computers

MDPI

Article

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

Duo Peng and Weiqi Yan *

Department of Computer and Information Science, Auckland University of Technology, Auckland 1010, New Zealand; pnt7686@autuni.ac.nz
* Correspondence: 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. While distribution shift is a common challenge in machine learning, Transformer-based models—despite their ability to capture long-term dependencies—often lack mechanisms for dynamic adaptation during inferencing. In this paper, we propose a TTT-Enhanced Transformer that incorporates Test-Time Training (TTT), enabling the model to refine its parameters during inferencing through a self-supervised auxiliary task. To further boost performance, an Adaptive Memory Layer (AML), a Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss are integrated to address multi-scale, long-term, and imbalance-related challenges. Our experimental results show that our model achieved an overall accuracy of 96.86% and a severe accident recall of 95.8%, outperforming the strongest Transformer baseline by 5.65% in accuracy and 9.6% in recall. The results of our confusion matrix and ROC analyses confirm our model's superior classification balance and discriminatory power. These findings highlight the potential of our approach in enhancing real-time adaptability and robustness under shifting data distributions and class imbalances in intelligent transportation systems.

Keywords: test-time training; traffic accident prediction; transformer network; self-supervised learning; adaptive memory; class-balanced learning

check for updates

Academic Editors: Behrouz Zolfaghari, Amin Ramezani, Mohsen Hadian and Firooz Saghezchi

Received: 17 February 2025
Revised: 6 May 2025
Accepted: 6 May 2025
Published: 10 May 2025

Citation: Peng, D.; Yan, W. Test-Time Training with Adaptive Memory for Traffic Accident Severity Prediction. *Computers* **2025**, *14*, 186. https://doi.org/10.3390/computers14050186

Copyright: © 2025 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

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 USD 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 inferencing 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 show 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 is as high as 93.2:1, highlighting the extreme disparity between frequent minor incidents and rare but critical severe accidents. To further characterize the Severity (levels 1–4) distribution and rule out any inadvertent train–test shifts, we computed skewness, kurtosis, and Jensen–Shannon divergence on the labels in both splits, as summarized in Table 1. These statistics confirm that the training and testing severity labels share virtually identical higher-order moments (JS divergence ≈ 0.00085), indicating that the imbalance challenge stems from the inherent long-tail distribution rather than any split mismatch. 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.

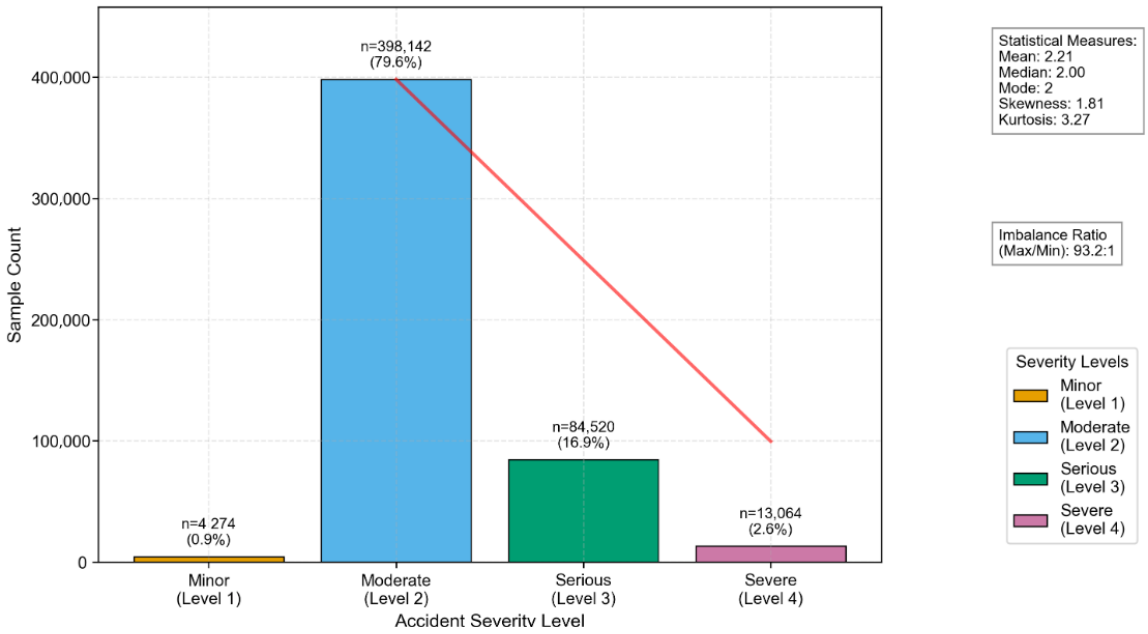

Figure 1. Distribution of accident severity levels in our dataset.

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 in 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].

Table 1. Severity label distribution statistics.

Feature	Skew Train	Skew Test	Kurt Train	Kurt Test	JS Divergence
Severity (1–4)	1.812	1.808	6.271	6.26	0.000085

Recent advancements in self-supervised learning and adaptive training provide promising solutions to addressing distribution shifts and class imbalance [5,6,16]. While Test-Time Training (TTT) has demonstrated significant potential to dynamically adapt models during inferencing, the existing research has mainly focused on structured environments and standard classification tasks, with limited exploration of real-time traffic accident prediction [16–18]. To date, there is no unified framework explicitly integrating TTT with advanced memory mechanisms, multi-scale feature extraction, and class-imbalance-aware strategies tailored specifically to traffic accident severity prediction. Such integration is crucial, as existing models continue to struggle to accurately differentiate between frequent minor accidents and rare, severe cases, resulting in biased predictions and insufficient recall on critical accident scenarios [5,11,15]. Therefore, there is still an open research gap regarding a comprehensive approach that simultaneously addresses these challenges.

It should be noted that this study focuses on predicting the severity level of reported accidents rather than forecasting accident occurrences. This formulation is consistent with prior work on probabilistic risk modeling and supports more actionable safety interventions based on the severity of traffic incidents.

To address these challenges, in this paper, we introduce a unified Transformer-based framework that incorporates Test-Time Training (TTT) as the core adaptation mechanism for traffic accident severity prediction. TTT dynamically refines model parameters during inferencing, improving real-time adaptability and robustness to distribution shifts. Additionally, we integrate an Adaptive Memory Layer (AML) to retain long-term dependencies [19,20] and a Feature Pyramid Network (FPN) to enhance multi-scale feature extraction [21,22]. To specifically mitigate class imbalance, the proposed model combines Class-Balanced Attention (CBA), class-weighted cross-entropy, Focal Loss, and SMOTE-based oversampling [15,23]. Unlike prior approaches, in which these strategies are treated independently, our cohesive architecture simultaneously addresses distribution shifts and class imbalance at t data level, loss level, and model level. Our experimental results and ablation studies confirm that our model achieved significant improvements in robustness, inference efficiency, and recall for severe and rare accident scenarios, highlighting its suitability for real-time intelligent transportation systems.

