Type of the Paper (Article, Review, Communication, etc.)

Test-Time Training with Adaptive Memory for Traffic Accident Severity Prediction

Duo Peng [1], Wei Qi Yan[2]

Auckland University of Technology, Auckland 1010 New Zealand
* Correspondence: pnt7686@autuni.ac.nz, weiqi.yan@aut.ac.nz

Abstract: Traffic accident prediction is essential for improving road safety and optimizing intelligent transportation systems. However, deep learning models often struggle with distribution shifts and class imbalance, leading to degraded performance in real-world applications. The existing Transformer-based models, despite their ability to capture long-term dependencies, lack mechanisms to adapt dynamically to new environments. In this paper, we introduce Test-Time Training (TTT) as a strategy to enhance Transformer-based accident prediction by allowing the model to refine its parameters during inference through a self-supervised auxiliary task. To further improve performance, Adaptive Memory Layer (AML), Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss are incorporated, addressing challenges related to long-term dependencies, multi-scale feature extraction, and imbalanced accident severity classifications. Our experimental results demonstrate that the proposed TTT-Enhanced Transformer outperforms standard Transformers and LSTMs, achieving higher accuracy, recall, and F1-score, particularly for severe accidents (Level 3 & 4), which are historically difficult to predict due to data imbalance. Confusion matrix and ROC curve confirm that TTT significantly reduces misclassification errors and enhances prediction reliability. These findings highlight the potential of TTT-Enhanced Transformer in mitigating real-world challenges in traffic accident prediction, improving model adaptability under shifting data distributions and class imbalances.

Keywords: Test-Time Training, Traffic Accident Prediction, Transformer Network, Self-Supervised Learning, Adaptive Memory, Class-Balanced Learning

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

1. Introduction

Traffic accidents remain a critical public safety concern, resulting in significant human and economic losses worldwide. According to the World Health Organization (WHO), more than 1.35 million people lose their lives in road accidents annually, with economic damages exceeding $800 billion in the United States alone. The increasing complexity of urban transportation systems, driven by rising population densities, evolving infrastructure, and unpredictable human behavior, necessitates advanced, data-driven solutions for improving traffic safety. Deep learning models, particularly Transformer architectures, have demonstrated remarkable performance in modeling sequential data[1–3], However, their effectiveness in real-world applications is often hindered by

distribution shifts and class imbalance, two fundamental challenges that degrade prediction reliability[4,5].

Most existing traffic accident prediction models are trained on static datasets and struggle to generalize when deployed in dynamic, real-world scenarios. As a result, their predictions deteriorate under distribution shifts, where unseen environmental conditions lead to significant accuracy degradation. To address this, Test-Time Training (TTT) has been introduced, enabling Transformer-based models to dynamically refine their parameters during inference using self-supervised auxiliary tasks[6]. This approach has shown success in various applications, such as handwritten document recognition, where an auxiliary branch continuously updates model parameters for enhanced adaptability[7]. Recent advancements like Test-Time Self-Training (TeST) have further improved test-time adaptation by employing a student-teacher framework, allowing models to learn robust and invariant representations under distribution shifts[8]. Comprehensive studies have emphasized the significance of TTT for handling such shifts across multiple domains[9].

Traffic accident prediction models often struggle with class imbalance, where severe but infrequent accidents are underrepresented in training data, leading to biased predictions that favor more common, less severe events while failing to recognize high-risk scenarios[5,10]. An analysis of the dataset in this paper reveals a highly skewed long-tail distribution of accident severity, as illustrated in ***Figure 1.*** The data shows that moderate accidents (Level 2) constitute 79.6% of all cases, whereas severe accidents (Level 4) represent only 2.6%. The maximum-to-minimum imbalance ratio reaches 93.2:1, highlighting the extreme disparity between frequent minor incidents and rare but critical severe accidents. This imbalance causes conventional models to prioritize majority classes, resulting in poor recall and frequent misclassification of severe accidents—the most critical category for traffic safety interventions[11,12]. Addressing this issue requires models that can adapt to the underlying distribution shifts while ensuring fair representation of minority accident types.

Previous studies address this challenge using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which improved temporal modeling but suffered from vanishing gradient issues and failed to capture long-range dependencies effectively[13]. The emergence of Transformer models revolutionized sequence modeling by leveraging self-attention mechanisms, enabling better long-term dependency modeling compared to RNNs and LSTMs[2,14]. However, even Transformer-based models remain vulnerable to class imbalance, leading to biased predictions toward frequent accident categories[5,11,15].

Recent advancements in self-supervised learning and adaptive training have provided promising solutions for mitigating both distribution shifts and class imbalance[5,6,16]. TTT, along with techniques such as meta-learning and continual learning, has shown potential in enhancing model generalization. While prior research has validated

TTT's effectiveness in structured environments, its capability in real-time traffic accident prediction remains underexplored[16–18]. The existing models struggle to differentiate between severe and minor accidents due to imbalanced training data, leading to overconfidence in frequent classes and poor generalization to rare but critical cases[5,11,15].

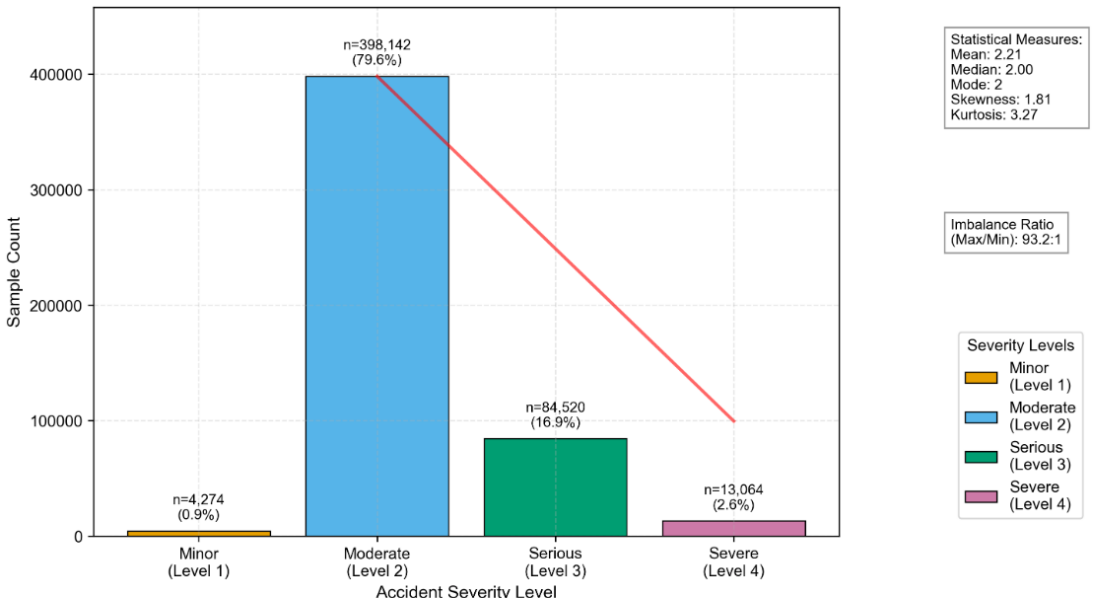

Figure 1. Distribution of accident severity levels in our dataset

To address these challenges, in this paper, we introduce TTT as the core adaptation mechanism in a Transformer-based framework for traffic accident prediction. TTT dynamically refines model parameters during inference, enabling the model to mitigate distribution shifts and improve real-time adaptability. To further enhance predictive performance, in this paper, we integrate Adaptive Memory Layer (AML), Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss into a Transformer-based TTT framework for traffic accident prediction. AML enhances the model's ability to retain long-term dependencies[19,20], while FPN improves multiscale feature extraction[21,22]. CBA and Focal Loss are designed to mitigate class imbalance, ensuring that severe accidents receive adequate representation during model training[15,23]. Our experimental results confirm that combining TTT with imbalance-aware learning strategies significantly improves accident severity classification, particularly for rare but high-risk cases.

