

Article

Attention-Based Residual Dilated Network for Traffic Accident Prediction

Ke Zhang  and Yaming Guo *

Department of Civil Engineering, Tsinghua University, Beijing 100080, China

* Correspondence: guoyaming123@tsinghua.edu.cn

Abstract: Traffic accidents directly influence public safety and economic development; thus, the prevention of traffic accidents is of great importance in urban transportation. The accurate prediction of traffic accidents can assist traffic departments to better control and prevent accidents. Thus, this paper proposes a deep learning method named attention-based residual dilated network (ARDN), to extract essential information from multi-source datasets and enhance accident prediction accuracy. The method utilizes bidirectional long short-term memory to model sequential information and incorporates an attention mechanism to recalibrate weights. Furthermore, a dilated residual layer is adopted to capture long term information effectively. Feature encoding is also employed to incorporate natural language descriptions and point-of-interest data. Experimental evaluations of datasets collected from Austin and Houston demonstrate that ARDN outperforms a range of machine learning methods, such as logistic regression, gradient boosting, Xgboost, and deep learning methods. The ablation experiments further confirm the indispensability of each component in the proposed method.

Keywords: traffic risk; attention mechanism; dilated convolution; accident prediction; multi-source fusion

MSC: 68T07



Citation: Zhang, K.; Guo, Y. Attention-Based Residual Dilated Network for Traffic Accident Prediction. *Mathematics* **2023**, *11*, 2011. <https://doi.org/10.3390/math11092011>

Academic Editors: Xinchao Zhao, Xingquan Zuo, Yinan Guo and Kunpeng Kang

Received: 19 March 2023

Revised: 11 April 2023

Accepted: 21 April 2023

Published: 24 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Road traffic crashes have a significant impact on public health, causing the loss of 1.3 million lives per year worldwide, as reported by the World Health Organization [1]. Furthermore, these accidents also have substantial economic consequences, accounting for approximately 3% of gross domestic product in most countries. Therefore, it is crucial to establish a safe transportation system that takes the vulnerability of road users to serious injuries in daily life into consideration. However, comprehensively considering all components can be challenging, and despite efforts to address this issue, traffic accidents remain a significant public safety issue. To reduce the number and severity of these incidents, researchers have been working on accident analysis and prediction for several decades. By identifying the factors that contribute to accidents and developing strategies to prevent them, governments, communities, and individuals can improve road safety and reduce the number of fatalities and injuries [2,3]. Centralized traffic management refers to a system where traffic control and management activities are coordinated and controlled from a central location [4,5]. Advanced technologies in centralized traffic management can facilitate the identification of traffic accident hotspots, analyze accident data to understand accident patterns and causes, and develop targeted interventions to address those issues. For example, by investing in better road design, traffic signals, and pedestrian crossings, governments can reduce the risk of accidents and create safer environments for all road users. Additionally, advanced driver assistance systems, such as automatic emergency braking and lane departure warning systems, can help prevent accidents caused by human error. Governments can also mandate the use of safety features to further reduce the risk of injuries and fatalities in the event of an accident.

As a result, increasing the accuracy of accident prediction in transportation systems is expected to eliminate fatal crashes and reduce serious injuries through warnings and preventive measures. In recent years, there have been two primary approaches to addressing this problem: causal analysis methods and statistical methods. Statistical methods focus on identifying correlations or associations between variables based on historical data, while causal analysis methods seek to understand the underlying mechanisms or factors that contribute to traffic risks. In other words, statistical methods primarily make predictions based on patterns and trends observed in historical data. In contrast, causal analysis methods aim to establish cause-and-effect relationships between variables, and their predictions are based on understanding the causal mechanisms that drive the observed outcomes. As this study focuses on improving the accuracy of traffic risk prediction, statistical methods are summarized in related works.

However, the data attributes that are related to traffic accidents are rather complex and heterogeneous. Factors such as geographical location, road network, traffic patterns, weather conditions, and specific areas of interest need to be considered in traffic accident prediction. Road characteristics, such as road types, road design, and road infrastructure, also influence the occurrence and severity of traffic accidents. Additionally, traffic conditions, such as traffic volume, traffic density, traffic speed, time period, and traffic congestion, can impact the likelihood of traffic accidents. Simply combining multiple deep learning modules might not effectively handle multi-source information. While recurrent neural network (RNN) architectures, including long short-term memory (LSTM), are commonly used for time series prediction, the problems of gradient explosion and vanishing remain unresolved. Moreover, the model output is primarily influenced by the nearest time series input. In contrast, a dilated convolutional network can build deep convolutional connections to extract higher-order hidden information and large convolutional kernels to extract features over long periods. This allows for the indirect expansion of convolutional kernels without increasing computing costs [6,7]. To improve accident prediction accuracy, we propose a novel deep learning method called attention-based residual dilated network (ARDN). This method utilizes LSTM to model sequence relationships and an attention mechanism to focus on key information. The dilated residual network is designed to extract complex feature information. Finally, the dataset of natural language descriptions and points of interest is fused through information encoding. We conducted experiments on datasets collected from Austin and Houston, and compared this method with a series of machine learning and deep learning methods to verify its validity. The highlights of this work are as follows:

- (1) To address the complexity of traffic accident data, this approach considers diverse data sources to capture complex patterns.
- (2) To overcome the limitations of traditional RNN architectures, the dilated residual network is adopted to capture long-range dependencies. Introducing an attention mechanism for key information further enhances prediction accuracy.
- (3) Through the comparison of experiments conducted on two datasets, the results show superiority over baselines.

The remainder of the paper is organized as follows: Section 2 presents a detailed overview of the model's architecture and its components; Section 3 provides information about the experiments, including dataset description, baselines, algorithms comparison, and ablation study; and Section 4 summarizes this paper and conceptualizes future work.