2. Materials and Methods

We designed our study according to a structured methodology inspired by established workflows used in traffic accident prediction and machine learning research [1,24], aiming to develop a scalable and adaptive prediction framework. Specifically, our approach integrates deep learning models within a systematic workflow to enhance predictive performance. The subsequent subsections detail our research design, data-processing pipeline, model architecture, training procedures, evaluation metrics, and ablation studies conducted to assess the contributions of the different components of the model.

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 (accessed on 1 January 2025)), 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; these records included information on 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 pre-processing workflow, covering data cleaning, feature engineering, standardization, and class balancing, is illustrated in Figure 2.

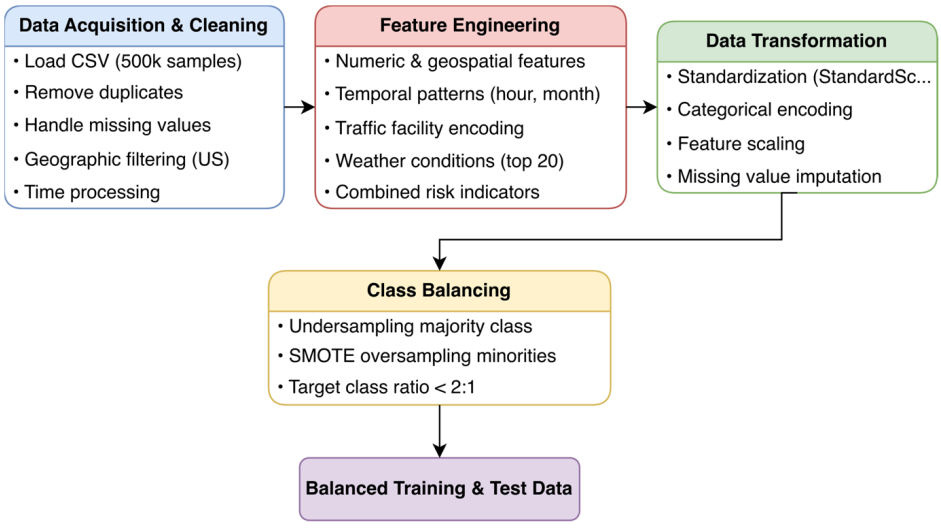

Figure 2. Overview of the data pre-processing pipeline, detailing data cleaning, filtering, feature engineering, and balancing steps.

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.

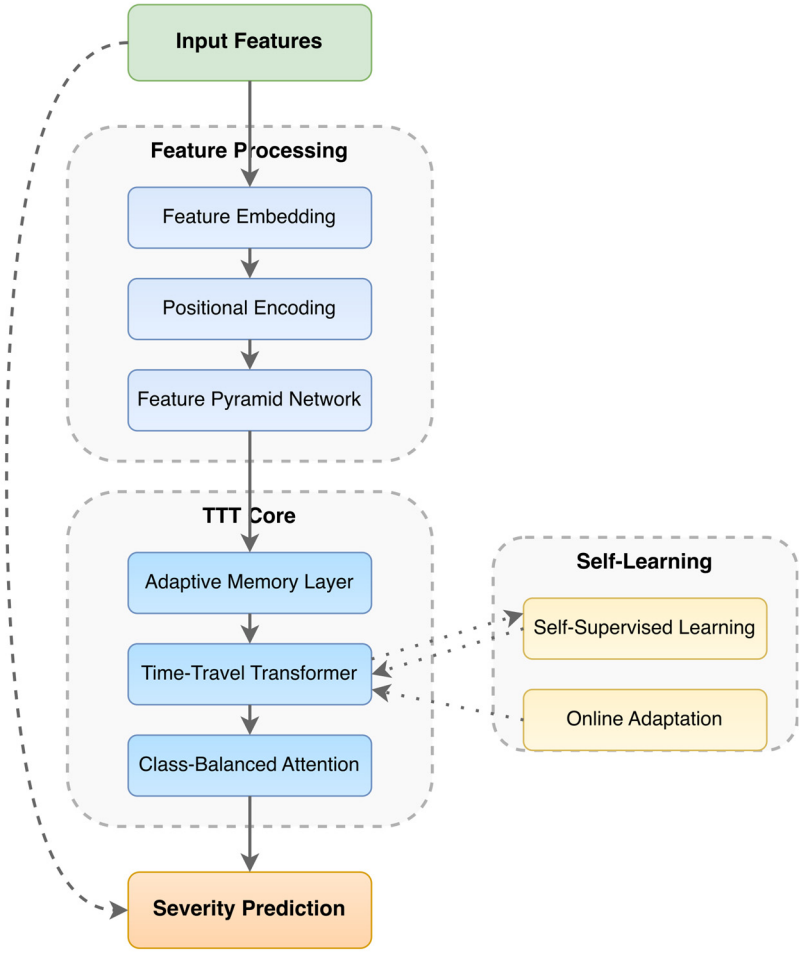

Figure 3. Model architecture of and training optimization process for the TTT-Enhanced Transformer for traffic accident prediction.

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

Traffic patterns exhibit hierarchical structures, wherein 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 [21,25,26].

An 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}, and 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, and 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.

