

Article

A Physics-Constrained Multi-Task Learning–Semi-Markov Framework for Bridge Condition Assessment, Deterioration Forecasting, and Risk-Aware Maintenance Prioritization

Zhihui Feng ^{1,2}, Yuchen Zhao ¹, Liangqi Zhang ², Fulei Xu ¹, Xiaojun Li ², Zhiqiang Liang ³, Yufeng Guo ^{1,*} and Hui Zhang ^{1,*}

¹ College of Artificial Intelligence, Henan Agricultural University, Zhengzhou 450046, China; hafzhh@henau.edu.cn (Z.F.); zycc28@163.com (Y.Z.)

² Detong Intelligent Technology Co., Ltd., Xuchang 461000, China; zhq6196@outlook.com (L.Z.)

³ School of Highway, Chang'an University, Xi'an 710064, China; liang-zhiqiang@chd.edu.cn

* Correspondence: gyfzhp@126.com (Y.G.); huizi@henau.edu.cn (H.Z.)

Abstract

Bridge health assessment and deterioration prediction are essential for traffic safety and maintenance planning. However, conventional bridge evaluation still relies heavily on expert judgment and heuristic rules, limiting objectivity and long-term forecasting capability. This study proposes a physics-constrained multi-task learning–semi-Markov framework for bridge condition assessment, multi-year deterioration forecasting, and risk-aware maintenance prioritization. Guided by the Highway Bridge Technical Condition Rating Code, the proposed model jointly predicts health grade, defect type, and severity from inspection item/defect entry records, improving assessment robustness through cross-task information sharing. The predicted health states are then incorporated into a physics-constrained semi-Markov model to forecast bridge deterioration over a three-year horizon, and the current and predicted states are further used for maintenance prioritization. Experiments on real bridge inspection data show that the proposed model achieves test accuracies of 95.20%, 99.67%, and 99.72% for health grade, defect type, and severity, respectively, with a Cohen's kappa of 0.872, while the proposed semi-Markov model achieves a three-year prediction accuracy of 95.4% and a weighted accuracy of 99.1%. Comparative and ablation studies further demonstrate the effectiveness of the proposed framework for bridge lifecycle management.

Keywords: bridge condition assessment; deterioration prediction; multi-task learning; semi-Markov modeling; risk-aware maintenance prioritization



Academic Editor: Joan R. Casas

Received: 6 June 2026

Revised: 14 July 2026

Accepted: 18 July 2026

Published: 22 July 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Transportation infrastructure is a critical foundation for socioeconomic development [1], and bridges, as indispensable components of highway networks, play an essential role in public welfare and economic stability [2,3]. According to the Ministry of Transport's 2024 Statistical Bulletin, approximately 28,700 bridges were added to the national highway network in that year alone. As bridges age, material degradation [4], connection damage [5], and other forms of deterioration gradually accumulate, posing increasing challenges to structural safety and serviceability. Nevertheless, bridge maintenance practice still largely follows a reactive “detect-and-fix” paradigm [6], which is often associated with delayed intervention [7], inefficient resource allocation [8], and limited ability to satisfy differentiated

maintenance needs [9]. In this context, reliable bridge condition assessment, deterioration forecasting, and maintenance prioritization are essential for intelligent and fine-grained infrastructure lifecycle management.

Existing bridge studies have gradually evolved from structural behavior analysis to condition evaluation, deterioration modeling, and maintenance decision support. Earlier research mainly focused on the mechanical behavior of specific bridge types; for example, Jia et al. investigated the dynamic behavior and seismic performance of the Panzhong Bridge under different auxiliary pier configurations [10]. More recent studies have shifted toward component-level deterioration analysis, such as the work of Greco et al. on reinforced concrete bridge decks [11]. At the same time, artificial intelligence has been increasingly introduced into bridge structural health monitoring and condition assessment. Prakash et al. reviewed AI-integrated technologies for structural health monitoring of concrete bridges [12]. Yu et al. developed a hybrid approach for automated bridge deck defect identification using ground penetrating radar data [13], while Jiang et al. proposed a lightweight object detection network for multi-damage recognition of concrete bridges in complex environments [14]. Image-based deep learning methods have also been applied to more specific bridge inspection tasks. For example, Ameli et al. used Mask R-CNN and YOLOv8 for steel bridge corrosion segmentation and condition rating [15], Huang et al. proposed BridgeNet for multiclass bridge surface damage detection and segmentation [16], and Pokhrel et al. used UAS-collected images with CNN and Vision Transformer models for automated concrete bridge deck damage detection [17]. Beyond image-based inspection, Wang et al. developed a deep learning-based bridge condition assessment method through fusion of multi-type inspection data [18].

In parallel with these data-driven inspection studies, other researchers have focused on bridge reliability analysis, deterioration forecasting, and maintenance decision support. Chen et al. proposed a hybrid framework combining the extended finite element method, artificial neural networks, and Monte Carlo simulation for fatigue crack and reliability analysis [19], and Miao et al. developed a bridge performance assessment method based on long short-term memory networks, temporal pattern attention, and a hybrid hidden Markov model [20]. In addition, maintenance prioritization methods such as the fuzzy analytic hierarchy process [21] and multi-attribute decision-making [22] have been applied to support bridge maintenance planning. To further clarify the research gap, representative studies related to bridge defect detection, condition assessment, deterioration prediction, uncertainty modeling, and maintenance decision support are systematically compared in Table 1.

As summarized in Table 1, although existing studies have contributed to bridge structural analysis, defect detection, deterioration modeling, and maintenance decision support, several limitations remain. First, many studies focus on current bridge condition assessment or defect identification, while the explicit integration of current-state evaluation with future-state forecasting remains insufficient. Second, although data-driven and deterioration-oriented models have improved condition prediction, their outputs are not always directly connected with engineering transition constraints and maintenance decision processes. Third, maintenance prioritization is often based on current condition indicators or expert-defined weights, while predicted deterioration trajectories, uncertainty, traffic exposure, and cost–benefit considerations are not fully incorporated. Therefore, a unified framework that can jointly support bridge condition assessment, deterioration forecasting, and risk-aware maintenance prioritization is still needed.

To address the above limitations, this study proposes a physics-constrained multi-task learning–semi-Markov framework for bridge condition assessment, deterioration forecasting, and risk-aware maintenance prioritization [23,24]. Specifically, a Transformer-based

multi-task learning module is constructed to jointly predict health grade, defect type, and severity level from bridge inspection item/defect entry records, thereby enhancing cross-task information sharing and improving the consistency of bridge condition evaluation [25–28]. On this basis, a physics-constrained semi-Markov chain (SMC) model is established by incorporating historical inspection information, deep-learning-derived initial states, Bayesian sojourn-time modeling, and engineering transition constraints so as to forecast bridge health trajectories over a three-year horizon. Furthermore, to support maintenance planning, the current and predicted bridge states are incorporated into a risk-aware priority evaluation framework whose composite weights are determined by combining the analytic hierarchy process and entropy weighting [29–31], together with uncertainty adjustment and cost–benefit considerations [32,33]. Finally, comparative experiments, ablation studies, temporal validation, and sensitivity analysis are conducted to evaluate the effectiveness and practical applicability of the proposed framework [34]. The proposed framework is expected to provide practical decision support for bridge condition assessment, deterioration forecasting, and lifecycle maintenance management.

The main contributions of this study are summarized as follows:

- (1) A multi-task deep learning framework is developed for bridge condition assessment, in which health grade, defect type, and severity are jointly predicted from bridge inspection data. By exploiting cross-task information sharing, the proposed framework improves the robustness and consistency of bridge condition evaluation.
- (2) A physics-constrained semi-Markov chain model is established for bridge deterioration prediction. By incorporating deep learning-derived initial states, Bayesian sojourn time modeling, and engineering transition constraints, the proposed model enables multi-year forecasting of bridge health trajectories.
- (3) An integrated risk-aware maintenance prioritization framework is constructed by linking current condition assessment with future deterioration prediction. The predicted results are further incorporated into a maintenance prioritization scheme that combines the analytic hierarchy process and entropy weighting, thereby supporting more rational and forward-looking bridge maintenance planning.

Table 1. Systematic comparison of representative studies on bridge condition assessment, deterioration forecasting, and maintenance decision support.

Study	Input Data	Prediction Target	Deterioration Model	Uncertainty Handling	Maintenance Prioritization	Main Limitation/Remaining Challenge
Jia et al. [10]	Structural parameters of a cable-stayed bridge with different auxiliary pier configurations	Dynamic behavior and seismic performance	Not considered	Not explicitly modeled	Not considered	Focuses on structural behavior of a specific bridge type rather than condition assessment or forecasting.
Greco et al. [11]	Reinforced concrete bridge deck/component-level bridge damage information	Component-level deterioration or damage behavior	Not considered as a lifecycle deterioration model	Not explicitly modeled	Not considered	Focuses on component-level deterioration, without integrated multi-task assessment and maintenance prioritization.
Prakash et al. [12]	Published studies on AI-based structural health monitoring of concrete bridges	Review of AI-integrated bridge SHM technologies	Not applicable	Discussed conceptually	Discussed generally	Provides a review of AI-based SHM but does not propose an integrated assessment–forecasting–decision framework.
Yu et al. [13]	Ground penetrating radar signals from bridge decks	Automated bridge deck defect identification	Not considered	Mainly reflected through classification performance and model robustness	Not considered	Focuses on current defect identification, without bridge-level deterioration forecasting or maintenance prioritization.
Jiang et al. [14]	Bridge damage images collected under complex environments	Multi-damage recognition of concrete bridges	Not considered	Mainly evaluated through detection performance under complex backgrounds	Not considered	Focuses on image-based damage detection under complex environments, without future-state forecasting.
Ameli et al. [15]	Steel bridge corrosion images	Corrosion segmentation and condition rating	Not explicitly modeled	Not explicitly considered	Not included	Performs image-based corrosion rating but does not integrate deterioration forecasting and maintenance decision support.
Wang et al. [18]	Multi-type inspection data	Bridge condition assessment	Not modeled	Limited	Not included	Integrates multi-type inspection data for condition assessment but does not include transition forecasting or maintenance prioritization.
Chen et al. [19]	Crack and reliability simulation data	Fatigue crack and reliability	XFEM–ANN framework	Monte Carlo reliability analysis	Not considered	Addresses fatigue reliability but not multi-task condition assessment or maintenance prioritization.
Miao et al. [20]	Sparse bridge inspection data	Bridge performance assessment and hidden condition–state inference	Hybrid LSTM–attention–HMM model	Partly considered through hidden-state probabilistic inference	Not explicitly considered	Includes temporal modeling but does not integrate risk-aware maintenance prioritization.

Table 1. *Cont.*

Study	Input Data	Prediction Target	Deterioration Model	Uncertainty Handling	Maintenance Prioritization	Main Limitation/Remaining Challenge
Shen et al.; Lu et al. [21,22]	Expert judgment, safety assessment indicators, and multi-attribute engineering decision indicators	Bridge safety assessment and engineering decision support	Not considered	Partly reflected through fuzzy evaluation or indicator weighting	Partly considered through safety-oriented or multi-attribute ranking	Supports maintenance ranking but relies mainly on current indicators rather than predicted deterioration trajectories.
Liu S.-S. et al. [24]	Bridge lifecycle maintenance information and maintenance-related constraints	Lifecycle maintenance scheduling of bridge infrastructure	Optimization-oriented lifecycle maintenance model	Considered indirectly through optimization setting	a two-stage maintenance scheduling model	Does not fully integrate multi-task assessment, uncertainty-aware forecasting, and cost–benefit prioritization.
This study	Official bridge inspection item/defect entry records, defect descriptions, traffic exposure, temporal descriptors, and technical condition information	Health grade, defect type, severity, multi-year health state, and maintenance priority	Physics-constrained semi-Markov model guided by multi-task deep learning	Bayesian sojourn-time modeling, prediction confidence, and entropy-based uncertainty adjustment	Risk-aware prioritization using AHP	Case-study dataset; external validation with broader bridge inventories is still needed.

2. Materials

2.1. Data Sources

The bridge inspection dataset used in this study was provided by the Henan Provincial Department of Transportation and covers 58 large and extra-large bridges from 2018 to 2024. The original data consist of two source tables: a bridge-level condition rating table and an inspection item/defect detail table. The bridge-level table records one bridge at one inspection date and includes bridge attributes, traffic volume [35], superstructure, substructure, and deck-system condition ratings, overall bridge technical condition grade, and overall technical condition score evaluated according to the Highway Bridge Technical Condition Rating Code (JTG/T H21-2011) [36]. The defect detail table records inspection items or defect entries associated with a specific bridge, inspection date, component type, component number, component defect type [37], defect count [38], defect location, and textual defect description [39]. In this context, it should be noted that the Highway Bridge Technical Condition Rating Code (JTG/T H21-2011) was used because the official inspection records in this study were evaluated under this Chinese bridge inspection standard. However, the proposed framework is not limited to this specific code. For applications in other countries or regions, the health-grade definitions, severity thresholds, and engineering transition constraints can be recalibrated according to the corresponding local bridge inspection standards, while the overall multi-task learning–semi-Markov modeling framework remains applicable.

In this study, one “inspection record” refers to an inspection item/defect entry-level observation after linking the two source tables, rather than an independent bridge-level sample. Thus, multiple records may correspond to the same bridge and inspection date when several components or defects are recorded during one inspection. After data cleaning and linking, the final unified dataset contained 30,635 inspection item/defect entry records. These bridges were selected because they have relatively complete official inspection records over a seven-year period and provide the key information required for multi-task condition assessment, longitudinal deterioration forecasting, and maintenance prioritization. It should be noted that the dataset is not intended to represent all bridge types or regions but rather serves as a practical case set for assessing the feasibility and applicability of the proposed integrated workflow using official bridge inspection data.

2.2. Data Processing

The two source tables were linked using the bridge identifier and inspection time, after converting these keys into consistent numeric and datetime formats. Completely duplicated rows were removed, whereas records with the same bridge identifier and inspection time but different components, defect types, defect locations, or defect descriptions were retained to avoid losing valid inspection information.

The bridge-level condition rating table was left-joined to the inspection item/defect detail table so that bridge-level attributes, traffic volume, component ratings, overall technical condition grade, and overall score were attached to each corresponding inspection item/defect entry record. Therefore, the merged dataset remained organized at the inspection item/defect entry level. To reduce potential data leakage, all subsequent data splitting was conducted by bridge identifier, ensuring that records from the same bridge were assigned to only one subset.

The detailed record counts during data cleaning and joining are summarized in Table 2. No records were removed due to missing bridge identifier or inspection time. After removing exact duplicate rows, 28,473 inspection item/defect detail records and 140 bridge-level condition rating records were retained. Repeated bridge identifier–inspection time keys were not automatically treated as duplicates. In the inspection item/defect detail table,

such records generally represented different components, defect types, defect locations, or defect descriptions from the same bridge inspection and were therefore retained. In the bridge-level condition rating table, the core target label fields, including the overall bridge technical condition grade and overall bridge score, were checked and found to be consistent among records with similar keys. Therefore, no conflicting target labels were introduced, and the left join produced 30,635 linked inspection item/defect entry records.

Table 2. Summary of source tables, data cleaning, and joining process.

Data Table/Step	Observational Level	Main Contents	Original Records	Removed Due to Missing Keys	Exact Duplicate Rows Removed	Records Retained/Generated
Inspection item/defect detail table	Inspection item/defect entry level	Bridge identifier, inspection time, component type, component number, defect type, defect location, defect description, rating scale, deduction score, crack width, affected area, and geometric measurements	29,914	0	1441	28,473
Bridge-level condition rating table	Bridge-inspection-date level	Bridge attributes, traffic volume, superstructure, substructure, and deck-system ratings, overall technical condition grade, and overall bridge score	146	0	6	140
Linked unified dataset	Inspection item/defect entry level after joining	Defect detail records linked with bridge-level condition rating and traffic information	—	—	—	30,635

Note: Same-key non-identical records refer to records sharing the same bridge identifier and inspection time but not being completely identical across all fields. In the bridge-level condition rating table, the core target label fields were checked and found to be consistent among same-key records. Therefore, these records did not introduce conflicting target labels.