2. Materials and Methods

This paper follows a structured methodology to develop a scalable and adaptive traffic accident prediction framework. The workflow integrates deep learning models to enhance predictive performance. The following subsections detail the research design, data processing pipeline, model architecture, training procedures, evaluation metrics, and ablation studies conducted to assess the impact of different model components.

2.1. Data Collection

The dataset, sourced from Kaggle (DOI: 10.34740/kaggle/ds/199387, available at: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data), comprises traffic accident records from 49 states across the United States, spanning the years 2016 to 2023. A subset of 500,000 records was selected for model training and evaluation, including meteorological conditions, traffic density, road types, and accident severity. Accident severity, the target variable, is categorized into four levels: minor (minimal impact), moderate (traffic delays without major disruption), severe (significant congestion and possible injuries), and extreme (major disruptions with serious injuries or fatalities). As illustrated in ***Figure 1***, the dataset exhibits a highly imbalanced distribution, where moderate accidents dominate, while severe and extreme cases are significantly underrepresented. This predefined classification serves as the foundation for training models to distinguish varying levels of accident severity.

2.2. Data Preprocessing

A structured data pre-processing pipeline was developed to ensure consistency and enhance predictive performance. The features with more than 30% missing values were removed, while those with lower missing rates were imputed using mean or median values. To improve feature representation, temporal attributes such as hour, weekday/weekend, and seasonality were extracted to capture variations in accident risk, while geospatial factors, including proximity to highways, intersections, and traffic signals, were incorporated to identify accident hotspots. Interaction terms, such as temperature-visibility and humidity-wind speed, were introduced to account for environmental dependencies. Numerical attributes were standardized using Min-Max Scaling, and categorical variables, including weather conditions and accident locations, were encoded using one-hot encoding to prevent artificial ordinal relationships. Additionally, composite features, such as traffic density and weather impact scores, were derived to better capture patterns associated with accident severity. The complete preprocessing workflow, covering data cleaning, feature engineering, standardization, and class balancing, is illustrated in ***Figure 2***.

2.3. Model Architecture

The proposed model extends the baseline Transformer by incorporating multi-scale feature extraction, adaptive memory, and TTT, as shown in ***Figure 3***. It consists of four key components: Feature Pyramid Network (FPN) for capturing hierarchical traffic patterns,

Adaptive Memory Layer (AML) for retaining long-term dependencies, Class-Balanced Attention (CBA) for mitigating class imbalance, and TTT for real-time adaptation.

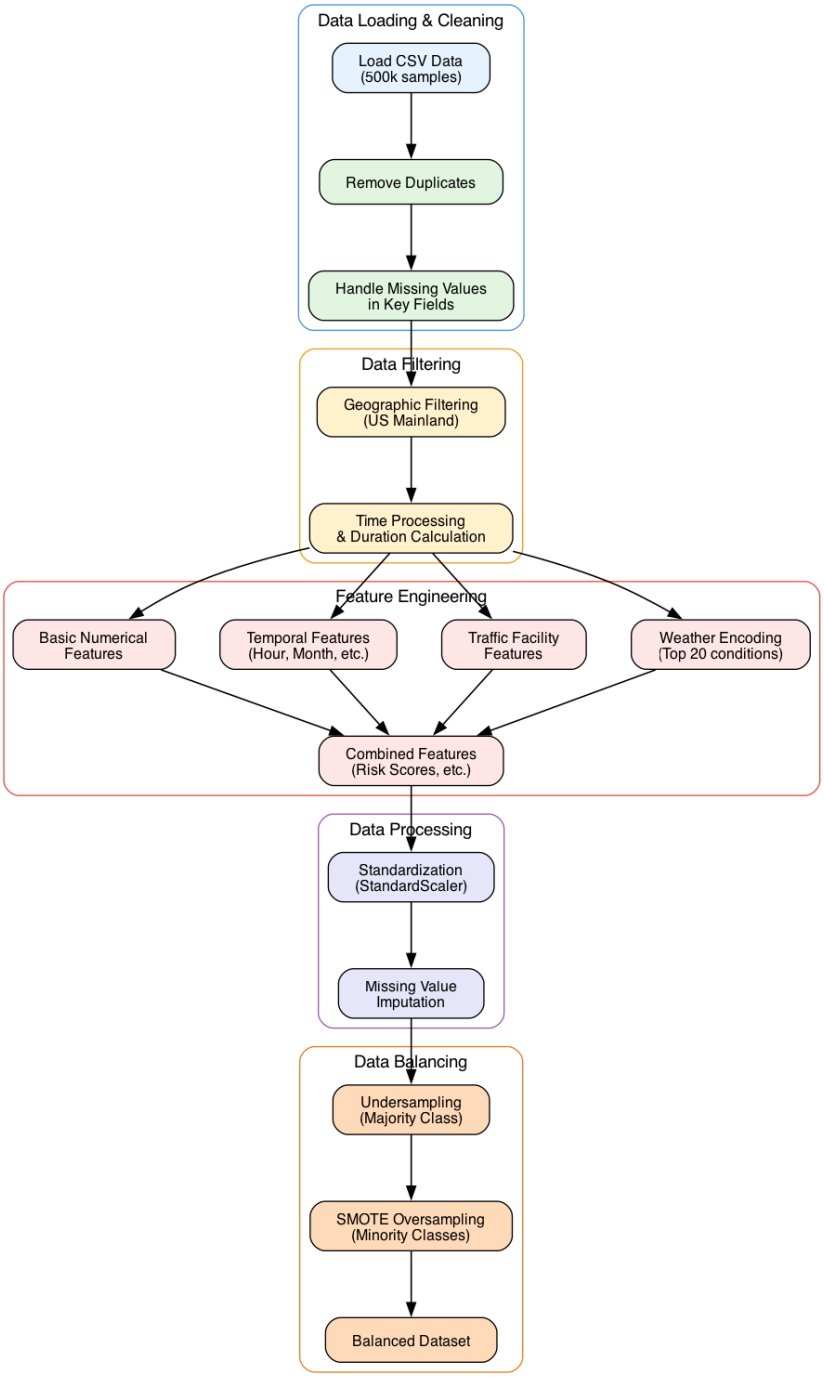

Figure 2. Overview of the data pre-processing pipeline, detailing data cleaning, filtering, feature engineering, and balancing steps.

