This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

Explainable Traffic Accident Severity Prediction with Attention-Enhanced Bidirectional GRU-LSTM

Auwal Sagir Muhammad*, Rufai Yusuf Zakari‡, Abdullahi Baba Ari§, Cheng Wang*, and Longbiao Chen*
*Fujian Key Laboratory of Sensing, and Computing for Smart Cities, Xiamen University, Xiamen, China.
†Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Xiamen University, Xiamen, China.
‡School of Digital Science, Universiti Brunei, Darussalam, Bandar Seri Begawan, Brunei.
§ Computer Science Department, Brunel University, Uxbridge, London, UK.

***Abstract*—This study aims to improve the accuracy and interpretability of traffic accident severity nowcasting by introducing a stacked Recurrent Neural Network (RNN) deep learning model. Accurately predicting traffic accident severity is crucial for enhancing traffic management and reducing the impact of accidents. We employed a stacked Bidirectional Gated Recurrent Unit (GRU) - Long Short Term Memory (LSTM) model with an attention mechanism, integrating multivariate accident data to capture complex temporal dynamics. The use of SHapley Additive exPlanations (SHAP) values enhances the interpretability of the model. The model demonstrates high reliability and effectiveness, achieving an accuracy of 88.06% and an F1-score of 0.867 in real-time applications. It provides valuable insights into the factors influencing predictions, making the decision-making process transparent. This framework not only advances predictive performance but also aligns with ethical AI deployment, making it a valuable tool for traffic management and policy formulation.**

***Index Terms*—Traffic Accident Severity, Deep Learning, Bi-GRU-LSTM, Attention Mechanism**

I. Introduction

Traffic accidents pose a significant threat to public safety and urban development, resulting in substantial economic losses, injuries, and fatalities worldwide. Effective prediction and mitigation of these accidents are crucial for the advancement of smarter and safer cities. Traditional accident prediction methods primarily rely on historical data and static models, which often fail to capture the dynamic and complex nature of traffic systems [1]. This shortfall underscores the necessity for more sophisticated approaches that can integrate diverse real-time data sources and adapt to the ever-changing traffic environment.

Recent advancements in machine learning, particularly deep learning, offer promising solutions to these challenges. Recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, excel in modeling time series data due to their capability to capture temporal dependencies and sequential patterns [2]. These models can leverage multivariate data, encompassing various traffic-related factors, to enhance the accuracy and reliability of accident risk predictions. However, the application of such models in traffic accident nowcasting to forecast accidents shortly remains underexplored. In addition to prediction accuracy, the interpretability of these models is critically important, especially in high-stakes domains such as traffic management [3]. Understanding the factors driving model predictions can lead to actionable insights and informed interventions. Therefore, integrating explainability frameworks with deep learning models is essential to bridge the gap between model performance and practical applicability.

This study proposes a framework for nowcasting traffic accident severity using stacked Bidirectional GRU and LSTM networks enhanced with an attention mechanism. The model improves prediction accuracy by capturing temporal dependencies in both forward and backward directions while also enhancing interpretability through SHAP values. The attention mechanism helps the model focus on key features or time intervals, providing insights into the factors influencing accident severity, such as vehicle speed or road conditions, and making the decision-making process more transparent for stakeholders. Overall, this hybrid architecture enhances predictive performance by effectively capturing complex temporal dependencies and interactions in the data and facilitates better interpretability, making it easier to understand the underlying reasons for predictions. Specifically, the main contributions of this research are:

- Development of a Stacked Bidirectional GRU-LSTM Model: We propose a stacked recurrent neural network architecture combining bidirectional GRU and LSTM units with an attention mechanism to capture the temporal dynamics of multivariate traffic accident data effectively.
- Application of Bidirectional GRU-LSTM with Attention: We propose a novel use of Bidirectional GRU-LSTM layers with an attention mechanism to capture both short-term and long-term dependencies in traffic accident data, improving the model's ability to predict accident severity accurately.
- Integration of Explainability Techniques: By incorporating SHAP values, we provide an interpretable framework that reveals the key features influencing accident risk severity predictions, facilitating better traffic management and policy formulation decision-making.

Manuscript received August 4, 2024; revised August 4, 2024. Corresponding author: Auwal Sagir Muhammad (email: auwalsm@stu.xmu.edu.cn).

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 describes the nowcasting model development, Section 4 discusses model evaluation, and Section 5 concludes the paper.

II. Related Works

Accurate nowcasting of traffic accident risk levels is crucial for enhancing traffic safety and management. This task involves leveraging real-time data sources such as traffic flow, weather conditions, and historical accident records. In recent years, deep learning techniques have made significant strides in improving traffic accident risk assessment. For instance, [4] introduced the BCDU-Net framework, and [5] proposed the GSNet framework, both emphasizing the need for accurate predictions to support public safety and urban planning. These studies highlight the effectiveness of integrating multiple data sources to improve prediction accuracy and help prevent accidents.