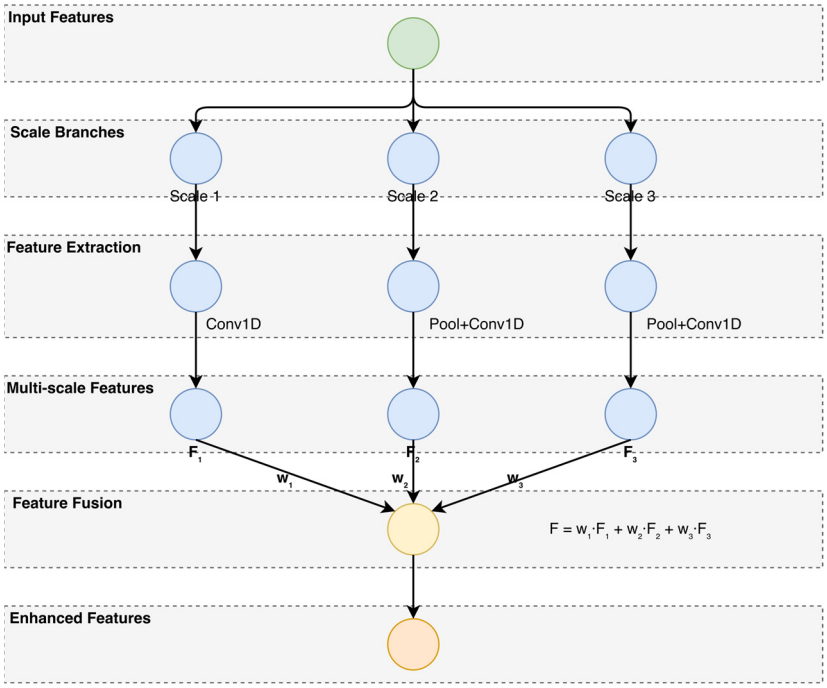

Figure 4. Schematic representation of the Feature Pyramid Network (FPN), showing the multi-scale processing pipeline, mathematical formulation, and theoretical guarantees.

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, constituting a particularly problematic shortcoming for traffic accident prediction, where past incidents influence future risks [3,19,20]. To address this, an Adaptive Memory Layer (AML) introduces an external memory module that dynamically retains and updates contextual information, ensuring that historical patterns are effectively incorporated into inferencing [19,20]. At each timestep t, the memory state M_t is updated recursively to maintain temporal continuity:

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

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.

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

$$X_{\text{AML}} = A_{\text{AML}} \cdot M_t \tag{4}$$

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], the proposed Adaptive Memory Layer (AML) is designed to maintain a dedicated memory state so that essential historical information remains accessible during inferencing [19]. The learnable decay γ allows the model to adapt dynamically to traffic conditions, balancing recent and historical accident data [27,28]. We expect this mechanism will help the model recognize recurring traffic patterns and potentially improve prediction reliability, especially in accident-prone areas, where past incidents can offer useful predictive cues. Figure 5 de-

picts the AML architecture, including its three key components—Memory Representation, Multi-Head Attention for Retrieval, and the Memory Update Mechanism [19,20,29].

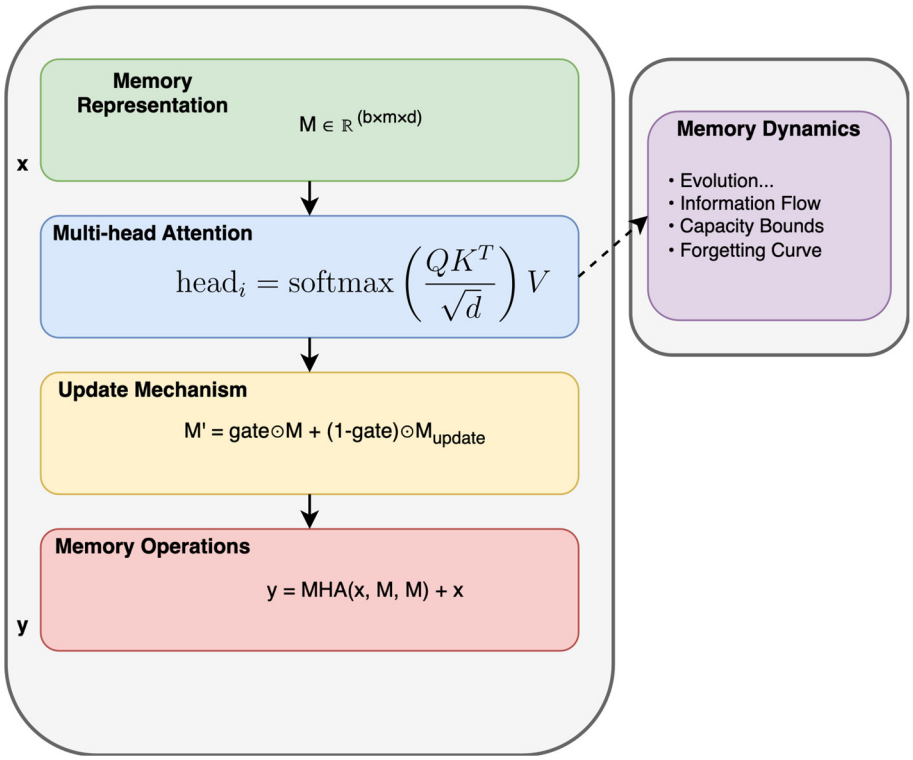

Figure 5. Adaptive Memory Layer (AML) architecture.

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 [30]. Conventional Transformers tend to focus on frequent accident types, leading to biased predictions. To counteract this, we 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

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

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 more attention, thereby improving the model's 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 rates of 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, which remain unchanged after training, TTT continuously refines model parameters during inferencing, mitigating distribution shifts and enhancing predictive robustness [18]. The

optimization objective consists of classification loss L_{CE} and self-supervised loss L_{SSL}, encouraging better generalization beyond the training set [16]:

$$L_{TTT} = L_{CE} + \lambda L_{SSL} \tag{6}$$

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

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

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:

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

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 [29]. By leveraging this memory, the model improves predictive accuracy in non-stationary environments where traffic risks evolve over time [25].

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]. For reproducibility, the pseudocode of the proposed method is provided in Appendix A (Algorithm A1).

