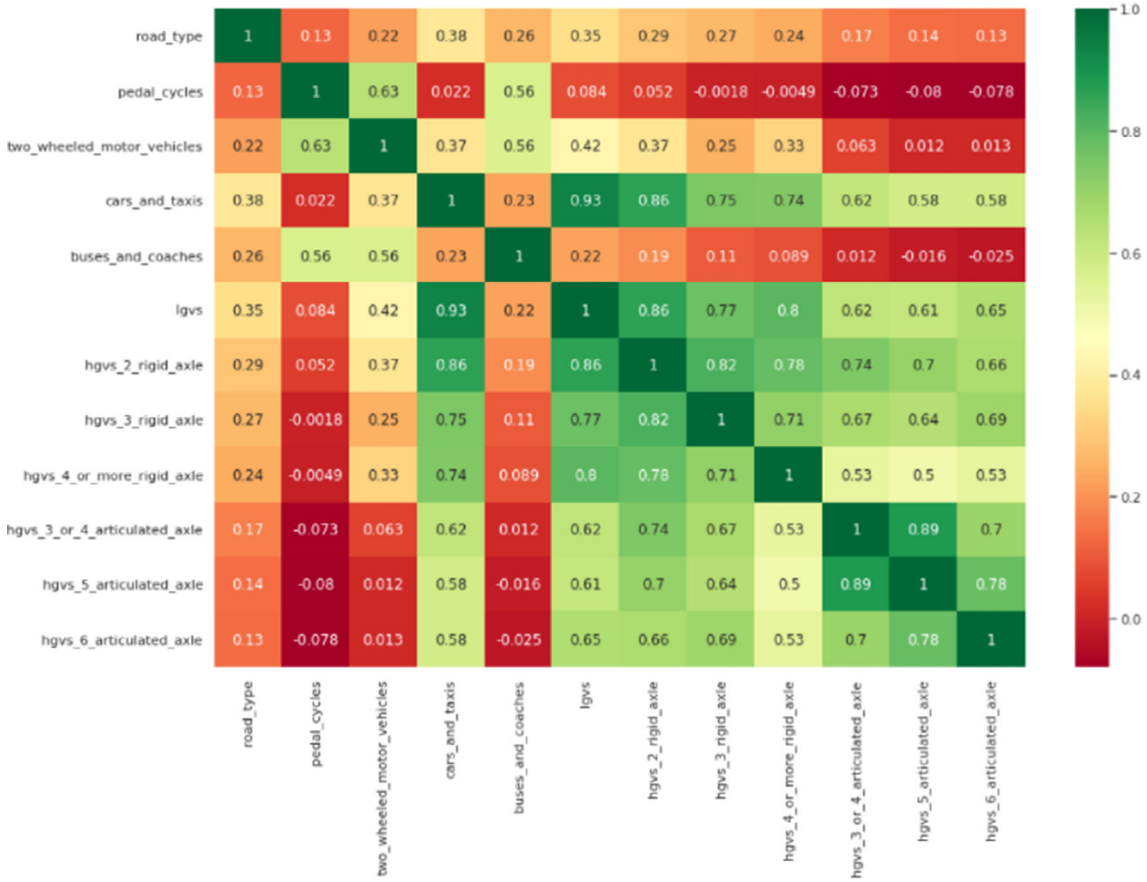

FIGURE 37 Correlation matrix.

After understanding the relationship between the attributes, the dataset was then divided into training and testing data. 60% of the data was used to train the model and 40% of the dataset was used to test its performance. Following that, due to the class imbalance, the training and testing data were oversampled and normalized separately before training the algorithm.

5.2 | Machine learning and analysis of results

The dataset is then exposed to the machine-learning classification procedure after it has been encoded and separated. Machine-learning algorithms are trained and developed using divided data. Logistics Regression, Decision Tree, K-Nearest Neighbors, Random Forest, and Gradient Boosting are the methods employed.