Deep learning has become transformative in time-series prediction, particularly within traffic accident analysis, due to its ability to uncover complex dependencies over time [6]. Models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) have consistently achieved state-of-the-art results in traffic-related forecasting tasks [7]. Moreover, innovations like attention-based LSTMs and Transformer models have enhanced performance by allowing the model to focus on the most relevant data points and overcome limitations in handling long-term dependencies [8] [9].

Explainability in machine learning has emerged as a critical area, particularly in domains such as traffic accident severity prediction where understanding model decisions is vital. Techniques like SHAP values and attention mechanisms have been successfully employed to increase the interpretability of these models by emphasizing the contributions of individual features and focusing on key parts of the input sequence [10]. This improves transparency, making predictions more understandable and actionable, thereby supporting more informed decision-making in traffic management [8].

Recent works in traffic accident severity prediction have further contributed to the integration of real-time data and advanced machine-learning models. For example, studies by Zhang et al. [1] and Li et al. [11] explored the application of Transformer architectures to predict accidents, highlighting the importance of capturing both temporal patterns and contextual data. Our study builds upon these advancements by incorporating multivariate data sources and attention mechanisms, with a focus on interpretability through SHAP values. Unlike previous studies, our approach prioritizes not only high prediction accuracy but also insights into the underlying factors that drive these predictions.

In summary, advancements in deep learning and explainability techniques have significantly improved traffic accident risk assessment and prediction, enabling more effective and safer traffic management strategies.

III. Methodology

In this section, we detail the processes and techniques used to develop our traffic accident severity nowcasting model. Our methodology includes data preprocessing, feature extraction, and model training using a stacked Bidirectional GRU-LSTM with an attention mechanism, alongside explainability techniques as illustrated in Figure 1. These steps are essential for ensuring the accuracy and interpretability of the predictions. Each phase is described in the following subsections.

A. Data Preprocessing

In the first phase, we meticulously prepare the dataset for modeling by performing several crucial steps. The raw traffic accident multivariate time-series dataset underwent extensive preprocessing, including cleaning, normalization, and feature extraction, to ensure its suitability for model training. To ensure consistency and improve the model's performance, we standardize the features using the StandardScaler. This transformation scales the features such that they have a mean of 0 and a standard deviation of 1, which can be represented mathematically as

$$X' = \frac{X - \mu}{\sigma} \tag{1}$$

where X is the original feature, μ is the mean, and σ is the standard deviation. After standardization, we adjust the target variable y by decrementing it by 1 to start the class labels from 0. Subsequently, we split the dataset into training and testing sets using a 70/30 split ratio. We employed SMOTEENN to address the class imbalance inherent in traffic accident datasets. This hybrid approach first generates synthetic samples for the minority class using SMOTE, followed by ENN, which removes misclassified samples. This dual approach improves the model's ability to generalize from imbalanced data, ensuring more robust predictions. SMOTE generates synthetic samples for the minority class, mathematically described by

$$x_{\text{new}} = x_i + \lambda \cdot (x_j - x_i) \quad \text{for} \quad \lambda \in [0, 1] \tag{2}$$

where x_i and x_j are minority class samples. ENN then removes samples that are misclassified by their nearest neighbors, balancing the dataset effectively. Furthermore, we compute class weights to mitigate the impact of class imbalance during model training. These weights are calculated using the formula:

$$w_i = \frac{N}{k \cdot n_i} \tag{3}$$

where N is the total number of samples, k is the number of classes, and n_i is the number of samples in class i. Combining both methods can provide a more robust solution to class imbalance, as it tackles the problem from two angles: data-level (SMOTE) and algorithm-level (class weights).

B. Feature Extraction

We employ Sequential Forward Selection (SFS) to select the top features for training the logistic regression model. SFS was chosen for feature selection due to its efficiency in selecting the most relevant features without introducing