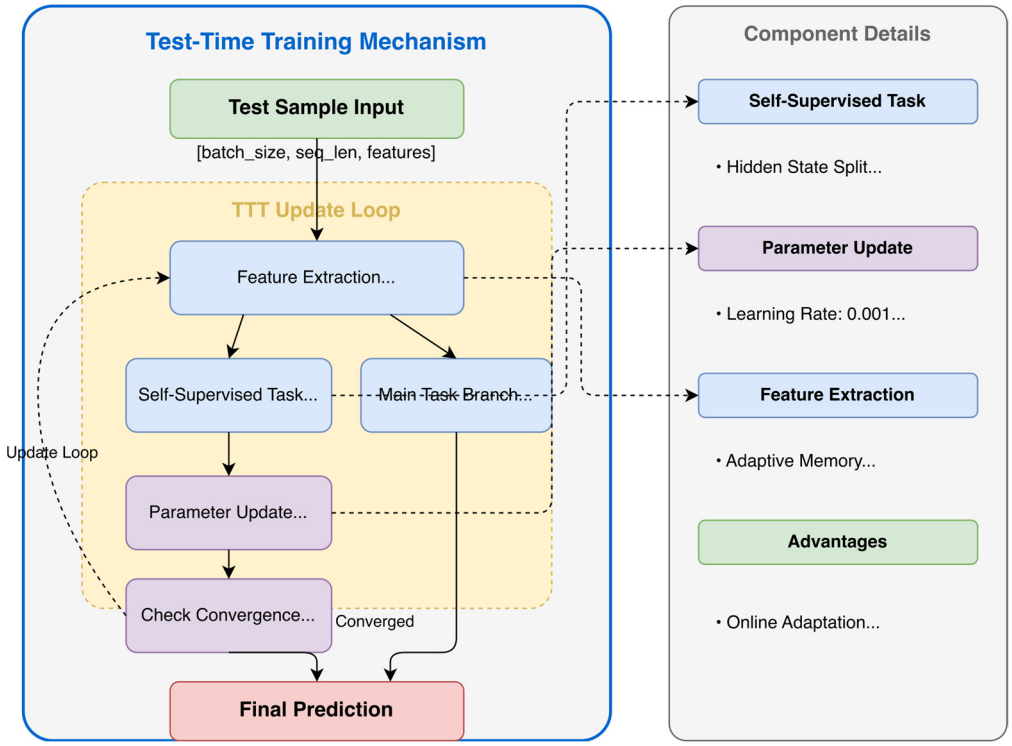

Figure 6. TTT framework for online adaptation.

2.3.5. Loss Function for Imbalanced Classification

In real-world accident data, severe accidents (Levels 3 and 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 follows:

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

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, which 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. In these experiments, we compared the proposed approach against existing models, investigated the contribution of key components through ablation studies, and analyzed performance using multiple evaluation metrics [6,16,31].

2.4.1. Baseline Model Comparisons

The evaluation framework includes multiple baselines for comparison. A Long Short-Term Memory (LSTM) network was selected due to its strong ability to model temporal dependencies in sequential data [32–34], serving as a traditional deep learning benchmark. Additionally, we implemented a Standard Transformer baseline to quantify the impact of our proposed enhancements [1,2,14]. To ensure this baseline was sufficiently competitive, we applied a comprehensive three-step class imbalance mitigation pipeline comprising (i) random undersampling of the majority class (Severity 2), (ii) SMOTE-based oversampling of minority classes (Severities 1, 3 and 4), and (iii) cost-sensitive training using class-weighted cross-entropy loss with label smoothing. This setup ensures that the baseline addresses class imbalance at the data level, loss level, and training strategy level, providing a strong foundation for fair comparison with our TTT-Enhanced framework.

To further isolate the contributions of each individual mitigation technique, we introduced two additional variants: Transformer + Only SMOTE, which applies synthetic oversampling without undersampling or weighted loss [5], and Transformer + Only Class-Weighted CE, which implements cost-sensitive learning without data-level balancing [35]. These ablation models help quantify the incremental benefit of each component.

In addition, we compare the results against those obtained using TabM, a recent state-of-the-art tabular deep learning architecture designed for structured data [36]. All models are evaluated using standard metrics such as overall accuracy, weighted precision, weighted recall, and weighted F1-score [11,12].

2.4.2. Ablation Studies

Ablation studies were conducted to quantify the contribution of individual components and training strategies within the proposed TTT-Enhanced Transformer [18,31]. In addition to systematically removing key modules—such as Test-Time Training (TTT), Adaptive Memory Layer (AML), Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss—we included two additional configurations to assess the isolated impact of CBA under data imbalance conditions, specifically without applying any sampling methods. All ablation variants were trained and evaluated under consistent experimental settings to enable fair comparison.

2.4.3. Evaluation Metrics

The performance of all models is assessed using multiple evaluation metrics to ensure a comprehensive understanding of predictive capabilities [6,29]. Overall accuracy measures the proportion of correct predictions [6,37], while weighted precision accounts for class imbalance by ensuring performance is evaluated fairly across different accident severity levels [5]. Weighted recall evaluates a 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 there is a balanced assessment of model performance across different severity levels [5,11]. The ROC-AUC score provides insight into a 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

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 comprehensively evaluate the effectiveness of the proposed TTT-Enhanced Transformer, we compared its performance against several baseline models and the recent state-of-the-art (SOTA) TabM architecture.

As shown in Table 2, the TTT-Enhanced Transformer consistently outperforms all baseline and comparison models across every evaluation metric. Compared to the LSTM baseline, which suffers from limited temporal modeling capabilities and heightened sensitivity to class imbalance (overall accuracy: 0.4798, F1-score: 0.55), the TTT-Enhanced Transformer achieved substantial improvements—particularly in recall (0.96 vs. 0.47)—demonstrating its effectiveness in identifying rare and severe accident cases [5,33].

Table 2. Comparison of the performances of deep learning baselines, imbalance-aware variants, a SOTA tabular model (TabM), and the proposed TTT-Enhanced Transformer.

Model	Overall Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
LSTM	0.4798	0.82	0.47	0.55
Standard Transformer (Undersampling + SMOTE + Weighted CE)	0.9120	0.93	0.91	0.92
Transformer + Only SMOTE	0.905	0.91	0.93	0.915
Transformer + Only Class-Weighted CE	0.920	0.94	0.89	0.915
TabM (SOTA)	0.925	0.94	0.92	0.93
TTT-Enhanced Transformer	0.9686	0.97	0.96	0.96

The Transformer (Undersampling + SMOTE + Weighted CE) performed markedly better than LSTM due to its superior sequence modeling and multi-level class imbalance handling (overall accuracy: 0.9120, F1-score: 0.92), yet it lacks adaptation during inferencing time, making it vulnerable to real-world distributional shifts [3,5,17]. The ablated variants—Transformer + Only SMOTE and Transformer + Only Class-Weighted CE—exhibited partial improvements but fell short of the balanced performance achieved by the full imbalance-aware baseline, which yielded the highest F1-score (0.920) among all Transformer-only configurations [5].

TabM, a recent state-of-the-art tabular deep learning model, performed competitively (F1-score: 0.93) [36]. However, its lack of temporal modeling and test-time adaptation limits its generalization capability under dynamic traffic conditions [16,17].

In contrast, the proposed TTT-Enhanced Transformer synergistically combines Focal Loss [15] and class-balanced attention [23] to address class imbalance, while employing memory-augmented encoding [19,20] and test-time training [6–8,16] to dynamically refine predictions during inferencing. This holistic design—spanning data-, loss-, and model-level strategies—enabled the model to achieve the highest overall accuracy, recall, and F1-score, particularly excelling in high-risk, low-frequency accident scenarios.

These findings strongly suggest that robust and generalizable performance in traffic accident severity prediction is best achieved through the integrated use of imbalance-aware learning techniques and real-time test-time adaptation [5,11,15–17].

