Automatic Accident Detection, Segmentation and Duration Prediction Using Machine Learning

Dr Pramila Rani K, Associate Professor,
Department Of Computer Science and Engineering,
New Horizon College Of Engineering,
Bengaluru, India

Vamsinath Reddy Renati
Department Of Computer Science and Engineering
New Horizon College Of Engineering
Banglore,India

Hanirudh Ravulapalli
Department Of Computer Science and Engineering
New Horizon College Of Engineering
Banglore,India

Naga Hima Sai Kushal Polisetty
Department Of Computer science and engineering)
New horizon college of Engineering
Banglore,India

Tejeswar Reddy Siddareddygari
Department Of Computer science and engineering
New Horizon College Of Engineering
Banglore,India

***Abstract*— Traffic accidents are often inaccurately reported, with incorrect location and disruption duration due to various external factors. This can result in imprecise predictions and inaccurate decision-making in data-driven models. To address these challenges, our study presents a comprehensive framework for traffic disruption segmentation from traffic speed data (obtained from Caltrans Performance Measurements system) in the time-space proximity of reported accidents (from Countrywide Traffic Accident dataset). Furthermore, we evaluate multiple machine learning models on reported, estimated, and manually marked disruption intervals, and demonstrate that our enhanced modelling approach reduces the root mean squared error (RMSE) of traffic accident duration prediction while providing higher similarity with disruptions observed in traffic speed. Our algorithm yields higher disruption detection precision than reported accident timelines. Although using multiple segments offers a slight decrease in the quality of results, it highlights more disruptions. Future research could explore expanding the algorithm's complexity and applying it to improve traffic incident impact predictions.**

***Index Terms*— Traffic management, traffic operations, traffic safety, accidents, accident detection, performance evaluation, traffic simulation, level of services, machine learning.**

I. Introduction

TRAFFIC accidents are a significant concern worldwide, causing fatalities, injuries, and economic losses. The number of vehicles has been substantially increasing during the past decades, which currently leads to an increase in the number of traffic accidents [1]. The National Highway Traffic Safety Administration (NHTSA) reported more than 5 million traffic accidents happening in the United States during the year 2013 [2]. Traffic Management Agencies usually rely on Traffic Incident Management Systems (TIMS) to collect data on traffic accidents, including information on various accidents, traffic states and environmental conditions. Accurately predicting the total duration of an incident shortly after it is being finished, will help in improving the effectiveness of accident response by providing important

information to decide the required resources to be allocated (response team size, equipment, traffic control measures) [3]. A traffic accident is a rare event with stochastic nature. The effect of the accident can be observed as an anomalous state in the time series of traffic flow [4].

Various terms and concepts are employed in the field of traffic accident duration prediction. Key terms include the Incident duration - The time between the occurrence of an incident and its clearance [5] and Predictive modelling - the process of developing data-driven models to forecast future outcomes, such as accident duration [6]. Important road safety concepts encompass that related to traffic incident duration prediction are the following:

Human factors: Elements related to driver behavior, such as attention, fatigue which affect decision-making Vehicle factors: aspects related to the vehicle itself, including design, maintenance, and safety features Infrastructure factors The design, con- struction, and maintenance of roads and their surroundings,

Traffic management [10]: Measures and strategies implemented to improve the efficiency, safety, and sustainability of road networks. In our research, we focus on a possible contribution to the field of traffic management by employing traffic speed disruption detection and performing traffic incident duration prediction with higher accuracy.

It allows to eliminate user-input errors from reports and improve the accident duration prediction performance in many traf- fic management centres around the world. To help address this issue, in our paper we propose various methods for a correct traffic disruption segmentation, the method for an association between vehicle detector stations and accident reports.

Another important challenge is that many incident data sets around the world are private and not shared for public investigation; for those open data sets, there are several missing information fields, or even worse, incomplete information regarding the traffic conditions in the vicinity of the accidents. Even often publish crash data sets are limited in size as well and contain a very small number of records. This represents a tight constraint when testing one framework over multiple countries with different traffic rules and regulations. For our studies we have oriented our attention towards two big open data sets - CTADS (Countrywise traffic accident data set) which contains 1.5 million accident reports and the Caltrans Performance Measurement System (PeMS) which provides data on traffic flow, traffic occupancy and traffic speed across California. Despite both being extensive data sets, vehicle detector station readings from PeMS are not associated with traffic accident reports from CTADS either by time, location or coverage area. The lack of such association makes it impossible to analyse the relation between accidents and their effects on traffic flow and speed. To address this challenge, in our paper we introduce the following mapping algorithm which will secure several steps such as:

- an association of Vehicle Detection Stations (VDS) with reported accidents in their proximity,
- a segmentation of traffic speed disruptions from detector readings,
- an association of detector stations with reported accidents (we will further show that this step is necessary due to many detected user-input errors in accident reports).

As a result, we obtain traffic disruptions segmented by the traffic speed associated with reported accidents. This association makes it possible to perform various important tasks of the accident analysis: prediction of the traffic accident impact on the traffic speed based on accident reports, prediction of the traffic accident duration derived directly from the effect of disruption on the traffic speed (impact-based duration), analysis of disruption propagation (each detected disruption can be studied for spatial-temporal impact within the traffic network). Through this work, we will focus on the prediction of the impact-based accident duration and lay the foundation for a further research.

Overall, the main contributions (summarised in Figure 1) of our paper are as follows:

1. We propose a fusion methodology of two large data sets (CTADS and PeMS) for a detailed traffic accident analysis. To the best of our knowledge, this is the first research study proposing the methodology for merging of these two large data sets, which allows an association between observed disruptions in traffic flow and the reported accidents.