Missing traffic volumes were imputed using the median value, with an additional missing-value flag retained. The median was used because traffic volume data may be affected by extreme values, and median imputation provides a more robust substitute than mean imputation for incomplete continuous records. Missing defect attributes were filled using group-wise means stratified by health grade and defect type.

According to the Highway Bridge Technical Condition Rating Code, the overall bridge scores are mapped to five health grades: Excellent (95–100), Good (80–95), Attention (60–80), Poor (40–60), and Dangerous (<40). Defect-type labels were assigned using a manually curated keyword lexicon, as summarized in Table 3. The lexicon was constructed by first enumerating the unique defect-type terms appearing in the original inspection item/defect detail table and then mapping them into broad engineering categories, including structural component defects, functional component defects, apparent surface defects, no defect, and other defects. Ambiguous terms were resolved using contextual keywords and a priority rule of structural component defects > functional component defects > apparent surface defects. The lexicon was designed to cover the defect-type terms observed in the present official inspection dataset rather than to serve as an exhaustive universal bridge defect ontology. For applications to other regions, bridge types, or inspection-code systems, the same enumeration-mapping procedure can be used to extend or recalibrate the lexicon. In line with the Urban Bridge Inspection and Evaluation Code, severity is classified into three levels (light, moderate, severe) using a weighted combination of crack width (≤ 0.1 mm, $0.1\text{--}0.3$ mm, >0.3 mm), affected area (≤ 1 m², $1\text{--}5$ m², >5 m²), and rating scale (≤ 2 ,

3–4, >4), yielding a final weighted severity score. To clarify the relationship between label construction and input variables, the overall bridge technical condition score and grade were used only to construct the health-grade labels and were not included as input variables for health-grade prediction; therefore, direct label leakage from the technical condition score was avoided. For the severity task, the labels were derived from measured defect attributes, including crack width, affected area, and rating scale, according to engineering threshold rules in relevant bridge condition assessment standards. These measured attributes were retained as input variables because practical severity assessment is inherently measurement-based. Therefore, the severity task is interpreted as a rule-consistent auxiliary assessment task rather than an entirely independent prediction task detached from these measurements.

Table 3. Keyword lexicon and priority rules for defect-type categorization.

Defect-Type Category	Keyword Lexicon/Matching Terms	Ambiguous-Case Handling
Structural component defects	Crack, breakage, damage, deterioration, corrosion, spalling, carbonation, void, hole, exposed reinforcement, anchorage, prestress, masonry, strength deterioration, deflection, fire damage, concrete, reinforcement, deformation, structural displacement, settlement, displacement, scouring, undermining, erosion, protective layer, failure, defect	Terms such as “missing” or “collision damage” were assigned to this category when they were associated with concrete, reinforcement, structural components, or protective layers.
Functional component defects	Bearing, expansion joint, drainage, railing, protection facility, coating, blockage, hanger, sheath, drainpipe, guide channel, sign, sliding plate, PTFE plate, wear, collision damage, missing	Terms such as “missing” or “collision damage” were assigned to this category when they were associated with railings, bearings, drainage facilities, signs, expansion joints, or other functional components.
Apparent surface defects	Honeycombing, pitting, stain, discoloration, surface contamination, uneven surface, asphalt bleeding, seepage	“Seepage” was assigned to this category when it appeared as a surface or appearance-related defect without additional structural-damage keywords.
No defect	Empty or missing defect type description	Empty strings or missing defect-type entries were assigned to the no-defect category.
Other defects	Terms not matched by the above lexicon	Unmatched terms were assigned to the other-defect category for diagnostic checking.

Note: The listed terms are matching keywords used for lexicon-based categorization, not an exhaustive list of all raw defect descriptions. For compound descriptions containing multiple keywords, the priority rule structural component defects > functional component defects > apparent surface defects was applied.

Each numerical feature (e.g., crack width, affected area, traffic volume) undergoes nonlinear transformations, interaction construction, and severity scoring, followed by 1st–99th percentile clipping to suppress outliers. The textual “defect type” field is vectorized with TF-IDF [40] and fused with the numerical features to form a multimodal input. The maximum vocabulary size was set to 3000, with a minimum document frequency of 2 and a maximum document frequency ratio of 0.95. After fitting on the training set, the resulting vocabulary contained 102 terms. The fitted vectorizer was subsequently applied unchanged to the validation, test, and new inspection records. Terms contained in the training vocabulary are encoded normally, whereas previously unseen terms are ignored. For applications involving substantial new defect terminology, the TF-IDF vocabulary should be updated or refitted using representative local inspection records, followed by model retraining. The resulting text features are fused with the numerical features to form

a multimodal input. In this study, the TF-IDF text vector was treated as a static inspection item/defect entry context rather than as a temporally varying text sequence. Complete per-timestep textual histories were not consistently available for all inspection records; therefore, temporal variation was represented by the numerical inspection sequence and temporal descriptors, including inspection interval, days since first inspection, calendar year, and state-change flags. We further introduce temporal descriptors [41]—inspection interval, days since first inspection, calendar year, and state-change flags—aggregating discontinuous yearly records, interpolating missing years, and constructing four-step temporal windows to feed the downstream deep-learning and semi-Markov models. Labels for the three tasks are standardized: health grades map “Excellent → Dangerous” to 0–4; defect types map structural, functional, surface, and no-defect categories to 0–3; severity encode light, moderate, and severe as 0–2.

A three-dimensional stratified grouping scheme based on bridge ID was used for splitting, with an approximate 6:2:2 ratio for training/validation/testing [42]. All records from the same bridge were assigned to a single subset, and the split was optimized to minimize discrepancies in the joint distribution of the three labels; distributional consistency and class coverage were prioritized over strict ratio matching, so minor deviations in totals are expected. Given the severe class imbalance in the original dataset (Excellent 69, Good 18,536, Attention 11,324, Poor 706, Dangerous 0), minority classes were augmented according to bridge condition assessment standards and the joint distribution of health grade, defect type, and severity. Because the original dataset contained no real Dangerous samples, the generated Dangerous samples were treated as rule-constrained safety-critical training scenarios rather than empirically observed Dangerous bridge records. These synthetic records were assigned virtual bridge identifiers and were used only for training-stage augmentation, while the validation and test sets were preserved as real inspection data [43]. Therefore, the Dangerous samples were intended to expose the model to plausible extreme deterioration patterns and reduce majority-class bias during training but were not used to validate Dangerous-class prediction performance. After augmentation, class counts were as follows: Excellent 2000; Good 18,536; Attention 11,324; Poor 2000; Dangerous 1000. The final split statistics for each task are summarized in Tables 4–6. Although the grouped split reduces bridge-level data leakage, it may lead to uneven class distributions in the validation and test subsets because rare classes are concentrated in a limited number of bridge-level units. Therefore, classes absent from the validation or test subsets were not interpreted as independently validated. The corresponding limitation was considered when interpreting the class-specific evaluation results.

Table 4. Distribution of the health level dataset.

Class	Training	Validation	Test
Poor	1663	0	0
Excellent	1481	7	13
Dangerous	780	0	0
Attention	5345	4446	1533
Good	12,366	1941	4229

Table 5. Distribution of the defect type dataset.

Class	Training	Validation	Test
Functional component	2855	321	1015
No defect	1018	0	0
Structural component	15,838	5873	4362
Surface defect	1924	200	398

Table 6. Distribution of the severity dataset.

Class	Training	Validation	Test
Moderate	4030	317	331
Mild	16,778	6077	5444
Severe	827	0	0

To provide a clearer overview of the dataset organization, Figure 1 summarizes the data schema and preprocessing workflow, including data joining, cleaning, feature construction, label generation, grouped data splitting, and training-stage synthetic expansion.

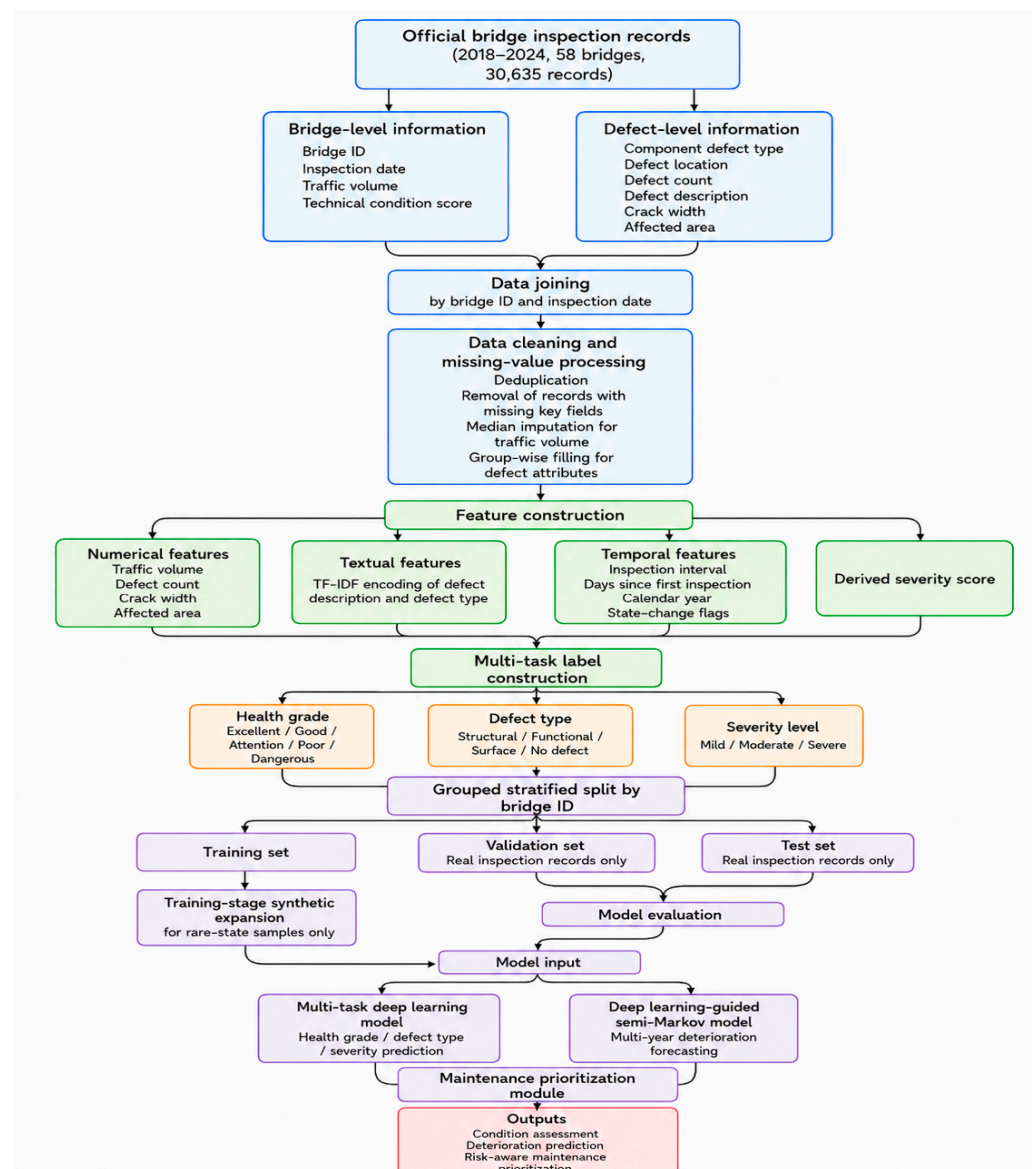


Figure 1. Data schema and preprocessing workflow of the bridge inspection dataset.

The scarcity of bridge inspection samples in rare states reflects an inherent characteristic of real-world bridge management data. Bridges in service rarely remain in an “Excellent” state for extended periods, whereas bridges assessed as “Dangerous” are usually closed, traffic-restricted, or subjected to immediate maintenance intervention, limiting the accumulation of such observations in routine inspection records [44]. As a result, the original dataset is highly imbalanced, with most samples concentrated in the Good and

Attention states. To reduce majority-class bias during training, rare-state samples were synthetically expanded according to the Highway Bridge Technical Condition Rating Code and the joint distribution of health grade, defect type, and severity. The synthetic records were labeled as virtual bridges and used exclusively for training-stage class balancing, while the validation and test sets contained only real inspection records. Although real Poor records were present in the original dataset, no Poor samples were included in the validation or test subsets after bridge-grouped splitting. In addition, the original dataset contained no real Dangerous samples, while the No-defect and Severe classes were also absent from the validation and test subsets. Therefore, the performance of the model on the Poor, Dangerous, No-defect, and Severe classes was not empirically validated in the present evaluation. Synthetic augmentation cannot substitute for validation using real rare-state inspection records, and additional external validation based on broader bridge inventories is required.

3. Methods

3.1. Multi-Task Deep Learning Architecture

The multi-task deep learning model adopts a three-layer architecture—multi-source feature backbone, multiscale semantic fusion, and multi-task decoding—which jointly outputs health grade, defect type, and severity within a single network. The overall framework is illustrated in Figure 2.

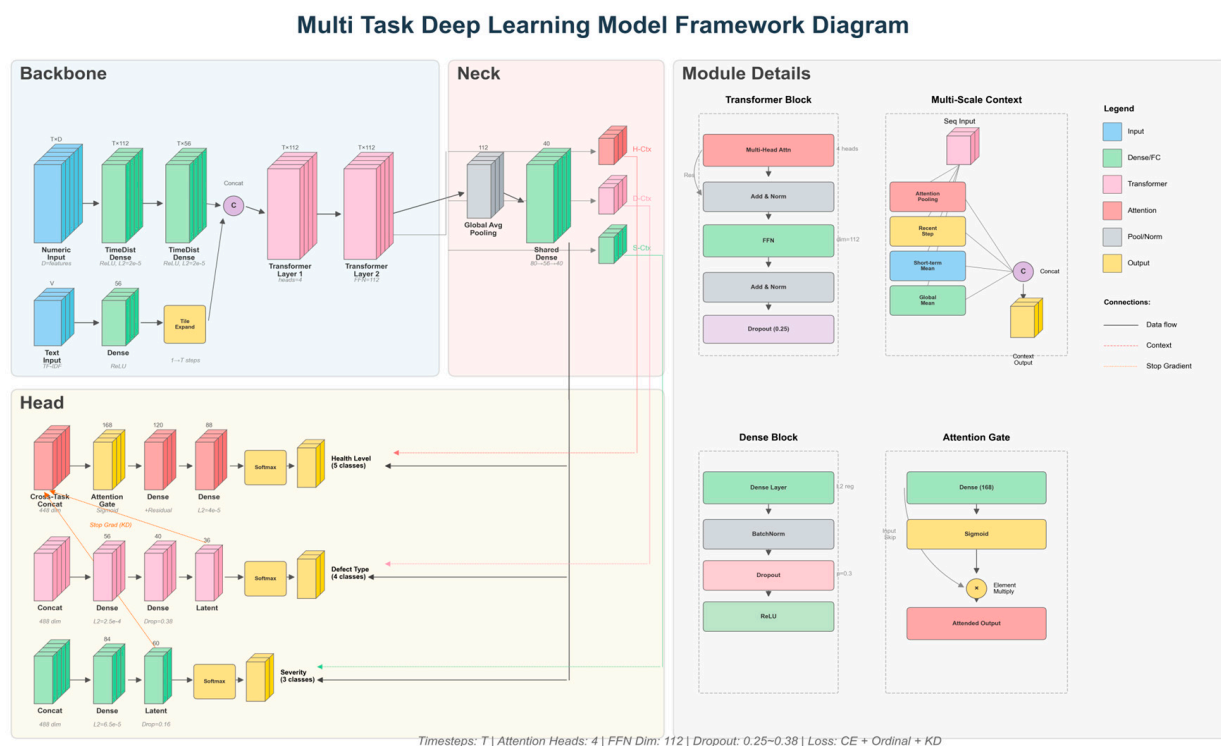


Figure 2. Multi-task deep learning model architecture.