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

A comparison of the performance for underrepresented severe accident classes is summarized in Table 3. The TTT-Enhanced Transformer maintained high accuracy for minor accidents (92.9%), achieved best-in-class performance for moderate accidents (97.4%), and yielded significantly improved severe (95.8%) and extreme accident prediction (87.9%). The improved performance regarding severe accidents suggests that the integration of TTT and memory-augmented learning provides substantial advantages in handling high-risk cases.

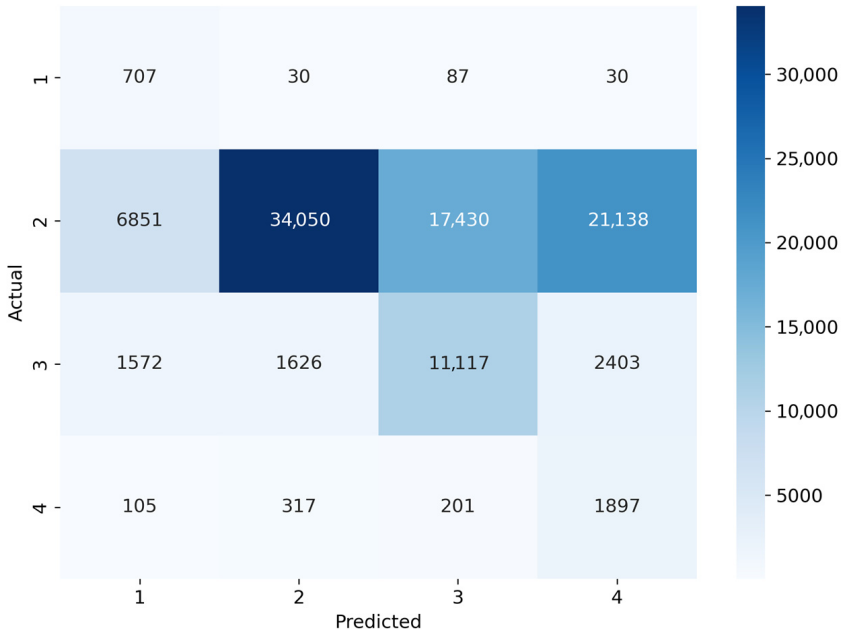

Figure 7. LSTM confusion matrix representation for accident prediction.

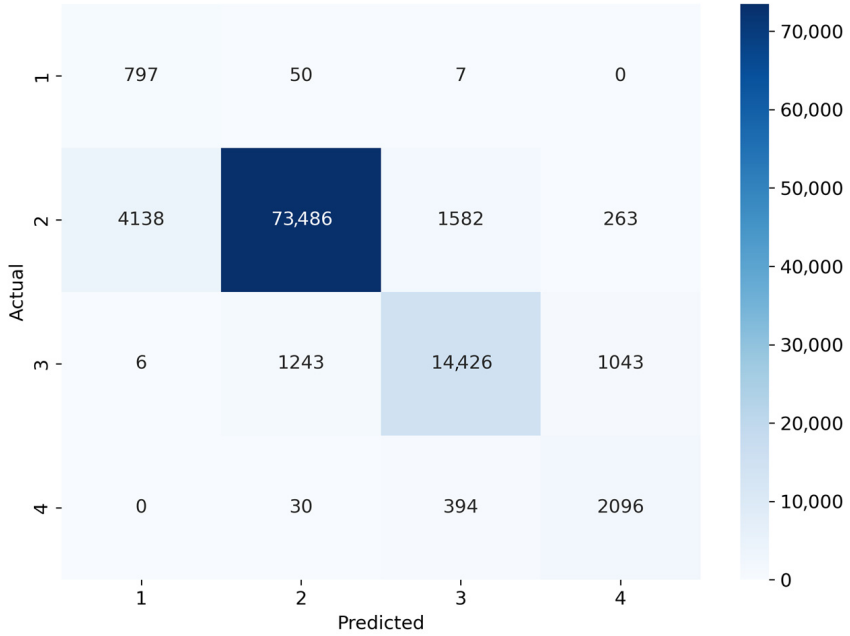

Figure 8. Standard Transformer confusion matrix representation for accident prediction.

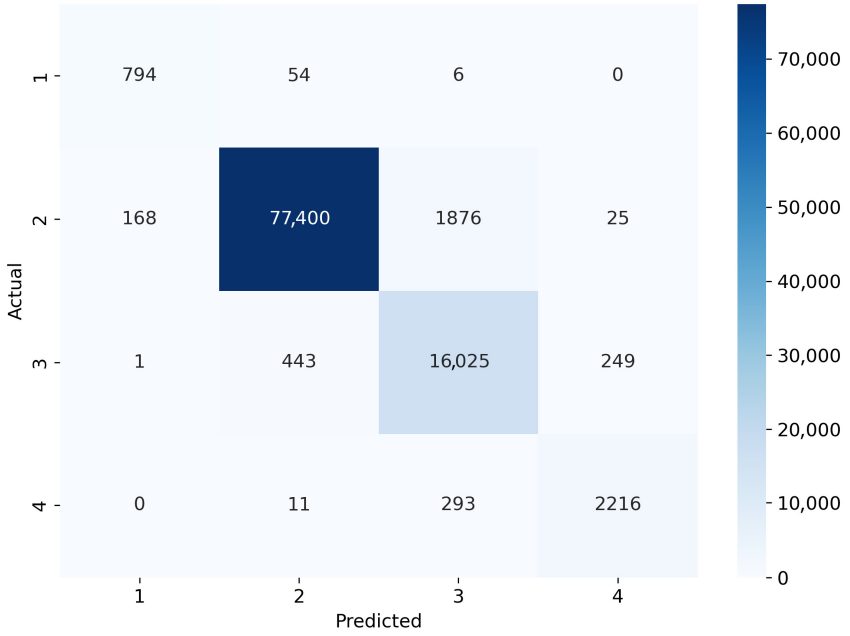

Figure 9. TTT-Enhanced Transformer confusion matrix representation for accident prediction.

Table 3. Performance comparison for underrepresented severe accident classes.

Severity Level	TTT-Enhanced Transformer	Standard Transformer (Undersampling + SMOTE + Weighted CE)	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

Prior studies have demonstrated that memory-aware architectures enhance long-term dependency retention, improving classification performance for underrepresented data distributions [3,29]. 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 [8,38]. 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 [23,39].

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 achieved 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].