The research of this nature (fusion of traffic flow and accident reports) has been performed before [13], [14], but our methodology has the following advantages: Our disruption segmentation model can be fine-tuned via hyper-parameter search to find optimal disruption detection rate, The method produces difference estimates proportional to the degree of observed disruption, which allows for control of false positives rate via threshold choice, We evaluate multiple compar- ison metrics for traffic speed difference estimation, The segmentation algorithm is more complicated and includes pre-processing convolution, test of multiple difference metrics, adjustment to selectivity and cyclic shift for difference window, our methodology is modular, where each logical part can be further refined and studied in a separate research.

2. We propose a novel methodology for the disruption mining using a combination of different metrics (which we further find to have properties important for disruption segmentation): a) the Wesserstein metric, which allows us to measure the disruption severity and b) the Chebyshev metric, which provides a higher selectivity for the disruption mining and a rectangular shape of the disrupted segments, allowing an automated disruption segmentation. We detail all unique properties of both metrics utilized together to allow an accurate disruption segmentation.

3. We perform the estimation of traffic accident disruption duration from traffic speed via the above metrics which allows us to alleviate user-input errors in accident reports.

4. We evaluate multiple machine learning models by comparing both the reported and the estimated accident duration predictions extracted from traffic speed disruptions.

5. We introduce a new modelling approach which focuses on the amount and shape of the the disruption associated with an accident, which allows a further analysis and modelling of accident impact.

In contrast to one of the previous studies [14], which utilized Fuzzy Modelling, Multi-layer Perceptron, Weibull Regression, and Log-logistic Regression, our methodology that offers a higher degree of complexity. We rely on advanced machine learning models with the use of a disruption segmentation algorithm, which relies on multiple hyper-parameters. This design allows fine-tuning to find the optimal disruption detection rate.

Overall, this research forms the foundation for a newearly traffic accident disruption detection, traffic disruption speed impact analysis and the use of observed traffic accident durations for correcting errors in user reports. Moreover, this work contributes to our ongoing objective to build a real-time platform for predicting traffic congestion and to evaluate the incident impact (see our previous works published in the figures.

The paper is further organised as follows: Section II discusses related works, Section III-A presents the data sources available for this study, Section III showcases the methodology, Section IV presents the disruption segmentation results, showcases the result of data set fusion, Section V presents the ablation study and Section VII provides conclusions and future perspectives.

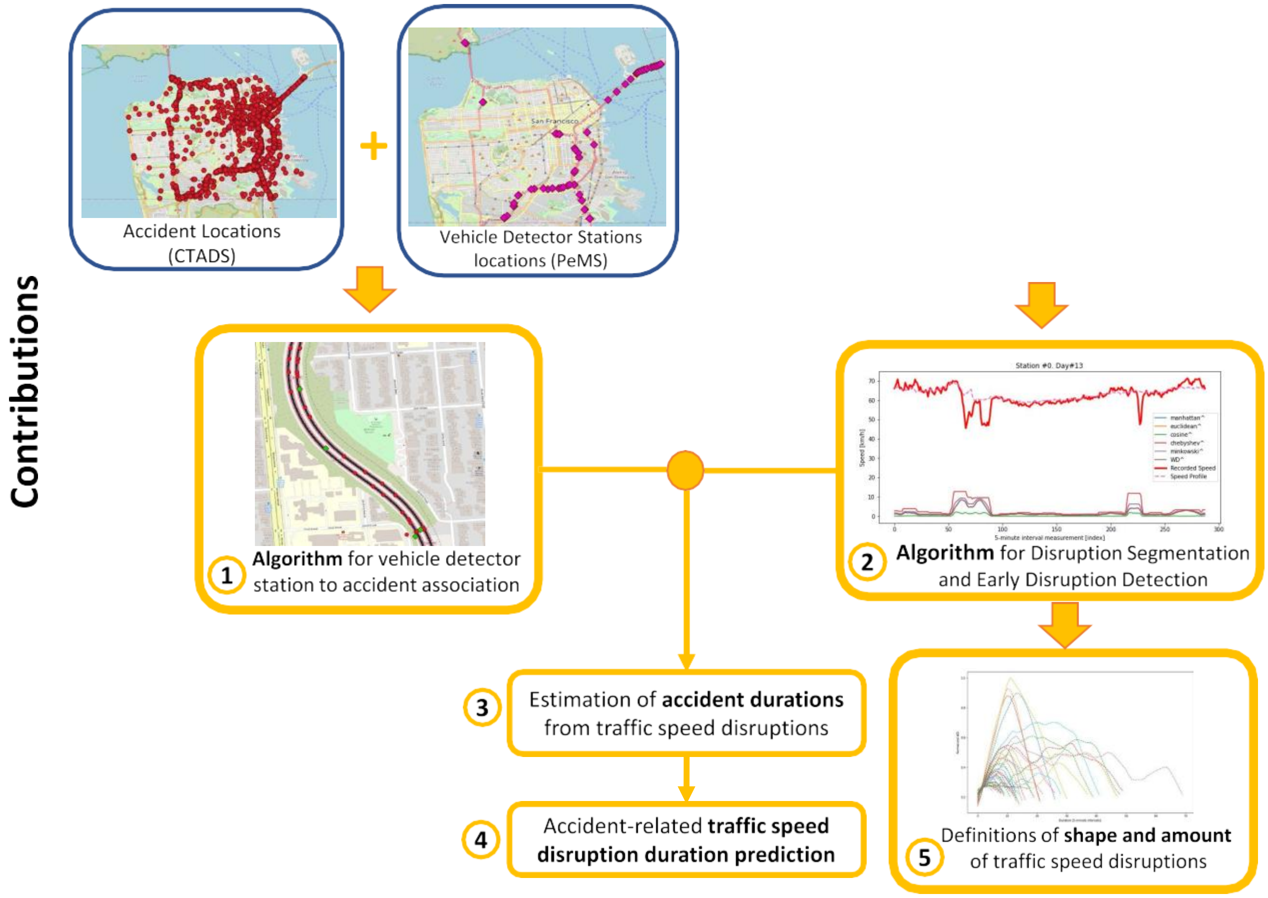

Fig. 1. Contributions and data-flow schema for association of traffic speed readings with accident reports.