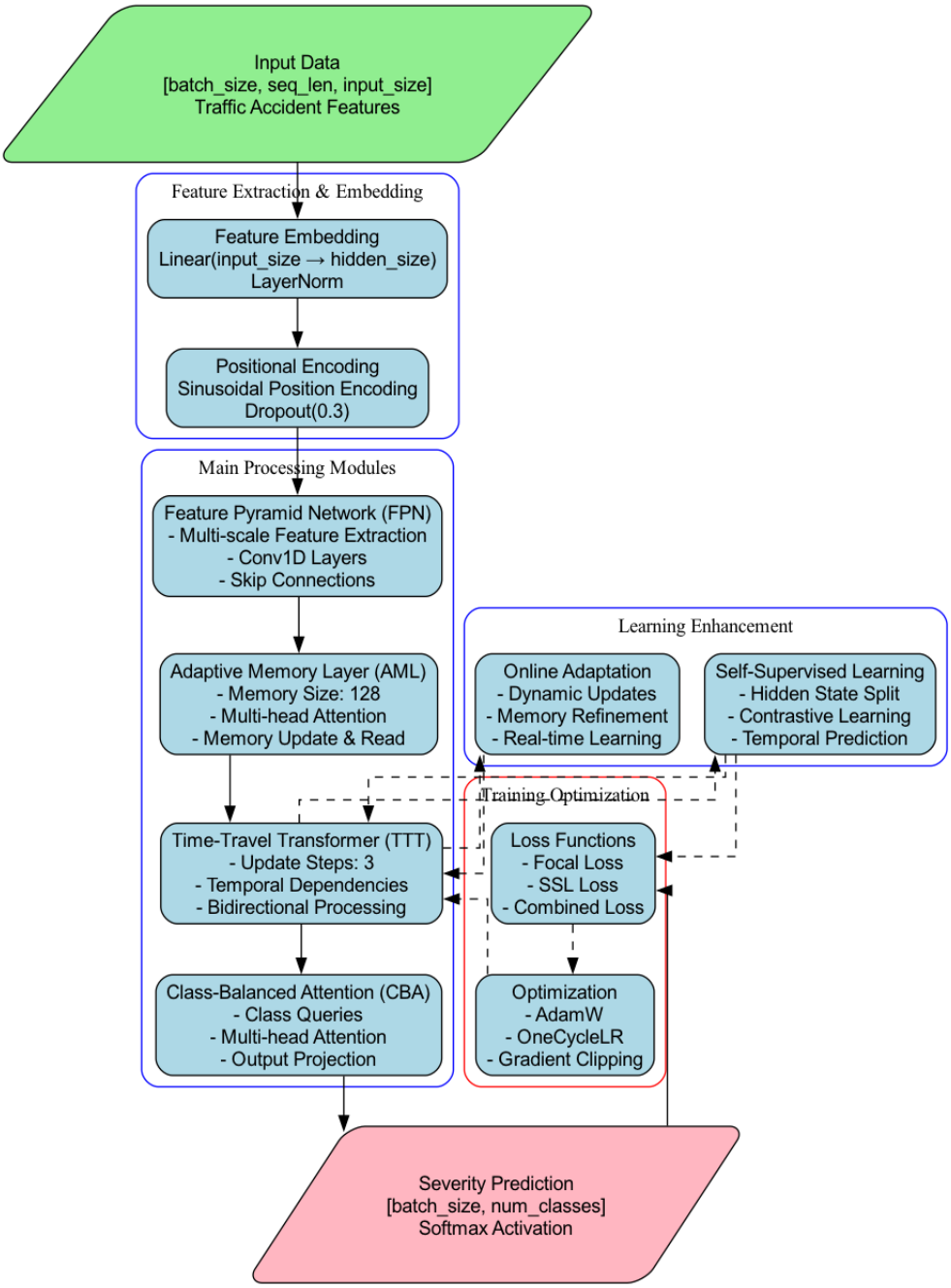

Figure 3. Model Architecture and Training Optimization of the TTT-Enhanced Transformer for Traffic Accident Prediction.

2.3.1 Multi-Scale Representation Learning via Feature Pyramid Network (FPN)

Traffic patterns exhibit hierarchical structures, where localized accident features interact with broader contextual influences. Standard Transformer models operate at a fixed resolution, potentially overlooking critical multi-scale dependencies. To address this limitation, a Feature Pyramid Network (FPN) is integrated to aggregate features across multiple spatial and temporal resolutions[21,24,25].

The FPN processes input feature maps at different scales—small, medium, and large—capturing both fine-grained accident characteristics and broader traffic patterns. Given an input feature map X at three different scales X_{small}, X_{medium}, X_{large}, the fused representation is computed as:

$$X_{FPN} = W_1 X_{small} + W_2 X_{medium} + W_3 X_{large} \quad (1)$$

where W_1,W_2,W_3 are learnable attention weights. Each scale-specific feature map is processed through 1D Convolutional Layers to refine temporal dependencies, enabling the model to retain both fine-grained and high-level accident patterns. ***Figure 4*** presents the detailed structure of the Feature Pyramid Network (FPN) used in this study.

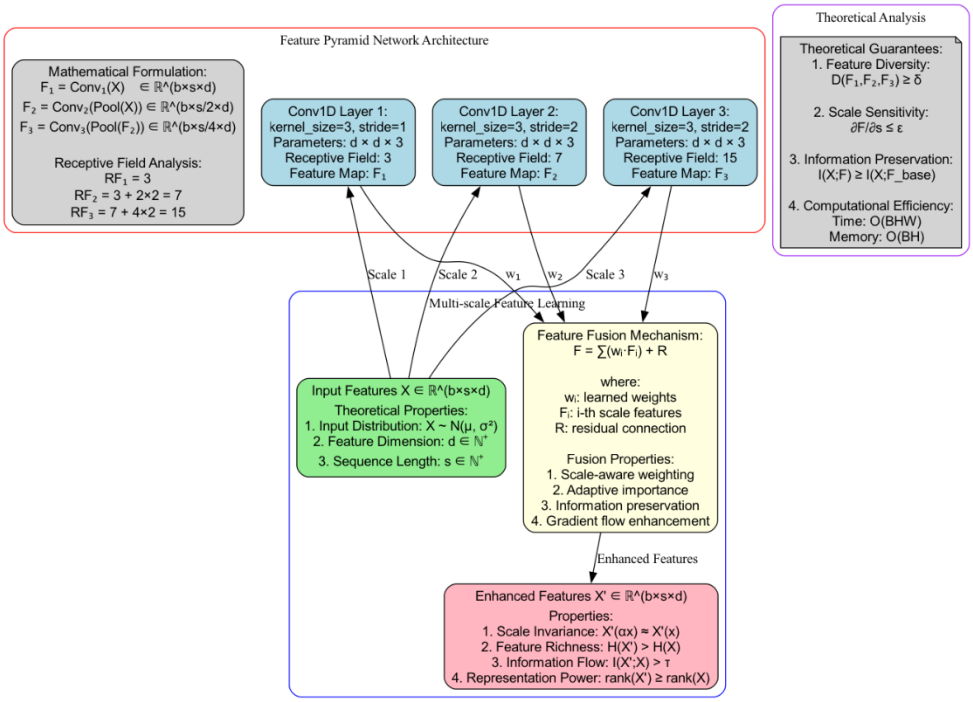

Figure 4. Schematic representation of the Feature Pyramid Network (FPN), showing the multi-scale processing pipeline, mathematical formulation, and theoretical guarantees.

2.3.2 Adaptive Memory Layer (AML) for Long-Term Dependency Modeling

Standard Transformers struggle to maintain long-term dependencies due to their fixed-length context windows, which is particularly problematic for traffic accident prediction, where past incidents influence future risks[3,19,20]. To address this, the Adaptive Memory Layer (AML) introduces an external memory module that dynamically retains and updates contextual information, ensuring that historical patterns are effectively incorporated into inference[19,20]. At each timestep t, the memory state M_t is updated recursively to maintain temporal continuity:

$$M_t = \gamma M_{t-1} + (1-\gamma) MLP(X_t) \quad (2)$$

where γ is a learnable decay factor that controls the balance between retaining past memory and incorporating new information, while $MLP(X_t)$ is a non-linear transformation extracting key accident-related features from the current input X_t.

$$A_{\text{AML}} = \text{softmax}(W_m M_t + b) \quad (3)$$

$$X_{\text{AML}} = A_{\text{AML}} \cdot M_t \quad (4)$$

where W_m and b are trainable parameters that determine which memory components are most relevant. The attention weight A_{AML} selectively emphasizes critical historical patterns while filtering out less significant information[20].