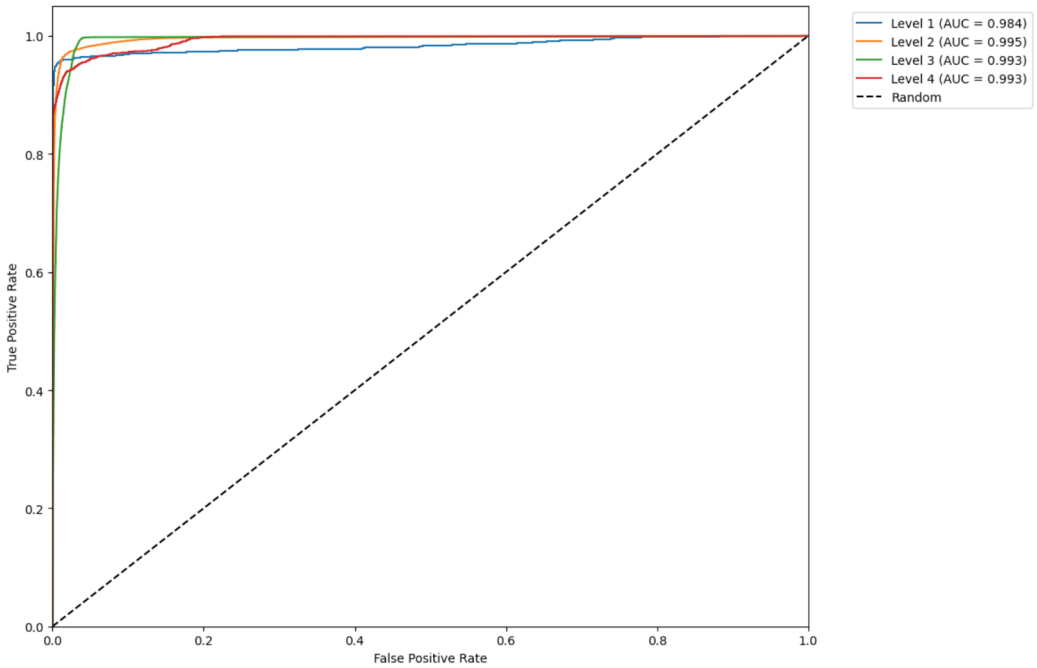

Figure 10. ROC curves for TTT-Enhanced model.

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 responses and suboptimal resource allocation [5,16].

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, the additional inference overhead must be defined as follows:

$$\mathrm{EC} = \frac{T_{TTT}}{T_{Base}} - 1 \tag{10}$$

where T_{TTT} is the inference time with TTT enabled, and T_{Base} is the inference time of the baseline Transformer model.

To provide a realistic benchmark, we measured inference latency on a consumer-grade Apple MacBook Pro (M1 Pro, 16 GB RAM)—a device whose computation budget is closer to typical edge-deployment or field-laptop scenarios than to high-end servers. As shown in Figure 11 and summarized in Table 4, the TTT mechanism introduces an average overhead of 3.3%, with latency remaining consistently low for typical deployment batch sizes (128–512), ranging between 2.1% and 3.43%. This confirms that TTT is both lightweight and scalable under realistic operational conditions.

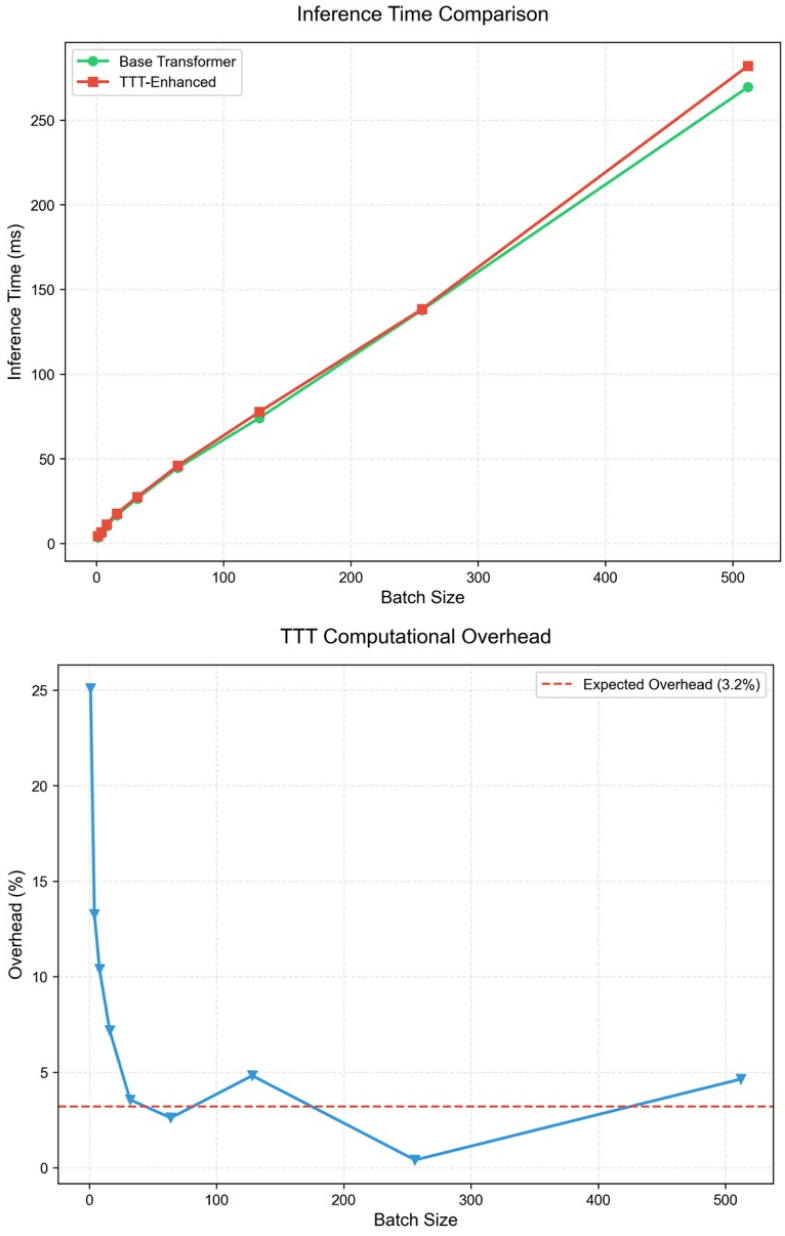

Figure 11. Comparison of inference time and computational overhead for TTT-Enhanced Transformers.