II. Related Works

Multiple studies rely on user-input-based incident reports from Traffic Management Centers (TMC) with different machine learning models to predict the traffic incident duration [17]. The use of traffic flow features is found to be rare and mostly specific - incident detection and incident impact prediction by using traffic flow [18]. In other words, traffic flow data is rarely combined with actual incident reports since it requires a higher system complexity and extensive data collection.

A. Anomaly Detection Related Works

There are numerous studies related to the accident detection problem from traffic flow using anomaly detection techniques [19]. Various methods used for anomaly detection in time series are applicable for the task of traffic disruption detection. The ability to perform the detection of an actual disruption, should give us the actual shapes of disruptions and time intervals and allows an in-depth analysis of usual accident statistics, including the effect of the type of accident on the pattern of disruption in traffic flow. By integrating data on the traffic state with accident reports we are able to further connect traffic flow disruption patterns to various accident characteristics (hour of the day, weather conditions, crash type, type of vehicle involved - truck/car [20], the effect of road pavement types [21], the road design and the road operation [22], etc).

Anomaly detection in time series data is a critical problem in various applications, such as finance and transportation. The data generated by many transportation applications (e.g. vehicle trajectory or vehicle loop data acquisition) is a continuous temporal process [23]. The detection of unusual events performed in a time-critical manner, is known as streaming outlier detection. There are two main aspects of the anomaly detection from traffic speed time series: continuity - traffic accident can be characterised by performing an abrupt change (which can be reformulated as a lack of continuity [24]) in traffic speed with steady or also abrupt return of traffic state back to normal condition after the accident elimination, and novelty - traffic accidents can also generate unusual unobserved earlier patterns of change in the traffic speed.

There are multiple approaches for time series anomaly detection:the sliding window technique, enabling contin- uous monitoring and timely detection of outliers, Offline Outlier Detection (OOD) using predictive and statistical mod- els, which processes data collected and analyzed later. This approach includes: a) ARIMA Models [25] for capturing temporal dependencies, b) Seasonal Hybrid ESD (S-H-ESD) [26], which combines ESD and STL techniques for high accuracy and robustness in detecting anomalies in time series with strong seasonal patterns, and c) LSTM networks [27], a type of recurrent neural network, for capturing long-range dependencies and estimating miss-prediction costs using moving window predictions.

Offline Outlier Detection using anomaly detection models perform using the following models: a) Isolation Forest [28] which is an unsupervised learning algorithm specifically designed for anomaly detection. It works by recursively partitioning the dataset using randomly selected features and split values, constructing multiple isolation trees in the process. The rationale behind this approach is that anomalies are generally more susceptible to isolation when compared to regular data points. Consequently, the path length from the root node to an anomalous point in the isolation tree is expected to be shorter than that for a regular data point. The average path length across all trees is then used as an anomaly score, with shorter path lengths indicating a higher likelihood of being an outlier. The method was also previously used for time series anomaly detection [29], b) One-Class SVM [30] which is a variant of the Support Vector Machine (SVM) algorithm tailored for unsupervised anomaly detection. It aims to find the smallest hyperplane that separates normal data points from the origin in the feature space, thereby constructing a boundary around the normal data. This is achieved by solving a quadratic optimization problem that maximizes the margin between the data and the origin. Any data point that falls outside the boundary is considered an anomaly.

Streaming outlier detection is important for timely detection of unusual events, such as traffic accidents. Continuity and novelty are the two main aspects of anomaly detection in traffic speed time series. There are multiple approaches to perform anomaly detection from time series, including the sliding window technique, offline outlier detection using predictive and statistical models, and offline outlier detection using anomaly detection models.

B. *Data Sets for Incident Duration Prediction*

Analysis of the effect of traffic incidents has been performed previously using Caltrans PeMS data, where the measure of incident impact was represented as a cumulative travel time delay [43], which is an aggregated value. However, traffic state recovery from disruptions is not necessarily following a single pattern - it may be slowly dissipating, we may observe secondary crashes, it may have a high or low impact, etc. Traffic accident duration prediction methodology relies on reported traffic accidents, but actual reports may contain user-input errors and be misaligned with the actual shape of disruption produced by the accident. Therefore, the approach for disruption segmentation may provide the accident duration estimated from the actual shape of disruption in traffic flow.

III. Methodology

The new framework we propose in this paper is represented in Figure 1 which we support across some initial definitions for our modelling approach (see next sub-section). First, we associate the road segments with their corresponding Vehicle Detector Stations (VDS) from the Caltrans PeMS data set, as well with the locations of reported accidents (see Algorithms 1 and 2 proposed in sub-section III-D). The main outcome of this algorithm is that traffic accidents will get associated with the traffic flow, speed and occupancy readings from the VDS stations.

Second, we propose a new algorithm for early disruption detection and segmentation, detailed in sub-section III-E. By detecting disruptions that occurred in time-space proximity of reported traffic accidents, we obtain the estimated traffic accident duration. This gives us much more information to include in the model training than just the simple accident duration.