The three-layer architecture was adopted to balance feature representation, temporal semantic modeling, and task-specific prediction. The first layer integrates heterogeneous numerical, textual, and temporal inspection features into a unified representation. The second layer captures multiscale bridge condition semantics, including the latest inspection state, short-term deterioration pattern, and long-term historical tendency. The third layer generates task-specific predictions for health grade, defect type, and severity while maintaining cross-task information sharing. This design is more suitable than a single-stage

structure for modeling heterogeneous bridge inspection records, while keeping the model capacity within a moderate scale, as quantified below. The capacity of the final model was quantified by parameter counting. The proposed model contains 763,189 total parameters, including 760,997 trainable parameters and 2192 non-trainable parameters, corresponding to approximately 2.91 MB. Thus, although the model includes Transformer-based temporal fusion and multi-task output heads, its parameter scale remains moderate. Dropout, batch normalization, layer normalization, early stopping, and bridge-level grouped splitting were also used to reduce overfitting risk.

3.1.1. Multi-Source Feature Backbone Layer

The multi-source feature backbone targets numerical time-series inputs and textual semantics, delivering an end-to-end unified representation while modeling long-range temporal dependencies; its architecture is shown in Figure 3 [45].

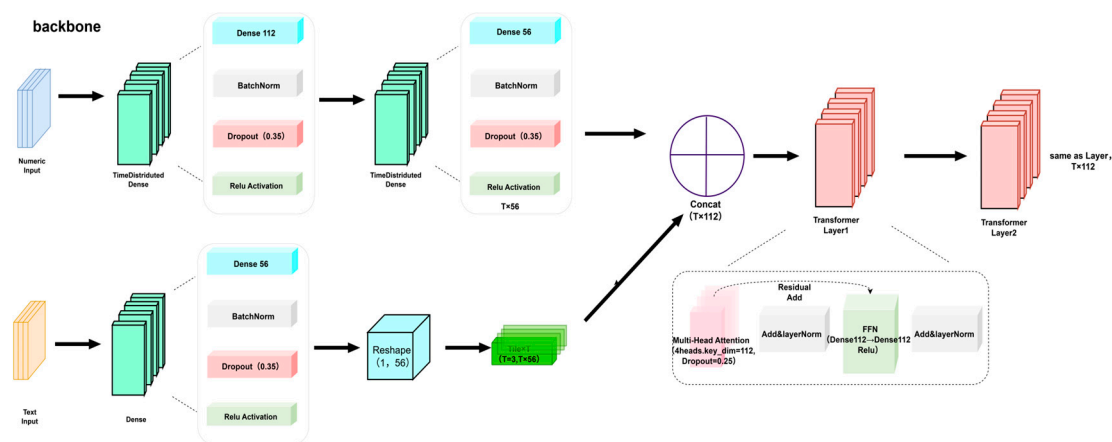


Figure 3. Architecture of the multi-source feature backbone layer.

The backbone ingests both numerical time-series and textual semantic features: the two modalities are processed in parallel branches to extract per-step representations, after which the TF-IDF-encoded text branch is compressed through fully connected layers, tiled across timesteps, and aligned with the numerical branch at each step, producing a $T \times 112$ cross-modal sequence. In the fusion layer, the processed text vector was broadcast to each numerical time step before concatenation with the temporal numerical features. This tiling operation was used as static-context conditioning, not as temporal text replication. It allows each numerical time step to be interpreted under the same defect-description context but does not imply that the same textual description was observed at every historical time step. Temporal variation is modeled through the numerical time-series branch.

The concatenated sequence is successively processed by two Transformer encoder layers, each structured as “Multi-Head Attention → Add & Layer Normalization → Feed-forward Network → Add & Layer Normalization.” Attention is computed following Equation (1), and dropout is applied before every residual connection to suppress overfitting. This backbone layer harmonizes the heterogeneous components and multiscale temporal dependencies, thereby establishing a unified semantic foundation for the subsequent multi-task collaboration.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

In this formulation, Q , K , and V denote the query, key, and value matrices, respectively, while d_k represents the feature dimension of each attention head (32 in this study).

3.1.2. Multiscale Semantic Feature Fusion Layer

The multiscale semantic feature fusion layer (Figure 4) receives the backbone output and aligns semantics across temporal scales and task dimensions. One branch of the Transformer sequence undergoes global average pooling followed by a shared dense stack to compress the representation into a 40-dimensional shared semantic vector. The complementary branch constructs a multiscale contextual module that preserves four aggregation strategies: “attention snapshots, most recent timestep, short-term average, and global average,” where the attention snapshot is computed according to Equation (2). To examine the individual contribution of each temporal aggregation strategy, an ablation study was conducted by retaining only one aggregation branch at a time, including attention pooling, recent timestep, short-term average, and global average. The results are reported in Section 4.6.2.

$$\text{catt}_n = \sum_{t=1}^T \alpha_t h_t, \alpha_t = \frac{\exp(f(h_t) + \beta_t)}{\sum_{k=1}^T \exp(f(h_k) + \beta_k)} \quad (2)$$

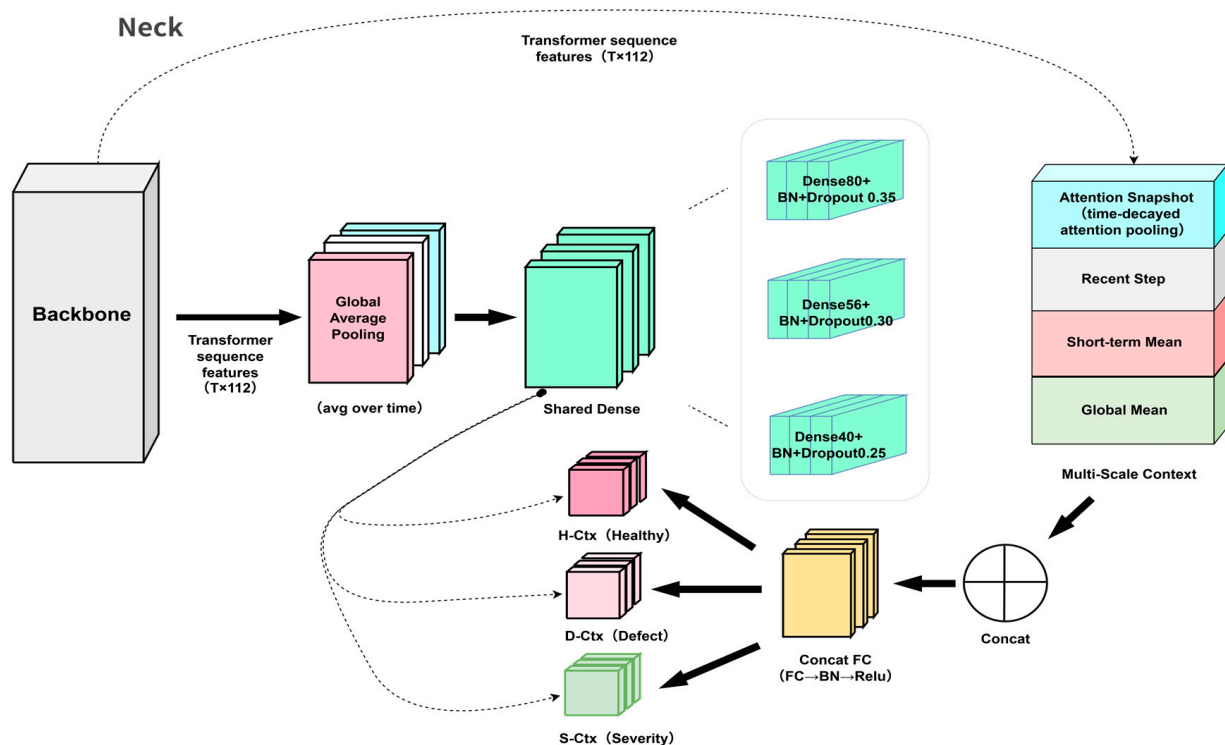


Figure 4. Architecture of the multiscale semantic feature fusion layer.

In this formulation, catt_n denotes the aggregated global attention feature vector, h_t represents the feature vector at time step (t), α_t is the attention weight for time step (t), and β_t is the learnable bias associated with time step (t). The most recent timestep channel captures the instantaneous health state, as formulated in Equation (3):

$$c_{\text{recent}} = h_T \quad (3)$$

Here, c_{recent} denotes the aggregated feature vector from the most recent timestep, while the short-term and global means—defined by Equations (4) and (5), respectively—characterize the pathology evolution trend and its long-term baseline.

$$c_{\text{short}} = \frac{1}{K} \sum_{t=T-T+1}^T h_t \quad (4)$$

Here, c_{short} denotes the aggregated feature vector over the short temporal window, and $\frac{1}{K}\sum$ represents the arithmetic mean pooling operation.

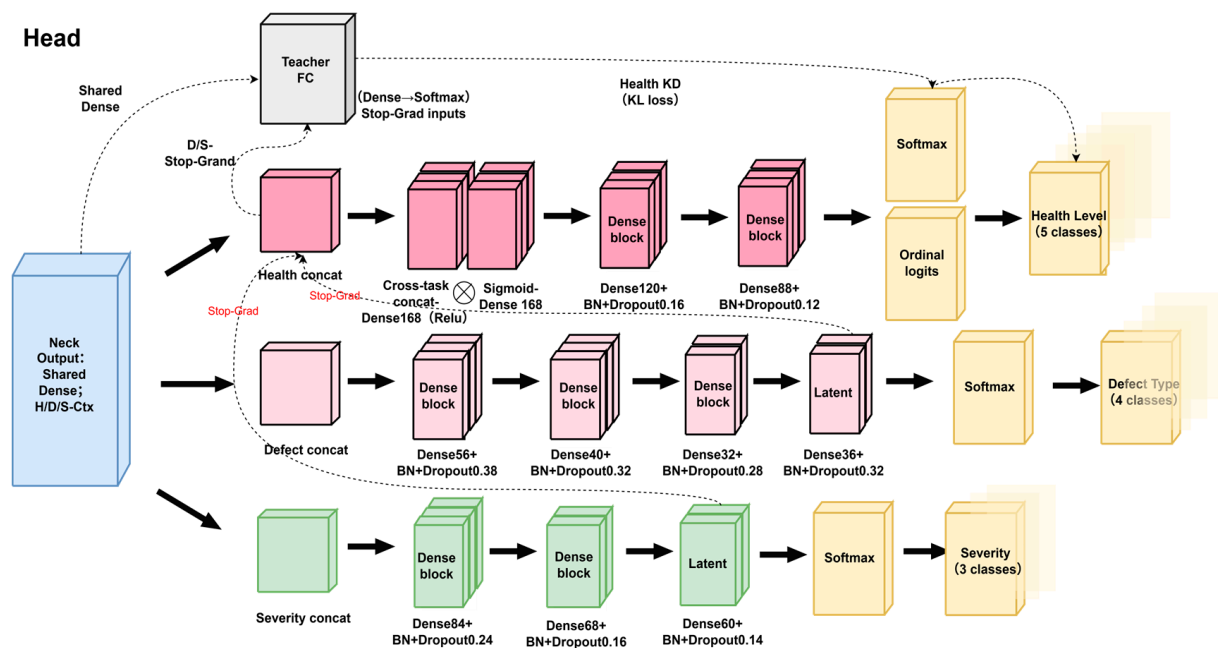
$$c_{\text{global}} = \frac{1}{T} \sum_{t=1}^T h_t \quad (5)$$

Here, c_{global} denotes the globally aggregated temporal feature vector, and $\frac{1}{T}\sum$ corresponds to the global arithmetic mean pooling operation.

After concatenation, the four feature branches are passed through a fully connected layer to yield three semantic categories—health (H-Ctx), defect (D-Ctx), and severity (S-Ctx). These task-specific semantics, together with the shared semantic representation, feed into the decoding layer to enable cross-task collaboration.

3.1.3. Multi-Task Decoding Layer

The multi-task decoding layer (architecture shown in Figure 5) adopts the health grade as the primary branch, with defect type and severity as auxiliary branches. After concatenating the backbone output with the multiscale temporal context, the joint representation feeds into the three branches separately, while Stop-Gradient and knowledge distillation are employed to facilitate cross-task knowledge sharing [46].



Training schedule: Warmup Health→Joint Balance→Disaese Focus→Health Refine

Figure 5. Architecture of the multi-task decoding layer.

The defect type and severity branches each employ a fully connected stack to directly output the four defect categories and three severities, respectively. Before feeding their signals into the health branch, Stop-Gradient truncates the gradients so that only semantic cues are retained.

The health-grade branch fuses the shared semantics, the health-specific context, and the two auxiliary latent vectors, then passes them through a high-capacity dense block and a channel attention gate Equation (6) to obtain recalibrated features. The subsequent residual dense module produces two outputs: a Softmax probability head and an ordinal-logits head. In the teacher–student consistency term, the teacher is not a separate pre-trained model but an internal auxiliary branch generated within the multi-task network. Specifically, the detached latent vectors from the defect-type and severity branches are concatenated with

the shared representation and transformed through a fully connected layer followed by Softmax to obtain the teacher probability distribution. The student probability distribution is produced by the main health-grade branch. The KL-divergence term in Equation (7) is used as an auxiliary consistency regularization term, while the Stop-Gradient operation prevents the distillation loss from directly modifying the defect type and severity branches.

$$\hat{h} = \sigma(W_a h + b_a) \odot h \quad (6)$$

$$\mathcal{L}_{KD} = \lambda \cdot \text{KL}(P_{\text{teacher}} || P_{\text{student}}) \quad (7)$$

where P_{teacher} denotes the auxiliary teacher probability distribution generated from the detached defect type and severity latent representations and the shared feature vector, and P_{student} denotes the student probability distribution produced by the main health-grade branch.

A staged training schedule of Warmup Health → Joint Balance → Defect Focus → Health Refine dynamically adjusts both the learning rate and the task weights [47], preventing imbalance between the primary and auxiliary objectives while enhancing overall convergence stability. The four-phase training schedule was designed as a heuristic curriculum-style strategy to improve training stability under severe class imbalance and heterogeneous task difficulty, rather than being obtained through exhaustive validation-set hyperparameter search. The phase lengths were selected to allow initial stabilization of the primary health-grade branch, subsequent joint representation learning, auxiliary task strengthening, and final health grade refinement. The learning rate was gradually decreased across phases to reduce optimization oscillation, and the loss weights were assigned according to task roles: the health grade and ordinal auxiliary losses were emphasized during the warm-up and refinement stages, while the defect type and severity losses were strengthened in the middle stages to improve auxiliary semantic representation. The validation set was used only for early stopping and training monitoring, and the test set was kept untouched until final evaluation. The maximum number of training epochs was set to 80 only as an upper bound; training could terminate earlier through early stopping, and the best validation loss weights were restored for final testing. The same predefined schedule was applied consistently across the proposed model and related ablation experiments.

3.2. Deep Learning-Guided Semi-Markov Chain Model

This study couples the initial condition estimates produced by the multi-task deep learning module with a physics-constrained semi-Markov chain (SMC) model for multi-year bridge deterioration forecasting, as illustrated in Figure 6. The SMC was adopted instead of a conventional Markov chain because bridge deterioration is influenced not only by the current health state but also by the duration for which a bridge remains in that state. Therefore, explicitly modeling state dwell time helps better represent the time-dependent nature of bridge deterioration.

In the proposed framework, the deep learning module provides bridge-specific initial state probabilities and risk-related information derived from health grade, defect type, and severity prediction. These outputs are used to guide the SMC inference process. Physical constraints are further incorporated into the transition matrix to ensure that the predicted state evolution follows realistic engineering pathways, including state preservation, gradual deterioration, and possible recovery after maintenance intervention. Dwell time is modeled using a Bayesian Weibull distribution to characterize the variability of degradation duration across health states [48,49]. Under these components, the proposed model iteratively forecasts bridge health states over a three-year horizon.