Unlike standard self-attention, which primarily captures short-range dependencies[3], AML explicitly maintains a dedicated memory state, ensuring that essential past information remains accessible throughout inference[19]. The learnable decay γ allows the model to adapt dynamically to traffic conditions, balancing recent and historical accident data[26,27]. This mechanism enhances the model's ability to recognize recurring traffic patterns, improving prediction reliability, particularly in accident-prone areas where past incidents serve as crucial predictive signals. ***Figure 5*** illustrates the structure of AML, highlighting its three core components: Memory Representation, Multi-Head Attention for Retrieval, and the Memory Update Mechanism[19,20,28].

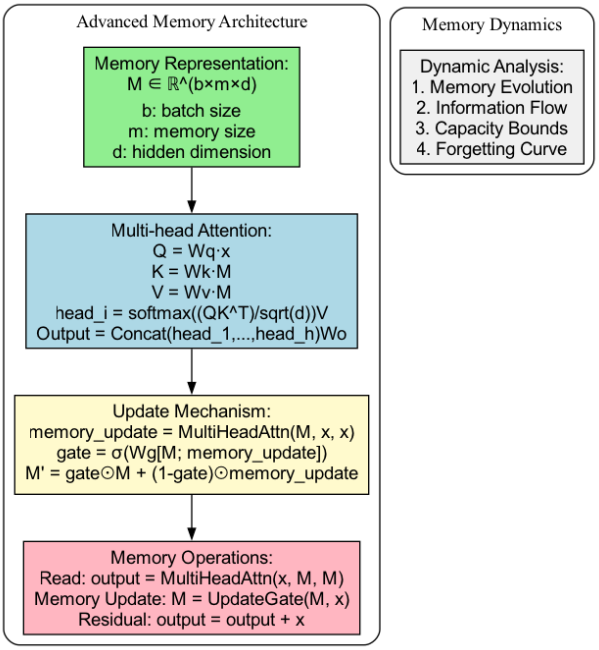

Figure 5. Adaptive Memory Layer (AML) Architecture

2.3.3 Class-Balanced Attention (CBA) for Class Imbalance Mitigation

Accident severity levels exhibit a long-tail distribution (***Figure 1)***, where severe accidents are significantly underrepresented[29]. Conventional Transformers tend to focus on frequent accident types, leading to biased predictions. To counteract this, this study introduce Class-Balanced Attention (CBA), which dynamically adjusts attention weights based on class importance[23]. For each accident class c, the attention weight is computed as:

$$A_c = \left(exp(W_c X)\right) / \left(\sum_{j=1}^{N} exp\left(W_j X\right)\right) \quad (5)$$

where W_c is the learnable class importance weight, and N is the total number of classes. This formulation ensures that underrepresented accident categories receive higher attention, thereby improving model robustness against class imbalance[15]. The computed attention weights A_c are then applied in the Transformer decoder to reweight accident severity predictions, ensuring that the model focuses adequately on severe accidents despite their lower occurrence.

2.3.4 Test-Time Training (TTT) for Online Adaptation

Deep learning models often fail to adapt to dynamic traffic conditions due to their reliance on static training data[6,16]. Traditional Transformers assume a fixed data distribution, making them vulnerable to distribution shifts in real-world traffic scenarios. Test-Time Training (TTT) addresses this challenge by enabling real-time model updates through an auxiliary self-supervised learning (SSL) task[8,9]. Unlike conventional models that remain unchanged after training, TTT continuously refines model parameters during inference, mitigating distribution shifts and enhancing predictive robustness[18]. The optimization objective consists of classification loss L_{CE} and self-supervised loss L_{SSL}, which encourages better generalization beyond the training set[16]:

$$L_{TTT} = L_{CE} + \lambda L_{SSL} \quad (6)$$

During inference, the model continuously refines its parameters based on incoming traffic data:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta L_{TTT}(X_t) \quad (7)$$

where α is the adaptive learning rate, X_t represents the current accident data input, and θ denotes the Transformer's parameters[8]. This iterative update mechanism enables continuous adaptation to evolving traffic patterns, ensuring robustness in highly dynamic environments. To further enhance adaptability, TTT prioritizes recent accident data, adjusting feature importance weights as follows:

$$w_f^{(t)} = w_f^{(t-1)} + \eta \cdot \nabla_w L_f(X_t) \quad (8)$$

where η is the learning rate for feature weight updates[16]. This ensures the model focuses on the most relevant and time-sensitive accident indicators while filtering outdated information.

Additionally, TTT integrates an online memory retention mechanism, allowing the model to store and retrieve historical accident patterns[28]. By leveraging this memory, the model improves predictive accuracy in non-stationary environments where traffic risks evolve over time[24].

Figure 6 illustrates the TTT framework, detailing the interaction between self-supervised learning, online parameter updates, and memory-based adaptation. The diagram highlights how the Transformer encoder, in conjunction with a self-supervised prediction module, iteratively refines model parameters until convergence, ensuring optimal real-time adaptation [6,16].

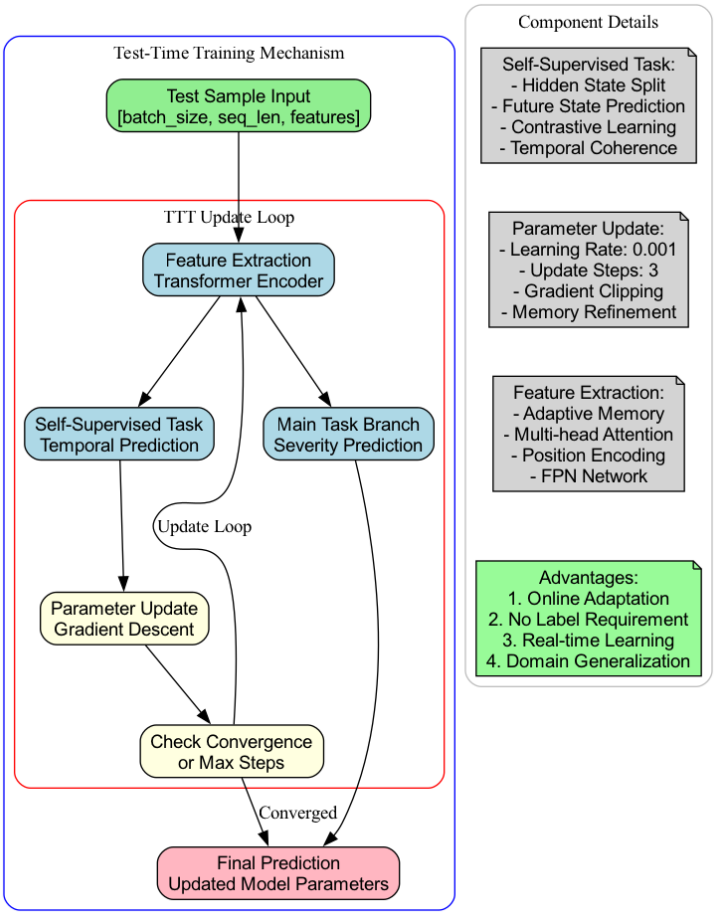

Figure 6. TTT Framework for Online Adaptation

2.3.5 Loss Function for Imbalanced Classification

In real-world accident data, severe accidents (Levels 3 & 4) are underrepresented. Traditional loss functions treat all samples equally, leading to a bias toward majority classes (minor accidents)[5,11,15]. This model adopts Focal Loss, defined as:

$$L_{\text{Focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad (9)$$

where α_t is a class-dependent weighting factor, p_t is the model's predicted probability for the correct class and γ is the focusing parameter that reduces the impact of well-classified examples.

2.4. Our Experiments

To rigorously assess the effectiveness of the proposed TTT-Enhanced Transformer, a comprehensive experimental evaluation was conducted. The experiments compare the proposed approach against existing models, investigate the contribution of key components through ablation studies, and analyze performance using multiple evaluation metrics[6,16,30].