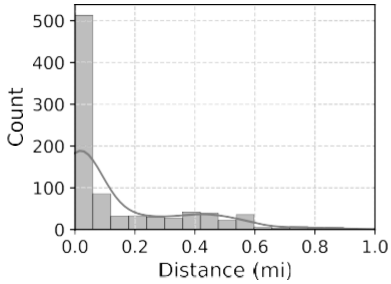

Fig. 3. CTADS: Bing - Histogram for recorded accident extent.

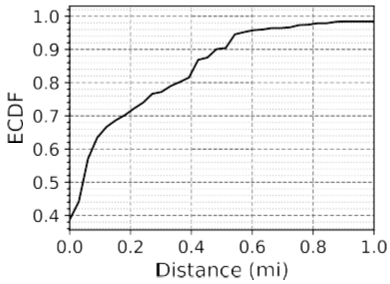

Fig. 4. CTADS: Bing - ECDF for recorded accident extent.

TABLE I
STATISTICS ON ACCIDENT EXTENT FROM CTADS (BING PART) DATA SET

Statistic	Value [km]
Mean	0.16
Median	0.04
0.95 Quantile	0.57
0.05 Quantile	0
Standard Deviation	0.27
Variance	0.08
Interquartile Range [0.25, 0.75]	0.27

aggregated 5-minute measurements of traffic flow, speed and occupancy across California. We decided to extract the data for the area of San-Francisco (see Figure 5a), which contains 83 Vehicle Detection Stations (VDS) placed in that area (see 5b), and we try to associate each traffic accident occurred with each of San-Francisco VDS in their 500m proximity using the algorithm detailed in the following section. In total, from 9,275 accidents in the area (extracted from CTADS) we

can be used to construct a detailed accident timeline that highlights the key events and their corresponding degrees of

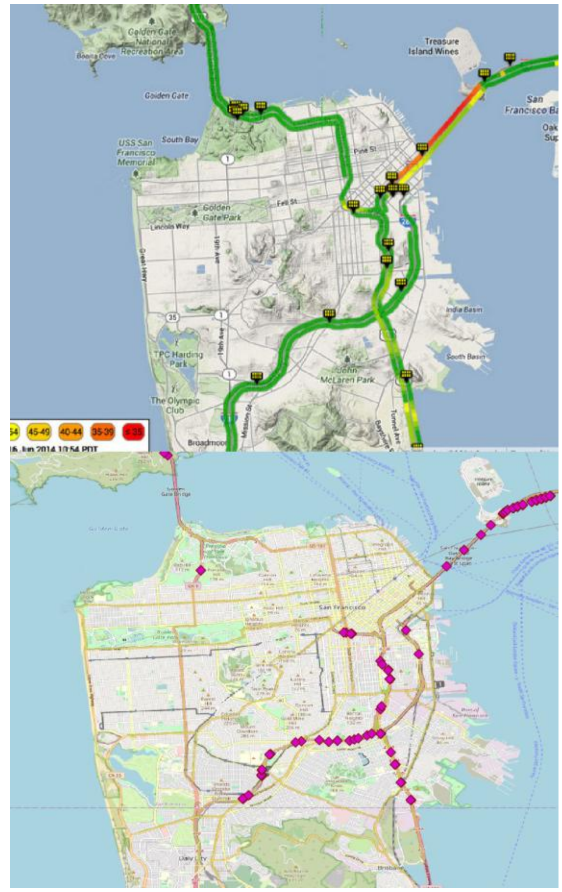

Fig. 5. 1) PeMS data set area coverage for San-Francisco (the map is available at https://pems.dot.ca.gov/) 2) Mapping of the Vehicle Detection Stations from PeMS data set. OpenStreetMap excerpt showing San Francisco. Available at:https://www.openstreetmap.org/#map=12/37.7612/-122.4395.

IV. RESULTS

A. Data Exploration and Setup

CTADS data set contains traffic accident reports, which after an initial data mining investigation, we found to contain several user-input errors; for example, a lot of traffic accident durations have been rounded to 30 or 360 minutes (see Fig. 7d)); or the incident start time which was reported is unrelated to any disruptions observed by the vehicle detector stations in the proximity - see Figure 7 in which we have provided two different examples of speed recorded during two different accidents A-5198 and A-4490; the red lines indicate the official reported start and end time of the accidents, while in reality the accidents have had a long lag in spreading across the network - see Fig. 7a) or were reported much later that the official speed drop was recorded - see Fig. 7b).

At this step we observed a significant amount of user-input errors in accident reports, which affect the accident combined with the semantic segmentation methods to create a joint machine learning model. This integrated model could leverage both the event-driven dynamic context and the mathematical metrics for segmentation to improve its predictions for incident duration and severity. By connecting these two research approaches, a more comprehensive framework for analyzing traffic accidents and predicting disruption durations can be developed. This integrated approach would benefit from the strengths of both methods, enabling more accurate and reliable predictions for incident durations. Ultimately, this could lead to improvements in road safety, emergency response, and traffic management.

In conclusion, image segmentation methods can be employed to not only segment the accident scene but also to quantify the degree to which an accident is observed in an image. This information can be used to create an accident timeline that reflects the progression of the accident, the severity of the event, and the critical moments when interventions or safety measures could have been taken. This approach in combination with our proposed disruption segmentation method can potentially contribute to better accident analysis, road safety improvements, and more effective emergency response strategies.