2. Related Works

Statistical methods have been widely used to predict traffic accidents by examining patterns, trends, and underlying factors [8]. Regression models have been commonly used to analyze the risk factors associated with traffic accidents [9–11]. However, these models rely on pre-defined assumptions regarding the underlying probability data distribution and the dependent relationships among various variables which may not always align with the actual circumstances [12]. Machine learning methods offer a more flexible and

robust approach as they can adapt to evolving circumstances without rigid pre-assumed relationships among variables. By using machine learning algorithms, transportation systems can analyze vast amounts of data to identify patterns and relationships that might be challenging to uncover using traditional statistical methods.

Thus, a series of machine learning methods are listed as follows. To identify the factors that influence the severity of accidents at intersections, a logistics model with a heavy-tail distribution was adopted to fit the dataset [13]. The goal of this study was to develop countermeasures that can help mitigate the risk of severe crashes. A support vector machine (SVM) model was adopted to analyze the severity of injuries sustained during accidents and predict the severity of individual crashes [14]. The results verified that the SVM model provides better prediction results than the ordered probit model. To further enhance the prediction ability, a fuzzy C-means based SVM has been proposed to predict the severity level of crashes [15]. The experiments demonstrate that this model outperforms the SVM model in terms of accuracy and F1 scores. This is because the fuzzy c-means clustering algorithm helps to enhance the prediction ability. Then, a comparison between Bayesian network and regression models has been conducted to analyze accident severity [16]. By comparing the results, it was concluded that the Bayesian network model is more effective in predicting the severity of accidents than the regression model. To analyze traffic accidents, researchers have utilized decision trees to extract decision structures from available datasets [17]. The results indicated that the decision tree is capable of extracting more relevant rules from the same dataset. In addition, a comparison between binary logistic regression and decision tree was conducted using a motorcycle accident dataset [18]. The results suggest that the performances of the two models are similar in predicting the identified factors. To explore the relationship between risk factors and accident severity, a severity causation network has been proposed that integrates information entropy and the Bayesian network [19]. This method can significantly enhance the feasibility of predicting crash severity. Furthermore, the effectiveness of various models, such as classification and regression tree models [20,21], log binomial regression [22,23], and the rule-based mining method [24], were also verified through real-world datasets. These studies confirm the potential of machine learning models to contribute to the prediction of traffic accidents.

With the prosperity of artificial intelligence, deep learning methods have emerged as a powerful tool in a wide range of applications, including time series prediction [25,26]. By leveraging the capabilities of deep neural networks, deep learning models have demonstrated improved accuracy and more precise prediction results [27]. The artificial neural network (ANN) has been widely used in this domain through employing a multi-layer perceptron architecture with sigmoid activation and linear output neurons to fit multi-dimensional mapping relationships. Comparative studies using different evaluation metrics, such as the R-value, root mean square error, mean absolute error, and sum of square error, have consistently shown that ANN provides superior prediction results compared with other methods, such as genetic algorithms and pattern search [28]. Additionally, a convex combination algorithm has been proposed to train the neural network model for accident severity prediction. A modified pruning method for function approximation algorithms has also been developed to optimize the network structure [29]. The experimental results have demonstrated that this approach achieves comparable classification accuracy when compared to fully connected networks and ordered logit models. To capture temporal correlations from traffic accident records, recurrent neural networks (RNN) have been employed to predict traffic accident severity. The RNN model has demonstrated better performance when compared with other models, such as multilayer perceptron and Bayesian logistic regression [30]. In recent years, a feature matrix to gray image algorithm has been designed to convert traffic accident features into gray images containing combination relationships. A convolutional neural network-based approach has then been proposed to predict traffic accident severity [31]. This method has demonstrated improved accuracy when compared with other methods. To leverage multi-source datasets, a spatio-temporal convolutional long short-term memory network has been proposed to predict citywide

short-term crash risk [32]. The results indicated that this approach surpasses the benchmark techniques when it comes to predicting various crash risks. Additionally, deep accident prediction (DAP) encompasses several components that capture interrelationships within multiple data sources, leading to a better performance than several baseline models [33]. Overall, these learning-based methods have shown promising results regarding predicting the severity of traffic accidents and identifying the factors that contribute to them. These findings can help improve traffic safety by enabling the development of effective countermeasures [34,35].

3. Methods

3.1. The Overall Architecture

This paper proposes a novel learning-based approach to extract spatial–temporal information and generate highly accurate predictions, as shown in Figure 1. The proposed model includes an attention-based dilated residual layer that effectively extracts time series information from input traffic data.

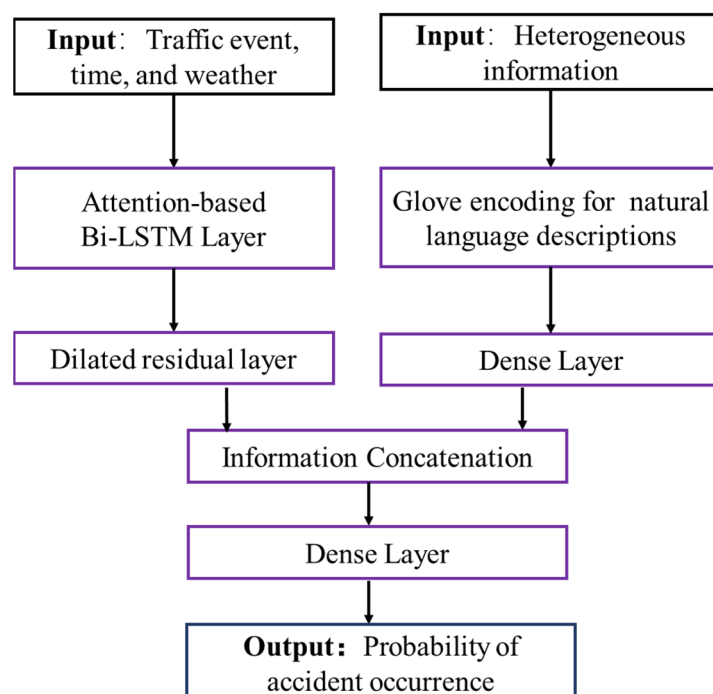


Figure 1. Attention-based residual dilated network.