The accuracy and consistency of an ML model can be evaluated by comparing the output to the dataset and analysing the results utilizing a confusion matrix. A score of 1 means the model is great at predictions, while 0 means the model has no predictive power. The most used assessment metric is accuracy. It determines how well a model can anticipate outcomes. Its disadvantage is that when there is a class imbalance, the accuracy will compute for only one class while disregarding the other, making it a less-than-ideal metric for class imbalance. The F1 score reflects the balance of accuracy and recall. It is a measure of the model that is particularly useful where the dataset has imbalanced classes. The accuracy and recall values obtained from the machine-learning method are used to calculate the F1 score.

$$\text{F1 score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} +_{\text{Recall}}}$$

Table 3 and Figure 38 show that the classifier with the highest F1-score is Random Forest, followed by Gradient Boosting, K-Nearest Neighbors, Decision Tree, and Logistic Regression.

The ROC curve is created by plotting the true-positive rate against the false-positive rate at various threshold levels. The true-positive rate is also known as sensitivity, memory, or likelihood of identification in machine learning. The false-positive number, which may be specificity, is also

TABLE 3 ML model performance evaluation.

Models	Accuracy	F1-score	Roc-Auc score
Logistics Regression	0.935870	0.958587	0.982139
Random Forest	0.962566	0.976364	0.990056
K-Nearest Neighbors	0.947467	0.966461	0.953562
Decision Tree	0.942045	0.963532	0.905952
Gradient Boosting	0.950846	0.968432	0.989399

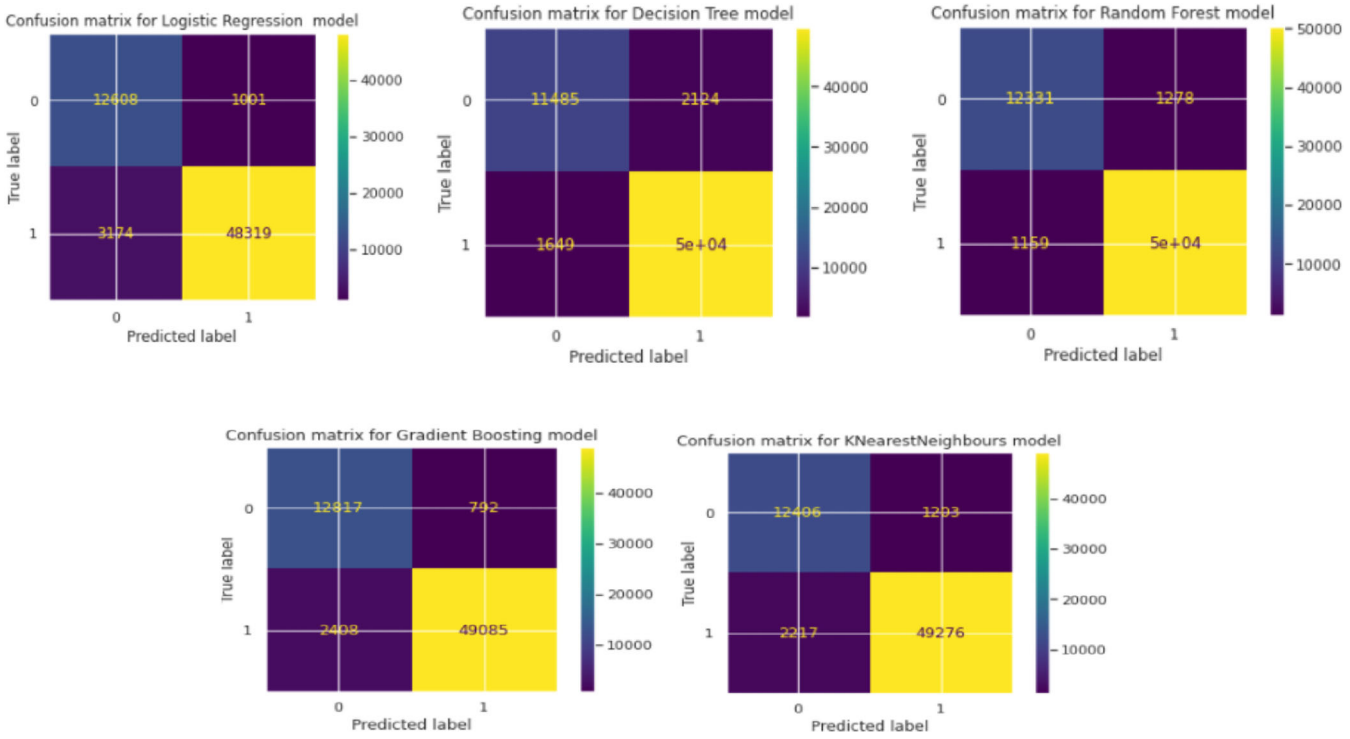

FIGURE 38 Confusion matrices for ML models.

known as the fall-out or probability of a false alarm. This graph depicts how effectively the model distinguishes between the target labels. The ROC-AUC score indicates how effectively the model predicts the appropriate classes for the problem. Based on this ROC curve, the Random Forest classifier is the best model, with the biggest area under the curve, whereas the Decision Tree model has the lowest performance. See Figure 39.