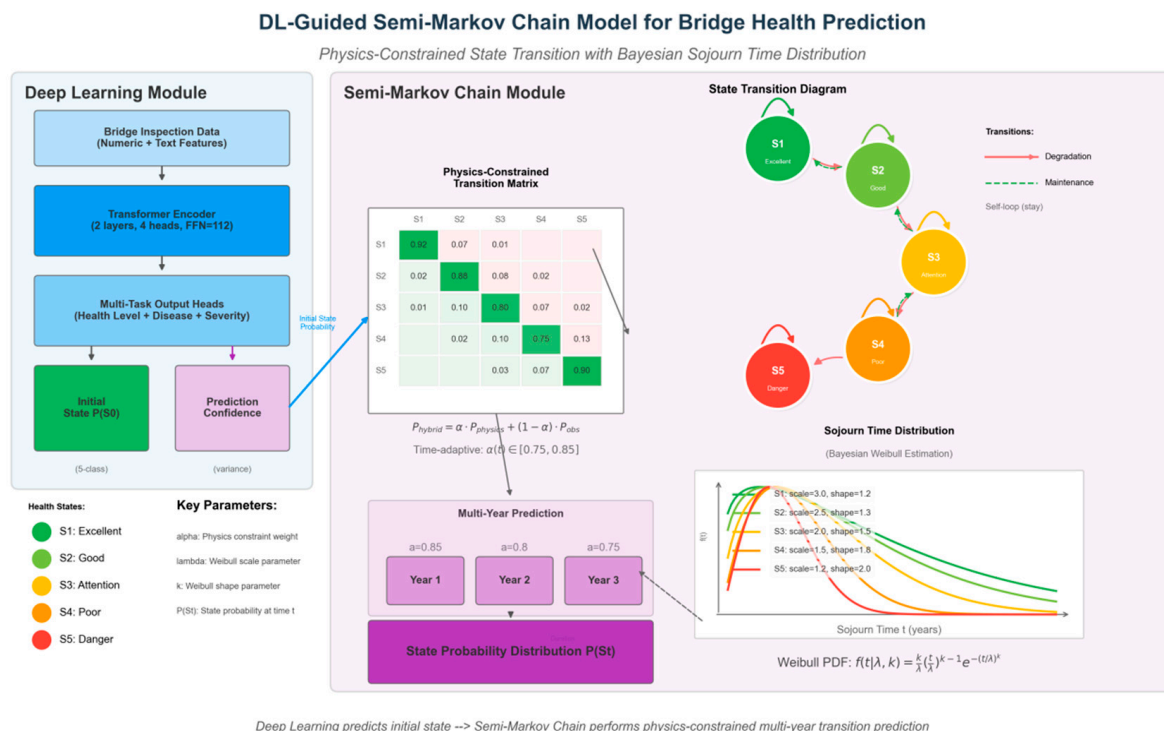


Figure 6. Architecture of the deep learning-guided semi-Markov chain model.

3.2.1. Deep Learning-Guided Initial Conditions

In addition to outputting the probabilities for health grade, defect type, and severity, the multi-task deep learning module provides intermediate features from the health branch and temporal context, which are used to initialize the SMC's state distribution, transition biases, Bayesian dwell time parameters, and prediction confidence (adaptive threshold/conservative fallback).

With the guidance provided by these deep learning-derived quantities, the SMC is equipped with individualized lifespan priors and confidence calibration prior to inference, which substantially reduces initial bias and suppresses error amplification during long-horizon rolling prediction.

3.2.2. State Dwell Time Statistics and Bayesian Weibull Fitting

To characterize the dwell properties of each health grade, this study retains only the genuinely reliable state intervals from longitudinal inspection sequences, excluding interpolated and abnormally short segments (<0.1 years), as trustworthy inputs for time decay weighting and the transition matrix. The dwell time (T_i) for any health state (S_i) is modeled using a Weibull distribution [50], as expressed in Equation (8).

$$f(t | \lambda_i, k_i) = \frac{k_i}{\lambda_i} \left(\frac{t}{\lambda_i}\right)^{k_i-1} \exp \left[-\left(\frac{t}{\lambda_i}\right)^{k_i} \right] \quad (8)$$

Here, λ_i is the scale parameter, and k_i is the shape parameter. When sufficient samples are available, conjugate Bayesian inference with a Gamma prior is used to estimate the posterior mean and variance of (λ_i) and (k_i); for limited samples (<3), empirical priors are adopted to ensure a reasonable distribution for each state. During SMC inference, the initial dwell time is either sampled directly or set to the expected value of the distribution, and in Monte Carlo simulation, the remaining dwell time is sampled conditionally based on the elapsed duration.

3.2.3. Robust Transition Matrix and Engineering-Informed Physical Constraints

Building on the reliable dwell-time statistics, each state change is weighted by its dwell duration, timestamp, and transition direction; extended dwell periods are amplified with $w_{dur} = 1 + \beta \cdot \log(1 + d)$, while short dwells receive a minimal weight. When the sample size is adequate, a temporal decay term is further incorporated to emphasize recent observations, as formulated in Equation (9).

$$w_{time} = \exp(-\gamma \Delta t) \quad (9)$$

γ denotes the temporal decay coefficient, from which the observation matrix is derived, as shown in Equation (10).

$$P_{obs}(i, j) = \frac{\sum_{k \in D_{i \rightarrow j}} w_{dur}^{(k)} w_{time}^{(k)}}{\sum_{j'} \sum_{k \in D_{i \rightarrow j'}} w_{dur}^{(k)} w_{time}^{(k)}} \quad (10)$$

$P_{obs}(i, j)$ denotes the observed probability of transitioning from state i to state j , and $D_{i \rightarrow j}$ is the set of valid samples corresponding to that transition.

To correct the artificial inflation of deterioration probabilities caused by inspection records that primarily document observed state changes, an engineering-informed transition prior was introduced. In this study, the term “physical constraint” refers to engineering-consistent state evolution rules, including state persistence, predominantly gradual deterioration, and possible recovery after maintenance intervention, rather than to a transition matrix derived directly from first-principles structural mechanics.

With the health states ordered as $S = [E, G, A, P, D]$, representing Excellent, Good, Attention, Poor, and Dangerous, respectively, the engineering-informed prior matrix is explicitly defined as shown in Equation (11):

$$P_{eng} = \begin{bmatrix} 0.920 & 0.068 & 0.0096 & 0.0024 & 0 \\ 0.018 & 0.880 & 0.0840 & 0.0144 & 0.0036 \\ 0.010 & 0.100 & 0.8000 & 0.0700 & 0.0200 \\ 0 & 0.025 & 0.1000 & 0.7500 & 0.1250 \\ 0 & 0.010 & 0.3000 & 0.0600 & 0.9000 \end{bmatrix} \quad (11)$$

The diagonal entries represent state retention probabilities, while the remaining probability mass is assigned primarily to adjacent deterioration pathways and, where applicable, to limited recovery pathways representing maintenance intervention [51]. These entries are engineering-informed calibration values rather than universal physical constants. They were introduced to correct the change-recording bias of the observed transition matrix and to preserve gradual and interpretable state evolution.

The annual hybrid transition matrix is defined as shown in Equation (12):

$$P^{(t)} = \alpha_t P_{eng} + (1 - \alpha_t) P_{obs}, \quad \alpha_t = \begin{cases} 0.85, & t = 1 \\ 0.80, & t = 2, 3 \end{cases} \quad (12)$$

Here, P_{obs} is the transition matrix estimated from the historical inspection sequences, and α_t controls the contribution of the engineering-informed prior. The first-year value is slightly higher to suppress implausible short-term state changes caused by sparse transition observations, whereas the lower values in Years 2 and 3 allow the observed deterioration information to contribute more strongly to the accumulated multi-year trajectory. These values were treated as empirically selected engineering parameters rather than as parameters derived from first principles.

Bridge age is incorporated through the following bounded piecewise correction factor shown in Equation (13):

$$g(a) = \begin{cases} 1.0, & a < 20 \\ 1 + 0.015(a - 20), & 20 \leq a < 40, g(a) \in [1.0, 1.8] \\ 1.3 + 0.01(a - 40), & a \geq 40, \end{cases} \quad (13)$$

where a denotes the bridge age at the corresponding forecasting year. Before row normalization, the age-adjusted transition entries are calculated as shown in Equation (14):

$$\tilde{p}_{ij}^{(t)} = \begin{cases} p_{ij}^{(t)} [1 - 0.1(g(a) - 1)], & j = i, \\ p_{ij}^{(t)} g(a), & j > i, \\ p_{ij}^{(t)} / g(a), & j < i, \end{cases} \quad p_{ij}^{(t,a)} = \frac{\tilde{p}_{ij}^{(t)}}{\sum_k \tilde{p}_{ik}^{(t)}} \quad (14)$$

Thus, increasing bridge age moderately reduces state retention, increases transitions toward poorer states, and decreases recovery probabilities. The 20- and 40-year breakpoints, the associated slopes, and the upper bound of 1.8 are engineering calibration settings used to impose a gradual and bounded age effect; they are not intended as universal deterioration thresholds and should be recalibrated when the framework is applied to other bridge inventories [52]. After row normalization, the resulting matrix preserves observed differences while embedding the physical constraints of “retention–slow deterioration–potential recovery.”

3.2.4. Time-Dependent Transition Probabilities and Deep Learning Correction

Based on the hybrid matrix P_{hybrid} , this study further makes the transition probabilities explicitly dependent on dwell time, forecast year, and deep learning-derived risk features. The baseline transition probability from any state S_i to candidate state S_j is denoted as p_{ij} . A temporal factor $w_{\text{time}} = 1 + \phi(h_i(d))$ is defined according to the Weibull hazard rate Equations (15) and (16), whereby high-probability transitions are slightly enhanced and low-probability transitions are substantially amplified.

$$h_i(d) = \frac{k_i}{\lambda_i} \left(\frac{d}{\lambda_i} \right)^{k_i-1} \quad (15)$$

$$w_{\text{time}} = 1 + \phi(h_i(d)) \quad (16)$$

The defect type, severity, and traffic weights are then combined to form a correction factor Equation (17).

$$c_{ij} = \underbrace{(a_1 w_{\text{dis}} + a_2 w_{\text{sev}} + a_3 w_{\text{traf}})}_{\text{Linear term (baseline contribution)}} \cdot \underbrace{[1 + \rho_1 \tanh((w_{\text{dis}} - 1)(w_{\text{sev}} - 1))]}_{\text{Nonlinear interaction (defect} \times \text{severity)}} \cdot \underbrace{[1 + \rho_2 \tanh((w_{\text{sev}} - 1)(w_{\text{traf}} - 1))]}_{\text{Nonlinear interaction (severity} \times \text{traffic)}} \quad (17)$$

The coefficients a and ρ are adaptively adjusted according to the forecast year. The final time-dependent transition probability is given by $\tilde{p}_{ij} = p_{ij} \cdot w_{\text{time}} \cdot c_{ij}$.

3.2.5. Adaptive Decision-Making and Monte Carlo Inference

To balance the accuracy and uncertainty of multi-year rolling forecasts, the model iterates year by year: the initial state and dwell time for the first year are provided by the deep learning outputs, while in subsequent years, the probability distribution is propagated

from the previous year. A threshold is set based on the predicted confidence, as shown in Equation (18).

$$\theta_t = f(t, \text{conf}_t, S_t) \quad (18)$$

Here, conf_t denotes the prediction confidence for the current year. The threshold is set highest at $t = 1$ (conservative, favoring status quo) and lowest at $t = 3$ (allowing reasonable deterioration forecasts). If a cross-grade transition lacks sufficient confidence support or the severity weight is low, a conservative fallback is triggered.

The SMC performs Monte Carlo simulation to infer future state distributions: for each run, it starts from the current bridge condition, samples a dwell time from the posterior distribution, and then recursively applies the transition probability matrix to reach the target year [53]. Repeating this process hundreds of times yields the empirical distribution, most frequent state, and weighted accuracy for each forecast year. If the Bayesian mode is enabled, the Weibull scale and shape parameters are randomly drawn prior to simulation to capture dwell time uncertainty.

3.3. Maintenance Prioritization Assessment

After obtaining deterioration predictions from the multi-task deep learning–SMC model, this study develops a risk-aware maintenance prioritization module to translate the current condition assessment and multi-year predicted deterioration states into maintenance decision support. In the prioritization process, the current health state and the predicted deterioration states over the 1-, 2-, and 3-year horizons are jointly considered, rather than using only the current condition state or a single predicted future state. This module was introduced because bridge maintenance planning should consider not only the current condition state but also future deterioration risk, defect severity, traffic exposure, prediction uncertainty, and maintenance cost. AHP and entropy weighting are combined to balance expert judgment and data-driven indicator variability, while uncertainty adjustment and cost–benefit analysis are incorporated to improve the practical relevance of the final maintenance priority scores.

3.3.1. Integration of Subjective and Objective Weights

The assessment integrates the analytic hierarchy process (AHP) and the entropy weight method to compute the comprehensive weights of criteria such as health grade, defect type, severity, and traffic volume, thereby coupling subjective and objective factors. AHP provides expert-driven subjective evaluation, while the entropy weight method yields objective weights; the final weights are obtained through a weighted fusion of the two approaches Equation (19).

$$w = \alpha \cdot w_{\text{AHP}} + (1 - \alpha) \cdot w_{\text{entropy}} \quad (19)$$

α denotes the proportion of subjective weight.

3.3.2. Standardization of States and Impact Factors

After determining the comprehensive weights, the predicted outputs related to health grade, defect type, and severity level are mapped to numerical state scores and subjected to safety-oriented min–max normalization. In the maintenance prioritization process, this mapping can be applied to both the current assessment results and the multi-year predicted deterioration states, so that the final priority score reflects not only the current bridge condition but also the future deterioration tendency. Continuous indicators such as traffic volume and bridge importance are normalized within their reference intervals, with missing values imputed as unknown or by the median. The resulting feature vector

is denoted as $x_i = [h_i, d_i, s_i, t_i, m_i]^T$, and the basic maintenance prioritization is defined by Equation (20).

$$\text{BasePriority}_i = w^T \cdot x_i \quad (20)$$

3.3.3. Cost–Benefit Extended Model

Building upon the basic maintenance prioritization, the assessment incorporates three types of costs: materials/labor, traffic interruption, and delay risk. The direct cost is calculated by multiplying the number of defects by the unit price of each component and adding a 30% labor coefficient (based on engineering experience); interruption losses are estimated from traffic volume and the expected duration of maintenance; delay costs are derived by mapping the health grade to a risk coefficient. The total cost C_i is the sum of these three components, and the final priority is computed as shown in Equation (21).

$$\text{PriorityCostRatio}_i = \frac{\text{BasePriority}_i \times \text{RiskAdjust}_i}{C_i} \quad (21)$$

The risk adjustment coefficient is derived from the subsequent uncertainty assessment.

3.3.4. Risk Perception and Uncertainty Adjustment

The assessment employs an entropy-based adjustment mechanism on the predicted probabilities to address overly optimistic maintenance prioritization for samples with high uncertainty. The entropy value for each bridge is calculated from the deep learning outputs Equation (22).

$$H_i = -\sum_{j=1}^5 p_{ij} \ln p_{ij} \quad (22)$$

It is then normalized as $\hat{H}_i \in [0, 1]$. The risk aversion coefficient ($\gamma > 1$) maps this value to a risk amplification factor Equation (23):

$$\text{RiskAdj}_i = 1 + \hat{H}_i \cdot (\gamma - 1) \quad (23)$$

This factor is then multiplied by the static state score to yield the risk-adjusted basic priority. When entropy is high, the factor increases, prompting earlier maintenance for potentially risky bridges; in the absence of probability inputs, it defaults to 1, preserving the original ranking.

3.3.5. Normalization and Traffic Weight Adjustment

After obtaining the risk-adjusted priority-to-cost ratio, the values are interval-normalized together with historical reference samples to yield a maintenance prioritization score ranging from 0 to 100. High-priority bridges are identified based on a threshold. Additionally, group-wise traffic weights are applied, and the scores are re-normalized to ensure that bridges with higher traffic volumes receive higher priority under equivalent costs.

4. Results

4.1. Experimental Environment

The experimental environment and hardware for the multi-task deep learning model include Windows 10 (Microsoft Corporation, Redmond, WA, USA), Python 3.10 (Python Software Foundation, Wilmington, DE, USA), and an Intel i5-10300H CPU (Intel Corporation, Santa Clara, CA, USA). TensorFlow version 2.17.0 (Google LLC, Mountain View, CA, USA) was used, with both training and inference performed entirely in CPU mode. All models were trained and evaluated using the same training, validation, and test partitions and preprocessing pipeline, while model-specific hyperparameters were selected according

to the characteristics of each algorithm; detailed training parameter configurations are provided in Table 7.

Table 7. Model hyperparameter settings.