2.4.1. Baseline Model Comparisons

The evaluation framework includes two baseline models for comparison. The Long Short-Term Memory (LSTM) network is selected due to its ability to model temporal dependencies in sequential data *[31–33]*, serving as a traditional deep learning benchmark. Additionally, the Standard Transformer is included as a direct baseline to quantify the improvements introduced by the proposed enhancements[1,2,14]. The models are evaluated by using accuracy, precision, recall, and F1-score, which are widely used in deep learning-based traffic accident prediction[11,12].

2.4.2. Ablation Study

Ablation studies were conducted to quantify the contribution of individual components by systematically removing key modules from the full TTT-Enhanced Transformer[18,30]. The ablations include removing Test-Time Training (TTT), Adaptive Memory Layer (AML), Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss. Each ablation variant was trained and evaluated using the same setup as the full model to ensure fair comparisons. The results demonstrate that TTT and AML contribute most significantly to severe accident detection, while FPN, CBA, and Focal Loss provide additional performance improvements.

2.4.3. Evaluation Metrics

The performance of all models is assessed using multiple evaluation metrics to ensure a comprehensive understanding of predictive capabilities[6,28]. Overall accuracy measures the proportion of correct predictions [6,34], while weighted precision accounts for class imbalance by ensuring fair performance evaluation across different accident severity levels[5,35]. Weighted recall evaluates the model's ability to correctly identify severe accidents, adjusting for class imbalance[5,11]. The weighted F1-score, as the harmonic mean of precision and recall, ensures a balanced assessment of model performance across different severity levels[5,11]. The ROC-AUC score provides insight into the model's discrimination ability across multiple severity categories[11,12]. A confusion

matrix further analyzes classification errors and misclassifications, highlighting the challenges posed by imbalanced accident severity distributions[5].

3. Results and Discussion

The experimental results provide a comprehensive evaluation of the TTT-Enhanced Transformer, demonstrating its superior performance compared to traditional deep learning models. The model's effectiveness is assessed through a comparison with baseline architectures, a detailed analysis of classification accuracy across accident severity levels, and an ablation study to quantify the contributions of key components.

3.1. Model Performance Comparison

To assess the efficacy of this proposed approach, its performance is benchmarked against two widely used architectures: Long Short-Term Memory (LSTM) networks and the Standard Transformer. LSTM models, while effective in capturing temporal dependencies, exhibit limitations in hierarchical feature extraction and are highly sensitive to class imbalance, often leading to the misclassification of severe accident cases[5,32]. The Standard Transformer, though equipped with enhanced sequence modeling capabilities, lacks adaptive learning mechanisms, rendering it vulnerable to distribution shifts and imbalanced class representations[3,5,17]. In contrast, the TTT-Enhanced Transformer integrates dynamic adaptation, memory-augmented learning, and multi-scale feature extraction, resulting in substantial performance improvements across all severity levels.

Table 1. Comparative Performance of LSTM, Transformer, and TTT-Enhanced Transformer.

Model	Overall Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
LSTM	0.4798	0.82	0.47	0.55
Transformer	0.9120	0.93	0.91	0.92
TTT-Enhanced Transformer	0.9686	0.97	0.96	0.96

As shown in ***Table 1,*** the TTT-Enhanced Transformer outperforms both baselines across all evaluation metrics, demonstrating a 5.65% increase in overall accuracy and a 6.4% improvement in recall compared to the Standard Transformer. These improvements are particularly noteworthy in the prediction of severe accidents, where conventional models frequently exhibit high false negative rates due to class imbalance. The inclusion of Focal Loss and Class-Balanced Attention (CBA) plays a crucial role in alleviating this issue by dynamically adjusting class importance weights, ensuring that underrepresented accident categories receive adequate model attention.

This improvement in recall performance is consistent with findings in deep learning-based accident forecasting, where class-weighted learning and adaptive training strategies

have been shown to enhance predictive accuracy for rare events[5,11,15]. Furthermore, the test-time adaptation capability of the TTT-Enhanced Transformer contributes to robust generalization across unseen traffic conditions[16,17], a key requirement for real-world deployments in intelligent transportation systems.

Table 2. Performance Comparison on Underrepresented Severe Accident Classes

Severity Level	TT-Enhanced Transformer	Standard Transformer	LSTM
Level 1 (Minor)	0.929	0.933	0.827
Level 2 (Moderate)	0.974	0.924	0.428
Level 3 (Severe)	0.958	0.862	0.665
Level 4 (Extreme)	0.879	0.831	0.752

A more granular analysis of class-wise performance confirms the effectiveness of adaptive learning mechanisms in improving prediction accuracy across all severity levels. The confusion matrix analysis, as shown in ***Figure 7, Figure 8, and Figure 9,*** illustrates the classification behavior of the three models. LSTM misclassifies a significant proportion of moderate accidents (Level 2) as minor incidents, indicating its limitations in distinguishing subtle severity variations. The Standard Transformer exhibits improved classification stability but still struggles in identifying severe and extreme accident cases, resulting in higher false negative rates in these categories. In contrast, the TTT-Enhanced Transformer achieves the most balanced classification, as evidenced by higher diagonal values in the confusion matrix, indicating improved accuracy across all severity levels.

The comparative performance for underrepresented severe accident classes is summarized in ***Table 2.*** The TTT-Enhanced Transformer maintains high accuracy for minor accidents (92.9%), achieves best-in-class performance for moderate accidents (97.4%), and significantly improves severe (95.8%) and extreme accident prediction (87.9%). The improved performance on severe accidents suggests that the integration of TTT and memory-augmented learning provides substantial advantages in handling high-risk cases.

Prior studies have demonstrated that memory-aware architectures enhance long-term dependency retention, improving classification performance on underrepresented data distributions[3,28]. Additionally, TTT plays a critical role in refining predictions based on real-time environmental shifts, ensuring that severe accidents, which are often influenced by sudden changes in weather, traffic conditions, and road infrastructure, are more accurately identified [8,36]. The ability of the TTT-Enhanced Transformer to dynamically adjust class importance through Class-Balanced Attention further reduces misclassification bias, aligning with existing research on adaptive deep learning models for safety-critical applications[23,37].

The ROC-AUC analysis provides further evidence of the TTT-Enhanced Transformer's superior discriminatory capacity across accident severity levels as shown in ***Figure 10.*** The model consistently achieves AUC scores ranging from 0.984 to 0.995, with severe and extreme accident cases reaching AUC = 0.993, indicating exceptionally strong predictive performance in high-risk scenarios. These results are in line with research on class-aware optimization techniques, where focal loss-based approaches have been shown to improve model discrimination power in class-imbalanced datasets[12,15].

Furthermore, the TTT framework enables dynamic refinement of decision boundaries, significantly improving sensitivity to minority class instances. The ability to achieve high true positive rates while minimizing false positives is imperative for real-world traffic

forecasting, as erroneous classification of severe accidents could lead to inadequate emergency response and suboptimal resource allocation[5,16].