The ROC curve is created by plotting the true-positive rate against the false-positive rate at various threshold levels. The true-positive rate is also known as sensitivity, memory, or likelihood of identification in machine learning. The false-positive number, which may be specificity, is also known as the fall-out or probability of a false alarm. The Random Forest classifier appears to be the best on this ROC curve since it has the greatest area under the curve. Because it just shaded the poorest-curve rows, the Decision Tree model is the worst performance.

6 | DISCUSSION

It was necessary to research road networks and the classification of local government road networks within each region. The study was narrowed to analyse the five major and minor roads within each region and the determinant factor of traffic flow. It was discovered that the most obstruction of traffic flow on major roads in London is caused by Buses and Coaches due to high commercial activities. High policies for parking cars, whereas in the Northeast, obstruction of traffic flow is caused by taxis and cars. Similarly, in the Northwest, studies show that these regions are more residential settlements than other regions in this study. In contrast, the Yorkshire obstructions are caused by HGVs, and this arises because of the regional knowledge of industrialization activities compared with other regions in the study. The study has broadened the understanding of the British transportation system, including the top five busiest local authorities in each region and certain driving precautions to be aware

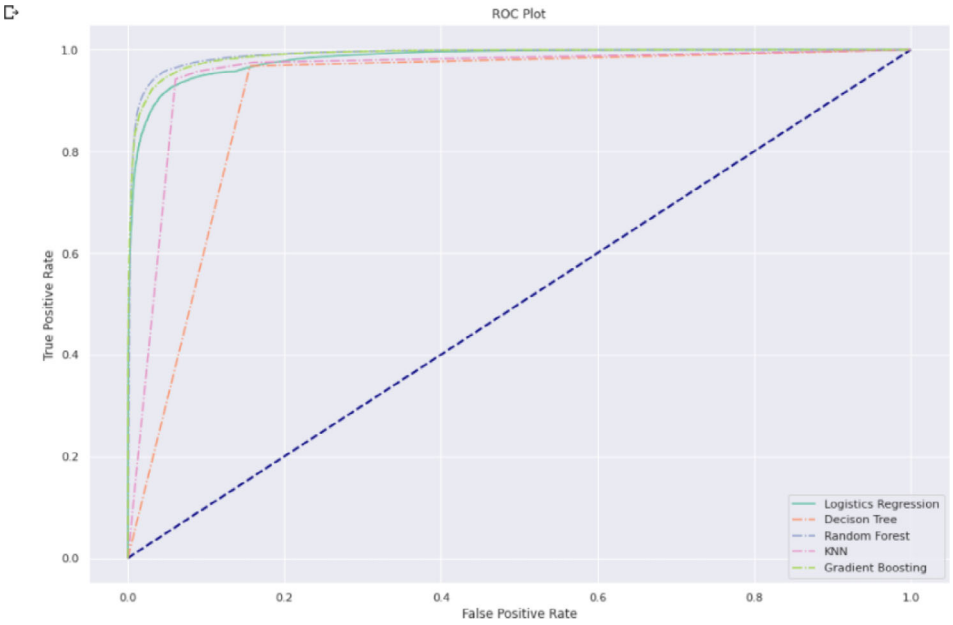

FIGURE 39 ROC curve.

of since they differ. The current traffic analysis research is focused on the frequency of usage and congestion status on the major and minor roads in each selected region. Although most vehicles tend to avoid usage of major roads in peak hours, avoidance of traffic congestion in some regions such as Yorkshire and the Humber is unlikely and improvements to improve traffic flows will be highly crucial. The futuristic algorithms should focus more on finding the paths that can get the least amount of time to the destination, rather than finding the shortest path that can result in congestion. Introducing policies similar to central London may help the current situation to some extent. A better long-term solution is to enable predictions on the traffic status and use intelligent algorithms to suggest how to use the roads to avoid congestion rather than funding the shortest path, which is widely used in current traffic analysis and GPS systems.