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

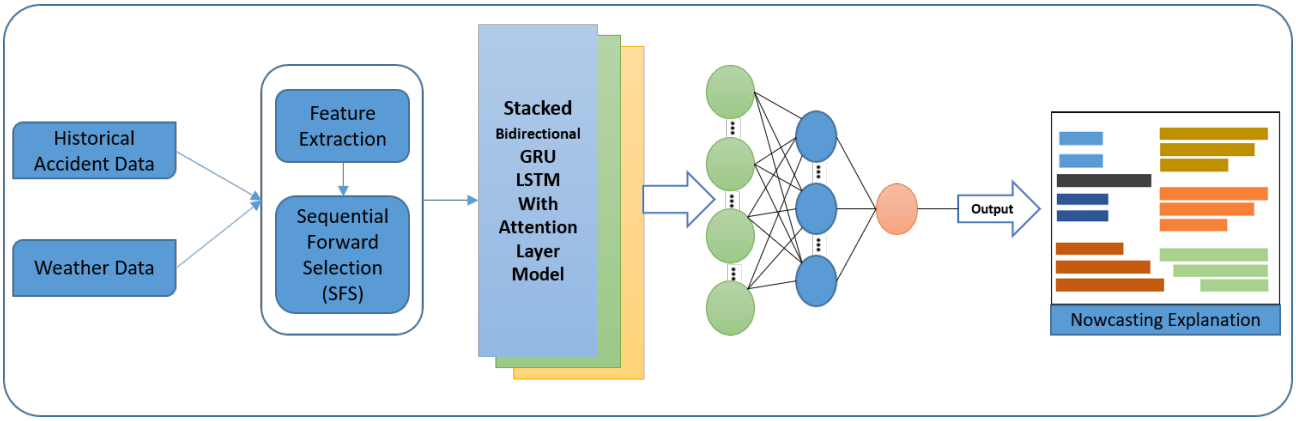

Fig. 1. Framework Overview

redundancy, optimizing the model's performance in terms of both accuracy and interpretability. The SFS process begins by initializing an empty set of selected features, $\mathcal{S}_0 = \emptyset$. For each iteration k from 0 to 9, the algorithm evaluates the performance $\mathcal{M}(\mathcal{S}_k \cup \{f\})$ of the logistic regression model for each feature f not already in $\mathcal{S}_k$. It selects the feature f_k that maximizes this performance and updates the set of selected features to $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{f_k\}$. This process repeats until 10 features are selected.

Algorithm 1: Feature Selection using SFS

Data: Feature set $\mathcal{F}$, Performance metric $\mathcal{M}$, Number of features to select $n = 10$
Result: Selected feature set $\mathcal{S}_{10}$

Initialization: $\mathcal{S}_0 \leftarrow \emptyset$;
for $k \leftarrow 0$ ***to*** $n - 1$ **do**
  $f_k \leftarrow \arg\max_{f \in \mathcal{F} \setminus \mathcal{S}_k} \mathcal{M}(\mathcal{S}_k \cup \{f\})$
  $\mathcal{S}_{k+1} \leftarrow \mathcal{S}_k \cup \{f_k\}$
end
Return $\mathcal{S}_{10}$

This comprehensive preprocessing phase ensures that the raw traffic accident and sensor data are cleaned, integrated, and transformed to extract useful features and handle class imbalance, thereby providing a solid foundation for developing robust deep learning models.

C. Nowcasting Model Development

This phase involves developing advanced deep learning models to forecast future accident severity levels using historical preprocessed data. We constructed a stacked recurrent neural network architecture to capture temporal dependencies in the accident data, integrating Bidirectional GRU and LSTM layers enhanced with an Attention mechanism for improved focus on relevant input sequence parts.

1) Model Architecture: The architecture of our stacked recurrent neural network (RNN) model is meticulously designed to capture complex temporal dependencies and patterns within traffic accident data. We employ a combination of Bidirectional GRU, LSTM layers, and an Attention mechanism to leverage their unique advantages in handling sequential data as illustrated in Figure 2.

- **Bidirectional GRU Layer**: The model begins with a Bidirectional GRU layer comprising 128 units. The return sequences parameter enables the GRU layer to output the full sequence of predictions to the next layer, allowing for further stacking of RNN layers. A GRU unit is composed of a reset gate r_t and an update gate z_t. The output h_t is determined by the current input x_t and the previous state h_{t-1}, under the control of these two gates. The equations for the GRU unit are as follows:

$$\begin{aligned} r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ \tilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned} \tag{4}$$

Where W_r, U_r, W_z, U_z, W_h, and U_h are the weight matrices, b_r, b_z, b_h are the bias vectors, σ is the logistic sigmoid function, $\tanh$ is the hyperbolic tangent activation function, and $\odot$ denotes the Hadamard product (element-wise multiplication).
In a Bidirectional GRU, there are two GRUs: one moving forward (from the start of the sequence) and the other moving backward (from the end of the sequence). The hidden state h_t at each time step is a concatenation of the forward and backward hidden states $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$.

$$\begin{aligned} \overrightarrow{h_t} &= GRU_{\text{fwd}}(x_t, \overrightarrow{h_{t-1}}) \\ \overleftarrow{h_t} &= GRU_{\text{bwd}}(x_t, \overleftarrow{h_{t+1}}) \\ h_t &= \overrightarrow{h_t} \oplus \overleftarrow{h_t} \end{aligned} \tag{5}$$

Where $\overrightarrow{h_t}$ is the hidden state from the forward GRU, $\overleftarrow{h_t}$ is the hidden state from the backward GRU, and $\oplus$ denotes the concatenation of the two hidden states.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