Importantly, this minor computational overhead is far outweighed by the substantial improvements in classification performance for rare and severe accident categories (Tables 3 and 5). In real-world traffic-monitoring applications, the ability to accurately identify high-risk incidents is far more critical than minor increases in inference time. These

findings demonstrate the practical feasibility of integrating TTT into intelligent transportation systems to achieve robust and real-time accident severity prediction [16,17,40,41].

Table 4. Inference latency and overhead comparison.

Batch Size	Baseline Inference Time (ms)	TTT-Enhanced (ms)	Overhead (%)
8	18.1	22.5	+24.3
32	38.7	41.1	+6.2
64	58.9	61.0	+3.5
128	61.3	63.4	+3.43
256	123.4	126.0	+2.1
512	268.1	276.4	+3.1

Table 5. 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
Without Sampling, Without CBA	0.863	0.512	0.573
Without Sampling, With CBA	0.889	0.674	0.707

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 impacts are summarized in Table 5. 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 [9,38].

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 6.6% and 6.3% declines 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 have a notable impact, particularly in enhancing model stability and multi-scale feature extraction. Removing the FPN resulted in a 4.0% recall drop for severe accidents, while removing AML led to a 3.7% decrease in recall, suggesting that hierarchical feature learning and memory-augmented processing are essential for accurate severity classification [21,26,42].

To further isolate the contribution of Class-Balanced Attention (CBA), we introduced two additional ablation variants in a no-sampling scenario. As shown in Table 5, adding CBA alone improved severe recall from 0.512 to 0.674 and the severe F1-score from 0.573 to 0.707, even without incorporating any data-level balancing strategies. These results demonstrate that CBA achieves significant gains in minority class recognition independently, highlighting its complementary role with respect to sampling techniques in mitigating class imbalance.

4. 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, which rely on static training data, TTT dynamically refines model parameters during inferencing, 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 costs 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.

5. Limitations and Ethical Considerations

Despite the strong empirical performance of the TTT-Enhanced Transformer, several technical limitations and ethical concerns must be acknowledged to ensure its responsible deployment.

From a technical standpoint, the model exhibits sensitivity to hyperparameter choices, particularly those governing test-time training—such as adaptation step size, learning rate, and the number of refinement iterations. Suboptimal settings may lead to unstable convergence or diminished generalization, particularly when there are severe distribution shifts. Moreover, while memory-augmented mechanisms improve long-term context retention, they can also amplify noise or outliers, especially in sparsely represented or mislabeled samples. Future work should explore more robust meta-learning strategies, uncertainty-aware adaptation, and automated hyperparameter optimization to enhance reliability under real-world conditions.

Ethically, traffic prediction models trained on historical and geospatial data should carefully manage risks related to privacy, bias, and accountability. Inadequate representation of certain regions or demographic groups may introduce algorithmic bias, resulting in

inequitable predictions or emergency resource allocation. Furthermore, the use of location-sensitive data may raise privacy concerns if such data are not appropriately anonymized. To mitigate these risks, predictive frameworks should incorporate bias audits, fairness metrics, and privacy-preserving techniques. It is important to note that such models should serve as decision-support tools rather than autonomous systems—particularly in critical applications such as emergency dispatch, traffic control, or policy planning.

In summary, while the TTT-Enhanced Transformer demonstrates strong performance, its deployment must be accompanied by comprehensive safeguards to ensure fairness, robustness, and ethical integrity across diverse deployment contexts.

6. Conclusions

In this paper, we introduced the TTT-Enhanced Transformer, a novel deep learning framework tailored to tackling distribution shifts and class imbalance in traffic accident severity prediction. By synergistically integrating Test-Time Training (TTT) for adaptive inference, an Adaptive Memory Layer (AML) for enhanced sequential modeling, a Feature Pyramid Network (FPN) for multi-scale feature extraction, and specialized class imbalance techniques (Class-Balanced Attention (CBA) and Focal Loss), the proposed model significantly improved prediction accuracy and robustness. Empirical evaluations demonstrated its clear superiority over conventional LSTM and standard Transformer models, achieving a 5.65% improvement in overall accuracy and a notable 9.6% enhancement in recall for severe accident categories. Ablation analyses further confirmed the effectiveness of each module, especially highlighting the critical role of test-time adaptation in ensuring robust generalization under dynamically evolving traffic conditions.

Future research should focus on expanding this model's applicability across diverse real-world settings. Specifically, we suggest exploring transfer learning strategies to adapt the model to traffic datasets from different geographic regions or infrastructure systems. This is essential for assessing the transferability and robustness of the framework beyond a single domain. In addition, integrating real-time sensor telemetry (e.g., GPS, weather, and traffic flow) may further improve predictive accuracy in evolving conditions. The development of lightweight variants optimized for edge computing could also facilitate real-time deployment in resource-constrained environments.

These findings suggest that incorporating adaptive learning mechanisms into traffic-forecasting pipelines can significantly enhance risk assessment, emergency response, and policymaking in intelligent transportation systems. Broader deployment, however, will require addressing domain adaptation, multi-modal data fusion, and fairness-driven evaluations to ensure reliable and equitable model performance at scale.

Author Contributions: Conceptualization, D.P. and W.Y.; Methodology, D.P.; Software, D.P.; Validation, D.P. and W.Y.; Formal analysis, D.P.; Investigation, D.P.; Resources, W.Y.; Data curation, D.P.; Writing—original draft preparation, D.P.; Writing—review and editing, W.Y.; Visualization, D.P.; Supervision, W.Y.; Project administration, W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study are available in the article.

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

Appendix A. TTT-Enhanced Transformer for Accident Severity Prediction

Algorithm A1. Pseudocode of the Proposed TTT-Enhanced Transformer

Input: Test sample x, model parameters θ
Output: Updated parameters θ', prediction y

// Feature Processing
$Z \leftarrow$ FeatureEmbedding(x)
$Z \leftarrow$ FPN(Z)
$Z \leftarrow$ AdaptiveMemory(Z)
$h \leftarrow$ TransformerEncoder(Z)
h_cba $\leftarrow$ ClassBalancedAttention(h)

// Test-Time Training Loop
for i = 1 to max_steps do
    [h_past, h_future] $\leftarrow$ TemporalSplit(h)
    h_pred $\leftarrow$ TemporalPredictor(h_past)
    L_ssl $\leftarrow$ MSE(h_pred, h_future)
    θ_ssl $\leftarrow$ θ_ssl $- \alpha \cdot \nabla$L_ssl
    if convergence_criteria met then break
end for