B. *Automated Disruption Segmentation Results*

Figure 9 presents the results obtained from our algorithm for the automated disruption segmentation. The segmentation line (dotted blue) represents the estimated disruption intervals represented as 0 and 1 to perform our visualisation investigation better. Figure 9a) shows that there may be multiple observed disruptions in a 300*5=1500 time interval. Due to errors in accident reports regarding the starting time and the duration of the accident, it is non-trivial to determine which disruption is associated with the accident. The situation may be easier in the case when only one disruption is observed during the day. According to our algorithm, we select the largest disruption on the day the accident was reported. Figures 9b) and 9c) highlight additional specific situations which need to be considered:higher traffic speed at the end of the day than observed from the monthly profile, unstable traffic speed approaching normal traffic conditions with high frequency, slight misalignment of disruption intervals with the visually observed disruption intervals. All these problems can be addressed by using manual segmentation with deployment of Deep Learning models since there are advanced computer vision methods proposed in recent years (e.g. autoencoders for segmentation).

C. *Comparison of Estimated, Reported and Manual Markup of Accident Durations*

There is a significant difference between the estimated and the reported accident durations that we would like to highlight: 1) the reported accident durations contain a large amount of 30 and 360 minutes duration values (nearly 40% of data - see Figure 10a)) while the estimated accident durations using our approach have an average duration of 58 minutes, while the reported is 108 minutes (which is by assumption skewed due to 360 placeholder values), the estimated accident dura- tions are distributed between 90 and 355 minutes (0.10 and 0.90 quantiles correspondingly) , while the reported durations are distributed between 29 and 360 minutes (see11 a) and manually detected disruptions distributed between 75 and 440 minutes), which highlights that disruptions observed from traffic speed are much shorter than reported in the original data set, 4) There is no noticeable correlation between observed and reported durations with high amount of horizontal anomalies in reported accident durations (see Figure11). Traffic accident duration is most common to follow log-normal or log-logistic distribution and on resulting plots, we see that accident reports are found to represent log-normal distribution to less extent than manual markup or estimated accident duration.

To perform the ablation study, we perform a manual markup of disruptions observed in traffic speed for 800 accidents, which will be discussed in the corresponding section.

D. *Extraction of Disruption Shapes*

In previous subsections we applied a Chebyshev metric to perform segmentation of disruptions. To analyse the disruption impact we apply the Wasserstein difference between monthly speed profile and daily traffic speeds and extract the corresponding disruption intervals. Wasserstein difference, originally named an Earth Mover distance, has an intuitive physical interpretation - the minimum “cost” of altering one pile of earth into the other, which is assumed to be the amount of earth that needs to be moved times the mean distance it has to be moved. In application to traffic state, it is the minimum amount of work necessary to alter the traffic state to disrupted condition, or in other words - the amount of disruption. We compare normalized metric values since every at every vehicle detector station there is a different average traffic speed. As in our proposed algorithm, we use a 12-units moving window (one hour) to estimate the Wasserstein difference between traffic speed measurements and provide the plot for the first 40 segmented disruptions, which allows for shape analysis of traffic disruption amount (see Figure 12): We observe the similarity between multiple disruptions - they have a ‘hill’ shape, there are secondary (double ‘hill’) and long-lasting disruptions. The observed shapes can be defined through the parametric equation to perform the classification of disruption effects and facilitate the prediction of disruption impact timeline since we observe that high-peak fast-ascending disruptions have a probability to end sooner than slowly ascending ones. The analysis of the speed of ascendance has potential to perform the early classification of disruptions, which is planned for further research.

TABLE II
MEAN ABSOLUTE ERROR (MAE) RESULTS

Model	Reported	Manual	Estimated
RandomForest [33]	26.52	21.89	17.21
XGBoost [34]	24.22	23.06	18.29
GBDT [37]	26.50	22.37	17.46
CatBoost [36]	23.96	21.58	17.55
LightGBM [35]	36.57	22.43	18.26
KNN [31]	44.11	26.22	19.73
LinearRegression	76.76	24.12	17.82
SVM [32]	84.82	23.70	17.55
NeuralNetwork [55]	55.34	24.33	19.27
RidgeRegression [56]	84.72	24.26	17.87
Target	(Reported)	(Manual)	(Estimated)

TABLE III
ROOT MEAN SQUARED ERROR (RMSE) RESULTS

Model	Reported	Manual	Estimated
GBDT [37]	73.14	30.46	22.11
CatBoost [36]	73.05	29.64	22.55
Random_Forest [33]	93.73	29.94	21.79
XGBoost [34]	82.67	31.75	23.61
LightGBM [35]	99.77	30.55	23.58
KNN [31]	142.97	35.27	24.45
Linear_Regression	117.53	32.54	22.35
SVM [32]	208.29	34.23	23.66
Neural_Network [55]	124.21	33.38	23.18
Ridge_Regression [56]	134.71	32.48	22.21
Target	(Reported)	(Manual)	(Estimated)

The highest RMSE is reported by the SVM model, with an estimated value of 208.29. As with the MAE results, the CatBoost model outperforms all the other models by a significant margin. All the methods use default parameters as they are presented in Scikit-learn [54] and corresponding modules.

When we are using accident reports to predict the estimated accident duration, we obtain a better performance using the RMSE metric across all the regression models, which may be connected to the lower amount of long accident durations than reported.

These best-performing models are all complex tree methods, which utilize multiple learners (via ensembles and boosting) to gain better predictive performance. They work well with mixed types of data (numeric and categorical), can capture non-linear relationships, and are less prone to overfitting. On the contrary, Linear Regression assumes a linear relationship between the input variables and the single output variable, KNN assumes that similar instances are near to each other, and SVM assumes that the data is linearly separable by a hyperplane in a feature space. Low performance of these methods shows that these assumptions may not align well with the data in case of traffic accident reports.