The input traffic data comprises several important factors, including descriptions of traffic events, time, and weather information [33,36,37]. The details are listed as follows:

- (1) Traffic event data. This includes the frequency of available traffic events during the current 15 min interval, such as accidents, broken vehicles, lane blockages, and flow incidents.
- (2) Time information. This is a critical component of the input traffic data and includes factors such as weekdays, hour of the day, and day or night. For example, weekdays can be represented as numerical variables (e.g., 1 for weekday, 2 for weekend, etc.), hour of the day can be represented as categorical variables (e.g., morning, afternoon, evening, and night), and day or night can be represented as binary variables (e.g., 0 for day and 1 for night).
- (3) Weather information. This is represented as numerical variables (temperature, pressure, humidity, visibility, wind speed, and precipitation amount) or binary variables (rain, snow, fog, and hail).

Overall, the input traffic data is formulated as a comprehensive vector denoted by x_t which includes all the aforementioned factors. In addition, a dense network is incorporated to enable the learning of information from heterogeneous data sources, including natural language descriptions of historical traffic events and points-of-interest data. To handle the heterogeneous information effectively, Glove encoding is utilized to encode the raw information into word vectors, which is described in Section 3.4. The proposed model has the capability to effectively extract information from these data sources and generate highly accurate predictions based on this information.

3.2. Attention-Based Bi-LSTM Layer

Long short-term memory is a popular type of neural network that is designed to capture temporal relationships within sequential data. It is composed of memory units that accumulate historical information through self-connections. Each memory unit consists of three gates: the input gate, the forget gate, and the output gate, as shown in Figure 2. The input gate controls the information flow into the memory unit, while the forget gate and output gate regulate the information flow within the memory unit. The details are listed as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \quad (2)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \quad (5)$$

$$h_t = o_t * \tanh(C_t). \quad (6)$$

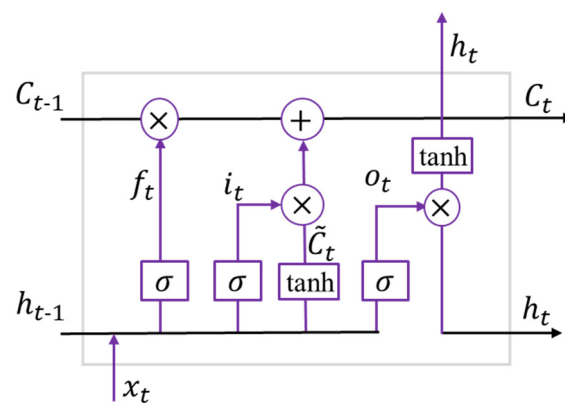


Figure 2. Long short-term memory.

The forget gate f_t determines which information from the previous cell state should be discarded. This gate concatenates the previous hidden state h_{t-1} and the current input x_t , then outputs information through the sigmoid function, which is multiplied with the previous cell state C_{t-1} . The next step is to update the cell state of the layer. A tanh layer is used to produce an updated value within the interval $[-1, 1]$, which can either strengthen or weaken some cell states. The output value of the sigmoid layer is multiplied by the output of the tanh layer, acting as a scaling function. The new cell state C_t is then generated according to the output of the forget gate f_t and the update gate $\tilde{C}_t$. Finally, the sigmoid layer determines the information of C_t to acquire the output h_t . Based on the given formula,

the memory unit allows for linear updates. Consequently, the current state of the hidden layer can be obtained by combining the nonlinear tanh function with the information from the output gate.

Although LSTM effectively models sequential information, it is limited to encoding information in only one direction, usually from front to back. This limitation can be overcome by using Bi-LSTM, which can capture the semantics of a sequence in both directions, as illustrated in Figure 3. Furthermore, when dealing with datasets that contain information regarding varying dimensions, the challenge lies in extracting important information that can contribute to improved model performance.

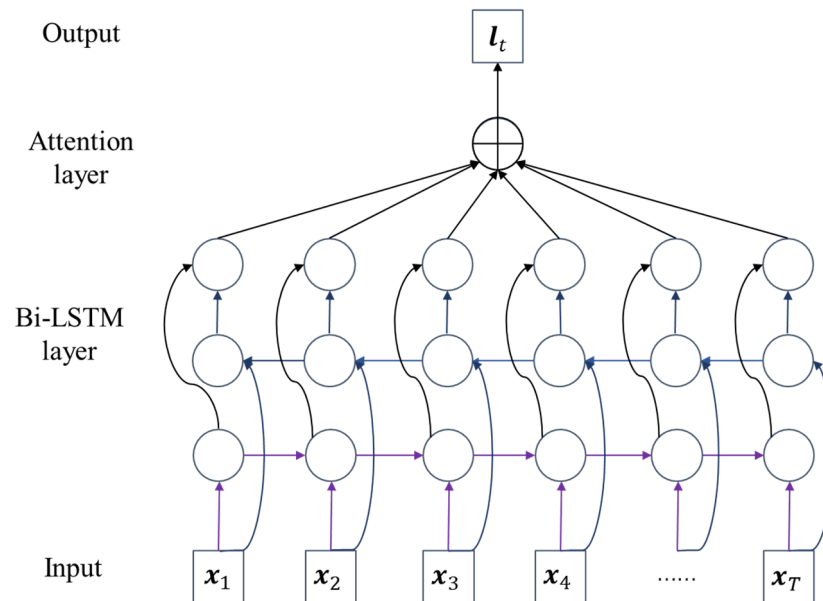


Figure 3. Attention-based Bi-LSTM.