Module	Key Settings
Optimizer	AdamW (initial lr = 5.4×10^{-5} , weight decay = 3.8×10^{-5} , gradient clip = 1.5)
Training	80 epochs/ batch size=128
Phase Schedule	① Warmup Health (0–24): lr = 5.4×10^{-5} , loss weights {4.8, 1.4, 0.3, 0.5} ② Joint Balance (25–44): lr = 4.6×10^{-5} , loss weights {3.4, 0.9, 1.7, 2.0} ③ Defect Focus (45–64): lr = 4.1×10^{-5} , loss weights {3.0, 0.8, 2.3, 2.2} ④ Health Refine (65–79): lr = 3.4×10^{-5} , loss weights {3.6, 1.0, 1.5, 2.0}
Loss Functions	Health grade: balanced Focal Loss ($\gamma = 1.5$, $\alpha = 0.28$, label smoothing 0.004) Defect type: cross-entropy + label smoothing 0.03 Severity: noise-robust loss (label smoothing 0.01, class_weight_factor 0.48, order_penalty 0.035) Health ordinal: Ordinal cross-entropy
Callbacks/Monitoring	Phase scheduler, EarlyStopping (patience = 60), ReduceLROnPlateau (monitor = loss, factor = 0.5, patience = 10, min_lr = 1×10^{-6}), task-balance monitor (patience = 120), train/validation subset monitor

4.2. Evaluation Metrics

For the deep learning [54] component, accuracy, balanced accuracy, macro recall, and macro/weighted F1 are consistently reported for health grade, defect type, and severity. Additionally, the kappa statistic is provided to assess agreement with manual evaluation.

The macro-average recall for health grade is used to evaluate the balance across the five health grades, as defined in Equation (24).

$$\text{MacroRecall} = \sum_{c=1}^C \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \quad (24)$$

TP_c and FN_c denote the number of true positives and false negatives for class c , respectively.

Macro and weighted F1 scores are used for defect type and severity to characterize the model's ability to recognize minority classes under class imbalance, as defined in Equations (25)–(27). The F1 score for a single class c is

$$\text{F1}_c = \frac{2 \cdot \text{TP}_c}{2 \cdot \text{TP}_c + \text{FP}_c + \text{FN}_c} = \frac{2 \cdot \text{P}_c \text{R}_c}{\text{P}_c + \text{R}_c} \left(\text{P}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \text{R}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \right) \quad (25)$$

Macro F1:

$$\text{MacroF1} = \frac{1}{C} \sum_{c=1}^C \text{F1}_c \quad (26)$$

Weighted F1:

$$\text{WeightedF1} = \sum_{c=1}^C \frac{n_c}{N} \cdot \text{F1}_c \quad (27)$$

FP_c denotes the number of false positives for class c ; P_c denotes the precision for class c ; and n_c denotes the number of samples in class c .

The kappa [55] coefficient is defined in Equation (28).

$$\kappa = \frac{\text{P}_o - \text{P}_e}{1 - \text{P}_e} \quad (28)$$

p_o denotes the observed agreement rate, and p_e is the expected agreement by chance. A kappa value closer to 1 indicates stronger alignment with manual judgment, while a value approaching 0 suggests the result is equivalent to random guessing.

For the semi-Markov chain component, probability distribution consistency, standard accuracy, and state distance-weighted accuracy are evaluated in parallel to comprehensively reflect the reliability and precision of deterioration predictions.

Probability distribution consistency: For each forecast year, the SMC predicts the most probable state and simultaneously outputs a state probability vector Equation (29).

$$\hat{p}_t = P_{\text{pred}}(S_t) \quad (29)$$

t denotes the prediction time; $\hat{p}_t$ represents the probability of each health state of the bridge in year t ; and S_t is the set of bridge health states at time t . During validation, the empirical frequency distribution of actual samples for that year is calculated Equation (30).

$$q_t = P_{\text{obs}}(S_t) \quad (30)$$

q_t represents the empirical frequency distribution of health states at time t . The distance between the predicted and observed distributions is then computed Equations (31) and (32).

$$D_t = \frac{1}{2} \sum_{c=1}^C \left| \hat{p}_{t,c} - q_{t,c} \right| \quad (31)$$

Alternatively,

$$D_{\text{KL}}\left(q_t \parallel \hat{p}_t\right) = \sum_c q_{t,c} \cdot \ln \left(\frac{q_{t,c}}{\hat{p}_{t,c}} \right) \quad (32)$$

D_t denotes the mean absolute deviation of the probability distribution at time t ; D_{KL} represents the KL divergence operator. A small discrepancy indicates that the predicted probability distribution aligns well with historical evolution, lending credibility to the risk and uncertainty estimates; conversely, a larger discrepancy suggests that the transition matrix or dwell time parameters require recalibration.

Standard accuracy: A prediction is scored as 1 if the predicted state matches the true state and is otherwise scored 0; the final accuracy is the average across all samples, as shown in Equation (33).

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N 1\{\hat{s}_i = s_i\} \quad (33)$$

$\hat{s}_i$ denotes the most probable state for the i -th prediction, and s_i is the true state.

For state distance-weighted accuracy, the five health states are ordered ($\text{rank}(\text{excellent}) = 0, \dots, \text{rank}(\text{Dangerous}) = 4$), and the score for a single sample is defined as in Equation (34).

$$\text{wAcc}(\hat{s}, s) = \max(0, 1 - 0.2 \cdot |\text{rank}(\hat{s}) - \text{rank}(s)|) \quad (34)$$

The overall weighted accuracy is then defined as in Equation (35):

$$\text{wAcc} = \frac{1}{N} \sum_{i=1}^N \text{wAcc}(\hat{s}_i, s_i) \quad (35)$$

For maintenance prioritization, the risk-adjusted priority-to-cost ratio is normalized to the range 0–100 to serve as the final maintenance prioritization score, and bridges with scores above a threshold of 70 are classified as high-priority.

To improve evaluation reliability under imbalanced data conditions, overall metrics, class-sensitive metrics, sliding-window validation, and historical replay were jointly used to assess model performance from different perspectives.

4.3. Multi-Task Deep Learning Performance Evaluation

After 80 epochs of training with unified hyperparameters, the multi-task deep learning model simultaneously outputs predictions for health grade, defect type, and severity, along with corresponding metrics such as kappa coefficients and macro-average recall. Tables 8–16 summarize the performance of the three tasks on the training, validation, and test sets, respectively. The overall and task-specific loss curves are presented in Figure 7.

Table 8. Health grade model training results.

Dataset	Accuracy	Macro Recall	Macro F1	Kappa	Loss
Training	0.9577	—	—	—	0.0703
Validation	0.9251	0.5699	0.3906	0.8397	0.2342
Test	0.9520	0.5649	0.5671	0.8720	—

Table 9. Health grade classification performance on the validation set.

Health Grade (Validation Report)	Precision	Recall	f1-Score	Support
Poor	N/A	N/A	N/A	0
Excellent	0.0159	1.0000	0.0313	7
Dangerous	N/A	N/A	N/A	0
Attention	1.0000	0.9253	0.9612	4446
Good	1.0000	0.9243	0.9606	1941
weighted avg	0.9989	0.9251	0.9600	6394

Note: Support denotes the number of real samples in the validation subset. N/A indicates that no real sample was available for the corresponding class; therefore, its precision, recall, and F1-score could not be empirically evaluated.

Table 10. Health grade classification performance on the test set.

Health Grade (Test Report)	Precision	Recall	f1-Score	Support
Poor	N/A	N/A	N/A	0
Excellent	0.9286	1.0000	0.9630	13
Dangerous	N/A	N/A	N/A	0
Attention	0.9937	0.8271	0.9028	1533
Good	0.9438	0.9972	0.9698	4229
weighted avg	0.9570	0.9520	0.9520	5775

Note: Support denotes the number of real samples in the test subset. N/A indicates that no real sample was available for the corresponding class; therefore, its precision, recall, and F1-score could not be empirically evaluated.

Table 11. Defect type model training results.

Dataset	Accuracy	Balanced Accuracy	Macro F1	Loss
Training	0.9936	—	—	0.1753
Validation	0.9992	0.9961	0.7465	0.1963
Test	0.9967	0.9980	0.7492	—

Table 12. Defect type classification performance on the validation set.

Defect Type (Validation Report)	Precision	Recall	f1-Score	Support
Functional component	0.9938	0.9938	0.9938	321
No defect	N/A	N/A	N/A	0
Structural component	1.0000	0.9997	0.9998	5873
Surface defect	0.9900	0.9950	0.9925	200
weighted avg	0.9994	0.9992	0.9993	6394

Note: Support denotes the number of real samples in the validation subset. N/A indicates that no real sample was available for the corresponding class; therefore, its precision, recall, and F1-score could not be empirically evaluated.

Table 13. Defect type classification performance on the test set.

Defect type (Test Report)	Precision	Recall	f1-Score	Support
Functional component	1.0000	0.9980	0.9990	1015
No defect	N/A	N/A	N/A	0
Structural component	0.9995	0.9961	0.9978	4362
Surface defect	1.0000	1.0000	1.0000	398
weighted avg	0.9997	0.9967	0.9982	5775

Note: Support denotes the number of real samples in the test subset. N/A indicates that no real sample was available for the corresponding class; therefore, its precision, recall, and F1-score could not be empirically evaluated.

Table 14. Severity model training results.

Dataset	Accuracy	Balanced Accuracy	Macro F1	Loss
Training	0.9863	—	—	0.0994
Validation	0.9573	0.9513	0.5524	0.1277
Test	0.9972	0.9801	0.6580	—

Table 15. Severity classification performance on the validation set.

Severity (Validation Report)	Precision	Recall	f1-Score	Support
Moderate	0.5311	0.9446	0.6800	317
Mild	0.9971	0.9579	0.9771	6077
Severe	N/A	N/A	N/A	0
weighted avg	0.9747	0.9573	0.9629	6394

Note: Support denotes the number of real samples in the validation subset. N/A indicates that no real sample was available for the corresponding class; therefore, its precision, recall, and F1-score could not be empirically evaluated.

Table 16. Severity classification performance on the test set.

Severity (Test Report)	Precision	Recall	f1-Score	Support
Moderate	0.9907	0.9607	0.9755	331
Mild	0.9976	0.9994	0.9985	5444
Severe	N/A	N/A	N/A	0
weighted avg	0.9972	0.9972	0.9972	5775

Note: Support denotes the number of real samples in the test subset. N/A indicates that no real sample was available for the corresponding class; therefore, its precision, recall, and F1-score could not be empirically evaluated.

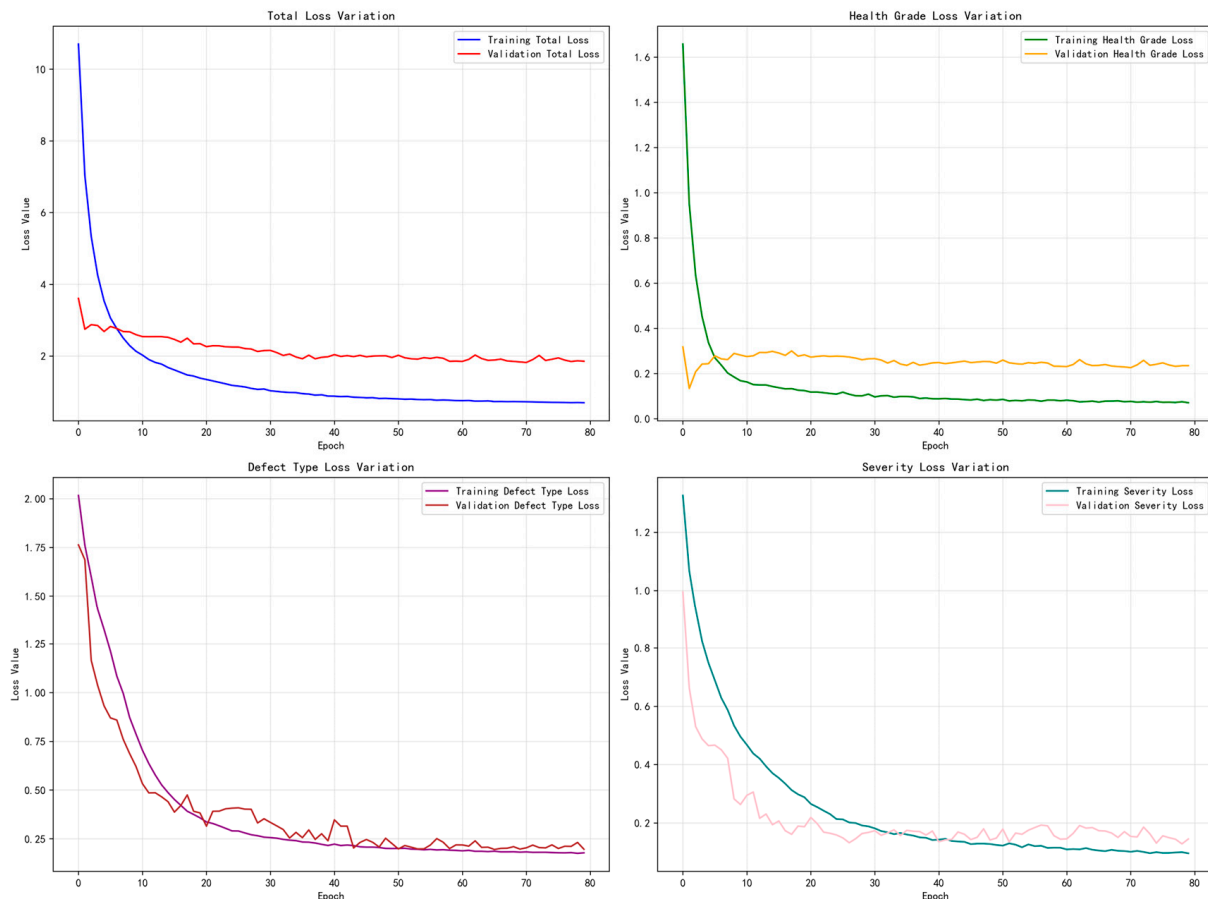


Figure 7. Total task and individual task loss curves.

Before interpreting the evaluation results, it should be noted that the validation and test sets were preserved as real inspection data and therefore retained the inherent class imbalance of the original dataset. As shown in Tables 4–6, several rare categories are absent or extremely underrepresented in the validation and/or test sets, including Poor and Dangerous health grades, the No defect category, and the Severe severity level. Although real Poor records were available in the original dataset, no Poor samples were included in the validation or test subsets after bridge-grouped splitting, whereas no real Dangerous samples were available in the original dataset. Therefore, the reported overall accuracy and weighted metrics mainly reflect the model performance on the observed and sufficiently represented categories and should be interpreted together with macro-level and class-wise metrics. For categories absent from the validation or test sets, the corresponding predictive performance cannot be fully verified and should not be regarded as validated model capability. In particular, because no real Dangerous samples were available in the original dataset or in the real validation/test subsets, Dangerous-class results should be interpreted only as the model’s exposure to rule-constrained safety-critical training scenarios, not as independently validated prediction capability for real Dangerous bridges. More generally, synthetic rare-class samples were used exclusively for training-stage class balancing and cannot substitute for empirical validation using real rare-state inspection records.

From the training results and loss curves, the model achieves primary convergence in the early epochs. During training, the health grade accuracy reaches approximately 0.95, while defect type and severity achieve about 0.99, with corresponding losses below 0.18. On the validation set, the health grade accuracy drops to around 0.92 due to the absence or scarcity of samples in the Poor, Excellent, and Dangerous states, yielding a macro recall of about 0.57; however, the metrics for the dominant states (good and attention) approach

near-perfect levels, and kappa is approximately 0.84. For defect type and severity, accuracy and balanced accuracy are about 0.99 and 0.95, respectively; the slightly higher validation loss may be partly associated with the severe class imbalance and the absence of several rare categories from the validation subset. Apart from the scarcity of defect-free samples, the accuracy, recall, and F1 for the three defect types are all around 0.99. The weighted average for severity reaches 0.97; although the moderate class, affected by sample scarcity, achieves only 0.53 precision, its recall remains at 0.94. This result further indicates that class-wise metrics should be considered when interpreting model performance under imbalanced data conditions.