7 | CONCLUSIONS

The goal of this study was to understand traffic analysis achieved by developing algorithms to demonstrate traffic flows in focused regions in 2021. This study gathered data relating to four geographical areas from the AADF traffic flow data in the United Kingdom and used it to train five machine-learning algorithms. When combined with the other algorithms, this study performed a deep analysis. Finally, the Random Forest method had the greatest results, scoring above 96% on all performance parameters. The computational analyses of this research can be over 96% certain of the predictions if it is put into production. As traffic congestion can be problematic, it is essential to understand the patterns of traffic flow and recommend ways to improve traffic conditions with the smarter utilization of different routes. In summary, the research contributions of this study are in two folds. First, this study provides comprehensive reviews of traffic analysis in the United Kingdom and present insights useful for planning, traffic improvement, and decision making. Second, this study develops intelligent algorithms to provide useful and up-to-date analysis on traffic analysis, so that drivers can decide on their routes, particularly in peak hours.

This study also has limitations. One potential limitation could be the generalizability of the study. The study focuses on traffic analysis in specific geographical areas within the United Kingdom. The findings and performance of the algorithms may not be directly applicable to other regions or countries with different road infrastructures, traffic patterns, and driving behaviours. Another limitation is that the execution time of the algorithms was not tested in this study to analyse their performance from the perspective of efficiency. The future work may include developing more advanced algorithms to understand traffic analysis better and present results in more interactive ways through analytics and visualization. In addition, future research may expand the scope of the study to include traffic analysis in various regions or countries, providing a broader understanding of the performance and applicability of the algorithms across different contexts. The real-time data sources will also be integrated into the analysis as well as the execution time of the selected algorithms will be tested to allow timely decision making.

ACKNOWLEDGEMENTS

This work is partly supported by VC Research (VCR 0000186) for Prof. Chang. The authors thank Xianghua Gao and Tommy Wong for their help for a short period of time.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in Road traffic bulk downloads at https://roadtraffic.dft.gov.uk/downloads, reference number https://roadtraffic.dft.gov.uk/downloads. These data were derived from the following resources available in the public domain: https://roadtraffic.dft.gov.uk/downloads, https://roadtraffic.dft.gov.uk/downloads.

ORCID

Victor Chang https://orcid.org/0000-0002-8012-5852
Karl Hall https://orcid.org/0000-0003-2863-3312

REFERENCES

Andersson, A., & Chapman, L. (2011). The use of a temporal analogue to predict future traffic accidents and winter road conditions in Sweden. *Meteorological Applications*, *18*(2), 125–136.

Azimjonov, J., & Özmen, A. (2021). A real-time vehicle detection and a novel vehicle tracking systems for estimating and monitoring traffic flow on highways. *Advanced Engineering Informatics*, *50*, 1–14.

Brownlee, J. (2022). *Data preparation for machine learning*. Machine Learning Mastery.

Cai, H., Zhu, D., & Yan, L. (2015). Using multi-regression to analyze and predict road traffic safety level in China. In *2015 International Conference on Transportation Information and Safety (ICTIS)* (pp. 363–369). IEEE.

Chen, C., Liu, B., Wan, S., Qiao, P., & Pei, Q. (2020). An edge traffic flow detection scheme based on deep learning in an intelligent transportation system. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1840–1852.

Chow, A. H., Santacreu, A., Tsapakis, I., Tanasaranond, G., & Cheng, T. (2014). Empirical assessment of urban traffic congestion. *Journal of Advanced Transportation*, *48*(8), 1000–1016.

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*, *27*, 1071–1092.

Díaz-Rodríguez, N., Lamas, A., Sanchez, J., Franchi, G., Donadello, I., Tabik, S., & Herrera, F. (2022). EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: The MonuMAI cultural heritage use case. *Information Fusion*, *79*, 58–83.

Fang, Y. R., & Shen, F. M. (2012). Development trend analysis and prediction of traffic accident. *Journal of Safety Science and Technology*, *8*(3), 141–146.

Garre, A., Ruiz, M. C., & Hontoria, E. (2020). Application of machine learning to support production planning of a food industry in the context of waste generation under uncertainty. *Operations Research Perspectives*, *7*, 100147.

Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Mining and Analytics*, *4*(2), 116–123.

Hussein, A. S., Khairy, R. S., Najeeb, S. M. M., & Alrikabi, H. T. (2021). Credit card fraud detection using fuzzy rough nearest neighbor and sequential minimal optimization with logistic regression. *International Journal of Interactive Mobile Technologies*, *15*(5), 24–42.

Ji, B., & Hong, E. J. (2019). Deep-learning-based real-time road traffic prediction using long-term evolution access data. *Sensors*, *19*(23), 1–16.

Jomnonkwao, S., Uttra, S., & Ratanavaraha, V. (2020). Forecasting road traffic deaths in Thailand: Applications of time-series, curve estimation, multiple linear regression, and path analysis models. *Sustainability*, *12*(1), 395.

Kalair, K., & Connaughton, C. (2021). Anomaly detection and classification in traffic flow data from fluctuations in the flow–density relationship. *Transportation Research Part C: Emerging Technologies*, *127*, 103178.

Li, K., Shi, Q., Liu, S., Xie, Y., & Liu, J. (2021). Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree. *Medicine*, *100*(19), e25813.

Lingras, P., Sharma, S. C., Osborne, P., & Kalyar, I. (2000). Traffic volume time-series analysis according to the type of road use. *Computer-Aided Civil and Infrastructure Engineering*, *15*(5), 365–373.

Lubbe, N., Jeppsson, H., Ranjbar, A., Fredriksson, J., Bärgman, J., & Östling, M. (2018). Predicted road traffic fatalities in Germany: The potential and limitations of vehicle safety technologies from passive safety to highly automated driving. In *Proceedings of IRCOBI conference, Athena, Greece, September 2018*.

Mehrannia, P., Bagi, S. S. G., Moshiri, B., & Al-Basir, O. A. (2023). Deep representation of imbalanced spatio-temporal traffic flow data for traffic accident detection. *IET Intelligent Transport Systems*, *17*(3), 606–619.

Monfared, A. B., Soori, H., Mehrabi, Y., Hatami, H., & Delpisheh, A. (2013). Prediction of fatal road traffic crashes in Iran using the box-Jenkins time series model. *Journal of Asian Scientific Research*, *3*(4), 425–430.

Pappalardo, G., Cafiso, S., Di Graziano, A., & Severino, A. (2021). Decision tree method to analyze the performance of lane support systems. *Sustainability*, *13*(2), 846.

Quek, C., Pasquier, M., & Lim, B. B. S. (2006). Pop-traffic: A novel fuzzy neural approach to road traffic analysis and prediction. *IEEE Transactions on Intelligent Transportation Systems*, *7*(2), 133–146. https://doi.org/10.1109/TITS.2006.874712

Ramesh, R., Divya, G., Dorairangaswamy, M. A., Unnikrishnan, K. N., Joseph, A., Vijayakumar, A., & Mani, A. (2019). Real-time vehicular traffic analysis using big data processing and IoT based devices for future policy predictions in smart transportation. In *2019 International Conference on Communication and Electronics Systems (ICCES)* (pp. 1482–1488). IEEE.

Shafiei, S., Mihăiţă, A. S., Nguyen, H., & Cai, C. (2022). Integrating data-driven and simulation models to predict traffic state affected by road incidents. *Transportation Letters*, *14*(6), 629–639.

Shepelev, V., Aliukov, S., Nikolskaya, K., & Shabiev, S. (2020). The capacity of the road network: Data collection and statistical analysis of traffic characteristics. *Energies*, *13*(7), 1765.

Thakkar, A., & Lohiya, R. (2021). Attack classification using feature selection techniques: A comparative study. *Journal of Ambient Intelligence and Humanized Computing*, *12*(1), 1249–1266.

Ulbricht, C. (1994). Multi-recurrent networks for traffic forecasting. In *AAAI-94 Proceedings, aaai.org* (pp. 883–888).

Wani, M. A., & Roy, K. K. (2022). Development and validation of consensus machine learning-based models for the prediction of novel small molecules as potential anti-tubercular agents. *Molecular Diversity*, *26*, 1345–1356.

Zhang, J., Liang, Q., Jiang, R., & Li, X. (2019). A feature analysis based identifying scheme using GBDT for DDoS with multiple attack vectors. *Applied Sciences*, *9*(21), 4633.