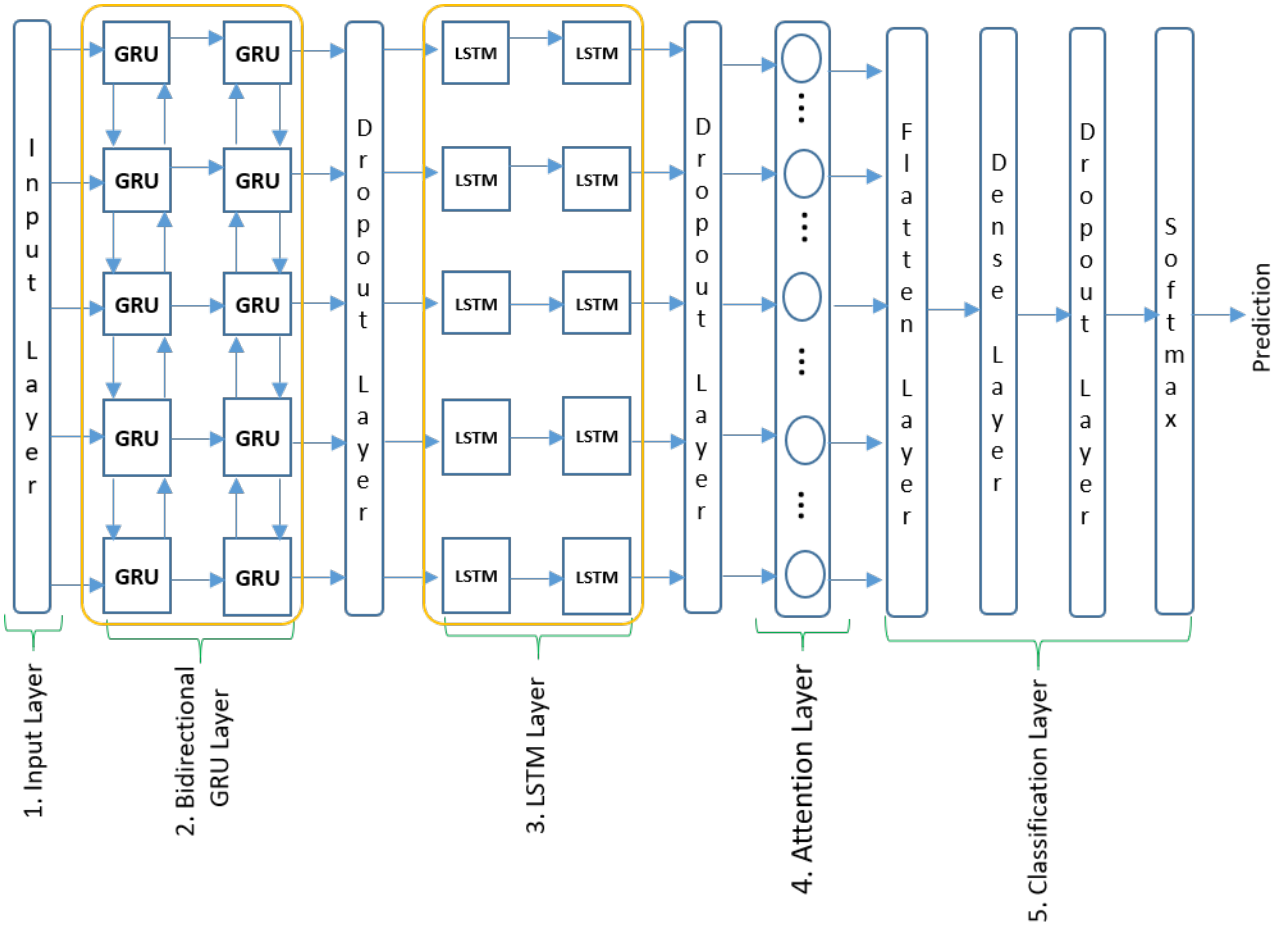

Fig. 2. Stacked Model Architecture

- **Dropout Layer**: A Dropout layer with a dropout rate of 0.2 follows, helping to regularize the model and prevent overfitting by randomly setting a fraction of input units to zero during training.
- **LSTM Layer**: The next layer is the LSTM layer with 64 unit which are particularly effective in learning long-term dependencies due to their gating mechanisms, which control the flow of information.

$$\begin{aligned} r_t &= \sigma(W_r \cdot (x_t \oplus S_{(t-1)}) + b_r); \\ z_t &= \sigma(W_z \cdot (x_t \oplus S_(t-1)) + b_z); \\ S_t &= tanh(W_s \cdot (x_t \oplus S_{(t-1)} \cdot r_t) + b_s); \\ S_t &= (1 - Z_t) \cdot S_{(t-1)} + Z_t \cdot S_t; and \\ y_t &= \sigma(W_y \cdot S_t + b_y), \end{aligned} \tag{6}$$

where $x_t \in R^m$ is the input vector of m input features at time t; $W_r, W_z, W_S \in R^{(n\times(m+n))}$ and $W_y \in R^{(n\times n)}$ are parameter matrices; n is the number of neurons in the GRU layer; $b_r, b_z, b_S, b_y \in R^n$ are bias vectors; σ is the sigmoid activation function, and S_t is the internal (hidden) state. The functions Z_t, r_t and (S_t) are implemented by the update gate, reset gate, and the third gate, respectively [12].