Originally proposed in the field of visual research, the attention mechanism simulates the way that the human brain focuses its attention on important regions and disregards less relevant parts [38]. The core idea behind the attention mechanism is to allocate more attention to key parts and less attention to other parts by rationally allocating brain resources [39]. This helps exclude the influence of non-key factors and improve the quality of feature extractions in the hidden layer. The attention mechanism achieves this by using a probability weight distribution to calculate the probability weights of feature vectors at different time steps in the sequence. By utilizing the attention mechanism, we can effectively extract the most important information from the sequence, leading to better model performance. The first step is to adopt a learnable matrix to calculate the similarity between the query and each key:

$$h_{t,t'} = \tanh(W_t \hat{x}_t^T + W_x \hat{x}_{t'}^T + b_t), \quad (7)$$

$$e_{t,t'} = \sigma(W_a h_{t,t'} + b_a), \quad (8)$$

where $\hat{x}_t^T$ represents the output of Bi-LSTM and W_t, W_x, W_a, b_t, b_a represent the learnable parameters. This produces a weight value $a_{t,t'}$, which is then normalized through the softmax function. Then, the corresponding values are multiplied by their weights and the sum is taken to generate the output:

$$a_{t,t'} = \text{softmax}(e_{t,t'}), \quad (9)$$

$$l_t = \sum_{t'} a_{t,t'} \hat{x}_{t'}^T. \quad (10)$$

In sum, this module is composed of Bi-LSTM and an attention mechanism. The former can effectively address the issue of gradient explosion and vanishment while maintaining high accuracy. The latter plays a critical role in mapping the query to a series of key-value pairs, assigning a higher weight to important parts of the data. This helps identify relevant information and improve the accuracy of the predictions. Finally, the output is transmitted to the next dilated residual layer for further refinement. In addition, as depicted in Figure 3, the flow of data through hidden layers allows information to pass through the network from input layers to output layers. The connections between nodes in different layers are associated with weights that determine the strength of the connection. These weights are adjusted during the training process through backpropagation, allowing the network to learn from input data and optimize the prediction performance by adjusting the weights.

3.3. Dilated Residual Layer

Although LSTM is designed to retain important information in its memory gate, convolutional neural networks (CNNs) have shown promise in retaining long-term memories in a more realistic manner. For instance, models such as WaveNet [6] and temporal convolutional networks (TCNs) [7] have demonstrated better performance than RNNs and LSTMs based on a range of datasets. The dilated residual layer is a key component that combines dilated convolution and residual connection to achieve high performance in sequence modeling tasks, as shown in Figure 4.

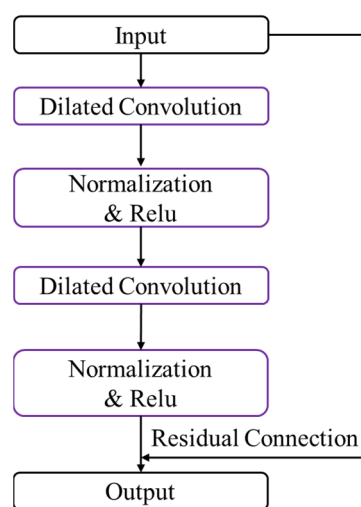


Figure 4. The dilated residual layer.

Dilated convolutional networks use a dilation rate parameter in addition to the convolution kernel size which indicates the degree of expansion that will be transmitted to the next dilated residual layer. This allows the network to capture long-range dependencies in the input sequence, which is particularly important in tasks such as speech recognition or language modeling. As the number of layers in the network increases, the convolution window becomes larger, and the number of empty holes in the convolution window also increases, leading to an even larger receptive field. This approach enables neural networks to retain long-term memories more effectively than traditional RNNs.

Residual connection is a powerful technique adopted in neural network architectures to improve training convergence [40]. It allows information from earlier layers to bypass the intervening layers and directly connect to later layers, enabling the network to learn more complex representations of the input sequence. This process is repeated for each layer, with the feature information of the previous layer being added to the feature graph of each subsequent layer.

To further enhance the performance of the network, this layer uses normalization and ReLU nonlinear functions. After each convolution operation, dropout is applied for regularization to prevent overfitting.

3.4. Heterogeneous Feature Encoding Layer

To improve the understanding of natural language descriptions of historical traffic events, the well-trained Glove model [41] is employed to encode raw information into word vectors through unsupervised learning. This approach measures the similarity between words by calculating the Euclidean distance between their respective vectors, and is designed to capture vector differences that correspond to linear structures through different vectors. The core procedure for the Glove model involves constructing a word co-occurrence matrix from the input text corpus, computing the probability distribution of word pairs from the matrix, and training the model using an iterative optimization process. Once trained, the word vectors can be used as feature representations for downstream machine learning tasks, capturing semantic and syntactic relationships between words.

Then, this method calculates the average values of these word vectors to represent each region. A dense network is then used to obtain a fixed-length feature vector that captures important information for classification. Point-of-interest information is also considered to enhance the understanding of traffic events. This approach enables effective processing and analysis of natural language descriptions of historical traffic events.

3.5. Output and Optimization

As illustrated in Figure 1, this method integrates information from the dilated residual layer and heterogeneous feature encoding. The fully connected layer, also known as the dense layer, is utilized to concatenate information as it is capable of learning complex patterns from input data. This method is differentiable, allowing us to use the back-propagation algorithm for training. Specifically, we employ the Adam optimizer to minimize the binary cross-entropy during several iterations of learning. The formulation is given below:

$$l(\theta) = -\frac{1}{N} \sum_{i=1}^N y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i)), \quad (11)$$

where N denotes the total number of samples in the training set, y_i represents the sign function to determine whether the sample is true or not, and $p(y_i)$ refers to the predicted probability of the occurrence of an accident. By minimizing this objective function, we can improve the performance of the model and enhance its ability to accurately predict the likelihood of accidents.

4. Experiment

In this section, we validate the effectiveness of the proposed ARDN method through experiments conducted on the datasets collected from two large cities. Python was chosen as the programming language to solve the traffic accident prediction problem, and the Keras framework was utilized to implement the deep learning algorithm. The experiment is organized as follows. Section 4.1 describes the datasets adopted in this work. Section 4.2 lists the baselines and evaluation criteria. Section 4.3 compares the proposed ARDN method with a series of baselines. Section 4.4 displays the implications for practice. Finally, the ablation experiment is conducted in Section 4.5 to verify the validity of the algorithm components.