// Parameter Propagation
θ_encoder, θ_memory $\leftarrow$ θ_ssl

// Final Prediction
$Z' \leftarrow$ AdaptiveMemory(Z, θ_memory)
$h' \leftarrow$ TransformerEncoder(Z')
y $\leftarrow$ SeverityClassifier(ClassBalancedAttention(h'))
return θ', y

References

1. Al-Thani, M.G.; Sheng, Z.; Cao, Y.; Yang, Y. Traffic Transformer: Transformer-based framework for temporal traffic accident prediction. *AIMS Math.* **2024**, *9*, 12610–12629. [CrossRef]
2. Pölz, A.; Blaschke, A.P.; Komma, J.; Farnleitner, A.H.; Derx, J. Transformer Versus LSTM: A comparison of deep learning models for Karst Spring discharge forecasting. *Water Resour. Res.* **2024**, *60*, e2022WR032602. [CrossRef]
3. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2019; Volume 32, pp. 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*; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 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*, 4845–4901. [CrossRef]
6. Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.A.; Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In Proceedings of the International Conference on Machine Learning, Proceedings of Machine Learning Research, Vienna, Austria, 12–18 July 2020; Volume 119, pp. 9229–9248.
7. Gu, W.; Gu, L.; Wang, Z.; Suen, C.Y.; Wang, Y. DocTTT: Test-time training for handwritten document recognition using meta-auxiliary learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 28 February–4 March 2025.

8. Sinha, S.; Gehler, P.; Locatello, F.; Schiele, B. TeST: Test-time self-training under distribution shift. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2758–2768.
9. Xiao, Z.; Snoek, C.G.M. Beyond model adaptation at test time: A survey. *arXiv* **2024**, arXiv:2411.03687. Available online: http://arxiv.org/abs/2411.03687 (accessed on 16 February 2025).
10. Sameen, M.I.; Pradhan, B. Severity prediction of traffic accidents with recurrent neural networks. *Appl. Sci.* **2017**, *7*, 476. [CrossRef]
11. Kim, S.; Lym, Y.; Kim, K.-J. Developing crash severity model handling class imbalance and implementing ordered nature: Focusing on elderly drivers. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1966. [CrossRef]
12. Fiorentini, N.; Losa, M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures* **2020**, *5*, 61. [CrossRef]
13. Noh, S.-H. Analysis of gradient vanishing of RNNs and performance comparison. *Information* **2021**, *12*, 442. [CrossRef]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, l.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Long Beach, CA, USA, 2017; pp. 5998–6008.
15. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
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? In Proceedings of the 35th Conference on Neural Information Processing Systems, Online, 6–14 December 2021; pp. 21808–21820.
17. Yan, H.; Ma, X.; Pu, Z. Learning dynamic and hierarchical traffic spatiotemporal features with transformer. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22386–22399. [CrossRef]
18. Park, D.; Jeong, J.; Yoon, S.H.; Jeong, J.; Yoon, K.J. T4P: Test-time training of trajectory prediction via masked autoencoder and actor-specific token memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 15065–15076.
19. Gülçehre, C.; Chandar, S.; Bengio, Y. Memory augmented neural networks with wormhole connections. *arXiv* **2017**, arXiv:1701.08718.
20. Sukhbaatar, S.; Grave, E.; Lample, G.; Jegou, H.; Joulin, A. Augmenting self-attention with persistent memory. *arXiv* **2019**, arXiv:1907.01470.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 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 Sens.* **2020**, *12*, 784. [CrossRef]
23. Zhuang, J.X.; Cai, J.; Zhang, J.; Zheng, W.; Wang, R. Class attention to regions of lesion for imbalanced medical image recognition. *Neurocomputing* **2023**, *555*, 126577. [CrossRef]
24. Ali, Y.; Hussain, F.; Haque, M.M. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accid. Anal. Prev.* **2024**, *194*, 107378. [CrossRef]
25. Wang, X.; Chen, F.; Jin, B.; Lin, M.; Zou, F.; Zeng, R. DSTSPYN: A dynamic spatial–temporal similarity pyramid network for traffic flow prediction. *Appl. Intell.* **2025**, *55*, 237. [CrossRef]
26. 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. [CrossRef]
27. Santoro, A.; Bartunov, S.; Botvinick, M.; Westra, D.; Lillicrap, T. Meta-learning with memory-augmented neural networks. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1842–1850.
28. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef]
29. Lee, H.; Jin, S.; Chu, H.; Lim, H.; Ko, S. Learning to remember patterns: Pattern matching memory networks for traffic forecasting. In Proceedings of the International Conference on Learning Representations(ICLR), Virtually, 25–29 April 2022.
30. McNulty, G. Severity curve fitting for long-tailed lines: An application of stochastic processes and Bayesian models. *Variance* **2017**, *11*, 118–132.
31. Amjad, R.A.; Geiger, B.C. Understanding neural networks and individual neuron importance via information-ordered cumulative ablation. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 1–5.
32. Zhang, Z.; Yang, W.; Wushour, S. Traffic accident prediction based on LSTM-GBRT model. *J. Control Sci. Eng.* **2020**, *2020*, 4206919. [CrossRef]
33. Li, P.; Abdel-Aty, M.; Yuan, J. Real-time crash risk prediction on arterials based on LSTM-CNN. *Accid. Anal. Prev.* **2019**, *135*, 105371. [CrossRef] [PubMed]
34. 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*, 398. [CrossRef]

35. Yeung, M.; Sala, E.; Schönlieb, C.B.; 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. [CrossRef]
36. Gorishniy, Y.; Kotelnikov, A.; Babenko, A. TabM: Few-Shot Fine-Tuning of Tabular Models for Small Datasets. In Proceedings of the 13th International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
37. Xu, Y.; Goodacre, R. On splitting training and validation set: A comparative study of cross-validation, bootstrap and system-atic sampling for estimating the generalization performance of supervised learning. *J. Anal. Test.* **2018**, *2*, 249–262. [CrossRef]
38. 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*, 102. [CrossRef]
39. Sarafianos, N.; Xu, X.; Kakadiaris, I.A. Deep imbalanced attribute classification using visual attention aggregation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 680–697.
40. Ghahremannezhad, H.; Shi, H.; Liu, C. Real-time accident detection in traffic surveillance using deep learning. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Virtual, 21–23 June 2022; pp. 1–6.
41. Al Falasi, H.A. Predictive Rescue System Through Real-time Accident Monitoring Leveraging Artificial Intelligence. Master's Thesis, Rochester Institute of Technology, Rochester, NY, USA, 2023.
42. Zhao, G.; Ge, W.; Yu, Y. GraphFPN: Graph feature pyramid network for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 2743–2752.

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.