- **Second Dropout Layer**: Another Dropout layer with the same dropout rate is applied to reduce the risk of overfitting further.
- **Attention Layer**: The Attention mechanism is employed next to enable the model to focus on relevant parts of the input sequence, improving interpretability and performance. The Attention mechanism can be mathematically represented as follows:

$$\begin{aligned} e_t &= \tanh(W_a \cdot h_t + b_a) \\ a_t &= \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \\ c &= \sum_{t=1}^{T} a_t \cdot h_t \end{aligned} \tag{7}$$

where W_a and b_a are the trainable weight and bias parameters, respectively. The attention scores a_t are computed

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

using a softmax activation, and the context vector c is obtained by a weighted sum of the hidden states.

- **Dense Layer**: A Dense layer with 50 units and ReLU activation with L2 Regularization is included to introduce non-linearity, address overfitting and capture complex patterns in the data.
- **Output Layer**: Finally, a Dense output layer with 3 units and a softmax activation function is used for multi-class classification. The softmax function ensures that the output probabilities sum to 1, suitable for categorical classification tasks.

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{8}$$

where $\hat{y}_i$ is the predicted probability for class i.

The model is compiled with the Adam optimizer, known for its efficient handling of sparse gradients and adaptive learning rate. The loss function used is sparse categorical cross-entropy, appropriate for multi-class classification problems with integer labels defined as:

$$\text{loss} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{9}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for class i. The training process is monitored using early stopping with a patience of 5 epochs to prevent overfitting by stopping training when the validation loss stops improving. This architecture effectively combines the strengths of Bidirectional GRU and LSTM layers with the Attention mechanism to learn hierarchical temporal representations, ensuring robust performance in classifying traffic accident severity.

D. Nowcasting Explainability

In the final phase, we focus on evaluating the model's performance and interpreting its predictions to ensure transparency and trustworthiness. We begin by predicting the probabilities for the test set and computing the accuracy, classification report, and various metrics such as ROC-AUC and F1 scores for each class. Additionally, we utilize SHAP (SHapley Additive exPlanations) to interpret the model's predictions. SHAP values are derived from Shapley values, a concept from cooperative game theory that fairly allocates the contribution of each feature to the prediction [13]. The Shapley value for a feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{10}$$

where N is the set of all features, S is a subset of N excluding feature i, $f(S)$ is the prediction for subset S, and $|S|$ is the number of features in subset S.

SHAP summary plots are generated to display feature importance, dependence plots to examine the relationship between features and predictions, and force plots to illustrate individual prediction explanations. These visualizations will provide insights into how the model makes decisions, highlighting key features and their influence.

IV. Evaluation

A. Dataset Description

The dataset is obtained from the UK Department of Transport, which provides a detailed road accident record collected between 2005 and 2015 [14]. The dataset consists of 1,22,636 instances and 32 attributes with Table 1 below presents some of the main attributes and their description.

TABLE I
Brief Description of the UK Road Accident Dataset

SN	Attribute	Description
1	Index	Index of the traffic accident.
2	Longitude	Longitude of the location of an accident scene.
3	Latitude	Latitude of the location of an accident scene.
4	Accident Severity	Severity of the accident: fatal, serious or slight.
5	Vehicles	Number of vehicles involved in the accident.
6	Casualties	The number of persons injured in the accident.
7	Date	Date of the accident.
8	Time	Timestamp of the accident.
9	Week Day	Day of the week that accident occurred.
10	Road Type	The type of road where the accident occurred.
11	Speed Limit	Speed limit of road where accident occurred.
12	Weather	Weather condition at the time of the accident.
13	Light Conditions	Light conditions at the time of the accident.
14	Rural / Urban Area	Area where the accident occurred.

B. Evaluation Metrics

A set of classification metrics was employed to evaluate the performance of the nowcasting model in predicting traffic accident severity. These metrics provide a comprehensive assessment of the model's accuracy, precision, recall, and overall effectiveness in handling the imbalanced nature of the dataset. The following metrics are particularly relevant for this classification problem:

- **Accuracy**: Accuracy is calculated as the ratio of correctly predicted instances to the total number of instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{11}$$

- **Precision, Recall, and F1-Score**: These metrics are crucial for imbalanced datasets where certain classes may be underrepresented.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

- **ROC-AUC**: The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across different threshold settings.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

The AUC provides a single measure of overall model performance, with higher values indicating better discriminatory ability.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{15}$$

By utilizing these metrics, the evaluation comprehensively assesses the model's ability to accurately predict traffic accident severity levels, handle imbalanced data, and provide reliable forecasts. This thorough evaluation ensures that the model's predictions are robust, interpretable, and actionable for traffic safety analysis and intervention planning.

C. Model Performance

1) Experiment Settings: All experiments were conducted on a Windows platform with an Intel(R) Core(TM) i7-8700K CPU at 3.70 GHz and an NVIDIA A40 GPU, ensuring robust computational power for model training and evaluation. The implementation of neural network-based models, including the stacked Bidirectional GRU-LSTM architecture, was carried out using TensorFlow, benefiting from its extensive library support and GPU acceleration.