On the test set, health grade accuracy reaches 0.95, with accuracy for the dominant states exceeding 0.92, recall above 0.82, and F1 above 0.9; kappa is approximately 0.87. For defect type and severity, the overall metrics remain high for the observed and sufficiently represented categories. Overall, these results indicate that the proposed multi-task model performs well for the dominant and sufficiently represented categories in the real inspection dataset, particularly the Good and Attention health states, common defect types, and mild-to-moderate severity levels. However, because several rare categories are absent or extremely underrepresented in the validation and test sets, the high overall accuracy and weighted metrics should not be interpreted as evidence of fully validated performance for all bridge condition states. Therefore, the model is currently more appropriate as a decision support tool for common bridge condition assessment, while further validation using larger datasets containing more rare safety-critical samples is required.

4.4. Semi-Markov Chain Performance Evaluation

To validate the deterioration forecasting capability of the deep learning-guided SMC model, two complementary validation protocols were adopted: time-sliding-window cross-validation and full historical replay. The sliding-window validation was used to examine whether the model maintains stable forward prediction performance under different historical training windows, while the historical replay was used to evaluate whether the learned transition and dwell-time mechanisms can reproduce observed bridge evolution patterns. Only authentic bridge inspection records were used in this validation stage to ensure that the SMC predictions were evaluated against real historical data. The prediction results are presented in Tables 17 and 18, while Table 19 summarizes the model performance validation outcomes.

Based on the deep learning-derived initial states of 58 bridges, SMC projections indicate that the dominant state over the next three years will remain Good, with a slight increase in the proportion of Attention states. The proportions of Poor and Dangerous states remain consistently low. The average probabilities further suggest a gradual transition toward Attention or Poor states, while overall conditions remain maintainable.

Table 17. SMC most probable state distribution (58 real bridges).

Year	Excellent	Good	Attention	Poor	Dangerous
Year 1	0.155	0.724	0.103	0.017	0.000
Year 2	0.155	0.707	0.121	0.000	0.017
Year 3	0.000	0.845	0.138	0.000	0.017

Table 18. SMC average probability distribution (reference values).

Year	Excellent	Good	Attention	Poor	Dangerous
Year 1	0.122	0.553	0.273	0.040	0.012
Year 2	0.089	0.455	0.354	0.072	0.030
Year 3	0.071	0.412	0.371	0.097	0.049

Table 19. SMC model performance validation results.

Validation Protocol	Metric	Standard Accuracy	Weighted Accuracy	
Sliding-window CV (overall)	3-fold average	0.909	0.982	Train/validation windows split by time
	Split date 2020-11-06	0.929	0.986	474 records, 2 bridges for training
	Split date 2021-07-20	0.918	0.984	3714 records, 10 bridges for training
	Split date 2021-12-02	0.900	0.980	5738 records, 17 bridges for training
Historical replay (overall)	Full reconstruction	0.954	0.991	
	Year 1	0.983	0.997	1 year ahead
	Year 2	0.965	0.993	2 years ahead
	Year 3	0.913	0.983	3 years ahead

Under time-sliding-window cross-validation, the standard and weighted accuracies are 0.909 and 0.98, respectively, with both metrics exceeding 0.90 and 0.98 across all split points, indicating that temporal weighting and physical constraints remain stable across different windows. In the historical data replay validation, the standard accuracy reaches 0.954, and the weighted accuracy approaches perfection. Year-by-year, the standard accuracies for years 1–3 are 0.983, 0.965, and 0.913, with weighted accuracies all above 0.983, demonstrating high engineering prediction reliability even as the forecast horizon extends. Furthermore, when samples are grouped by time since the latest inspection (<1 year, 1–3 years, 3–5 years, >5 years), the difference in standard accuracy between recent and >5-year samples is less than 0.05, indicating that temporal decay weighting does not introduce systematic bias and that predictions remain robust across different time spans.

To examine the influence of the engineering-prior weights, a grid-based sensitivity analysis was conducted using the same deep learning prediction outputs and fixed model settings as those adopted in the main analysis. The three annual weights, α_1 , α_2 , and α_3 , were independently varied from 0.70 to 0.90 at intervals of 0.05, yielding 125 combinations. Only the engineering-prior weights were varied, while the observed transition matrix, engineering-informed prior matrix, dwell-time distributions, bridge-age correction, historical replay samples, and all other SMC settings were kept unchanged. Representative results are summarized in Table 20.

Table 20. Sensitivity analysis of representative engineering-prior weight schedules.

α_1	α_2	α_3	Historical Standard Accuracy	Historical Weighted Accuracy	Historical Replay Distribution MAE	Forecast Retention Rate	Year-3 Deterioration Rate
0.70	0.70	0.70	−0.5130	−0.1038	0.1491	27.59%	87.93%
0.75	0.75	0.75	−0.0899	−0.0180	0.1382	55.75%	63.79%
0.80	0.80	0.80	−0.0116	−0.0023	0.1264	86.21%	18.97%
0.85	0.80	0.80	0.0000	0.0000	0.1182	89.08%	18.97%
0.85	0.85	0.85	+0.0319	+0.0064	0.1128	94.25%	13.79%
0.90	0.90	0.90	+0.0319	+0.0064	0.0969	98.85%	1.72%

Note: Accuracy changes are absolute differences relative to the adopted schedule ($\alpha_1, \alpha_2, \alpha_3$) = (0.85, 0.80, 0.80) in the same sensitivity analysis setting. The formal standard and weighted accuracies of the adopted model are 0.954 and 0.991, respectively, as reported in Table 19. Historical replay distribution MAE measures the difference between predicted and observed historical state distributions. Forecast retention rate and Year 3 deterioration rate denote, respectively, the proportions of unchanged bridge-year predictions and bridges predicted to be in poorer states by Year 3.

The sensitivity results show a clear trade-off between historical replay performance and predicted deterioration. Lower prior weights produced overly aggressive deterioration, whereas higher weights improved historical fitting but caused excessive state retention; for example, the (0.90, 0.90, 0.90) schedule retained 98.85% of bridge-year states and predicted Year 3 deterioration for only 1.72% of bridges. The adopted schedule, (0.85, 0.80, 0.80), maintained strong historical performance while yielding a retention rate of 89.08% and a Year 3 deterioration rate of 18.97%. No tested schedule produced abrupt transitions spanning two or more condition grades. Therefore, the adopted schedule was retained as an engineering-informed balance between historical consistency and gradual multi-year deterioration rather than as the setting that maximizes a single fitting metric.

4.5. Comparative Experiments

4.5.1. Comparative Experiments for the Multi-Task Deep Learning Model

To evaluate the comparative performance of the proposed multi-task deep learning model, representative baselines were selected from different modeling paradigms. Logistic Regression was used as a simple linear classification baseline; Temporal CNN was used as a convolution-based temporal modeling baseline; LSTM [56] was used as a sequential deep learning baseline for temporal inspection records; Transformer was used as an attention-based sequence modeling baseline; and Random Forest [57] and XGBoost [58] were included as representative traditional machine learning baselines for nonlinear classification on structured engineering data. In addition, a non-Transformer multi-task network was included to evaluate multi-task learning without Transformer-based temporal fusion. All experiments used the same exported data, preprocessing settings, and train/validation/test splits. The detailed hyperparameter settings of the comparative baseline models are provided in Appendix A Table A1, while those of the proposed model are reported in Table 7. A single-task variant was also evaluated for each prediction task to isolate the contribution of multi-task learning itself, where each variant used the same input features, shared backbone, and task-specific training settings as the proposed model, with the other task heads and cross-task information-sharing mechanisms removed. The comparative results are summarized in Tables 21–23.

Table 21. Comparative experimental results for health grade.

Model	Accuracy	Kappa	Macro Recall	Macro F1
LSTM	0.7653	0.1759	0.2811	0.2709
Transformer	0.7948	0.2892	0.3089	0.3147
Random Forest	0.9198	0.7811	0.5453	0.5634
XGBoost	0.7601	0.3685	0.5241	0.5130
Logistic Regression	0.7962	0.3462	0.4602	0.4388
Temporal CNN	0.7755	0.2065	0.2333	0.2310
Non-Transformer MTL	0.8074	0.3551	0.4277	0.4417
Single-task variant	0.9247	0.7792	0.5445	0.5258
Ours	0.9520	0.8720	0.5649	0.5671

Table 22. Comparative experimental results for defect type.

Model	Accuracy	Balanced Accuracy	Macro F1	Kappa
LSTM	0.6850	0.3447	0.2544	0.0117
Transformer	0.7348	0.3263	0.2872	−0.0227
Random Forest	0.7437	0.3307	0.2904	−0.0091
XGBoost	0.7438	0.3632	0.3519	0.0405
Logistic Regression	0.8997	0.8993	0.6996	0.7991
Temporal CNN	0.7487	0.4032	0.3979	0.1975
Non-Transformer MTL	0.9392	0.7151	0.7443	0.8275
Single-task variant	0.9638	0.9640	0.7168	0.9555
Ours	0.9967	0.9980	0.7492	0.9917

Table 23. Comparative experimental results for severity.

Model	Accuracy	Balanced Accuracy	Macro F1	Kappa
LSTM	0.9962	0.9685	0.6572	0.9604
Transformer	0.9522	0.6256	0.3781	0.3623
Random Forest	0.9968	0.9783	0.6501	0.9583
XGBoost	0.9943	0.9753	0.6467	0.9535
Logistic Regression	0.9942	0.9515	0.6482	0.9447
Temporal CNN	0.9950	0.9717	0.5774	0.8949
Non-Transformer MTL	0.9497	0.8740	0.6079	0.6168
Single-task variant	0.9546	0.9132	0.6145	0.9248
Ours	0.9972	0.9801	0.6580	0.9636

The comparative results show that the proposed multi-task model achieves better or competitive performance compared with the selected baseline models. For health-grade prediction, the proposed model clearly outperforms LSTM, Transformer, Random Forest, XGBoost, Logistic Regression, Temporal CNN, the non-Transformer multi-task network, and the single-task variant in terms of overall performance. In particular, compared with the single-task health-grade model, the multi-task model improves accuracy from 0.9247 to 0.9520, macro F1 from 0.5258 to 0.5671, and kappa from 0.7792 to 0.8720, indicating that defect type and severity provide useful auxiliary information for health-state assessment through cross-task representation sharing. For defect-type prediction, the multi-task model achieves slightly better performance than the corresponding single-task variant, with accuracy increasing from 0.9638 to 0.9967 and kappa from 0.9555 to 0.9917. For severity prediction, the proposed multi-task model remains highly competitive, achieving an accuracy of 0.9972, a balanced accuracy of 0.9801, a macro F1-score of 0.6580, and a kappa coefficient of 0.9636, while jointly producing all three outputs within a unified framework. Overall, these results suggest that the proposed model effectively integrates multi-source features and cross-task information, especially improving the primary health-grade prediction task. However, considering the class imbalance and the absence or scarcity of several rare categories in the validation and test sets, the comparative results should be interpreted mainly for the observed and sufficiently represented categories.

4.5.2. Comparative Experiments for the Semi-Markov Chain Model

To evaluate the deterioration forecasting performance of the proposed deep learning-guided SMC model, two stochastic baseline models were selected: a conventional Markov chain and a standard semi-Markov chain [59]. The conventional Markov chain was used as a memoryless transition baseline, while the standard semi-Markov chain was used as a dwell-time-based baseline without deep learning guidance or physical transition constraints. These two baselines were selected to examine whether the proposed integration of deep learning initial states, dwell-time modeling, and engineering constraints improves multi-year deterioration prediction. All three models were validated using the same time-sliding windows constructed from historical inspection sequences to ensure a consistent comparison. The comparative results are presented in Table 24.

The comparative results show that the proposed DL-SMC model achieves higher standard accuracy and weighted accuracy than the conventional Markov chain and the standard semi-Markov chain across the evaluated 1- to 3-year forecasting horizons. The conventional Markov chain tends to preserve the current bridge state and shows limited ability to represent long-term deterioration changes. The standard semi-Markov chain improves dwell-time representation, but its prediction stability decreases as the forecast horizon extends. By incorporating deep learning-derived initial states, probabilistic dwell-

time modeling, and physical transition constraints, the proposed DL-SMC model provides more stable long-horizon forecasting performance under the current validation setting.

Table 24. Comparative experimental results for the semi-Markov chain.

Model	Forecast Horizon	Accuracy	Weighted Acc.
DL-SMC (ours)	Year 1	0.9826	0.9965
	Year 2	0.9652	0.9930
	Year 3	0.9130	0.9826
MC	Year 1	0.9304	0.9861
	Year 2	0.9304	0.9861
	Year 3	0.7478	0.9478
SMC	Year 1	0.9810	0.9961
	Year 2	0.9739	0.9948
	Year 3	0.7652	0.9530

4.6. Ablation Experiments

4.6.1. Ablation Experiments for the Multi-Task Deep Learning Model

To further validate the contribution of each design module, four ablation experiments are conducted using the main experiment D0 data and preprocessing: D1 removes text input, D2 removes cross-task context, D3 removes distillation (kd), and D4 removes the ordinal relationship auxiliary head. The ablation results are shown in Table 25.

Table 25. Ablation experimental results for the multi-task deep learning model.

Ep	Health Acc.	Health Kappa	Health Macro Recall	DefectAcc.	Defect Kappa	Defect MacroF1	Severity Acc.	Severity Kappa	Severity MacroF1
D0	0.9520	0.8720	0.5649	0.9967	0.9917	0.7492	0.9972	0.9636	0.6580
D1	0.8109	0.3805	0.2663	0.7406	−0.0176	0.2868	0.9543	0.9167	0.6297
D2	0.9052	0.8196	0.5016	0.9951	0.9877	0.7429	0.9460	0.6575	0.4929
D3	0.9629	0.8887	0.3952	0.9944	0.9859	0.7353	0.9096	0.5508	0.4947
D4	0.7558	0.1954	0.2054	0.9958	0.9912	0.4987	0.9481	0.9104	0.6097

The ablation results indicate that text input and the ordinal auxiliary head are critical for performance improvement. Removing text input leads to a significant decline in accuracy, kappa, and macro metrics for both health grade and defect tasks, while eliminating the ordinal relationship head nearly renders the health task ineffective. The cross-task context primarily enhances the health and severity branches, and although knowledge distillation (KD) slightly reduces the accuracy of some dominant classes, it significantly improves the macro recall and F1 for health grade and severity, thus enhancing class balance. Overall, text input and ordinal constraints contribute the most, while cross-task context and KD support health classification and imbalance control, respectively.

4.6.2. Ablation Study of Multiscale Temporal Aggregation

To evaluate the contribution of each temporal aggregation strategy in the multiscale semantic feature fusion layer, five aggregation variants were compared under the same exported data, preprocessing settings, train/validation/test split, and training protocol. These variants are attention pooling only (E0), recent timestep only (E1), short-term average only (E2), global average only (E3), and the full multiscale fusion strategy used in the proposed model (E4). The results are summarized in Table 26.

Table 26. Ablation results of temporal aggregation strategies.

Ep	Health Acc.	Health Kappa	Health Macro Recall	DefectAcc.	Defect Kappa	Defect MacroF1	Severity Acc.	Severity Kappa	Severity MacroF1
E0	0.9276	0.8053	0.5457	0.9867	0.9816	0.7362	0.9555	0.4416	0.6071
E1	0.7165	0.4953	0.2463	0.6708	−0.0721	0.2041	0.9048	0.1697	0.3754
E2	0.7089	0.5137	0.2451	0.7061	0.1342	0.2973	0.8860	0.1313	0.3379
E3	0.7197	0.5608	0.2468	0.7174	−0.0210	0.2235	0.8781	0.1194	0.3356
E4	0.9520	0.8720	0.5649	0.9967	0.9917	0.7492	0.9972	0.9636	0.6580

Note: All variants used the same exported data, preprocessing settings, train/validation/test split, and evaluation metrics. The only changed component was the temporal aggregation strategy in the multiscale semantic feature fusion layer.