The reported duration, as provided directly from the source or via some other form of direct measurement is subject to more variability due to factors such as measurement errors (incorrectly reported duration), reporting biases ("rounded" 30 and 360 minute durations), or other uncontrolled external influences (late accident detection, disruption effects misaligned to reported accident timeline). We expect that correct estimation of the incident duration contributes to reduction in modelling complexity due to reduced effect of outliers, bias and errors on prediction performance.

In contrast, the manual and estimated durations are derived using more controlled processes and algorithms. The manual duration calculated by a consistent procedure, minimizing the room for error. The estimated duration, relies on parametric model, would also tend to have less variation due to the model fine-tuning to minimize prediction error based on the available data.

Overall, the CatBoost model consistently outperforms all the other models across all metrics.

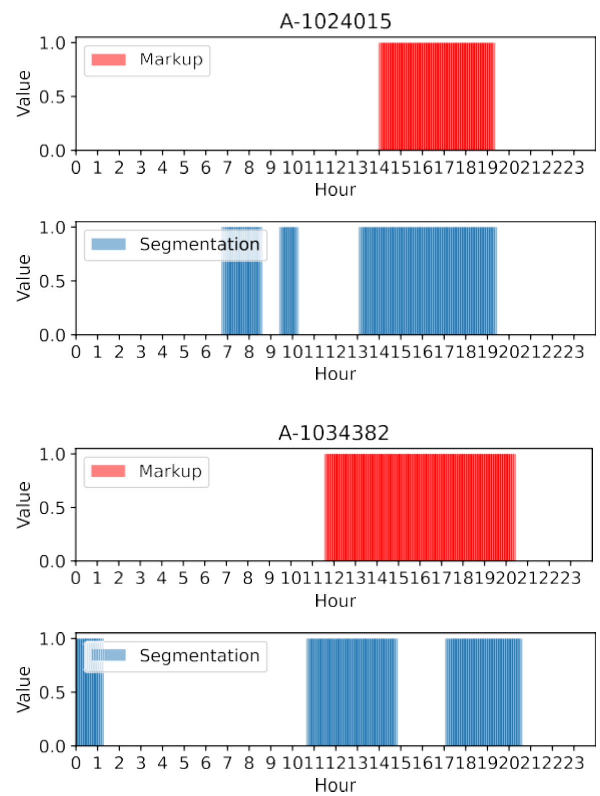

Fig. 13. Manual markup and algorithm segmentation comparison. Time series segments represented as binary values of 0 and 1.

V. ABLATION STUDY

In this paper, we propose using the F1 score to estimate the quality of time interval segmentation in binary time series (see Figure 13) in which we provide two different examples of different stations with both manual markups of the incidents - red markups- and our segmentation algorithms - blue markups- that is more efficient at detecting multiple incidents throughout the 24h time period and not only one single isolated event. The value on Y-axis shows a positive 1.0 value if the interval contains the disruption. Examples are provided for Accidents with ID A-1024015 and A-1034382 from CTADS data set.

Given a ground truth dataset with original reported accident duration, we perform a manual labelling of segments and obtain a set of predicted segments obtained from our automated segmentation algorithm, we compute the precision and the recall of the algorithm, and then combine them into a single F1 score.

F1-score is a popular metric used to evaluate the quality of binary classification models defined as follows:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$
$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

where true positives are the number of correctly classified positive instances, false positives are the number of negative instances classified as positive, and false negatives are the number of positive instances classified as negative.

F1-score is defined as the harmonic mean of precision and recall, given by:

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F1-score ranges from 0 to 1, with higher values indicating a better classification performance.

In the case where a time series is represented as a series of points with values of 1 for segmented intervals and 0 for intervals with no segments, F1-score can be applied to estimate the quality of the time interval segmentation.

To apply the F1-score, we need a ground truth dataset with manually labelled segments (and we obtain this manual markup for 820 accidents), and a set of predicted segments obtained from our automated segmentation algorithm. We can use these two sets to compute the precision and recall of the segmentation algorithm, and then combine them into a single F1-score.

Precision measures the proportion of true positives among all the predicted positives. In the context of time interval segmentation, the precision measures the accuracy of the algorithm in detecting the true segments. The Recall measures the proportion of true positives among all the actual positives. In the context of time interval segmentation, the recall measures the completeness of the algorithm in detecting all the true segments.

To apply the F1-score to estimate the quality of time interval segmentation, we can compute the precision and recall for each segment, and then compute the overall F1-score as the weighted average of precision and recall, weighted by the number of segments. This provides a single metric that reflects the quality of the time interval segmentation.

As a result (see Figure 14), the official reported incident segmentation is found to be very off (with a mean F1-score of 0.29 - Figure 14a)); next, the segmentation done by the algorithm while selecting only the interval closest to the reported timeline yields the highest average F1-score of 0.51 - Figure 14c)) with a peak at 0.3; lastly, when considering multiple segmented incident intervals detected from our algorithm, it produced a slightly lower F1 score of 0.47 - Figure 14b)), but more evenly distributed. Overall, the algorithm performance that we propose in this paper yields a higher precision in detecting disruptions from time series of traffic speeds than from the reported accident timeline. The use of multiple segments produced by the algorithm can highlight multiple disruptions while producing just a slight decrease in the quality of results. The error for multiple intervals segmentation increases because more additional intervals are considered in the evaluation of the metric, which may lay outside of originally marked intervals (see Figures 13 and 9).