We conducted extensive hyperparameter tuning using the RandomizedSearchCV technique from scikit-learn. This process explored various configurations, including GRU units (64 to 256), LSTM units (32 to 128), dropout rates (0.2 to 0.4), batch sizes (16, 32, 64), learning rates (0.001 to 0.00001), and training epochs (50 to 150). Through a randomized search with 20 iterations and 3-fold cross-validation, we identified the optimal hyperparameters that maximized model accuracy and mitigated overfitting.

To evaluate the performance of our proposed model, we conducted extensive experiments with the optimal hyperparameters using a comprehensive multivariate dataset of traffic accident records. We divided the dataset into training (70%) and testing (30%) sets to ensure the robustness of our evaluation. We trained the stacked Bidirectional GRU-LSTM model with an attention mechanism using the training set and evaluated its performance on the testing set. The model training was monitored using early stopping to prevent overfitting, with a patience of 10 epochs.

Figure 3 illustrates the training and validation loss curves, showing the model's convergence during training. Figure 4 presents the ROC curves for all classes, providing a graphical representation of the model's ability to distinguish between different severity levels of traffic accidents. The area under the ROC curve (AUC) was calculated for each class, offering a quantitative measure of the model's discriminatory power.

2) Evaluation Results: Table 2 presents the evaluation results of our model which was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

3) Ablation Study: To understand the contributions of individual components in our Bidirectional GRU-LSTM model, we conducted an ablation study examining the effects of bidirectional layers, attention mechanisms, and the combination of GRU and LSTM layers on model performance.

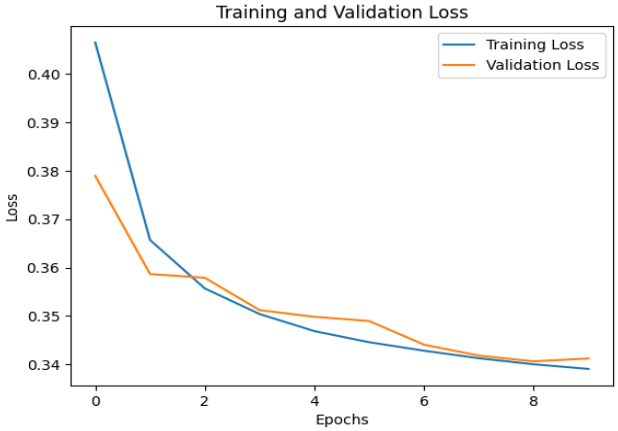

Fig. 3. Training and validation loss

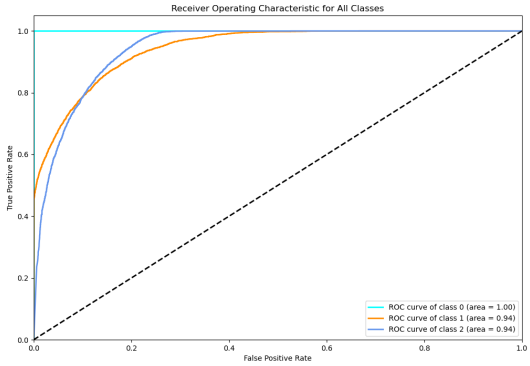

Fig. 4. ROC Curves for all classes

We tested various configurations, including GRU, LSTM, Bi_GRU, Bi_LSTM, and their attention-enhanced versions. The addition of bidirectional layers resulted in modest improvements in accuracy and F1-score, indicating better capture of temporal dependencies, while attention mechanisms provided slight gains by focusing on key input features. However, combining GRU and LSTM layers alone did not substantially improve performance. The proposed Bidirectional GRU-LSTM with attention mechanism achieved the highest accuracy (88.06%) and F1-score (0.867), demonstrating its effectiveness in capturing complex temporal patterns and making it the best model for predicting traffic accident severity. This study highlights the importance of integrating attention with bidirectional layers and combining GRU and LSTM for optimal performance.

D. Explainability Results

To ensure the transparency and interpretability of our traffic accident severity prediction model, we employed SHapley Additive exPlanations (SHAP) to analyze feature importance and individual prediction explanations. SHAP values provide