As shown in Table 26, attention pooling achieved the strongest performance among the four single-aggregation variants, especially for health grade prediction and defect type prediction. In contrast, using only the recent timestep, short-term average, or global average caused clear performance degradation, indicating that a single fixed temporal summary is insufficient to represent bridge inspection sequences. The full multiscale fusion strategy achieved the best health-grade accuracy and kappa and the highest severity accuracy, while maintaining competitive defect-type performance. These results suggest that adaptive temporal weighting is the most informative single strategy, and that combining attention-based, recent-state, short-term, and global temporal summaries provides a more comprehensive temporal representation for the proposed multi-task framework.

4.6.3. Ablation Studies for the Semi-Markov Chain Model

To quantify the contributions of deep learning guidance and physical constraints, five ablation experiments were conducted on the validation set from the comparative study: SMC-1 removed physical constraints, SMC-2 removed deep learning guidance, SMC-3 adjusted the temporal decay weight α to 0.8, SMC-4 fixed the Weibull shape parameter k to 1 (exponential distribution), and SMC-5 used point estimates for dwell time. Overall standard and weighted accuracy were used for evaluation. The ablation results are presented in Table 27.

Table 27. Ablation experimental results for the SMC.

Experiment	Standard Accuracy	Weighted Accuracy
Ours	0.950	0.990
SMC-1	0.887	0.976
SMC-2	0.901	0.980
SMC-3	0.945	0.987
SMC-4	0.944	0.986
SMC-5	0.887	0.975

The ablation results show that removing physical constraints, deep learning guidance, or using point estimates for dwell time leads to a significant decrease in SMC model prediction accuracy, whereas merely adjusting the temporal decay weight or fixing the Weibull shape parameter k results in only a slight performance decline. This indicates that the first three factors play a central role, while the latter two have a secondary impact.

4.7. Maintenance Prioritization Assessment Results

Based on the outputs of the multi-task deep learning model and the SMC deterioration forecasts, the proposed maintenance prioritization module was applied to rank the

58 bridges. The purpose of this assessment was to examine whether current condition assessment and future deterioration prediction could be translated into actionable maintenance priority information. The AHP consistency check yielded a CR value of 0.015, indicating acceptable consistency of the subjective weights. To present the state information used in maintenance prioritization more explicitly, the current condition states and the multi-year predicted deterioration states of all 58 bridges were further summarized. For each bridge and each forecasting horizon, the predicted state was defined as the state with the highest transition probability. As shown in Table 28, the prioritization analysis covers both the current condition's state and the 1-, 2-, and 3-year predicted states for all bridges, rather than relying only on a single current state or a single future prediction.

Table 28. Distribution of current and 1–3-year predicted condition states for all 58 bridges.

Condition State	Current State	Year-1 Predicted State	Year-2 Predicted State	Year-3 Predicted State
Excellent	9	9	9	0
Good	42	42	41	49
Attention	7	6	7	8
Poor	0	1	0	0
Dangerous	0	0	1	1
Total	58	58	58	58

Note: For each bridge and each forecasting horizon, the predicted condition state refers to the state with the highest transition probability.

The results show that the bridge inventory is currently concentrated mainly in the Excellent, Good, and Attention states. Over the three-year forecasting horizon, the distribution gradually shifts toward lower condition levels, with the Excellent state decreasing and the Good and Attention states becoming dominant. This does not indicate abrupt large-scale deterioration; rather, it suggests that some bridges initially classified as Excellent are predicted to move into the Good state, while only a small number of bridges are predicted to reach Poor or Dangerous states. These current and predicted states are then combined with deterioration probability, traffic exposure, prediction uncertainty, defect severity, and maintenance cost to calculate the final maintenance priority score. Therefore, the proposed prioritization module reflects both the present condition distribution and the predicted future deterioration tendency, supporting preventive maintenance planning before severe deterioration becomes dominant in the observed bridge inventory.

Based on the above current and predicted state distributions, we further conducted a rolling evaluation of the 58 bridges using SMC forecasts for the next 1–3 years. The results show that the maintenance prioritization range remains 0.01–100 over the next three years, with mean values of 13.25, 13.24, and 14.00 and the number of high-priority bridges being 6, 6, and 7, respectively. Notably, bridges 27, 18, 19, 26, and 48 see their maintenance prioritization surge to 86–100, allowing for early identification of bridges with rapid risk accumulation. Sensitivity analysis indicates that the average fluctuation of AHP, entropy, and combined weights is approximately 24% and 100%; when the risk aversion coefficient is switched among 1.0, 1.2, and 1.5, the fluctuation remains below 7%; and when maintenance costs vary by $\pm 20\%$, the fluctuation is about 21%. These results suggest that the prioritization outcomes are relatively stable under different risk preferences, while weight and cost assumptions still have noticeable effects on the final ranking. Therefore, the proposed module should be used as an interpretable decision support reference together with field inspection and engineering judgment. To further assess the credibility and engineering rationale of the maintenance prioritization results, the proposed rankings

were compared with three complementary baseline methods: condition-based ranking, Entropy–TOPSIS, and forecast risk-based ranking. Spearman’s ρ , Kendall’s τ , and the Top-5 and Top-10 overlap rates were used to quantify ranking consistency, as summarized in Table 29.

Table 29. Ranking consistency between the proposed maintenance prioritization method and baseline methods.

Forecast_Year	Baseline_Method	Spearman’s ρ	Kendall’s τ	Top5_Overlap_Rate	Top10_Overlap_Rate
1	Condition-based	0.8503	0.7551	0.6	0.7
2	Condition-based	0.8511	0.7552	0.8	0.8
3	Condition-based	0.8531	0.7587	0.8	0.8
1	Entropy–TOPSIS	0.8117	0.7660	0.8	0.8
2	Entropy–TOPSIS	0.8095	0.7632	0.8	0.8
3	Entropy–TOPSIS	0.8050	0.7589	0.8	0.7
1	Forecast risk-based	0.9090	0.8170	0.8	0.9
2	Forecast risk-based	0.9251	0.8838	1	1
3	Forecast risk-based	0.9470	0.8863	1	1

Note: The condition-based baseline uses the current bridge condition, the forecast risk-based baseline uses predicted deterioration risk, and Entropy–TOPSIS represents a conventional multi-criteria decision-making method. The Top-k overlap rate denotes the proportion of shared bridges in the two Top-k priority lists.

The proposed rankings showed consistently high agreement with all three baseline methods, with Spearman’s ρ ranging from 0.8050 to 0.9470 and Top-10 overlap rates ranging from 0.70 to 1.00. The strongest agreement was observed with the forecast risk-based baseline, including complete Top-10 overlap in Years 2 and 3. These results support the engineering reasonableness of the proposed rankings, while the proposed framework further integrates current condition, multi-year deterioration, prediction uncertainty, traffic exposure, and cost-effectiveness. In addition, the Top-10 high-priority bridges were qualitatively reviewed by a senior bridge engineering professor with extensive experience in bridge inspection and maintenance decision-making and were considered generally consistent with engineering judgment.

5. Discussion

The results show that the proposed framework has practical potential for bridge condition assessment, deterioration forecasting, and maintenance prioritization using official inspection records. The multi-task learning module jointly predicts health grade, defect type, and severity, while the physics-constrained semi-Markov model links current assessment with multi-year deterioration forecasting. The maintenance prioritization module further translates prediction results into decision support information for bridge management.

The proposed framework should be regarded as a decision support tool rather than a replacement for manual inspection or expert engineering judgment. It can assist bridge management agencies in identifying bridges that may require closer attention, earlier intervention, or more detailed inspection. The framework is particularly suitable for routine management of common and sufficiently represented bridge condition states, such as Good and Attention health grades, common defect types, and mild-to-moderate severity levels.

Several limitations should be acknowledged. First, the dataset includes 58 large and extra-large bridges from Henan Province, and the severe class imbalance prevented empirical validation of the Poor, Dangerous, No-defect, and Severe categories. Therefore, the reported performance mainly reflects the sufficiently represented classes, and broader generalizability requires validation using larger and external bridge inventories. Second, prediction performance may be affected by the quality and completeness of inspection records, while longitudinal text, real-time monitoring, environmental exposure, traffic variation, and detailed maintenance histories were not fully incorporated. Third, the engineering-informed transition prior, annual prior weights, and bridge age correction pa-

rameters are inventory-specific calibration settings rather than universal physical constants and should be recalibrated before application to other bridge inventories. Future work will address these limitations through larger cross-regional datasets, improved rare-state coverage, richer multimodal inputs, and external validation, including comparisons with independent expert assessments and actual maintenance intervention records.

Based on the results of this study, several practical suggestions can be drawn for bridge maintenance management. First, bridge agencies should establish standardized inspection databases that integrate bridge-level condition ratings, defect-level descriptions, traffic exposure, and temporal inspection information, because such integrated records are essential for data-driven condition assessment and deterioration forecasting. Second, bridges with high predicted deterioration risk, high traffic exposure, or high uncertainty should be prioritized for follow-up inspection and preventive maintenance rather than relying only on the current condition grade. Third, rare but safety-critical conditions should be continuously collected and independently verified in future bridge inventories, because the present dataset contains limited real observations of extreme deterioration states. Finally, the proposed framework should be recalibrated according to local bridge inspection standards before being applied to other regions or countries, and external validation using broader bridge inventories is recommended.

6. Conclusions

This study proposed an integrated framework for bridge condition assessment, deterioration forecasting, and maintenance prioritization by combining a multi-task deep learning model with a physics-constrained semi-Markov chain model. The framework links current inspection-based condition assessment with multi-year deterioration prediction and maintenance decision support.

The main research conclusions are as follows:

1. The multi-task deep learning model simultaneously outputs health grade, defect type, and severity, achieving test accuracies of 95.20%, 99.67%, and 99.72%, respectively, with a health-grade kappa of 0.8720. Compared with baseline models, it shows improved accuracy and stability for dominant and sufficiently represented categories, particularly Good and Attention health states, common defect types, and mild-to-moderate severity levels. However, its performance for rare safety-critical categories remains insufficiently verified due to their absence or scarcity in the validation and test sets.
2. The deep learning-guided semi-Markov chain model uses deep learning outputs as the initial state and risk weights, incorporating Bayesian dwell time distributions, temporal decay, and physical constraints. The overall and weighted accuracies for three-year forecasts reach 95.4% and 99.1%, respectively, both outperforming traditional Markov chain and semi-Markov chain models.
3. The maintenance prioritization assessment system, built upon the results of the above two models, enables ranking of bridge maintenance priorities for both current and future (1–3 years) states, identifying both currently high-priority bridges and key objects with significant future priority increases, thus providing a basis for bridge maintenance strategy formulation.

In summary, the proposed method shows practical potential for bridge health assessment, multi-year deterioration prediction, and maintenance prioritization. By integrating current bridge condition evaluation, future deterioration forecasting, and management-oriented decision support into a unified framework, it provides useful decision support value for refined bridge maintenance management and infrastructure lifecycle planning. Nevertheless, the reported overall performance should be interpreted as applying primarily

to the classes represented in the real validation and test subsets. Performance on the Poor, Dangerous, No-defect, and Severe classes was not empirically validated in the present study and requires further verification using additional real rare-state inspection records.

Author Contributions: Conceptualization, Z.F. and H.Z.; data curation, Z.F. and Y.Z.; formal analysis, L.Z.; funding acquisition, Y.G.; investigation, H.Z.; methodology, Z.F. and Y.Z.; project administration, H.Z.; resources, Y.G.; software, F.X.; supervision, X.L. and Z.L.; validation, Y.Z. and L.Z.; visualization, Y.Z.; writing—original draft, Z.F.; writing—review and editing, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 51878059) and the Science and Technology Research Project of Henan Province (No. 252102240037).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data, models, and code that support the findings of this study are available from the corresponding author upon reasonable request. These materials include the bridge inspection dataset, the multi-task deep learning model code, and the semi-Markov chain inference code. The bridge inspection dataset is not publicly available due to infrastructure security and data owner restrictions.

Acknowledgments: The authors thank the College of Artificial Intelligence, Henan Agricultural University; the School of Highway, Chang'an University; and Detong Intelligent Technology Co., Ltd.

Conflicts of Interest: Authors Zhihui Feng, Liangqi Zhang, Xiaojun Li were employed by company Detong Intelligent Technology Co., Ltd., other authors have no conflicts of interest. The authors declare no conflicts of interest.

Appendix A

Table A1. Hyperparameter settings of the comparative baseline models.

Model	Input and Architecture	Optimizer or Solver	Training Settings	Regularization and Other Settings
LSTM	Numerical temporal sequence; LSTM(64, return sequences)–Dropout(0.30)–LSTM(32)–Dropout(0.30)–Dense(32)–Softmax	Adam; learning rate = 1.0×10^{-3} ; gradient clipping norm = 1.0	Maximum epochs = 80; batch size = 128; sparse categorical cross-entropy	Early stopping monitored validation loss, patience = 15, minimum improvement = 1.0×10^{-4} ; learning-rate reduction factor = 0.5, patience = 6, minimum learning rate = 1.0×10^{-6}
Transformer	Numerical temporal sequence; Dense(64)–layer normalization–multi-head attention–residual connection–global average pooling–Dense(32)–Dropout(0.30)–Softmax; 4 attention heads; key dimension = 16	Adam; learning rate = 1.0×10^{-3} ; gradient clipping norm = 1.0	Maximum epochs = 80; batch size = 128; sparse categorical cross-entropy	Early stopping monitored validation loss, patience = 15, minimum improvement = 1.0×10^{-4} ; learning-rate reduction factor = 0.5, patience = 6, minimum learning rate = 1.0×10^{-6}
Random Forest	Flattened numerical temporal features	Gini-based Random Forest	Number of trees = 300; maximum depth = 20	Minimum samples for splitting = 2; minimum samples per leaf = 1; maximum features = square root; class weighting = balanced subsample; bootstrap sampling enabled; random seed = 42

Table A1. *Cont.*

Model	Input and Architecture	Optimizer or Solver	Training Settings	Regularization and Other Settings
XGBoost	Flattened numerical temporal features	Gradient boosting; learning rate = 0.05	Number of trees = 300; maximum depth = 6	Minimum child weight = 1; gamma = 0; subsample ratio = 0.90; column subsample ratio = 0.90; L ₁ regularization = 0; L ₂ regularization = 1.0; tree method = histogram; evaluation metric = multiclass log loss; random seed = 42
Logistic Regression	Standardized flattened numerical temporal features concatenated with text features	L-BFGS; C = 1.0	Maximum iterations = 5000; convergence tolerance = 1.0×10^{-4}	StandardScaler preprocessing; class weighting = balanced; random seed = 42
Temporal CNN	Numerical temporal sequence; Conv1D(64, kernel = 2)–batch normalization–Dropout(0.30)–Conv1D(32, kernel = 2)–batch normalization–Dropout(0.30)–global max pooling–Dense(32)–Dropout(0.30)–Softmax	Adam; learning rate = 1.0×10^{-3} ; gradient clipping norm = 1.0	Maximum epochs = 80; batch size = 128; sparse categorical cross-entropy	Early stopping monitored validation loss, patience = 15, minimum improvement = 1.0×10^{-4} ; learning-rate reduction factor = 0.5, patience = 6, minimum learning rate = 1.0×10^{-6}
Non-Transformer MTL	Numerical branch: BiLSTM(48)–Dropout(0.30)–BiLSTM(32)–batch normalization; text branch: Dense(64)–batch normalization–Dropout(0.30); shared Dense layers of 96 and 64 units; three task-specific Dense(48) heads	Adam; learning rate = 1.0×10^{-3} ; gradient clipping norm = 1.0	Maximum epochs = 80; batch size = 128; sparse categorical cross-entropy for each task	Shared-layer dropout = 0.35 and 0.30; task-head dropout = 0.25; early-stopping patience = 20; learning-rate reduction factor = 0.5, patience = 8, minimum learning rate = 1.0×10^{-6}
Single-task variant	Same input features, shared backbone, and corresponding task-specific decoder as the proposed model; the other task heads and cross-task information-sharing mechanisms were removed	AdamW; weight decay = 3.8×10^{-5} ; gradient clipping norm = 1.5	Maximum epochs = 80; batch size = 128; four-stage learning-rate schedule: 5.4×10^{-5} , 4.6×10^{-5} , 4.1×10^{-5} , and 3.4×10^{-5}	The corresponding task-specific loss and sample-weighting settings were identical to those of the proposed model; early-stopping patience = 60; learning-rate reduction factor = 0.5, patience = 10, minimum learning rate = 1.0×10^{-6}

Note: All models were trained and evaluated using the same training, validation, and test partitions, while model-specific hyperparameters were selected according to the characteristics of each algorithm. The hyperparameter settings of the proposed model are reported separately in Table 7.