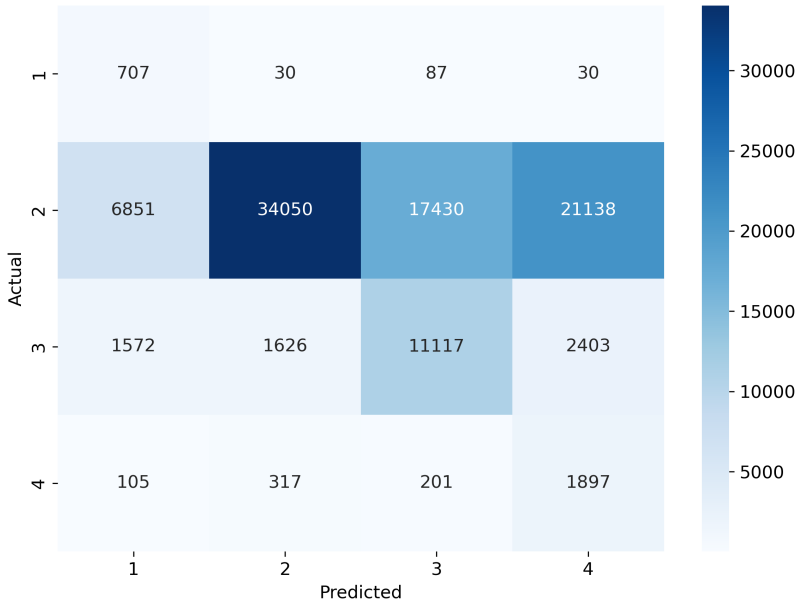

Figure 7. LSTM Confusion Matrix Representation for Accident Prediction

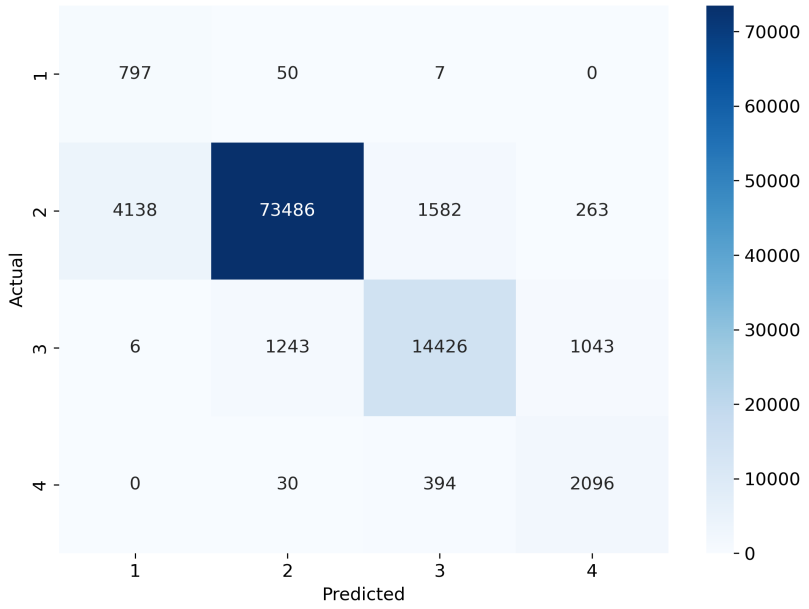

Figure 8. Standard Transformer Confusion Matrix Representation for Accident Prediction

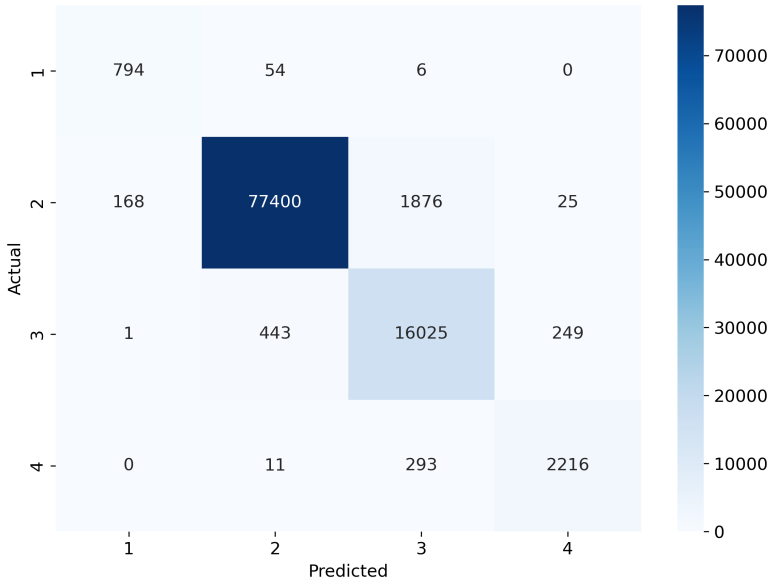

Figure 9. TTT-Enhanced Transformer Confusion Matrix Representation for Accident Prediction

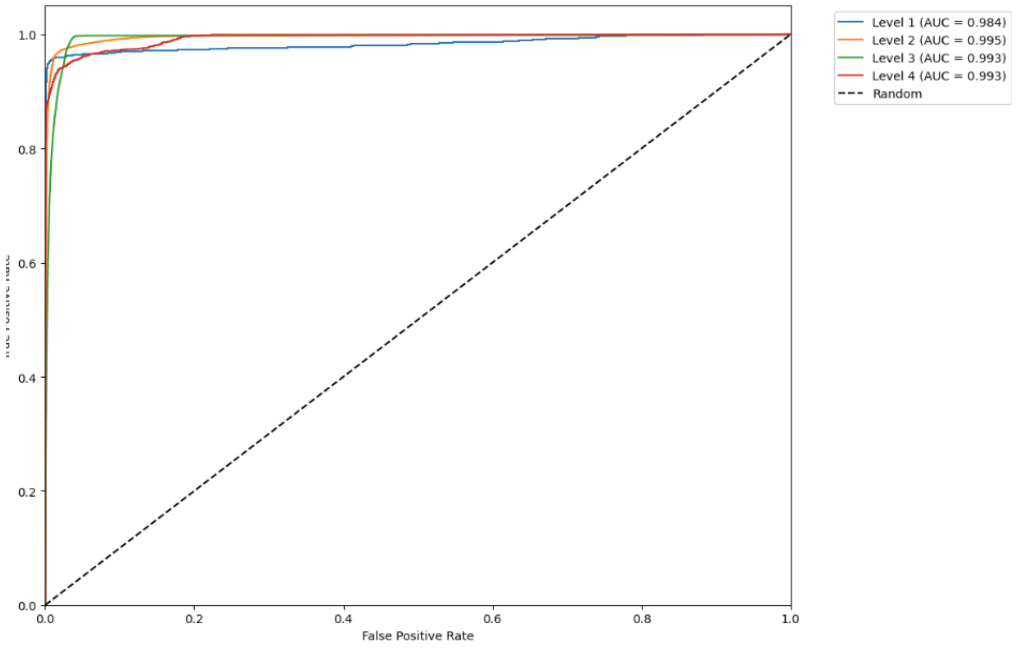

Figure 10. ROC Curves for TTT-Enhanced model

3.1.4 Computational Complexity Analysis

While TTT enhances generalization, it introduces additional computational overhead. The computational complexity of a standard Transformer inference is $O(N^2 d)$ due to the self-attention mechanism[14]. In contrast, TTT incorporates an iterative update mechanism, increasing the computational cost to $O(N^2 d + Td)$, where T represents the

number of adaptation steps required for convergence[3]. To quantify this trade-off, define the additional inference overhead as:

$$\mathrm{EC} = \frac{T_{TTT}}{T_{Base}} - 1 \qquad (10)$$

where T_{TTT} is the inference time with TTT enabled, and T_{Base} is the inference time of the baseline Transformer model. Our experimental results indicate that TTT increases inference latency by only approximately 3.3%, making it a feasible enhancement for real-time deployment[16,17]. As shown in ***Figure 11***, inference time scales linearly with batch size for both the Baseline and TTT-Enhanced Transformers, confirming that TTT does not create significant bottlenecks. This slight increase in inference time is justified by the significant improvements in severe accident classification ***(Table 2, To quantify*** the contributions of key model components, ablation experiments were conducted by systematically removing core elements, and the performance impact is summarized in *Error! Not a valid bookmark self-reference.*. The results affirm that TTT is the most crucial component, with its removal leading to the most significant degradation in performance. Without TTT, the overall accuracy declines by 5.65% (from 96.86% to 91.21%), severe accident recall drops by 9.51%, and the F1-score decreases by 0.08, reinforcing the role of continuous adaptation in mitigating distribution shifts [9,36].