A. False Positives Rate Analysis

The issue of false alarms in the incident detection task can be significant. Traffic authorities may need the control over incident detection specificity. Since our segmentation algorithm provides real values after applying a difference metric, the value of false positives can be controlled by selecting an appropriate threshold of binarization. We provide a receiver operating characteristic curve (ROC) curve for comparison across total merged timeline of incidents and represent manual and estimated segmentation procedures as a binary classification problem. Parameters like granularity and binarization threshold can be fine-tuned according to specific metric (e.g. Area under ROC curve, F1-score or heuristics of metrics) to increase the amount of true positives while reducing the amount of false positives. We utilized F1-score as it able to grasp both of these values in a single formula. As shown on Figure 15, our proposed methodology, even without the tuning of hyper-parameters, allows to maintain a high detection rate while keeping the false alarm rate low.

We further look into specifics of disruption detection for various accidents (see Figures 16 and 17). For some accidents, high detection rate cannot be achieved without increasing the false positives rate. It is important to note that the selection of the binarization threshold plays a crucial role in controlling algorithm performance. A lower threshold might increase the sensitivity, thereby escalating the detection rate, but at the cost of specificity, leading to more false positives. Conversely, a higher threshold might reduce false alarms but may also miss some real incidents, thus lowering the detection rate. Therefore, the end users can fine-tune the parameters according to their specific needs, demonstrating the flexibility and adaptability of our proposed methodology.

B. Parameter Importance Study

For our model, we have the following variables and their intervals of variation:

- **gran**: Granularity, an integer value controlling the level of detail (moving window size) in the metric estimation function. In the provided search space, the range of gran is [2, 40] with a step of 1. Default value is 12.
- **kernel_size**: A list of float values used as weights in the dilation convolution operation. The search space for the kernel is the size of the convolution [1, . . . 4], float values primarily intended to implement pre-processing operation for the day time-series. Default value is 3.
- **selectivity**: A float value between 0.01 and 4.0 that determines the power coefficient in post-processing difference estimations Default value is 2.0.
- **shift**: An integer value between -32 and +32 that represents a cyclic shift of the resulting time series to attribute

Fine-tuning of the disruption segmentation algorithm can be performed automatically using hyper-parameter search for best performance on alternative sources of data,Sensitivity-specificity control: maintaining high incident detection rates while minimizing false alarms is a key challenge. The disruption segmentation algorithm allows us to estimate the “degree” of disruption before applying the binarization threshold. This property allows for false-alarm control using fine-tuning of the detection threshold. Balancing sensitivity (identifying real incidents) and specificity (avoiding false alarms) often involves trade-offs and can be fine-tuned to specific data set.

In urban networks with hundreds of measurement locations, the data retrieval is a bottleneck, since each accident report will require a request for daily and fortnight measurements at specific detectors. Depending on the speed of VDS data retrieval, the amount of data that can be obtained in an acceptable amount of time can be limited.

The code for the paper can be found by the following link: https://github.com/Future-Mobility-Lab/AAA-toolkit/tree/main.

The Table IV presented below provides a summary of the key parts involved in the Accident Analysis& Association (AAA) Toolkit codebase. Each row represents a specific segment of the code, outlining the corresponding inputs required and outputs produced for each segment. The sequence of code parts indicates the flow of data and the transformation processes that occur from acquiring the initial raw data to ultimately applying segmentation algorithms on the compiled information.

VI. Conclusion

Our methodology aims to automatically detect, segment, and extract traffic disruptions and accidents using distance metrics. This approach improves incident prediction accuracy across multiple machine learning models and provides better fit to manual markup of observed traffic speed disruptions. By obtaining the intervals and shapes of traffic disruptions, we can model the impact of accidents with greater precision, using traffic state measurements rather than just reported parameters (duration, start time, etc). This approach provides more data on the accident and allows us to study accident impacts in greater detail.

A. Relevance of This Work Can be Summarized in Following Points

1) Enhancement of Traffic Management Systems: Integrate the proposed early detection and disruption segmentation algorithm into existing traffic management systems to improve and automate incident detection and corresponding data collection. This will help to minimize congestion and the overall impact of incidents on traffic flow, 2) highlight of reporting errors to standardize data reporting: Establish standardized guidelines and protocols for reporting traffic incidents, including the accurate reporting of the location, start and end times, number of lanes affected, and other relevant details; this will ensure that data-driven models can accurately predict incident severity and disruption length, 3) highlight the necessity of creating of data standards policies across countries for collecting necessary traffic accident information, 4) development of Incident Response Strategies by utilizing the improved incident prediction models to develop data-driven incident response strategies, including the dynamic traffic rerouting and real-time traffic guidance; this will help to mitigate the impact of traffic incidents on road users and reduce the risk of secondary incidents; 5) Data Fusion for a better traffic accident analysis: due to observed improvement in the quality of prediction arising from data fusion, traffic Authorities can consider integrating data sets from private companies for jointly analysing traffic datasets of various types to improve traffic safety by improving accuracy of traffic incident duration prediction.

B. Future Research in This Area

1) Algorithm’s complexity can be expanded by incorporating custom kernels, which can be found using hyper-parameter search, 2)Disruption measurements obtained over time can enable the prediction of traffic incident impact propagation with greater accuracy than relying solely on reported values, 3) The proposed methodology can be extended to include disruptions beyond accidents, such as construction or road closures, which can improve the accuracy of impact prediction, 4) Further improvement can also be achieved by performing data fusion and incorporating external data sources, such as weather and events, into the incident impact prediction models. We are currently modelling the cascading effect on traffic disruptions and how these can be automatically identified based on multiple incoming traffic state streams; the main challenge of detecting subsequent incidents lie in the time-span duration of the first incident which is normally stochastic in nature.

C. Limitations of This Work

The current modelling approach has been applied to a San Francisco data set due to its public availability and easiness to access. However, we would like to test the approach on multiple other countries and incident databases across the globe; the main challenge is the lack of both traffic states and traffic accidents logs to be released with synchronised timelines.