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

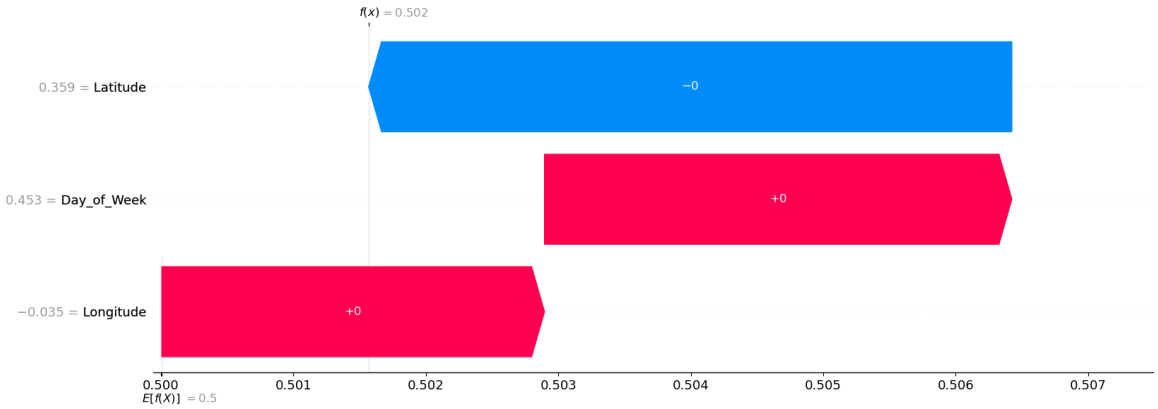

Fig. 5. SHAP Force Plot

TABLE II
EVALUATION RESULT

Severity	ROC-AUC	Precision	Recall	F1-Score
Level 1	1.00	1.00	0.99	0.99
Level 2	0.94	0.84	0.71	0.77
Level 3	0.94	0.75	0.87	0.84

TABLE III
ABLATION STUDY RESULT

Model	Accuracy	F1-Score
Bi_GRU	85.98	0.859
Bi_LSTM	86.24	0.862
Bi_GRU_Attention	85.98	0.860
Bi_LSTM_Attention	86.02	0.860
Bi_GRU_Bi_LSTM	85.88	0.859
Bi_GRU_LSTM_Attention	**88.06**	**0.867**

a unified measure of feature importance, enabling us to understand the impact of each feature on the model's predictions.

Figure 6 presents the SHAP summary plot, which illustrates the distribution of SHAP values for each feature across all predictions. This plot highlights the most influential features contributing to the model's predictions. Notably, 'Weather Conditions: Raining without high winds', 'Longitude', and 'Number of Casualties' emerged as key features affecting the prediction outcomes.

To gain insights into individual predictions, we generated SHAP force plots. Figure 5 shows a SHAP force plot for a specific prediction, depicting how each feature contributes to the final prediction of traffic accident severity. For this particular instance, 'Latitude' has a negative impact, while 'Day of Week' and 'Longitude' have positive impacts on the prediction.

Additionally, we analyzed the overall feature importance using a feature importance chart, as shown in Figure 7. This chart ranks the features based on their average impact on the model's predictions. 'Weather Conditions: Raining without high winds' is identified as the most critical feature, followed by 'Longitude' and 'Number of Casualties'.

These visualizations and analyses using SHAP enhance the explainability of our model, ensuring that stakeholders can trust and comprehend the factors influencing the predictions.

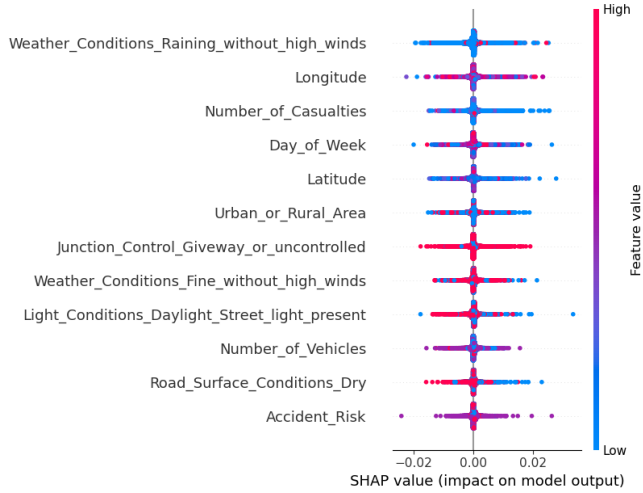

Fig. 6. SHAP summary plot

E. Discussion

The proposed Bidirectional GRU-LSTM model, enhanced with an attention mechanism and SHAP values for interpretability, significantly improves traffic accident severity prediction. Through our ablation study, we examined the individual contributions of bidirectional layers, attention mechanisms, and the combination of GRU and LSTM layers to understand their impact on model performance. The study revealed that while bidirectional layers (Bi_GRU and Bi_LSTM) offered modest improvements in accuracy and F1-score, they helped the model capture both forward and backward temporal dependencies more effectively than traditional GRU or

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/SWC62898.2024.00174, 2024 IEEE Smart World Congress (SWC)