4.1. Data Description

This study aims to predict the probability of traffic accidents occurring within a specific 15 min window. The datasets used in this study was obtained from the website: <https://smoosavi.org/datasets/> (accessed on 1 June 2022). The data were collected from multiple data providers, including several APIs that provide streaming traffic event data.

These APIs capture and broadcast traffic events from various sources, such as the US and state Departments of Transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks [37]. The datasets provide detailed information on each geographical region with a size of $5 \text{ km} \times 5 \text{ km}$ per 15 min interval in Austin and Houston [33]. Figure 5 depicts the road networks of the two cities, which was obtained from the website: www.openstreetmap.org/ (accessed on 15 April 2023). They include various traffic-related events occurring during the current interval, such as traffic events, time, and weather attributes. To predict the label of the next time interval, the time series information of the last eight time intervals, i.e., two hours, is adopted in this work.

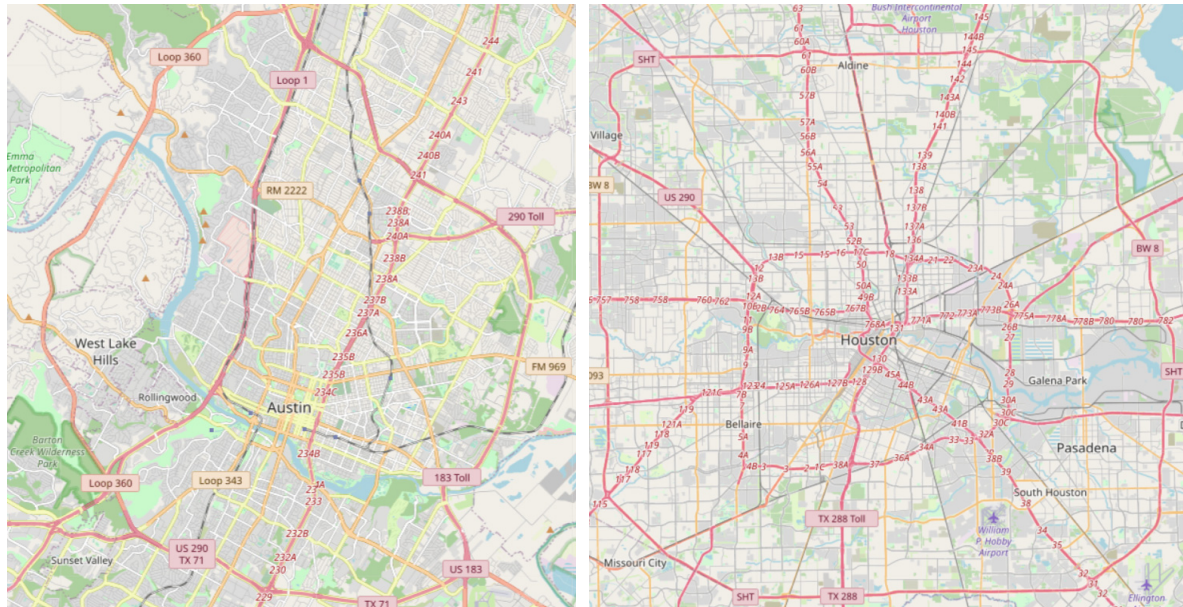


Figure 5. The road networks of Austin (on the left) and Houston (on the right).

4.2. Baseline and Evaluation Criteria

This section aims to compare the effectiveness of our proposed methods against five baselines:

- (1) Logistic regression is a statistical method used for binary classification where the goal is to predict the probability of an event occurring using a logistic function.
- (2) Gradient boosting [42] is a machine learning technique that builds an ensemble of weak predictive models, such as decision trees, in a sequential manner. It uses gradient descent optimization to minimize the loss function, combining the predictions of multiple weak models to make a final prediction.
- (3) Xgboost [43] is an optimized implementation of gradient boosting. It uses a variety of regularization techniques to prevent overfitting and provides additional features such as parallel processing and handling of missing values.
- (4) ANN consists of interconnected nodes organized in layers, where each node performs a mathematical operation on its inputs and passes the result to the next layer.
- (5) DAP [33] was proposed in a previous study [33] and incorporates data augmentation and preprocessing techniques to improve the accuracy of traffic accident prediction.

To ensure fair comparisons, we set several hyper parameters to fixed values, including a learning rate of 0.01 and an early stop threshold of 10. The early stop method was used to prevent overfitting by limiting the number of training iterations. In particular, during the implementation of ARDP, we used a dropout rate of 0.3, an embedding dimension of 128, and a kernel size of 3. The dropout rate of 0.3 was chosen to introduce regularization and prevent overfitting during training. The embedding dimension of 128 was selected as a

trade-off between model complexity and representation capacity. The kernel size of 3 was used to capture local spatial-temporal patterns in the input data.

To evaluate the performance of our proposed methods and the baselines, we adopted F1 scores as evaluation criteria which are computed based on precision and recall rate. Precision measures the proportion of positive examples predicted by the binary classifier that are real positive examples, while recall measures the proportion of real positive cases in the test set that are correctly identified by the binary classifier. Precision and recall rate are often in conflict with each other. Adopting F1 scores as evaluation criteria provides a more comprehensive measure of the performance of the different methods, defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (13)$$

$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (14)$$

where Precision is the ratio of true positive predictions to the total number of positive predictions and Recall is the ratio of true positive predictions to the total number of actual positive cases in the test set.

4.3. Discussion on the Results

Tables 1 and 2 present the experimental results obtained from the Austin and Houston datasets, respectively. The first column indicates the algorithm used for the experiment, while the remaining columns show the F1 score for each class, i.e., non-accident, accident, and weighted average. Based on the results presented in these tables, it can be concluded that the ARDN is the most effective method for predicting accident occurrences in both the Austin and Houston datasets. The following conclusions can be drawn.

Table 1. A comparison of criteria F1 scores among several baselines in the Austin dataset.