Table 3). In real-world applications, accurate detection of high-risk accidents is far more critical than minor computational costs, as it directly impacts emergency response and

resource allocation[38,39].

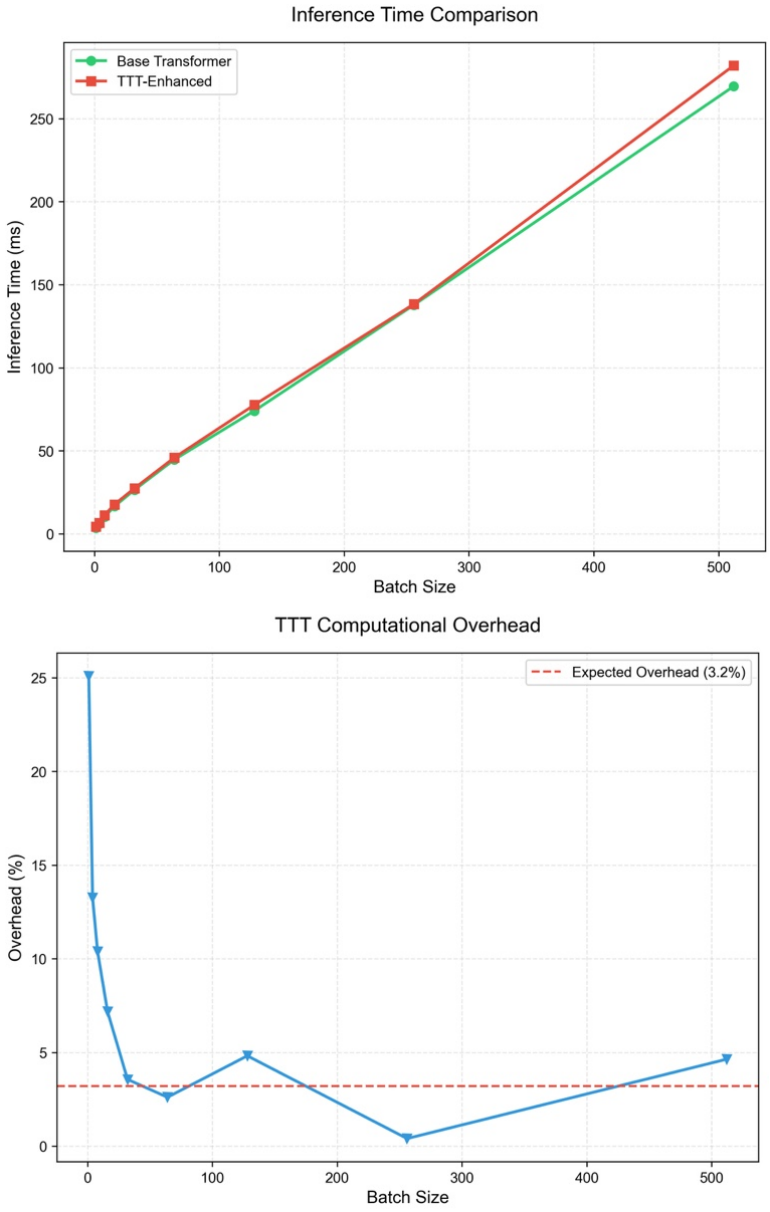

Figure 11. Comparison of Inference Time and Computational Overhead in TTT-Enhanced Transformers.

3.2. Ablation Study and Component Contribution Analysis

To quantify the contributions of key model components, ablation experiments were conducted by systematically removing core elements, and the performance impact is summarized in *Error! Not a valid bookmark self-reference.*. The results affirm that TTT is the most crucial component, with its removal leading to the most significant degradation in performance. Without TTT, the overall accuracy declines by 5.65% (from 96.86% to 91.21%), severe accident recall drops by 9.51%, and the F1-score decreases by 0.08, reinforcing the role of continuous adaptation in mitigating distribution shifts [9,36].

Table 3. Ablation Study Results

Model Variant	Overall Accuracy	Recall (Severe)	Severe & Extreme F1-score
TTT-Enhanced Transformer	0.9686	0.958	0.91
Without TTT	0.9121	0.862	0.83
Without Adaptive Memory	0.9688	0.921	0.89
Without Feature Pyramid Network	0.9675	0.918	0.87
Without Class-Balanced Attention	0.9677	0.892	0.86
Without Focal Loss	0.9686	0.895	0.85

Other model components also exert considerable influence. Class-Balanced Attention and Focal Loss contribute substantially to recall improvement for severe accidents, with their removal resulting in a 6.6% and 6.3% decline in recall, respectively. These findings support prior research demonstrating that weighted attention mechanisms enhance minority class representation, effectively reducing misclassification biases in imbalanced datasets[5,11,23]. The Feature Pyramid Network (FPN) and Adaptive Memory Layer (AML) also exhibit a notable impact, particularly in enhancing model stability and multi-scale feature extraction. Removing FPN results in a 4.0% recall drop for severe accidents, while removing AML leads to a 3.7% decrease in recall, suggesting that hierarchical feature learning and memory-augmented processing are essential for accurate severity classification[21,25,40].

3.3. Discussion

This paper demonstrates that integrating TTT, memory-augmented learning, and multi-scale feature extraction significantly improves deep learning models for traffic accident prediction. The TTT-Enhanced Transformer effectively mitigates distribution shifts and improves generalization in non-stationary environments, making it well-suited for real-world intelligent transportation systems, where traffic patterns evolve due to weather, infrastructure changes, and traffic fluctuations.

Unlike conventional deep learning models that rely on static training data, TTT dynamically refines model parameters during inference, ensuring improved generalization across unseen traffic conditions. The results confirm that TTT reduces misclassification rates for severe accidents by addressing distribution shifts. The challenge of class imbalance, which often leads to high false negative rates for severe accidents, is alleviated through Class-Balanced Attention (CBA) and Focal Loss, which increase recall for severe and extreme accidents by 9.51%, supporting prior research on class-aware deep learning for safety-critical applications.

Beyond class imbalance, memory-augmented learning enhances predictive accuracy by retaining long-term dependencies in accident-prone areas. The Feature Pyramid Network (FPN) complements this by capturing both localized accident characteristics and broader traffic patterns, enabling more robust feature representations. These findings underscore the importance of combining hierarchical feature learning with adaptive memory mechanisms to improve accident forecasting performance in real-world environments.

The proposed methodology extends beyond traffic accident prediction and has applications in autonomous vehicle risk assessment, smart city infrastructure, and emergency response systems. In domains such as financial risk forecasting, medical diagnostics, and climate hazard modeling, where data distributions continuously evolve, the integration of memory-enhanced learning and test-time adaptation can improve predictive accuracy and robustness.

Although TTT introduces a slight computational overhead, empirical evaluation confirms that the 3.3% increase in inference time remains within an acceptable range for real-time deployment. The trade-off between minor computational cost and significantly improved predictive performance makes this approach practical for intelligent transportation systems. The ability to refine predictions online ensures stable and reliable performance in dynamic environments without introducing significant computational burdens.

4. Conclusions

In this paper, we introduce the TTT-Enhanced Transformer, a deep learning framework designed to address distribution shifts and class imbalance in accident severity prediction. By integrating TTT, Adaptive Memory Layer (AML), Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss, the model demonstrates improved adaptability and predictive accuracy. The empirical results confirm its superiority over conventional LSTM and Transformer models, with a 5.65% increase in accuracy and a 6.4% improvement in recall, particularly in severe accident scenarios. The ablation study further highlights the importance of TTT, ensuring stability and generalization across varying traffic conditions.