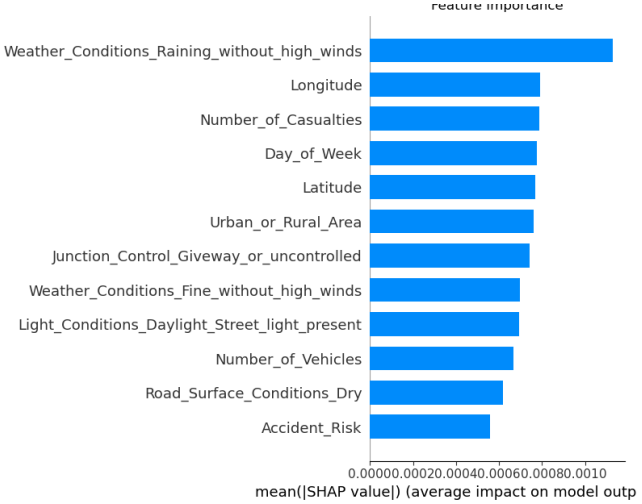

Fig. 7. Feature Importance Chart

LSTM models. The addition of attention mechanisms further enhanced the model's ability to focus on critical features of the input sequence, leading to slight improvements in performance metrics. However, combining GRU and LSTM layers alone did not result in substantial gains, indicating that the complementary strengths of these recurrent units are best realized when integrated with bidirectional and attention components. Our final model, the Bidirectional GRU-LSTM with Attention, achieved the highest performance across all configurations, with an accuracy of 88.06% and an F1-score of 0.867. This demonstrates that the combination of GRU, LSTM, attention, and bidirectional layers provides the most robust framework for accurately predicting traffic accident severity, as it effectively captures both short- and long-term dependencies while maintaining interpretability through SHAP values.

While the model performs well, several limitations must be acknowledged. The model's reliance on historical data may reduce its accuracy in predicting rare or unforeseen events, such as extreme weather conditions or sudden road closures. Additionally, the scalability of the model for real-time, large-scale traffic systems has not been fully tested. The model may also struggle with rare but severe accidents, despite the use of class imbalance techniques like SMOTEENN. Lastly, the current feature set could benefit from the inclusion of real-time driver behavior and road condition data, which could further enhance the model's adaptability to dynamic traffic environments.

Future work will focus on addressing these limitations by integrating additional data sources, improving scalability for real-time applications, and exploring alternative architectures such as Transformer models or ensemble learning techniques. This would further strengthen the model's applicability to real-world traffic management systems and its ability to predict traffic accident severity in more complex scenarios.

V. Conclusion

This study presents a novel framework for nowcasting traffic accident severity using a stacked Bidirectional GRU-LSTM model with an attention mechanism. The proposed model effectively integrates multivariate accident data, capturing the complex temporal dynamics and improving prediction accuracy. The incorporation of SHAP values enhances the model's interpretability, providing valuable insights into the factors influencing accident severity predictions. The results demonstrate the model's potential to significantly improve traffic management and policy formulation significantly, contributing to safer urban environments.

Future work will further refine the model by incorporating additional data sources and exploring advanced deep learning techniques. The integration of external factors such as road conditions, traffic regulations, and driver behavior can further enhance the model's predictive capabilities. Additionally, the deployment of the model in real-time traffic management systems will be explored to evaluate its practical applicability and impact on traffic safety.

References

[1] Amin Karimi Monsefi, Pouya Shiri, Ahmad Mohammadshirazi, Nastaran Karimi Monsefi, Ron Davies, Sobhan Moosavi, and Rajiv Ramnath. Crashformer: A multimodal architecture to predict the risk of crash. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Advances in Urban-AI*, pages 42–51, 2023.

[2] Xueli Zhang, Cankun Zhong, Jianjun Zhang, Ting Wang, and Wing WY Ng. Robust recurrent neural networks for time series forecasting. *Neurocomputing*, 526:143–157, 2023.

[3] Zhixuan Zeng, Xianming Tang, Yang Liu, Zhengkun He, and Xun Gong. Interpretable recurrent neural network models for dynamic prediction of the extubation failure risk in patients with invasive mechanical ventilation in the intensive care unit. *Biodata Mining*, 15(1):21, 2022.

[4] Zhenghua Hu, Jibiao Zhou, and Enyou Zhang. Improving traffic safety through traffic accident risk assessment. *Sustainability*, 15(4):3748, 2023.

[5] Noushin Behboudi, Sobhan Moosavi, and Rajiv Ramnath. Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques. *arXiv preprint arXiv:2406.13968*, 2024.

[6] Angelo Casolaro, Vincenzo Capone, Gennaro Iannuzzo, and Francesco Camastra. Deep learning for time series forecasting: Advances and open problems. *Information*, 14(11):598, 2023.

[7] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[8] Qing Kang, Elton J Chen, Zhong-Chao Li, Han-Bin Luo, and Yong Liu. Attention-based lstm predictive model for the attitude and position of shield machine in tunneling. *Underground Space*, 13:335–350, 2023.

[9] R Mohammadi Farsani and Ehsan Pazouki. A transformer self-attention model for time series forecasting. *Journal of Electrical and Computer Engineering Innovations (JECEI)*, 9(1):1–10, 2020.

[10] Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396, 2022.

[11] Mansoor G Al-Thani, Ziyu Sheng, Yuting Cao, and Yin Yang. Traffic transformer: Transformer-based framework for temporal traffic accident prediction. *AIMS Mathematics*, 9(5):12610–12629, 2024.

[12] J Shi, M Jain, and G Narasimhan. Time series forecasting (tsf) using various deep learning models. arxiv 2022. *arXiv preprint arXiv:2204.11115*.

[13] Meng Li, Hengyang Sun, Yanjun Huang, and Hong Chen. Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, 4(1):2, 2024.

[14] Department for Transport. Road safety data, 2023.