References

1. Rageh, M.O.; Elbeltagi, E.E.; Gabr, A.R.; El-Badawy, S.M.; Motawa, I.A. An intertwined triple-bottom-line rating system for highway sustainability in developing countries. *Sci. Rep.* **2026**, *16*, 5433. [\[CrossRef\]](#)
2. Lotfi Karkan, N.; Shakeri, E.; Sadeghi, N.; Banihashemi, S. Smart surveillance of structural health: A systematic review of deep learning-based visual inspection of concrete bridges using 2D images. *Infrastructures* **2025**, *10*, 338. [\[CrossRef\]](#)
3. Zemed, N.; Moulay Abdelali, H.; Cherradi, T.; Bouyahyaoui, A.; Mouzoun, K. Predictive maintenance based on the time-dependent reliability of RC bridges under material degradation and traffic growth. *J. Build. Rehabil.* **2025**, *10*, 147. [\[CrossRef\]](#)
4. de Oliveira, A.M.; Nicoletti, R.S.; de Souza, A.S.C.; Amorim, G.A.N. A numerical investigation of fire effects on the effective width of steel–concrete composite box girder bridges. *J. Braz. Soc. Mech. Sci. Eng.* **2025**, *47*, 366. [\[CrossRef\]](#)
5. Liu, B.; Meng, X.; Li, J.; Tu, Z. Load testing in bridge maintenance. *J. Archit. Res. Dev.* **2024**, *8*, 63–68. [\[CrossRef\]](#)
6. Liu, S.-S.; Huang, H.-Y.; Kumala, N.R.D. Two-stage optimization model for life cycle maintenance scheduling of bridge infrastructure. *Appl. Sci.* **2020**, *10*, 8887. [\[CrossRef\]](#)
7. Sun, B.; Xie, Y.; Zhou, H.; Li, R.; Wu, T.; Ruan, W. Hybrid WARIMA-WANN algorithm for data prediction in bridge health monitoring system. *Structures* **2024**, *69*, 107490. [\[CrossRef\]](#)
8. Ulansky, V.; Raza, A. A historical survey of corrective and preventive maintenance models with imperfect inspections: Cases of constant and non-constant probabilities of decision making. *Aerospace* **2024**, *11*, 92. [\[CrossRef\]](#)
9. Chen, C.-L.; Hung, C.C.W.; Zhang, W.-Y. Establishing severe deterioration and maintenance rules of bridge components: An empirical study in Taiwan. *Case Stud. Constr. Mater.* **2025**, *22*, e04907. [\[CrossRef\]](#)

10. Jia, B.Y.; Yu, X.L.; Huang, K.; Yan, Q.S. Study on dynamic behavior and seismic performance of cable-stayed bridge with different auxiliary pier positions. *Appl. Mech. Mater.* **2014**, *587–589*, 1655–1658. [\[CrossRef\]](#)
11. Greco, F.; Lonetti, P.; Pascuzzo, A.; Sansone, G. An analysis of the dynamic behavior of damaged reinforced concrete bridges under moving vehicle loads by using the moving mesh technique. *Struct. Durab. Health Monit.* **2023**, *17*, 457–483. [\[CrossRef\]](#)
12. Prakash, V.; Debono, C.J.; Musarat, M.A.; Borg, R.P.; Seychell, D.; Ding, W.; Shu, J. Structural health monitoring of concrete bridges through artificial intelligence: A narrative review. *Appl. Sci.* **2025**, *15*, 4855. [\[CrossRef\]](#)
13. Yu, Y.; Rashidi, M.; Dorafshan, S.; Samali, B.; Farsangi, E.N.; Yi, S.; Ding, Z. Ground penetrating radar-based automated defect identification of bridge decks: A hybrid approach. *J. Civ. Struct. Health Monit.* **2025**, *15*, 521–543. [\[CrossRef\]](#)
14. Jiang, T.; Li, L.; Samali, B.; Yu, Y.; Huang, K.; Yan, W.; Wang, L. Lightweight object detection network for multi-damage recognition of concrete bridges in complex environments. *Comput.-Aided Civ. Infrastruct. Eng.* **2024**, *39*, 3646–3665. [\[CrossRef\]](#)
15. Ameli, Z.; Nesheli, S.J.; Landis, E.N. Deep learning-based steel bridge corrosion segmentation and condition rating using Mask RCNN and YOLOv8. *Infrastructures* **2024**, *9*, 3. [\[CrossRef\]](#)
16. Huang, L.; Fan, G.; Li, J.; Hao, H. Deep learning for automated multiclass surface damage detection in bridge inspections. *Autom. Constr.* **2024**, *166*, 105601. [\[CrossRef\]](#)
17. Pokhrel, R.; Samsami, R.; Elmi, S.; Brooks, C.N. Automated concrete bridge deck inspection using unmanned aerial system (UAS)-collected data: A machine learning (ML) approach. *Eng* **2024**, *5*, 1937–1960. [\[CrossRef\]](#)
18. Wang, Y.; Cai, C.S.; Han, B.; Xie, H.; Bao, F.; Wu, H. A deep learning-based approach for assessment of bridge condition through fusion of multi-type inspection data. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107468. [\[CrossRef\]](#)
19. Chen, H.; Li, P.; Zhuo, Y. Fatigue life prediction and reliability analysis of reinforced concrete bridge decks based on an XFEM-ANN-Monte Carlo hybrid framework. *Appl. Sci.* **2026**, *16*, 209. [\[CrossRef\]](#)
20. Miao, P.; Zhou, C.; Wu, Y.; Hu, W.; Luo, D.; Ma, S.; Wang, W.; Srimahachota, T. Enhancing bridge performance assessment with a hybrid attention-based long short-term memory and hidden Markov model using sparse inspection data. *Structures* **2025**, *71*, 107995. [\[CrossRef\]](#)
21. Shen, P.; Chen, Y.; Ma, S.; Yan, Y. Safety assessment method of concrete-filled steel tubular arch bridge by fuzzy analytic hierarchy process. *Buildings* **2024**, *14*, 67. [\[CrossRef\]](#)
22. Lu, Y.; Nie, C.; Zhou, D.; Shi, L. Research on programmatic multi-attribute decision-making problem: An example of bridge pile foundation project in karst area. *PLoS ONE* **2023**, *18*, e0295296. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Abdelsamie, M.M.; Azab, S.S.; Hefny, H.A. Deep multi-task learning: A review of concepts, methods, and cross-domain applications. *Int. J. Data Sci. Anal.* **2026**, *21*, 77. [\[CrossRef\]](#)
24. Liu, S.; Bai, M.; Li, H.; Guo, L.; Li, D. Urban traffic health assessment based on multi-state semi-Markov process. *Reliab. Eng. Syst. Saf.* **2026**, *265*, 111586. [\[CrossRef\]](#)
25. Tran Viet, H.; Tran Ngoc, H.; Tien Thanh, B.; Vu Quang, T.; Bui Phuc, L. Time-series based bridge damage diagnosis using a novel graph transformer with vector quantization. *Knowl.-Based Syst.* **2026**, *336*, 115248. [\[CrossRef\]](#)
26. Hakimi, O.; Liu, H.; Abudayyeh, O. Deep learning-driven multi-level data fusion framework for predictive maintenance and structural health monitoring of concrete bridge decks. *Autom. Constr.* **2025**, *175*, 106180. [\[CrossRef\]](#)
27. Li, R.; Zhao, L.; Wei, H.; Hu, G.; Xu, Y.; Ouyang, B.; Tan, J. Multi-defect type beam bridge dataset: GYU-DET. *Sci. Data* **2025**, *12*, 1101. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Oyarhossein, M.A.; Sugiyama, G.; Rodrigues, F.; Rodrigues, H. Colour-coded BIM models for corrosion severity assessment in steel bridges. *CivilEng* **2025**, *6*, 67. [\[CrossRef\]](#)
29. Kulpa, M.; Siwowski, T.; Rajchel, M.; Błazik-Borowa, E.; Jukowski, M. Long-term performance evaluation of an FRP composite road bridge using DFOS monitoring system. *Sensors* **2025**, *25*, 7131. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Altamimi, S.; Fang, L.; Amleh, L. Identifying climatic hazard importance factors for bridges using expert-based fuzzy analytic hierarchy process. *Int. J. Disaster Risk Reduct.* **2026**, *133*, 105971. [\[CrossRef\]](#)
31. Li, Q.; Zhou, J.; Feng, J. Safety risk assessment of highway bridge construction based on cloud entropy power method. *Appl. Sci.* **2022**, *12*, 8692. [\[CrossRef\]](#)
32. Zhao, L.; Gao, Q.; Dong, Y.; Hou, Y.; Sun, L.; Wang, W. Vision-guided dynamic risk assessment for long-span PC continuous rigid-frame bridge construction through DEMATEL-ISM-DBN modelling. *Buildings* **2025**, *15*, 4543. [\[CrossRef\]](#)
33. Ma, R.; Su, R.; Liu, Y.; Zhu, Q.; Hu, X. Wind-induced collision of hangers in a network arch bridge: Reliability and cost-benefit analysis. *Structures* **2025**, *80*, 109947. [\[CrossRef\]](#)
34. Shahriver, F.; Sidiq, A.; Mahmoodian, M.; Jayasinghe, S.; Sun, Z.; Setunge, S. AI-based bridge maintenance management: A comprehensive review. *Artif. Intell. Rev.* **2025**, *58*, 135. [\[CrossRef\]](#)
35. Xin, S.; Wang, Z.; Su, H.; Shang, L.; Meng, K.; Wang, X.; Zhou, Z.; Zhao, Z.; Zhang, P. Structural prediction analysis of cross-sea cable-stayed bridges during operation based on existing traffic volumes. *J. Mar. Sci. Eng.* **2022**, *10*, 1758. [\[CrossRef\]](#)
36. JTG/T H21-2011; Standards for Technical Condition Evaluation of Highway Bridges. China Communications Press Co., Ltd.: Beijing, China, 2011.

37. Amirkhani, D.; Allili, M.S.; Hebbache, L.; Hammouche, N.; Lapointe, J.-F. Visual concrete bridge defect classification and detection using deep learning: A systematic review. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 10483–10505. [\[CrossRef\]](#)
38. Xu, B.; Qi, N.; Zhou, J.; Li, Q. Reliability assessment of highway bridges based on combined empowerment-TOPSIS method. *Sustainability* **2022**, *14*, 7793. [\[CrossRef\]](#)
39. Zhang, J.; Yang, H. Robust support vector machine based on the bounded asymmetric least squares loss function and its applications in noise corrupted data. *Adv. Eng. Inform.* **2025**, *65*, 103371. [\[CrossRef\]](#)
40. Wan, Q.; Xu, X.; Han, J. A dimensionality reduction method for large-scale group decision-making using TF-IDF feature similarity and information loss entropy. *Appl. Soft Comput.* **2024**, *150*, 111039. [\[CrossRef\]](#)
41. Dibiantara, D.P.; Adha, A.; Darmawan, M.S.; Imjai, T.; Russell, J.; Laory, I. Spatio-temporal enhanced anomaly detection in FRP bridge monitoring using MPCA, biGRU, and attention mechanisms. *J. Civ. Struct. Health Monit.* **2025**, *15*, 1837–1855. [\[CrossRef\]](#)
42. Zhu, Y.; Feng, K.; Hua, C.; Wang, X.; Hu, Z.; Wang, H.; Su, H. Model analysis and experimental investigation of soft pneumatic manipulator for fruit grasping. *Sensors* **2022**, *22*, 4532. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Wang, H. Towards intelligent bridge condition prediction with SMOTE resampling method. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence (ICAAI '24)*, London, UK, 17–19 October 2024; Association for Computing Machinery: New York, NY, USA, 2025; pp. 48–53. [\[CrossRef\]](#)
44. Feng, D.-C.; Wang, W.-J.; Mangalathu, S.; Sun, Z. Condition assessment of highway bridges using textual data and natural language processing (NLP-) based machine learning models. *Struct. Control Health Monit.* **2023**, *2023*, 9761154. [\[CrossRef\]](#)
45. Guo, S.; Guan, Y.; Li, R.; Li, X.; Tan, H. Frame-based multi-level semantics representation for text matching. *Knowl.-Based Syst.* **2021**, *232*, 107454. [\[CrossRef\]](#)
46. Li, P.; Chen, J.; Tang, C. Bridging knowledge distillation gap for few-sample unsupervised semantic segmentation. *Inf. Sci.* **2024**, *673*, 120714. [\[CrossRef\]](#)
47. Liao, Q.; Feng, Z.; Wu, H.; Chen, S.; Xue, X. Multi-objective self-organization scheduling of dynamic AAV-MD networks via two stage deep reinforcement learning. *IEEE Trans. Consum. Electron.* **2025**, *71*, 3826–3836. [\[CrossRef\]](#)
48. Wang, B.; Chen, K.; Wang, B. Hierarchical Bayesian fusion of inspection and monitoring data for probabilistic bridge deterioration assessment. *Sci. Rep.* **2026**, *16*, 5965. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Toasa Caiza, P.D.; Ummenhofer, T.; Correia, J.A.F.O.; De Jesus, A. Applying the Weibull and Stüssi methods that derive reliable Wöhler curves to historical German bridges. *Pract. Period. Struct. Des. Constr.* **2020**, *25*, 04020029. [\[CrossRef\]](#)
50. Fang, Y.; Sun, L. Developing a semi-Markov process model for bridge deterioration prediction in Shanghai. *Sustainability* **2019**, *11*, 5524. [\[CrossRef\]](#)
51. Goyal, R. *Bridge Condition Rating Forecast—Survival-Based Models*; Long-Term Bridge Performance InfoBridge, Federal Highway Administration: Washington, DC, USA, 2020.
52. Jia, G.; Li, M. *Development of Age and State Dependent Stochastic Model for Improved Bridge Deterioration Prediction*; Report No. MPC 24-560; Mountain-Plains Consortium, North Dakota State University: Fargo, ND, USA, 2024.
53. Zheng, W.; Chen, Y.T. Novel probabilistic approach to assessing barge-bridge collision damage based on vibration measurements through transitional Markov chain Monte Carlo sampling. *J. Civ. Struct. Health Monit.* **2014**, *4*, 119–131. [\[CrossRef\]](#)
54. Zhang, G.Q.; Wang, B.; Li, J.; Xu, Y.-L. The application of deep learning in bridge health monitoring: A literature review. *Adv. Bridg. Eng.* **2022**, *3*, 22. [\[CrossRef\]](#)
55. Ade, O.A.; Sindhu, T.N.; Dutta, S.; Shafiq, A.; Al-Mdallal, Q.M. Harriss generalized kappa distribution: Statistical modeling and practical applications in finance and engineering. *Eng. Rep.* **2025**, *7*, e70441. [\[CrossRef\]](#)
56. Jamshidi, S.; Mohammadi, M.; Bagheri, S.; Najafabadi, H.E.; Rezvanian, A.; Gheisari, M.; Ghaderzadeh, M.; Shahabi, A.S.; Wu, Z. Effective text classification using BERT, MTM LSTM, and DT. *Data Knowl. Eng.* **2024**, *151*, 102306. [\[CrossRef\]](#)
57. Abiona, Q.O.; Head, M.H. Determination of bridge elements' weights using the random forest algorithm. *J. Perform. Constr. Facil.* **2025**, *39*, 04024056. [\[CrossRef\]](#)
58. Nguyen, H.Q.; Tran, N.H.; Nguyen, N.L.; Bui, T.T. Damage detection for truss bridge structure using XGBoost. *Adv. Sci. Technol.* **2025**, *158*, 65–74. [\[CrossRef\]](#)
59. Moscoso, Y.F.M.; Rincón, L.F.; Leiva-Maldonado, S.L.; Campos e Matos, J.A.S.C. Bridge deterioration models for different superstructure types using Markov chains and two-step cluster analysis. *Struct. Infrastruct. Eng.* **2024**, *20*, 791–801. [\[CrossRef\]](#)

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.