AUTHOR BIOGRAPHIES

Prof. Victor Chang is a Professor of Business Analytics at Operations and Information Management, Aston Business School, Aston University UK, since mid-May 2022. He was previously a Professor of Data Science and Information Systems at the School of Computing, Engineering and Digital Technologies, Teesside University, UK, between September 2019 and mid-May 2022. He has deep knowledge and extensive experience in AI-oriented Data Science and has significant contributions in multiple disciplines. Within 4 years, Prof Chang completed PhD. (CS, Southampton) and PG Cert (Higher Education, Fellow, Greenwich) while working for several projects simultaneously. Before becoming an academic, he has achieved 97% on average in 27 IT certifications. He won 2001 full Scholarship, a European Award on Cloud Migration in 2011, IEEE Outstanding Service Award in 2015, best papers in 2012, 2015 and 2018, the 2016 European award: Best Project in Research, 2016-2018 SEID Excellent Scholar, Suzhou, China, Outstanding Young Scientist award in 2017, 2017 special award on Data Science, 2017-2022 INSTICC Service Awards, Talent Award Suzhou 2019, Top 2% Scientist 2017/2018, 2019/2020 & 2020/2021, the most productive AI-based Data Analytics Scientist between 2010 and 2019, Highly Cited Researcher 2021 and numerous awards mainly since 2011. Prof Chang was involved in different projects worth more than £14 million in Europe and Asia. He has published 3 books as sole authors and the editor of 2 books on Cloud Computing and related technologies. He published 1 book on web development, 1 book on mobile app and 1 book on Neo4j. He gave 38 keynotes at international conferences. He is widely regarded as one of the most active and influential young scientist and expert in IoT/Data Science/Cloud/security/AI/IS, as he has the experience to develop 10 different services for multiple disciplines. He is the founding conference chair for IoTBDS, COMPLEXIS and FEMIB to build up and foster active research communities globally with positive impacts.

Qianwen Ariel Xu is a PhD student at Operations and Information Management, Aston Business School, Aston University UK. She was previously a PhD student in Computer Science at the School of Computing, Engineering and Digital Technologies at Teesside University. Her dissertation research focuses on box-office prediction based on techniques of Artificial Intelligence and Data Science. She completed her master's degree in Business Analytics with Distinctions from the University of Liverpool, UK. She is also a member of Prof. Chang's research team. She is a hardworking, dedicated and resourceful student who can make things happen. She has published several publications in refereed academic journals, such as Technological Forecasting and Social change, Information Systems Frontiers, Expert Systems, Journal of Global Information Systems, etc.

Karl Hall is a first year PhD student, early career researcher and Part-time Lecturer in Computer Science at the School of Computing, Engineering and Digital Technologies at Teesside University. His doctoral thesis is concerned with discovering novel biomarkers linking COVID-19 and other diseases by developing an information retrieval framework and systems. His research is primarily focused on artificial intelligence and machine learning systems, focusing on disease diagnosis, bioinformatics, and healthcare analytics. He has authored a book chapter and several papers in academic journals, such as Future Internet, Healthcare Analytics, Decision Analytics and COMPLEXIS. He has also published work on Intrusion Detection systems for Fog and Cloud Computing, and social network analysis.

Olojede Theophilus Oluwaseyi completed MSc in AI and Data Analytics, Teesside University. He as worked with Prof Chang for a period of time for this research.

Dr Jiabin Luo is Lecturer in Business Analytics at the Department of Operations and Information Management, Aston Business School since September 2018. After completing her PhD in Management Science at the University of Southampton in 2013, she held postdoc researcher position at Coventry University, before joining De Montfort University as the Vice Chancellor's 2020 Lecturer in Business and Management in 2016. Her research interests are the optimisation and decision making in the area of logistics and supply chain management. She is Programme Co-Director of MSc Degree Apprenticeship: Digital & Technology Solutions Specialist. https://www2.aston.ac.uk/study/courses/digital-and-technology-solutions-specialist.

How to cite this article: Chang, V., Xu, Q. A., Hall, K., Oluwaseyi, O. T., & Luo, J. (2023). Comprehensive analysis of UK AADF traffic dataset set within four geographical regions of England. *Expert Systems*, e13415. https://doi.org/10.1111/exsy.13415