Method	Non-Accident	Accident	Weighted Average
Logistic Regression	0.93	0.58	0.87
Gradient Boosting	0.93	0.61	0.87
XGBoost	0.93	0.62	0.87
ANN	0.92	0.62	0.87
DAP	0.89	0.65	0.87
ARDN	0.93	0.67	0.89

Table 2. A comparison of criteria F1 scores among several baselines in the Houston dataset.

Method	Non-Accident	Accident	Weighted Average
Logistic Regression	0.94	0.49	0.88
Gradient Boosting	0.94	0.51	0.88
XGBoost	0.94	0.50	0.88
ANN	0.93	0.59	0.88
DAP	0.93	0.58	0.88
ARDN	0.94	0.60	0.89

Most algorithms perform well regarding predicting non-accidents as accidents tend to be special cases and algorithms often err on the side of non-accidents. This highlights the need for more accurate algorithms that predict accidents specifically.

Gradient boosting and Xgboost outperform logistic regression models as tree-based models can improve accuracy and efficiency by automatically processing missing values, outliers, feature selection, and feature fusion.

The F1 scores for accidents are generally lower than those of non-accidents, indicating that algorithms are better at identifying non-accident data points. This emphasizes the importance of developing algorithms that are specifically designed to accurately identify accidents.

Among all the comparison results, the ARDN method performs the best, achieving the highest F1 scores for both datasets. This is attributed to its ability to effectively capture relationships within time-series information using dilated residual networks and attention mechanisms. The attention mechanism allows the model to focus on the most relevant regions and time intervals, thus improving the accuracy of the predictions. By using dilated convolutions, the network can capture long-range dependencies across the input data. These results demonstrate the potential of the proposed model to improve the accuracy of accident prediction.

The proposed method addresses the complexity and heterogeneity of traffic accident data by proposing a deep learning approach that takes diverse data sources into account to capture the complexity of traffic accident patterns, including the geographical location, road characteristics, and traffic conditions. Compared with previous deep learning methods, this method overcomes limitations in several ways. Firstly, RNN architectures might suffer from issues like gradient explosion and vanishing and primarily rely on the nearest time series input. In contrast, our proposed dilated residual network is designed to extract complex features over long periods, achieving indirect expansion of convolutional kernels without increasing computing costs. Secondly, we introduce an attention mechanism that allows the model to focus on important information within the data. This attention-based approach enhances the model's ability to capture relevant features and relationships, leading to more accurate predictions. This attention mechanism helps overcome the limitations of traditional RNN methods that may not effectively capture important information within the data.

4.4. Implications for Practice

According to the proposed method, high accuracy traffic risk prediction has practical significance in various ways. Traffic accidents result in significant casualties and property losses. Accurate traffic risk prediction facilitates timely measures to prevent accidents, thereby reducing casualties and property losses. It enables traffic management departments and practitioners to understand the patterns and trends of accidents, leading to the implementation of preventive measures. For example, patrolling high-risk areas, improving road designs and infrastructure, adjusting traffic flow, and warning drivers about accident-prone areas can all contribute to reducing the occurrence of accidents.

Moreover, accurate traffic risk prediction allows traffic management departments to allocate resources more effectively and optimize safety management strategies. Prioritizing police and rescue resources in areas with high accident rates and high-risk roads and making timely adjustments to routes, speed limits, and intersection designs can improve the effectiveness of traffic safety management. Traffic accidents often cause congestion and delays on roads, impacting traffic flow and road capacity. Accurate traffic risk prediction can provide advance warnings about areas prone to accidents, enabling measures to reduce the impact on traffic flow and road capacity, thereby improving overall traffic efficiency.

Finally, an interactive dashboard can be utilized as a valuable tool to alert drivers to potential risks on the road ahead. This dashboard can display real-time information about traffic conditions, road characteristics, and other relevant data, which can help drivers make informed decisions and take appropriate precautions while driving. For example, the dashboard can provide visual cues or audible warnings to drivers when approaching areas with a high risk of accidents, such as accident-prone intersections or stretches of road with a history of frequent collisions. By proactively informing drivers about potential risks, the dashboard can assist in reducing the occurrence of accidents and enhancing overall road safety. Additionally, the dashboard can consider weather conditions, road construction, and other relevant factors to provide comprehensive and accurate risk alerts.

to drivers. Furthermore, the dashboard can be integrated with other advanced driver assistance systems or connected vehicle technologies to provide a holistic approach to traffic risk prediction and prevention. This integration can enhance the effectiveness of the dashboard in alerting drivers to potential risks and preventing accidents.

4.5. Ablation Experiment

Furthermore, we conducted ablation experiments to assess the effectiveness of the various components of our proposed method. For this purpose, we trained and evaluated two baseline models alongside our proposed method. The first baseline, named No-Hetero, is the proposed method without the natural language description of historical traffic events and point-of-interest information. The second baseline, named No-RDN, uses a fully connected layer instead of the dilated residual layer. The results of experiments are presented in Figures 6 and 7. It can be observed that the proposed method ADRN outperforms the two baselines in terms of F1 scores for accident prediction. These results highlight the importance of the various components within the ADRN method and demonstrate that the dilated residual layer contributes to the improved performance. A dilated convolutional network can cover a larger area of the input data which increases the effective receptive field of the filter. Thus, it can expand the effective receptive field of the convolutional filters and allow the network to consider a broader local region.

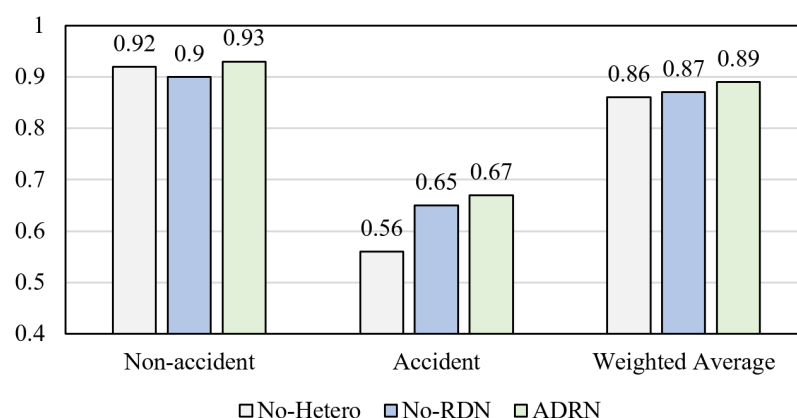


Figure 6. A comparison of F1 scores among several baselines in the Austin dataset.