REFERENCES

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

[2] *Traffic Safety Facts 2013, U.S. Department of Transportation*, Nat. Highway Traffic Saf. Admin., Washington, DC, USA, 2013.

[3] W. Kim and G.-L. Chang, "Development of a hybrid prediction model for freeway incident duration: A case study in Maryland," *Int. J. Intell. Transp. Syst. Res.*, vol. 10, no. 1, pp. 22–33, 2011, doi: 10.1007/s13177-011-0039-8.

[4] A. Theofilatos, G. Yannis, P. Kopelias, and F. Papadimitriou, "Predicting road accidents: A rare-events modeling approach," *Transp. Res. Proc.*, vol. 14, pp. 3399–3405, Jan. 2016, doi: 10.1016/j.trpro.2016.05.293. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235214651630299X

[5] A. J. Khattak, J. Liu, B. Wali, X. Li, and M. Ng, "Modeling traffic incident duration using quantile regression," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2554, no. 1, pp. 139–148, Jan. 2016.

[6] A. Mehdizadeh et al., "A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling," *Sensors*, vol. 20, no. 4, p. 1107, Feb. 2020.

[7] J. Wang, K. Li, and X.-Y. Lu, "Chapter 5. Effect of human factors on driver behavior," in *Advances in Intelligent Vehicles*. London, U.K.: Academic, 2013, pp. 111–155, doi: 10.1016/B978-0-12-397199-9.00005-7.

[8] M. S. Horswill and M. E. Coster, "The effect of vehicle characteristics on drivers' risk-taking behaviour," *Ergonomics*, vol. 45, no. 2, pp. 85–104, Feb. 2002, doi: 10.1080/00140130110115345.

[9] R. B. Noland and L. Oh, "The effect of infrastructure and demographic change on traffic-related fatalities and crashes: A case study of Illinois county-level data," *Accident Anal. Prevention*, vol. 36, no. 4, pp. 525–532, Jul. 2004.

[10] A. Winder, M. Brackstone, and P. Site, "Traffic management for land transport: Research to increase the capacity, efficiency, sustainability and safety of road, rail and urban transport networks," Transp. Res. Knowl. Centre (TRKC), Tech. Rep., 2023.

[11] 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," *Transp. Res. C, Emerg. Technol.*, vol. 141, Aug. 2022, Art. no. 103721, doi: 10.1016/j.trc.2022.103721. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X22001589

[12] A. S. Mihaita, Z. Liu, C. Cai, and M. Rizoiu, "Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting," 2019, *arXiv:1905.12254*.

[13] B. Ghosh, M. T. Asif, J. Dauwels, U. Fastenrath, and H. Guo, "Dynamic prediction of the incident duration using adaptive feature set," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4019–4031, Nov. 2019.

[14] L. Dimitriou and E. I. Vlahogianni, "Fuzzy modeling of freeway accident duration with rainfall and traffic flow interactions," *Analytic Methods Accident Res.*, vol. 5, pp. 59–71, Jan. 2015.

[15] S. Shafiei, A. Mihaita, H. Nguyen, C. Bentley, and C. Cai, "Short-term traffic prediction under non-recurrent incident conditions integrating data-driven models and traffic simulation," in *Proc. Transp. Res. Board 99th Annu. Meeting*, 2020, pp. 1–24.

[16] T. Mao, A.-S. Mihaita, F. Chen, and H. L. Vu, "Boosted genetic algorithm using machine learning for traffic control optimization," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 112–7141, Jul. 2022.

[17] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Overview of traffic incident duration analysis and prediction," *Eur. Transp. Res. Rev.*, vol. 10, no. 2, pp. 1–13, Jun. 2018.

[18] S. Fukuda, H. Uchida, H. Fujii, and T. Yamada, "Short-term prediction of traffic flow under incident conditions using graph convolutional recurrent neural network and traffic simulation," *IET Intell. Transp. Syst.*, vol. 14, no. 8, pp. 936–946, Aug. 2020.

[19] A. B. Parsa, H. Taghipour, S. Derrible, and A. Mohammadian, "Real-time accident detection: Coping with imbalanced data," *Accident Anal. Prevention*, vol. 129, pp. 202–210, Aug. 2019.

[20] L. Eboli, C. Forciniti, and G. Mazzulla, "Factors influencing accident severity: An analysis by road accident type," *Transp. Res. Proc.*, vol. 47, pp. 449–456, Jan. 2020, doi: 10.1016/j.trpro.2020.03.120. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146520303197

[21] T. Tsubota, C. Fernando, T. Yoshii, and H. Shirayanagi, "Effect of road pavement types and ages on traffic accident risks," *Transp. Res. Proc.*, vol. 34, pp. 211–218, Jan. 2018, doi: 10.1016/j.trpro.2018.11.034. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146518303235

[22] G. Yannis et al., "Use of accident prediction models in road safety management—An international inquiry," *Transp. Res. Proc.*, vol. 14, pp. 4257–4266, 2016, doi: 10.1016/j.trpro.2016.05.397. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146516304033

[23] C. C. Aggarwal and C. C. Aggarwal, "Probabilistic and statistical models for outlier detection," in *Outlier Analysis*. Cham, Switzerland: Springer, 2013, pp. 41–74.

[24] J. Takeuchi and K. Yamanishi, "A unifying framework for detecting outliers and change points from time series," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 4, pp. 482–492, Apr. 2006.

[25] G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,"

[26] Dr. Boda Saroja and Dr.Rajesh G ," A New Approximate Adder with Block-based Carry Speculation for High- Speed Applications".

[27] Mrs.K.Pramilarani and Dr. Vasanthi Kumari P," Analyzing Various Techniques to Safeguard user Sensitive Data".