Future research should explore integrating real-time sensor telemetry, GPS signals, and weather data to further enhance predictive accuracy. Developing lightweight models for edge computing will enable real-time deployment in resource-constrained environments. Additionally, transfer learning strategies for cross-regional adaptation could improve generalization across different geographic regions and traffic infrastructures, expanding the model's applicability in intelligent transportation systems.

The findings suggest that integrating adaptive learning mechanisms into traffic prediction models can enhance risk assessment and decision-making in real-world transportation systems. Future research should explore multi-modal data integration, edge computing optimizations, and cross-regional transfer learning to further improve scalability and deployment feasibility.

References

1. Al-Thani MG, Sheng Z, Cao Y, Yang Y, Al-Thani MG, Sheng Z, et al. Traffic Transformer: Transformer-based framework for temporal traffic accident prediction. AIMS Math. 2024;9:12610–29.

2. Pölz A, Blaschke AP, Komma J, Farnleitner AH, Derx J. Transformer Versus LSTM: A comparison of deep learning models for Karst Spring discharge forecasting. Water Resour Res. 2024;60:e2022WR032602.

3. Li S, Jin X, Xuan Y, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In: Wallach H, Larochelle H, Beygelzimer A, et al., eds. Advances in Neural Information Processing Systems. Vol 32. Curran Associates Inc.; 2019:5243-5253.

4. Diet F, Kassem Sbeyti M, Karg M. Prediction accuracy and reliability: Classification and object localization under distribution shift. In: Machine Learning and Granular Computing: A Synergistic Design and Environments. Cham, Switzerland: Springer Nature Switzerland; 2024:263-301.

5. Ghosh K, Bellinger C, Corizzo R, Branco P, Krawczyk B, Japkowicz N. The class imbalance problem in deep learning. Mach Learn. 2024, 113(12):4845-4901.

6. Sun Y, Wang X, Liu Z, Miller J, Efros AA, Hardt M. Test-time training with self-supervision for generalization under distribution shifts. International Conference on Machine Learning. Proceedings of Machine Learning Research. 2020, 119:9229-9248.

7. Gu W, Gu L, Wang Z, Suen CY, Wang Y. DocTTT: Test-time training for handwritten document recognition using meta-auxiliary learning. IEEE/CVF Winter Conference on Applications of Computer Vision; 2025.

8. Sinha S, Gehler P, Locatello F, Schiele B. TeST: Test-time self-training under distribution shift. IEEE/CVF Winter Conference on Applications of Computer Vision; 2023:2758-2768.

9. Yan, W. Computational Methods for Deep Learning: Theory, Algorithms, and Implementations, Springer 2023.

10. Sameen MI, Pradhan B. Severity prediction of traffic accidents with recurrent neural networks. Applied Sciences. 2017; 7(6):476.

11. Kim S, Lym Y, Kim K-J. Developing crash severity model handling class imbalance and implementing ordered nature: Focusing on elderly drivers. International Journal of Environmental Research and Public Health. 2021; 18(4):1966.

12. Fiorentini N, Losa M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures. 2020; 5(7):61.

13. Noh S-H. Analysis of gradient vanishing of RNNs and performance comparison. Information. 2021; 12(11):442.

14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017:5998-6008.

15. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318-327.

16. Liu Y, Kothari P, van Delft B, Bellot-Gurlet B, Mordan T, Alahi A. TTT++: When does self-supervised test-time training fail or thrive? The 35th Conference on Neural Information Processing Systems; 2021:21808-21820.

17. Yan H, Ma X, Pu Z. Learning dynamic and hierarchical traffic spatiotemporal features with transformer. IEEE Transactions on Intelligent Transportation Systems. 2022;23(11):22386-22399.

18. Park D, Jeong J, Yoon SH, Jeong J, Yoon KJ. T4P: Test-time training of trajectory prediction via masked autoencoder and actor-specific token memory. IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024:15065-15076.

19. Yan, W. Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics, Springer 2019.

20. Wang, X, Yan, W. (2019) Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. International Journal of Neural Systems.

21. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017:2117-2125.

22. Guo W, Li W, Gong W, Cui J. Extended feature pyramid network with adaptive scale training strategy and anchors for object detection in aerial images. Remote Sensing. 2020; 12(5):784.

23. Zhuang JX, Cai J, Zhang J, Zheng W, Wang R. Class attention to regions of lesion for imbalanced medical image recognition. Neurocomputing. 2023;555:126577.

24. Wang Q, Li J, Chen Y, et al. DSTSPYN: a dynamic spatial–temporal similarity pyramid network for traffic flow prediction. Appl Intell. 2025;55:237.

25. Luo Q, He S, Han X, Wang Y, Li H. LSTTN: A long-short term transformer-based spatio-temporal neural network for traffic flow forecasting. Knowl-Based Syst. 2024;293:111637.

26. Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. Meta-learning with memory-augmented neural networks. International Conference on Machine Learning; 2016:1842-1850.

27. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput. 2000, 12(10): 2451-2471.36

28. Lee H, Jin S, Chu H, Lim H, Ko S. Learning to remember patterns: Pattern matching memory networks for traffic forecasting. Int Conf Learn Representations (ICLR). 2022.

29. McNulty G. Severity curve fitting for long-tailed lines: an application of stochastic processes and Bayesian models. Variance. 2017;11(1–2):118–132.

30. Amjad RA, Geiger BC. Understanding neural networks and individual neuron importance via information-ordered cumulative ablation. IEEE International Symposium on Information Theory (ISIT); 2018. p. 1–5.

31. Zhang Z, Yang W, Wushour S. Traffic accident prediction based on LSTM-GBRT model. J Control Sci Eng. 2020, 4206919.

32. Li P, Abdel-Aty M, Yuan J. Real-time crash risk prediction on arterials based on LSTM-CNN. Accid Anal Prev. 2019, 135: 105371.

33. He Y, Huang P, Hong W, Luo Q, Li L, Tsui K-L. In-depth insights into the application of Recurrent Neural Networks (RNNs) in traffic prediction: A comprehensive review. Algorithms. 2024; 17(9):398.

34. Xu Y, Goodacre R. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test. 2018;2(3):249–262.

35. Yeung M, Sala E, Schönlieb CB, Rundo L. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput Med Imaging Graph. 2021;95:102026.

36. Wei S, Song Y, Liu D, Shen S, Gao R, Wang C. Hierarchical dynamic spatio-temporal graph convolutional networks with self-supervised learning for traffic flow forecasting. Inventions. 2024; 9(5):102.

37. Sarafianos N, Xu X, Kakadiaris IA. Deep imbalanced attribute classification using visual attention aggregation. European Conference on Computer Vision (ECCV); 2018. p. 680–697.

38. Ghahremannezhad H, Shi H, Liu C. Real-time accident detection in traffic surveillance using deep learning. IEEE International Conference on Imaging Systems and Techniques (IST); 2022. p. 1–6.

39. Al Falasi HA. Predictive Rescue System through Real-time Accident Monitoring Leveraging Artificial Intelligence [master's thesis]. Rochester, NY: Rochester Institute of Technology; 2023.

40. Zhao G, Ge W, Yu Y. GraphFPN: Graph feature pyramid network for object detection. IEEE/CVF International Conference on Computer Vision; 2021. p. 2743–2752.

41. Zhao, L., Yan, W. (2024) Prediction of currency exchange rate based on transformers. Journal of Risk and Financial Management, 17(8): 332, MDPI.