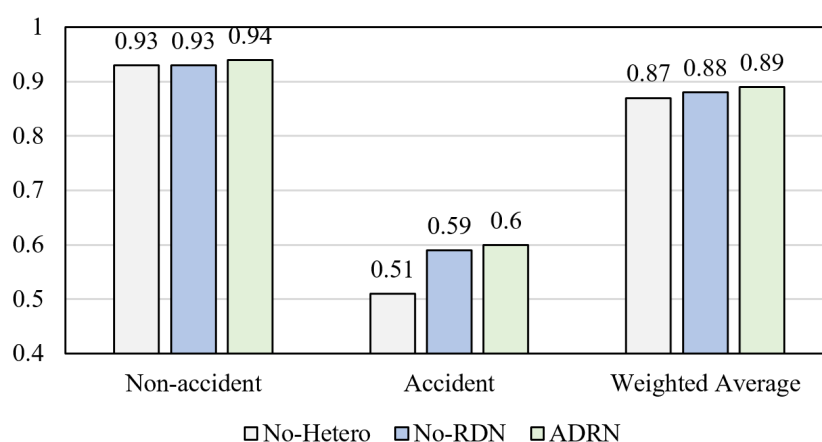


Figure 7. A comparison of F1 scores among several baselines in the Houston dataset.

In summary, ablation experiments confirm that the various components we have introduced are essential for achieving state-of-the-art performance in traffic accident prediction.

5. Conclusions

Providing reliable and precise predictions can help inform policymakers and decision-makers by analyzing the factors that contribute to traffic accidents. Therefore, we proposed a deep learning approach for traffic risk prediction using LSTM, attention mechanism, dilated residual network, and a Glove encoding technique. The experimental results of the datasets from Austin and Houston demonstrated the effectiveness of the proposed method in extracting important information from multiple sources and outperforming other baselines. Therefore, our approach has the potential to provide reliable and precise predictions that can inform policymakers and decisionmakers about the factors that contribute to traffic accidents.

This research builds upon previous studies in the field of traffic accident prediction by integrating multiple deep learning techniques and natural language processing. While previous research has focused on traditional RNNs, which might include the problems of gradient explosion and vanishing, we adopt the dilated residual network to extract complex features over long periods and utilize the attention-based approach to further enhance the model's ability. By accurately predicting traffic accidents, policymakers and decisionmakers can proactively allocate resources to high-risk areas, implement targeted preventive measures, and improve traffic management strategies. This can help reduce the number of accidents, save lives, and reduce economic losses associated with traffic accidents.

This research has limitations but suggests options for future studies. Firstly, the proposed method relies heavily on the availability and accuracy of data from various sources, which might have inherent biases or inaccuracies. Secondly, the study is based on data from Austin and Houston, which might not be fully representative of other regions. Further research could explore the generalizability of the proposed method to other locations and datasets. Additionally, machine learning methods, including deep learning techniques, may encounter challenges in delivering highly interpretable results. Interpretable models are important for policymakers and decisionmakers to understand the underlying factors that contribute to traffic accidents and make informed decisions. Therefore, future work could consider integrating causal inference techniques with machine learning approaches to gain a deeper understanding of the underlying causal mechanisms driving traffic accidents. This could potentially lead to the development of more robust and interpretable predictive models that can provide insights into causal relationships among various factors that influence traffic accidents.

Author Contributions: Conceptualization, K.Z.; methodology, K.Z.; validation, Y.G. and K.Z.; formal analysis, Y.G. and K.Z.; writing—original draft preparation, K.Z.; writing—review and editing, K.Z. and Y.G.; funding acquisition, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. Road Traffic Injuries. 2020. Available online: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries> (accessed on 10 December 2022).
2. Hosseini, S.M.; Gilani, V.N.M. Analysis of factors affecting urban road accidents in rasht metropolis. *Eng Trans.* **2020**, *1*, 1–4.
3. Čubranić-Dobrodolac, M.; Švadlenka, L.; Čičević, S.; Trifunović, A.; Dobrodolac, M. Using the interval type-2 fuzzy inference systems to compare the impact of speed and space perception on the occurrence of road traffic accidents. *Mathematics* **2020**, *8*, 1548. [\[CrossRef\]](#)
4. Zambrano-Martinez, J.L.; Calafate, C.T.; Soler, D.; Lemus-Zúñiga, L.G.; Cano, J.C.; Manzoni, P.; Gayraud, T. A centralized route-management solution for autonomous vehicles in urban areas. *Electronics* **2020**, *8*, 722. [\[CrossRef\]](#)
5. Hong, Z.; Sun, P.; Tong, X.; Pan, H.; Zhou, R.; Zhang, Y.; Han, Y.; Wang, J.; Yang, S.; Xu, L. Improved A-Star algorithm for long-distance off-road path planning using terrain data map. *Int. J. Geo Inf.* **2021**, *10*, 785. [\[CrossRef\]](#)

6. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 173–182.
7. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
8. Santos, K.; Dias, J.P.; Amado, C. A literature review of machine learning algorithms for crash injury severity prediction. *J. Saf. Res.* **2022**, *80*, 254–269. [[CrossRef](#)]
9. Milton, J.C.; Shankar, V.N.; Mannering, F.L. Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accid. Anal. Prev.* **2008**, *40*, 260–266. [[CrossRef](#)]
10. Kaplan, S.; Prato, C.G. Risk factors associated with bus accident severity in the United States: A generalized ordered logit model. *J. Saf. Res.* **2012**, *43*, 171–180. [[CrossRef](#)]
11. Vajari, M.A.; Aghabayk, K.; Sadeghian, M.; Shiwakoti, N. A multinomial logit model of motorcycle crash severity at Australian intersections. *J. Saf. Res.* **2020**, *73*, 17–24. [[CrossRef](#)]
12. De Oña, J.; Mujalli, R.O.; Calvo, F.J. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accid. Anal. Prev.* **2011**, *43*, 402–411. [[CrossRef](#)]
13. Kidando, E.; Kitali, A.E.; Kutela, B.; Karaer, A.; Ghorbanzadeh, M.; Koloushani, M.; Ozguven, E.E. Use of real-time traffic and signal timing data in modeling occupant injury severity at signalized intersections. *Transp. Res. Record.* **2022**, *2676*, 825–839. [[CrossRef](#)]
14. Li, Z.; Liu, P.; Wang, W.; Xu, C. Using support vector machine models for crash injury severity analysis. *Accid. Anal. Prev.* **2012**, *45*, 478–486. [[CrossRef](#)] [[PubMed](#)]
15. Assi, K.; Rahman, S.M.; Mansoor, U.; Ratrou, N. Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5497. [[CrossRef](#)] [[PubMed](#)]
16. Zong, F.; Xu, H.; Zhang, H. Prediction for traffic accident severity: Comparing the Bayesian network and regression models. *Math. Probl. Eng.* **2013**, *2013*, 475194. [[CrossRef](#)]
17. Abellán, J.; López, G.; De Oña, J. Analysis of traffic accident severity using decision rules via decision trees. *Expert. Syst. Appl.* **2013**, *40*, 6047–6054. [[CrossRef](#)]
18. Rezapour, M.; Molan, A.M.; Ksaibati, K. Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. *Int. J. Transp. Sci. Technol.* **2020**, *9*, 89–99. [[CrossRef](#)]
19. Zong, F.; Chen, X.; Tang, J.; Yu, P.; Wu, T. Analyzing traffic crash severity with combination of information entropy and Bayesian network. *IEEE Access* **2019**, *7*, 63288–63302. [[CrossRef](#)]
20. Chang, L.Y.; Chien, J.T. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. *Saf. Sci.* **2013**, *51*, 17–22. [[CrossRef](#)]
21. Chen, C.; Zhang, G.; Qian, Z.; Tarefder, R.A.; Tian, Z. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accid. Anal. Prev.* **2016**, *90*, 128–139. [[CrossRef](#)]
22. Asgarzadeh, M.; Fischer, D.; Verma, S.K.; Courtney, T.K.; Christiani, D.C. The impact of weather, road surface, time-of-day, and light conditions on severity of bicycle-motor vehicle crash injuries. *Am. J. Ind. Med.* **2018**, *61*, 556–565. [[CrossRef](#)]
23. Erhardt, T.; Rice, T.; Troszak, L.; Zhu, M. Motorcycle helmet type and the risk of head injury and neck injury during motorcycle collisions in California. *Accid. Anal. Prev.* **2016**, *86*, 23–28. [[CrossRef](#)] [[PubMed](#)]
24. Jiang, F.; Yuen, K.K.R.; Lee, E.W.M. Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology. *J. Saf. Res.* **2020**, *75*, 292–309. [[CrossRef](#)]
25. Crivellari, A.; Beinat, E. Forecasting spatially-distributed urban traffic volumes via multi-target LSTM-based neural network regressor. *Mathematics* **2020**, *8*, 2233. [[CrossRef](#)]
26. Zhao, S.; Xing, S.; Mao, G. An Attention and Wavelet Based Spatial-Temporal Graph Neural Network for Traffic Flow and Speed Prediction. *Mathematics* **2022**, *10*, 3507. [[CrossRef](#)]
27. Xie, Q.; Guo, T.; Chen, Y.; Xiao, Y.; Wang, X.; Zhao, B.Y. Deep graph convolutional networks for incident-driven traffic speed prediction. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual, 19–23 November 2020; pp. 1665–1674.
28. Kunt, M.M.; Aghayan, I.; Noii, N. Prediction for traffic accident severity: Comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. *Transport* **2011**, *26*, 353–366. [[CrossRef](#)]
29. Zeng, Q.; Huang, H. A stable and optimized neural network model for crash injury severity prediction. *Accid. Anal. Prev.* **2014**, *73*, 351–358. [[CrossRef](#)]
30. Sameen, M.I.; Pradhan, B. Severity prediction of traffic accidents with recurrent neural networks. *Appl. Sci.* **2017**, *7*, 476. [[CrossRef](#)]
31. Zheng, M.; Li, T.; Zhu, R.; Chen, J.; Ma, Z.F.; Tang, M.J.; Cui, Z.Q.; Wang, Z. Traffic accident's severity prediction: A deep-learning approach-based CNN network. *IEEE Access* **2019**, *7*, 39897–39910. [[CrossRef](#)]
32. Khan, M.N.; Ahmed, M.M. A novel deep learning approach to predict crash severity in adverse weather on rural mountainous freeway. *J. Transp. Saf. Secur.* **2022**. [[CrossRef](#)]
33. Moosavi, S.; Samavatian, M.H.; Parthasarathy, S.; Teodorescu, R.; Ramnath, R. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, IL, USA, 5–8 November 2019; pp. 33–42.

34. Ma, Z.; Mei, G.; Cuomo, S. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accid. Anal. Prev.* **2021**, *160*, 106322. [[CrossRef](#)]
35. Yuan, Z.; Zhou, X.; Yang, T. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 984–992.
36. Moosavi, S.; Samavatian, M.H.; Nandi, A.; Parthasarathy, S.; Ramnath, R. Short and long-term pattern discovery over large-scale geo-spatiotemporal data. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2905–2913.
37. Moosavi, S.; Samavatian, M.H.; Nandi, A.; Parthasarathy, S.; Ramnath, R. A countrywide traffic accident dataset. *arXiv* **2019**, arXiv:1906.05409.
38. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
39. He, W.; He, Z.; Wu, H.; Wang, H. Improved neural machine translation with SMT features. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision–ECCV, Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.
41. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
42. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
43. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2016; pp. 785–794.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.